AN APPROACH TO ANALYZING CASE-CONTROL SCREENING DATA WITH AN APPLICATION TO DIGITAL RECTAL EXAM AND METASTATIC PROSTATE CANCER

by

Eric Sayre
B.Sc., Simon Fraser University, 1999

PROJECT SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

In the
Department
of
Statistics and Actuarial Science

SIMON FRASER UNIVERSITY

Spring 2005

APPROVAL

Name:	**Eric Sayre**
Degree:	**Master of Science**
Title of Project:	**An Approach to Analyzing Case-Control Screening Data with an Application to Digital Rectal Exam and Metastatic Prostate Cancer**

Examining Committee:

Chair: **Dr. Richard Lockhart**

Dr. K. Laurence Weldon
Senior Supervisor

Dr. Charmaine Dean

Dr. Xiaoqiong Joan Hu
External Examiner

Date Defended/Approved:

ABSTRACT

Prostate cancer (PC) is a prominent killer; men over 40 should receive annual digital rectal exam (DRE) to prevent metastatic prostate cancer (MPC). For rare outcomes like MPC, case-control studies are common. Previous studies counted DREs as dichotomous and compared cases/controls using conditional logistic regression and the case-control odds ratio.

Observation time for DRE ends at PC since afterwards DREs are diagnostic. Cases and controls are matched on time; but when cases are screen-detected there is bias counting more case screenings. Excluding tests shortly before diagnosis only reverses the bias.

We begin instead with a life table for first DRE, stratified by case/control and propensity score (representing confounders), from which the Mantel-Haenszel odds ratio is calculated for case/control versus DRE on tables cross-classifying these, stratified by interval and propensity score. We find a mildly protective effect (OR=0.861, 95% CI=0.611,1.215), and find the approach superior to those used in previous studies.

ACKNOWLEDGEMENTS

I want to thank my senior supervisor, Larry Weldon, who encouraged and supported me in my study of statistics since my second undergraduate statistics course, which he made so interesting that it inspired me to make statistics my field of study.

I also wish to thank Charmaine Dean, Joan Hu and Richard Lockhart, whose well-taught courses provided me methods that were critical to the analysis performed in this project.

Finally, I wish to thank my supervisor at the Arthritis Research Centre of Canada, Jacek Kopec. Jacek provided the data, and worked with me in developing the methods in this project as we analyzed prostate specific antigen (PSA) screening.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 CHAPTER ONE: INTRODUCTION

Prostate cancer (PC) is the second leading cause of death due to cancer in North American men [1]. Men over the age of 40 are recommended to receive annual screening for PC. Digital rectal exam (DRE) has long been the standard method of screening for PC by general practitioners. The goal of DRE is not to prevent PC, but to detect it early enough to reduce the metastatic form of the disease (the spreading of the cancer throughout the body). Metastatic prostate cancer (MPC) almost always leads to death [2], and almost all deaths from prostate cancer result from the metastatic form. According to the U.S. National Cancer Institute in 2004, the estimated lifetime risk of PC is about 17.8%, and the lifetime risk of MPC is 3%.

One way to study the efficacy of a screening test is through a randomized controlled trial. Such trials have been used to study various screening tests including mammography, clinical examination of the breast, sputum cytology among smokers, and fecal occult blood testing [3]. However, for a rare outcome, randomized trials require prohibitively large sample sizes to detect a moderate screening effect [3]. Cohort studies face the same problem. The case-control study is an alternative design that avoids this problem; by selecting every case in a specified region and then sampling an appropriate number of controls from the same population, reasonable power can be achieved with much smaller sample sizes.

Previous case-control studies of DRE efficacy have found equivocal results in the protective direction. A 1998 study by Jacobsen et al [4] found a strong protective effect of DRE on prostate cancer mortality (OR=0.31; 95% CI=0.19-0.49). However, in studies by Richert-Boe et al [6] and Friedman et al [7], no significant association was found (OR=0.84 and 0.90, respectively), and in a 2004 case-control study of prostate cancer mortality and DRE screening by Weinmann et al [8], a protective but insignificant result was also seen (OR=0.70; 95% CI=0.46-

1.1). These studies counted screening/no screening as a dichotomous variable, and analyzed whether the prevalence of this exposure differed between cases and controls. Methods of analysis varied from conditional logistic regression to the case-control odds ratio for 2 by 2 tables. Cases and controls are matched on the basis of observation time to ensure that each group had an equal opportunity to be screened. Often subjects are also matched on age and region of residence. The observation time for a case ends at the diagnosis of PC, since beyond that time no test would be screening, but rather diagnostic. But the observation time may end sooner if for example there is the onset of symptoms related to PC (e.g., urinary difficulties), since symptoms also make it likely that any testing is diagnostic.

The problem with this approach arises when a case's PC is detected by a screening test (screen-detected case). The diagnosis of PC ends the observation time for the case—and also his matched controls, to produce a fair comparison. But the result is a bias favoring higher apparent screening rates in cases compared with controls [9]. To see this, consider an example in which screening will always detect PC if it's present, and cases and controls are screened at the same rate over time. The latter condition implies that there is no effect of screening on MPC, so an unbiased estimate should show no effect. Each matched set is observed exactly long enough to record the screening test for the case, and no longer. Whether the screening test occurred first for the case or for his matching control is a coin toss if they are screened at the same rate. The half of the controls who are screened after their matching cases will not have their test recorded, but all cases have theirs recorded. This bias is mitigated somewhat if screening is more frequent or if the sensitivity of the screening test is not perfect. Figure 1.1 and Figure 1.2 illustrate the bias.

Figure 1.1 In matched sets with no screen-detected case there is no dependence between screening tests in the cases and the end of the observation time in matched controls; if cases and controls are screened at the same rate all DREs are counted equally

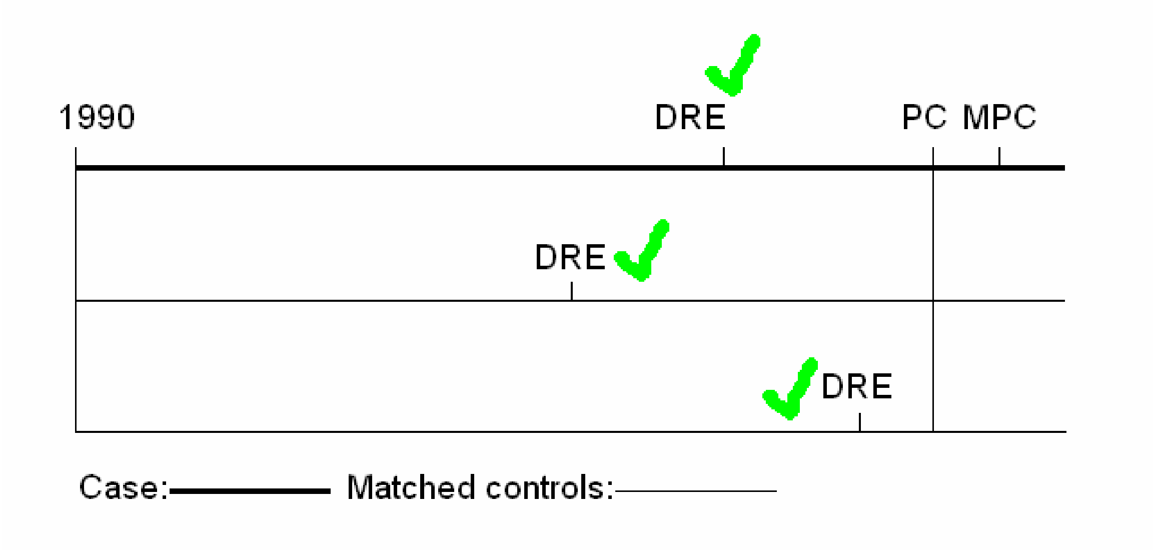
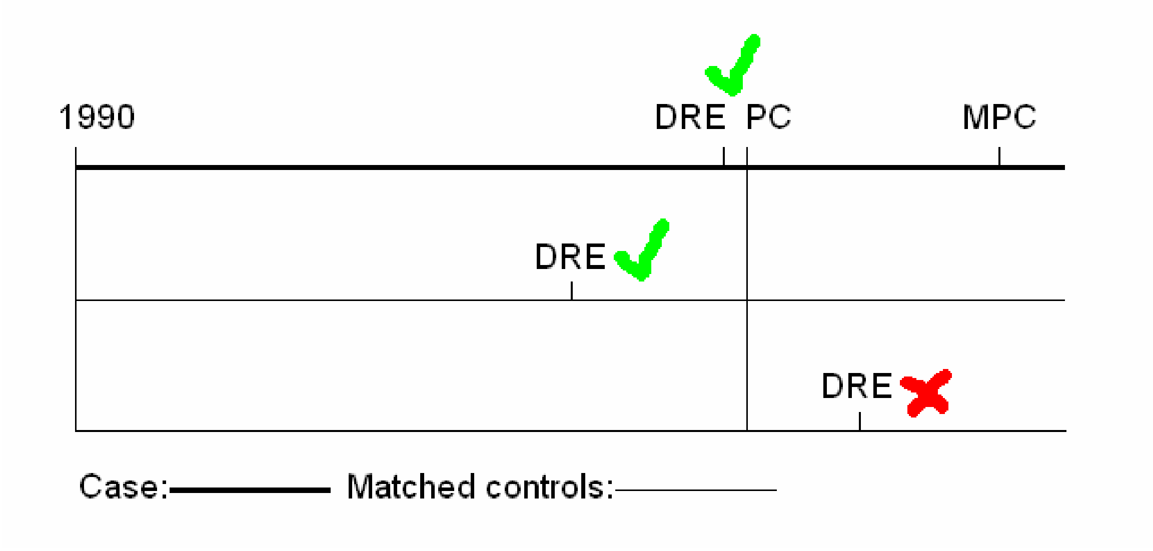
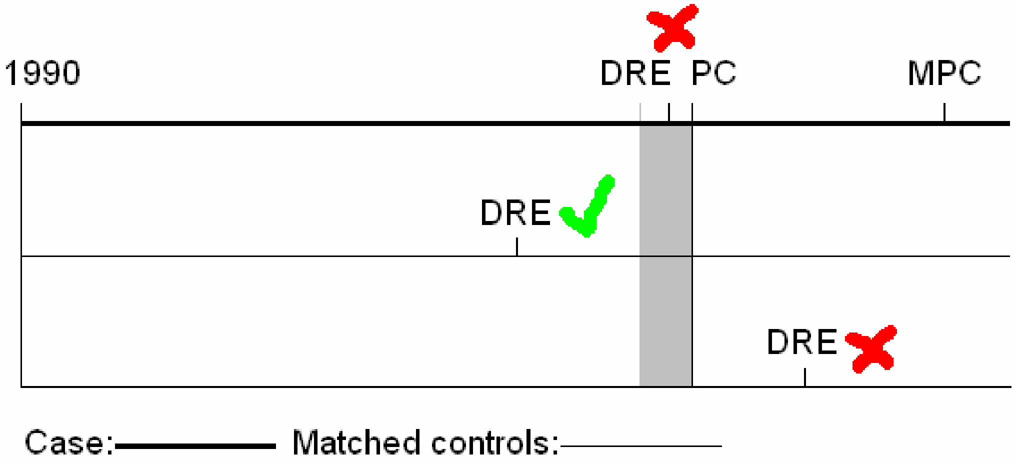


Figure 1.2 In matched sets with a screen-detected case there is dependence between screening tests in the cases and the end of the observation time in matched controls; if cases and controls are screened at the same rate all DREs are counted in cases but only half the DREs in controls are counted



One solution that has been proposed [7] in an effort to eliminate this bias is the exclusion of all tests done within some time before the diagnosis. However, this introduces a bias in the opposite direction; amongst screen-detected cases, the test will always precede the diagnosis by a short time, and all these tests will be excluded. The cases will be observed as long as possible to just miss recording the screening, which is as unfair as possible towards the cases having an opportunity to be screened. This strongly biases the result in favor of screening [7]. The strongly protective result found in the 1998 study by Jacobsen et al was based on this approach. Figure 1.3 illustrates this reverse bias.

Figure 1.3 In an effort to eliminate the dependence bias, all DREs a short time from diagnosis are excluded, but this reverses the bias; if cases and controls are screened at the same rate no DREs are counted in cases but half the DREs in controls are counted



In this study, we propose an alternative method of analyzing case-control screening data that is free of these biases. The data are initially analyzed in a stratified life table survival analysis where first screening DRE is the event, stratified by case/control status and propensity score [14], a single variable representing the combination of several confounding variables. The overall Mantel-Haenszel odds ratio for case/control status versus DRE is calculated from the life tables on the set of 2 by 2 tables cross-classifying DRE with case/control status, stratified by interval and propensity score. This estimates the relative odds of DRE screening in cases versus controls [10], but since odds ratios are symmetric, it also estimates the effect of having a screening DRE on the odds of developing MPC.

In this project we apply this method to the study of DRE efficacy in preventing MPC. We examine the method both mathematically and using Monte Carlo simulation.

2 CHAPTER TWO: DATA COLLECTION

2.1 Cases

In a case-control study, data are first collected on some fraction (for rare diseases, this may be 100%) of the disease cases occurring within the study population during some period of time. In this project, the study population consisted of all residents of Metropolitan Toronto and the 5 surrounding counties of Durham, Halton, Peel, York and Simcoe. This can be described in terms of the following 8 regions: Halton, Peel, Simcoe, Etobicoke, City of York, York Region, Old Toronto, Durham, North York, Scarborough, East York. This describes a population of 5.2 million (2001) and is served by two Regional Cancer Centres. Cases were men who developed metastatic prostate cancer between August 1, 1999 and May 31, 2002. Patients with metastases to the lymph nodes and those with local spread of the original tumor into adjacent organs were excluded. This is because this study is using MPC as a proxy for mortality from PC; thus MPC was strictly assessed as spread into bones and/or distant organs. Cases had to be 40 to 84 years of age when diagnosed with MPC, diagnosed with PC on or after January 1, 1990, living in the study area at the time of PC diagnosis and able to answer a questionnaire in English. The date of diagnosis of PC was the date of biopsy.

New cases of MPC were found by searching computerized lists of PC patients treated at the two Regional Cancer Centres. To ascertain cases that were not referred to the Cancer Centres, monthly contacts were maintained by telephone, mail, or email with the participating urologists (88 out of 90) and all the oncologists in the study area who would normally treat men with MPC. There were 235 cases in the analysis dataset.

Charts were searched for screening DREs only between January 1, 1990 and the date of diagnosis of PC (henceforth this date may be referred to as the "reference date"). Prostate-related symptoms (e.g., urinary difficulties) were ascertained from the beginning of the patient history. Data were also collected on the matching variables age and geographic region, as well as several other variables that would be considered as potential confounders, including smoking history, alcohol consumption "about ten years ago" (around 1990) measured on an 8-point index, body weight, frequency of routine doctor checkups from 1980 to 1984 and from 1985 to 1989, frequency of strenuous activity between the ages of 30 and 39 and "about ten years ago" (around 1990), whether a blood relative was ever diagnosed with prostate cancer, whether a blood relative was ever diagnosed with any other form of cancer, education, country of origin (born in Canada or elsewhere), and frequency of use of multivitamins during the past 20 years.

2.2 Controls

Controls were men without metastatic prostate cancer, selected randomly from the municipal tax records database for the study area and able to answer a questionnaire in English. There were 458 controls in total. Men known to have non-metastatic prostate cancer were eligible to be controls as long as their cancer was diagnosed after January 1, 1990. 11 controls developed prostate cancer during the study. Controls were sampled throughout the period of case accrual and frequency matched to cases for age group (5-year intervals) and region of residence (8 regions). Because of difficulties encountered obtaining charts, the relative number of controls matched to the cases is not constant across strata, but this should not adversely affect the analysis. Controls were further "matched" on the basis of observation time to individual cases; their charts were searched for screening DREs only between January 1, 1990 and the date of diagnosis of PC for the matching case (the reference date). However, we will see that in the analysis performed in this study, the data can be treated as having been matched only on age group and region. Table 2.1 shows the number of cases and controls in each stratum of the matching variables.

Table 2.1 Distribution of the matching variables age group (in 1990) and region in the study sample

Region	Age Group	Cases	Controls	Region	Age Group	Cases	Controls
Durham	40-44	1	3	Scarborough/East York	55-59	6	16
Durham	45-49	2	3	Scarborough/East York	60-64	8	15
Durham	50-54	4	8	Scarborough/East York	65-69	12	19
Durham	55-59	7	8	Scarborough/East York	70-74	6	12
Durham	60-64	4	17	Simcoe	35-39	1	2
Durham	65-69	5	7	Simcoe	45-49	4	7
Durham	70-74	3	7	Simcoe	50-54	3	4
Etobicoke/City of York	45-49	2	3	Simcoe	55-59	3	8
Etobicoke/City of York	50-54	3	10	Simcoe	60-64	2	6
Etobicoke/City of York	55-59	7	14	Simcoe	65-69	6	7
Etobicoke/City of York	60-64	6	16	York Region	35-39	2	4
Etobicoke/City of York	65-69	11	15	York Region	40-44	2	6
Halton/Peel	40-44	1	1	York Region	45-49	3	4
Halton/Peel	45-49	1	4	York Region	50-54	5	7
Halton/Peel	50-54	2	8	York Region	55-59	3	12
Halton/Peel	55-59	14	19	York Region	60-64	9	12
Halton/Peel	60-64	6	18	York Region	65-69	6	17
Halton/Peel	65-69	9	17	York Region	70-74	3	3
Halton/Peel	70-74	3	4	Old Toronto	40-44	1	2
North York	45-49	1	2	Old Toronto	45-49	1	2
North York	50-54	2	3	Old Toronto	50-54	4	6
North York	55-59	5	9	Old Toronto	55-59	6	9
North York	60-64	8	18	Old Toronto	60-64	9	20
North York	65-69	8	17	Old Toronto	65-69	6	14
Scarborough/East York	45-49	3	5	Old Toronto	70-74	4	5
Scarborough/East York	50-54	2	3				

Besides the matching variables age group and region, the same potential confounding variables that were collected on the cases were also collected on the controls.

3 CHAPTER THREE: STATISTICAL ANALYSIS

3.1 Propensity score

3.1.1 Background

Propensity scoring is a method of combining several variables into a single continuous measure for the purpose of adjusting for the original variables as potential confounders in the analysis of some treatment [14]. This approach is useful when the data are too sparse to adjust for the confounding variables simultaneously via inclusion in the model. The propensity score is the estimated probability of a subject having the treatment level of interest given his set of potential confounders. It is often estimated with logistic regression.

When using logistic regression to analyze matched data, failure to condition on the matching variables can lead to bias or instability [19]. The former problem can arise if one applies an unstratified analysis, while both problems can arise if one uses an unconditional likelihood and represents strata with design variables, in part because there are often too few subjects within each stratum to facilitate the estimation of stratum-specific parameters.

The correct approach for matched data analysis is conditional logistic regression (CLR). Hosmer and Lemeshow [19] explain CLR maximum likelihood estimation as maximizing a likelihood that is, for the k^{th} stratum, the probability of the observed data conditional on the Stratum k total n_k and number of cases n_{1k}, which are the sufficient statistic for the stratum-specific (nuisance) parameters. This is the likelihood of the data in Stratum k divided by the sum of the probabilities over the $\binom{n_k}{n_{1k}}$ possible assignments of cases and controls within Stratum k under the observed number of cases. Conditioning on the sufficient statistic removes the dependence of the likelihood on the nuisance parameters while allowing the estimation of the slope. The CLR likelihood is:

$$L_C(\beta \mid \mathbf{X}) = \prod_k \left[\prod_{i=1}^{n_{1k}} e^{\beta' \mathbf{x}_i} \Bigg/ \sum_{j=1}^{\binom{n_k}{n_{1k}}} \prod_{i_j=1}^{n_{1k}} e^{\beta' \mathbf{x}_{ji_j}} \right]$$

which can be expressed as

$$L_C(\beta \mid \mathbf{X}) = \prod_k \left[\frac{\prod_{i=1}^{n_{1k}} P(\mathbf{x_i} \mid y_i = 1) \prod_{i=n_{1k}+1}^{n_k} P(\mathbf{x_i} \mid y_i = 0)}{\sum_{j=1}^{\binom{n_k}{n_{1k}}} \left[\prod_{i_j=1}^{n_{1k}} P\left(\mathbf{x_{ji_j}} \mid y_{ji_j} = 1\right) \prod_{i_j=n_{1k}+1}^{n_k} P\left(\mathbf{x_{ji_j}} \mid y_{ji_j} = 0\right) \right]} \right]$$

Note that the probabilities in this expression are of the observed covariate vectors conditional on the outcome, case/control status. This is the right direction for case-control data, in which the disease outcome is fixed and the covariates are random.

Goodness of fit for a regression model can be tested with the likelihood ratio test (LRT) comparing the likelihood under a p-parameters restricted model L_R with that under a larger m-parameters model that is presumed correct, L_F. The smaller model should be nested within the larger model. The LRT test statistic is:

$$\chi^2_{m-p} = -2\log(L_R / L_F)$$

which is chi-square with m-p degrees of freedom under the null hypothesis that the smaller model fits the data as well as the larger.

A measure of predictive utility for a regression model is Akaike's information criterion (AIC). It is defined as

$$AIC = -2\log(L) + 2p$$

where there are p parameters in the model. AIC is maximum log likelihood adjusted for over-fitting—it penalizes the likelihood for including too many variables in a model. The idea of AIC is to estimate how well a model will predict the outcome in new data. Lower AIC indicates better predictive utility.

3.1.2 Calculating propensity score in the DRE study

In the DRE study, the first stage of analysis will involve considering case/control status as the explanatory variable and screening DRE as the outcome; the distribution of time to first screening DRE will be compared between cases and controls in a stratified life table. For this reason, the propensity score must be the estimated probability of caseness given a set of potential confounders. This is estimated in a binary conditional logistic regression model with case/control status as the dependent variable, conditioning on the matching variables age group and region. Recall that the controls were "matched" to cases also on observation time, that is, controls were only observed for DRE up to the reference date for a case. This need not be considered matching for this part of the analysis, however, because case/control status was determined at the end of the study regardless of reference date, and the potential confounders were measured at the start of the study. Therefore for this stage of the analysis, all subjects were observed equally. Age group and region are to be conditioned on as they were true matching variables guiding sample selection.

The variables to be considered as potential confounders were selected from a larger pool of available variables under the guidance of the lead epidemiologist. They were those variables that had some scientific basis for possibly acting as confounders in this analysis. For example, in order to act as a confounder, a variable needs at least to be related to both the explanatory variable (screening DRE) and the outcome (caseness, or MPC). The set of potential confounders selected for consideration in this study included smoking history, alcohol consumption "about ten years ago" (around 1990) measured on an 8-point index, body weight, frequency of routine doctor checkups from 1980 to 1984 and from 1985 to 1989, frequency of strenuous activity between the ages of 30 and 39 and "about ten years ago" (around 1990), whether a blood relative was ever diagnosed with prostate cancer, whether a blood relative was ever diagnosed with any other form of cancer, education, country of origin (born in Canada or elsewhere), and frequency of use of multivitamins during the past 20 years.

The propensity score model is a conditional logistic regression model with case/control status as the outcome and the confounders as the explanatory effects, conditioning the likelihood on the matching variables age group and region. First the full model was fit with all potential confounders. The selection strategy proceeded by dropping each variable from the full model in turn and assessing AIC for each smaller model. The one variable that had the most beneficial (or least detrimental) effect on the AIC when dropped from the full model was chosen to be dropped in the step 1 model. The reduced (by one variable) set of confounders then became the reference set from which each remaining variable was again dropped one at a time. AIC formed the basis for deciding which single variable to drop from that reduced set, to form the step 2 model. This process was repeated until all the variables were eventually dropped from the model. At each step a LRT was performed to test whether the sub-model fit the data as well as the original, full model. The final set of confounders selected came from the selected model with the lowest AIC.

Table 3.1 Model AIC and likelihood ratio tests in the development of the propensity score

Step	Dropped	AIC	-2Log(L)	Model d.f.	LRT Chi-square	LRT d.f.	LRT p-value
0		639.58	579.58	30			
1	Strenuous activities 10 years ago	633.36	581.36	26	1.78	4	0.776
2	Education	627.97	585.97	21	6.39	9	0.701
3	Routine checkups 1985-1989	623.03	589.03	17	9.45	13	0.738
4	Strenuous activities age 30-39	618.12	592.12	13	12.54	17	0.766
5	**Smoking**	**614.73**	**592.73**	**11**	**13.15**	**19**	**0.831**
6	Multivitamins	615.88	597.88	9	18.30	21	0.630
7	Relative with other cancer	617.13	601.13	8	21.55	22	0.487
8	Born in Canada	618.77	604.77	7	25.19	23	0.341
9	Routine checkups 1980-1984	624.12	618.12	3	38.54	27	0.070
10	Body weight	631.19	627.19	2	47.61	28	0.012
11	Alcohol	646.87	644.87	1	65.29	29	0.000

Figure 3.1 Model AIC at each step in the development of the propensity score

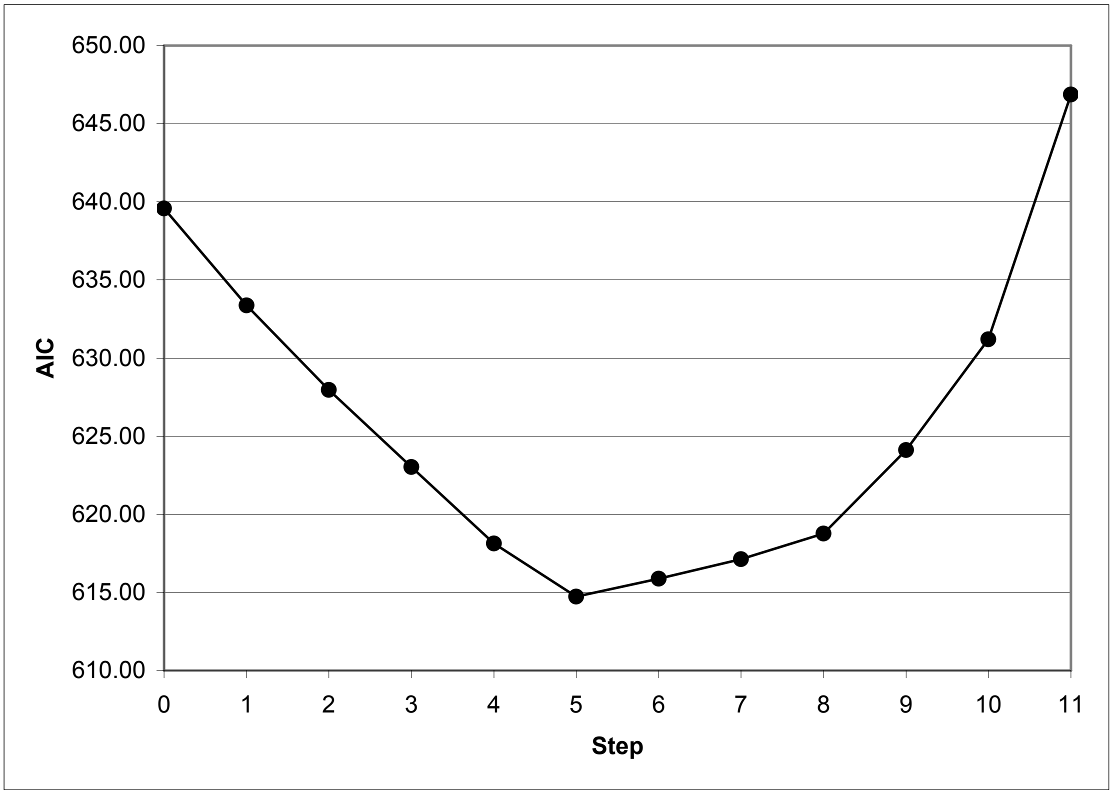


Table 3.1 shows which variable was dropped at each step of the model selection, and provides the AIC and LRT results. Figure 3.1 plots the AIC for each step of the model selection. Step 5 is the model that produces the lowest AIC. The LRT shows that the model fits as well as the full model. The final set of confounders is therefore whether a blood relative was ever diagnosed with prostate cancer, whether a blood relative was ever diagnosed with any other form of cancer, frequency of use of multivitamins during the past 20 years, country of origin (born in Canada or elsewhere), frequency of routine doctor checkups from 1980 to 1984, body weight, and alcohol consumption "about ten years ago" (around 1990) measured on an 8-point index.

Table 3.2 Final propensity score model

Variable	Level	Estimate	Standard error	Wald chi-square	P-value	Odds Ratio
Body weight		0.009	0.003	7.577	0.006	1.009
Relative with other cancer		-0.001	0.002	0.215	0.643	0.999
Relative with prostate cancer		-0.006	0.002	6.393	0.011	0.994
Born in Canada		-0.147	0.187	0.616	0.433	0.864
Alcohol		-0.001	0.041	0.000	0.988	0.999
Multivitamins during past 20 years (reference=never)	Yes irregularly	0.450	0.225	3.986	0.046	1.569
	Yes regularly	-0.107	0.235	0.208	0.648	0.898
Routine checkups 1980-1984 (reference=never)	Once in 5 years	-0.283	0.324	0.760	0.383	0.754
	Once each 2-4 years	-0.172	0.293	0.345	0.557	0.842
	Once per year	-0.804	0.270	8.866	0.003	0.447
	More than once per year	-0.566	0.342	2.742	0.098	0.568

Table 3.2 shows the conditional logistic regression model coefficients for the final model. Note that in conditional logistic regression no intercept terms are estimated. In a case-control study, since it is under the investigators' control how many cases and controls to sample, the data contain no information about baseline risk; the propensity score from this model for each subject is the probability of caseness only within the selected sample, not the general population. As a propensity score, this will suffice; it is the relative probabilities between subjects with different confounder vectors that is the important aspect of a propensity score that allows it to provide an adjustment for the set of confounders, when stratified upon. The relative probabilities are approximately preserved even without information on baseline risk.

Fifty-three subjects had one or more missing covariates. These subjects were assigned the median value of the missing variable(s) in order to calculate a propensity score for them. In most cases there was only a single missing variable.

Finally, the propensity score was categorized into three levels according to tertiles. This would allow it to be stratified upon in subsequent analyses without problems of sparse strata.

3.2 Life table analysis

Life table analysis is a method of analyzing survival (or time to event) data. The next step in the DRE study is to compare the distributions of time to first DRE screening between cases and controls using the life table method. The reason "first DRE" is the event rather than, for example, "first DRE that detects PC" is that the difference in exposure distributions is desired between cases and all controls, not just between cases and those controls who were diagnosed with prostate cancer during the course of the study. The question is whether cases have different rates of screening DRE than controls, and in advance of knowing the results of a screening DRE, all screening DREs may be considered equal.

3.2.1 Background

In survival data, there are observed event times and censored observations. A right-censored observation is one in which a subject ceases to be observed before the event is recorded; for such a subject it is only known that the event occurred sometime after the censoring time. A left-censored observation is one in which a subject does not begin to be observed until after the event has already occurred; for such as subject it is only known that the event occurred sometime before the censoring time. In a life table analysis the data are made up of observed event times and right-censored times.

The first step is to split the time axis into k intervals of generally equal length. If data are overly sparse then later intervals may be made longer. Selecting intervals is somewhat subjective.

In deciding the number and width of the intervals, the precision of the estimates can be a guide; intervals that are too short may not contain enough events to produce precise estimates of survival probability. On the other hand, if the intervals are too broad one can lose important information about changes in the survivor function over time.

Lawless [16] provides an overview of the life table method and theory for usual, prospective data. The following explanation is drawn largely from that source.

Let D_j denote the number of events in interval j, N_j the number of subjects still being observed at the start of interval j (not yet failed or censored by that point), and W_j the number of subjects censored in interval j. Then the likelihood of the data is:

$$\prod_{j=1}^{k} P\left(D_j \mid \mathbf{H}(j)\right) P\left(W_j \mid D_j, \mathbf{H}(j)\right)$$

where $\mathbf{H}(j) = \left(D_1, W_1, \ldots, D_{j-1}, W_{j-1}\right)$ is the vector containing all the information about what happened prior to interval j. The likelihood is developed under a special case where all censoring occurs at interval endpoints. The number of events in each interval j is assumed to be Binomial and not depend on censoring within the interval:

$$P\left(D_j \mid \mathbf{H}(j)\right) \sim \text{Binomial}\left(N_j, q_j\right)$$

This is reasonable since all censoring occurs at interval endpoints, and can only occur among those subjects who did not fail by that point. If the terms $P\left(W_j \mid D_j, \mathbf{H}(j)\right)$ do not contain information about the parameters q_j, then they can be dropped from the likelihood to give:

$$L\left(q_1, \ldots, q_k\right) = \prod_{j=1}^{k} \binom{N_j}{D_j} q_j^{D_j} \left(1 - q_j\right)^{N_j - D_j}$$

In a stratified life table analysis, the likelihood is maximized within each stratum.

This likelihood was developed for prospective data, but we wish to apply it to our retrospective data, and stratify the life table according to case/control status. The question is whether this is a valid approach. Were we to have analyzed the time to disease outcome amongst the cases stratified by one or more (in this case random) covariates, this might prove more difficult to answer. However, our life table is stratified by the fixed quantity case/control status, and the event time being modeled is the random "covariate" time to first screening DRE. Therefore this likelihood, like the conditional logistic regression likelihood in the previous section, is already in the right direction for our retrospective data.

Continuing Lawless's derivation, we take the log and solve the likelihood equations to yield the MLE:

$$\hat{q}_j = D_j / N_j$$

Conditional expectation can be used to obtain the (diagonal) Fisher information matrix:

$$\left(I(\mathbf{q})\right)_{jj} = \frac{E(N_j)}{q_j(1-q_j)}$$

Finally, taking the inverse and estimating $E(N_j)$ with N_j gives an asymptotic diagonal covariance matrix for $\hat{\mathbf{q}}$:

$$\left(\mathrm{AS}\hat{\mathrm{VAR}}(\hat{\mathbf{q}})\right)_{jj} = \frac{\hat{q}_j(1-\hat{q}_j)}{N_j}$$

The above was developed under the special case where all censoring occurs at interval endpoints. In reality, there is usually a censoring process that is assumed to run independently of and compete with the survival process, and therefore censoring times can occur anywhere in an interval. The standard life table estimator for q_j is actually

$$\hat{q}_j = D_j / N_j^{'}$$

where $N_j^{'} = N_j - 0.5W_j$ is the "effective sample size" in interval j. This is based on the idea that if the average censoring time within an interval is approximately at the midpoint, then each censored observation only had about half the time to fail as did one that was observed throughout the interval. In general, $\hat{q}_j$ is not a consistent estimator of q_j. However, according to Lawless, if censoring is fairly evenly distributed throughout an interval and not too heavy, and the intervals are not too wide, then the life table estimates can be considered acceptable.

Now let P_j denote the probability of survival beyond interval j (not having an event before the end of interval j). This is the survivor function evaluated at the endpoint of interval j. To survive beyond interval j, one must not experience an event in any interval up to and including interval j. Therefore:

$$P_j = \prod_{i=1}^{j} (1 - q_i) \text{ and } \hat{P}_j = \prod_{i=1}^{j} (1 - \hat{q}_i).$$

Since $\hat{q}_j$ is not a consistent estimator of q_j, neither in general is $\hat{P}_j$ a consistent estimator of P_j. The life table variance estimate is based on limiting multivariate normal distributions with mean **0** of $\sqrt{n}\left(\hat{\mathbf{P}} - \mathbf{P}^*\right)$ and $\sqrt{n}\left(\hat{\mathbf{q}} - \mathbf{q}^*\right)$, application of the multivariate delta formula

$$\text{ASCOV}(g_1(T_1,\ldots,T_k), g_2(T_1,\ldots,T_k)) = \sum_{i=1}^{k} \sum_{j=1}^{k} \frac{\partial g_1}{\partial \theta_i} \frac{\partial g_2}{\partial \theta_j} \text{ASCOV}(T_i, T_j),$$

where T_i are statistics and

$$\sqrt{n}(T_1 - \theta_1, \ldots, T_k - \theta_k) \xrightarrow{D} N(\mathbf{0}, \sum \ ),$$

and substitution of $\mathbf{q}^*$ and $\mathbf{P}^*$ with $\hat{\mathbf{q}}$ and $\hat{\mathbf{P}}$. This gives Greenwood's formula:

$$\text{ASV}\hat{\text{A}}\text{R}\left(\hat{P}_j\right) = \hat{P}_j^2 \sum_{i=1}^{j} \frac{\hat{q}_i}{N_i^{'}(1 - \hat{q}_i)}$$

This result will be used to provide confidence intervals for the life table estimates.

3.2.2 Setting up the life table in the DRE study

To begin the life table analysis in the DRE study, 2-year intervals were selected for time to first screening DRE, with left boundaries ranging from 0 to 4 years, and a final longer interval from 6 to 12 years. The latter interval was made longer in order to ensure adequate data in the fully stratified analysis.

In this study, event times were censored if any of the following conditions occurred before a subject had a screening DRE: diagnosis of prostate cancer in the subject, or for a control, diagnosis of PC in his matching case (the reference date); onset of prostate-related symptoms; positive PSA (prostate specific antigen) test result (defined as PSA level >4.0); or any TRUS (transrectal ultrasound) performed. Event times for controls were censored at the reference date because during the data collection they were only observed up to that point in their charts, while the latter three conditions were considered censoring events because any DRE test performed after their occurrence was likely to be diagnostic instead of screening. It is reasonable to view these censoring processes as independent of the screening DRE event process.

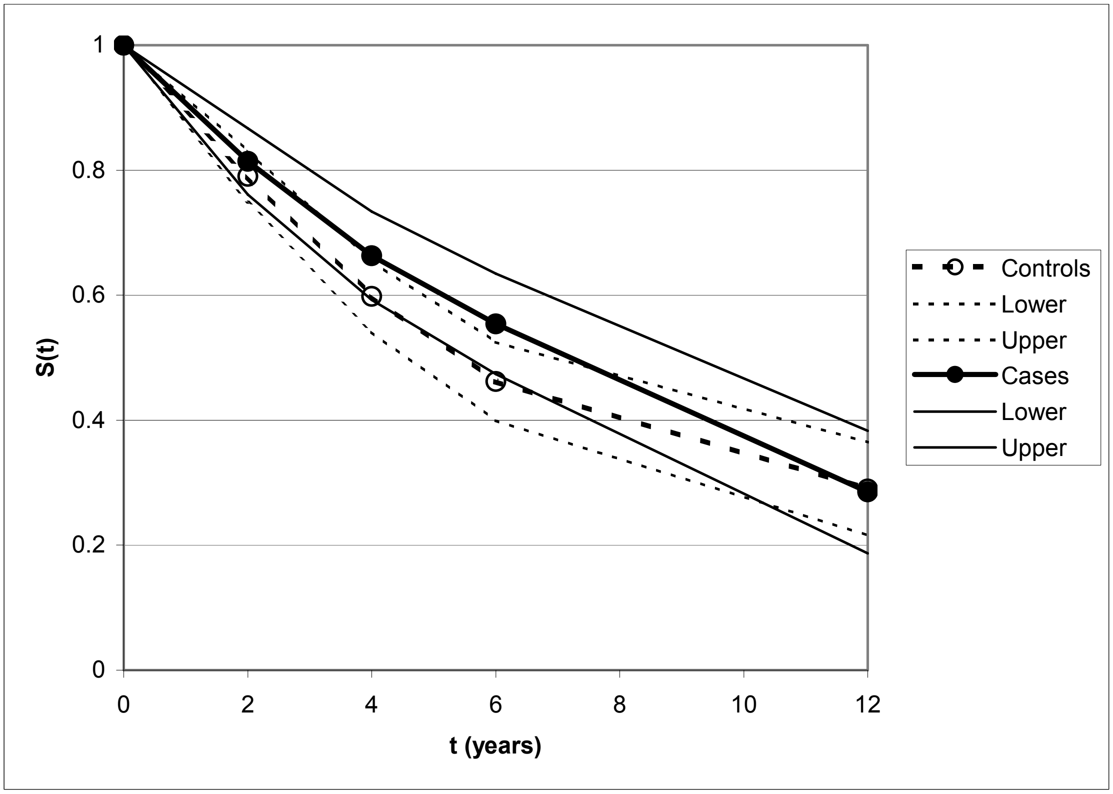
3.2.3 Life table stratified only by case/control status

The first life table is stratified only by case/control status. Survivor functions for cases and controls are listed in Table 3.3 and plotted along with lower and upper 95% confidence limits in Figure 3.2. The higher survivor function in cases suggests that they had a lower rate of screening DREs than controls, indicating a protective effect of DRE screening. The case survivor function lies partially above and partially below the upper 95% confidence limit for the controls, suggesting that the difference between survivor functions might be only borderline statistically significant. However, this life table is not adjusted for the matching variables or potential confounders, so we must expand the stratification.

Table 3.3 Life table analysis of time to first screening DRE stratified by case/control status

Case/ Control	Lower Time	Upper Time (UT)	# Failed	# Censored	Effective Sample Size	Survivor Function S(UT)	Standard Error
Controls	0	2	79	163	376.5	0.79	0.021
	2	4	48	37	197.5	0.598	0.029
	4	6	27	25	118.5	0.462	0.032
	6	12	18	61	48.5	0.29	0.038
Cases	0	2	38	62	204	0.814	0.027
	2	4	23	22	124	0.663	0.036
	4	6	13	21	79.5	0.554	0.041
	6	12	18	38	37	0.285	0.05

Figure 3.2 Life table survivor functions for time to first screening DRE stratified by case/control status, with 95% confidence limits

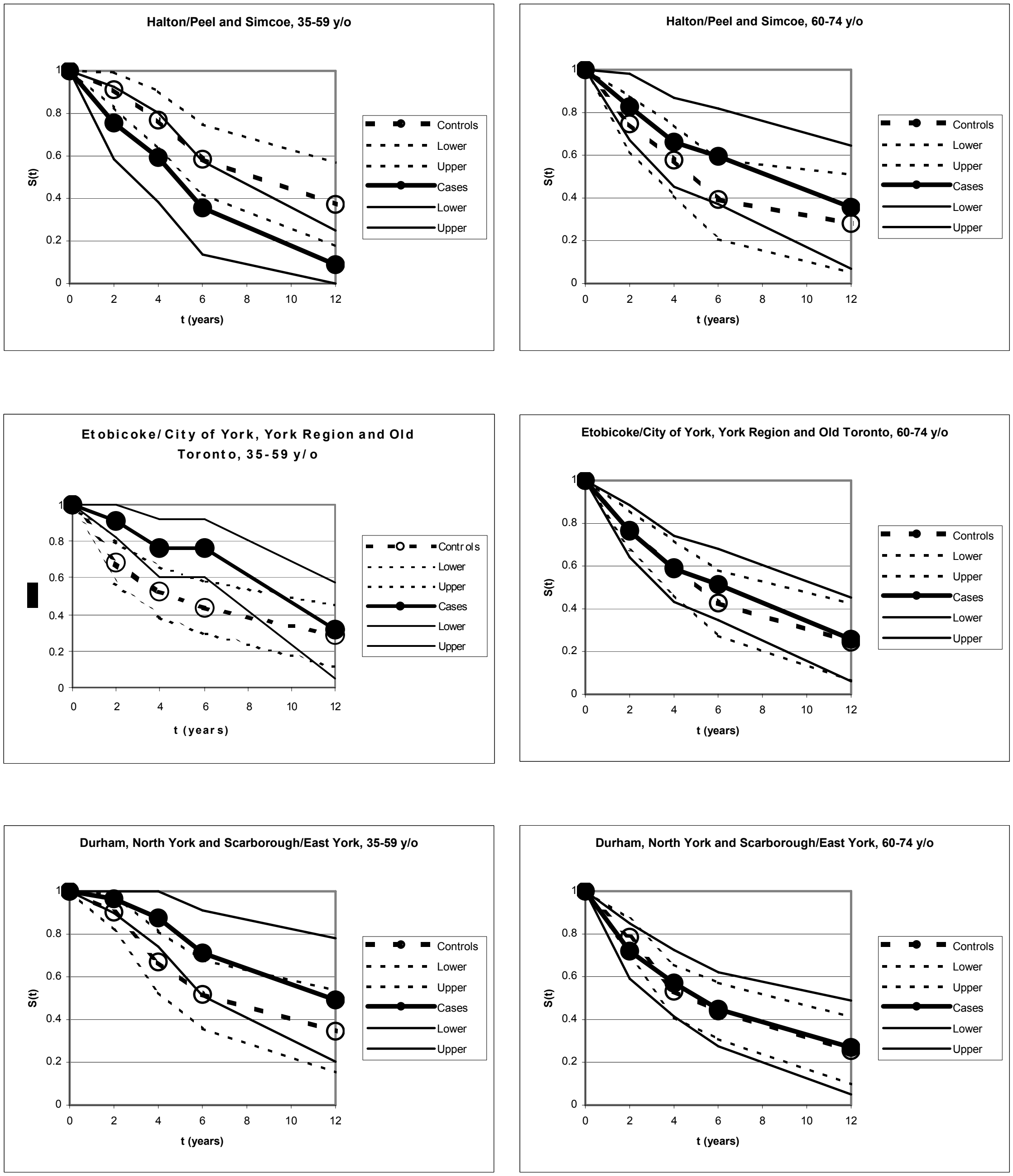


3.2.4 Life table stratified by case/control status and collapsed matching variables age group and region

Next we adjust the life table for the matched design. Recall that cases and controls were matched on age group with 5-year intervals and region with 8 regions. Subsequent stages of the DRE analysis will require data for both cases and controls in each stratum of the stratified life table. For this reason, the stratification by the matching variables had to be done on collapsed variables. Age group was dichotomized into two groups, 35-59 (n=293) and 60-74 (n=400). Region was collapsed into three geographically contiguous areas: Halton/Peel and Simcoe (n=160); Etobicoke/City of York, York Region and Old Toronto (n=274); and Durham, North York and Scarborough/East York (n=259). There are 6 strata besides case/control status.

Survivor functions for cases and controls within each of the 6 strata are plotted along with lower and upper 95% confidence limits in Figure 3.3. Cases show lower or equal rates of first screening DRE than controls in all but one stratum; in Halton/Peel and Simcoe ages 35-59, the effect is reversed, and the confidence limits suggest that the difference might be borderline significant. In Halton/Peel and Simcoe ages 60-74, the graph suggests borderline statistical significance for the opposite and prevailing effect (controls showing higher rates of screening). Possible statistical significance of the prevailing effect can also be seen in the stratum of Durham, North York and Scarborough/East York ages 35-59, and the stratum of Etobicoke/City of York, York Region and Old Toronto ages 35-59. Overall, the graph suggests a protective effect of DRE screening against MPC. However, the life table has not been adjusted for the propensity score. The stratification must be taken one step further.

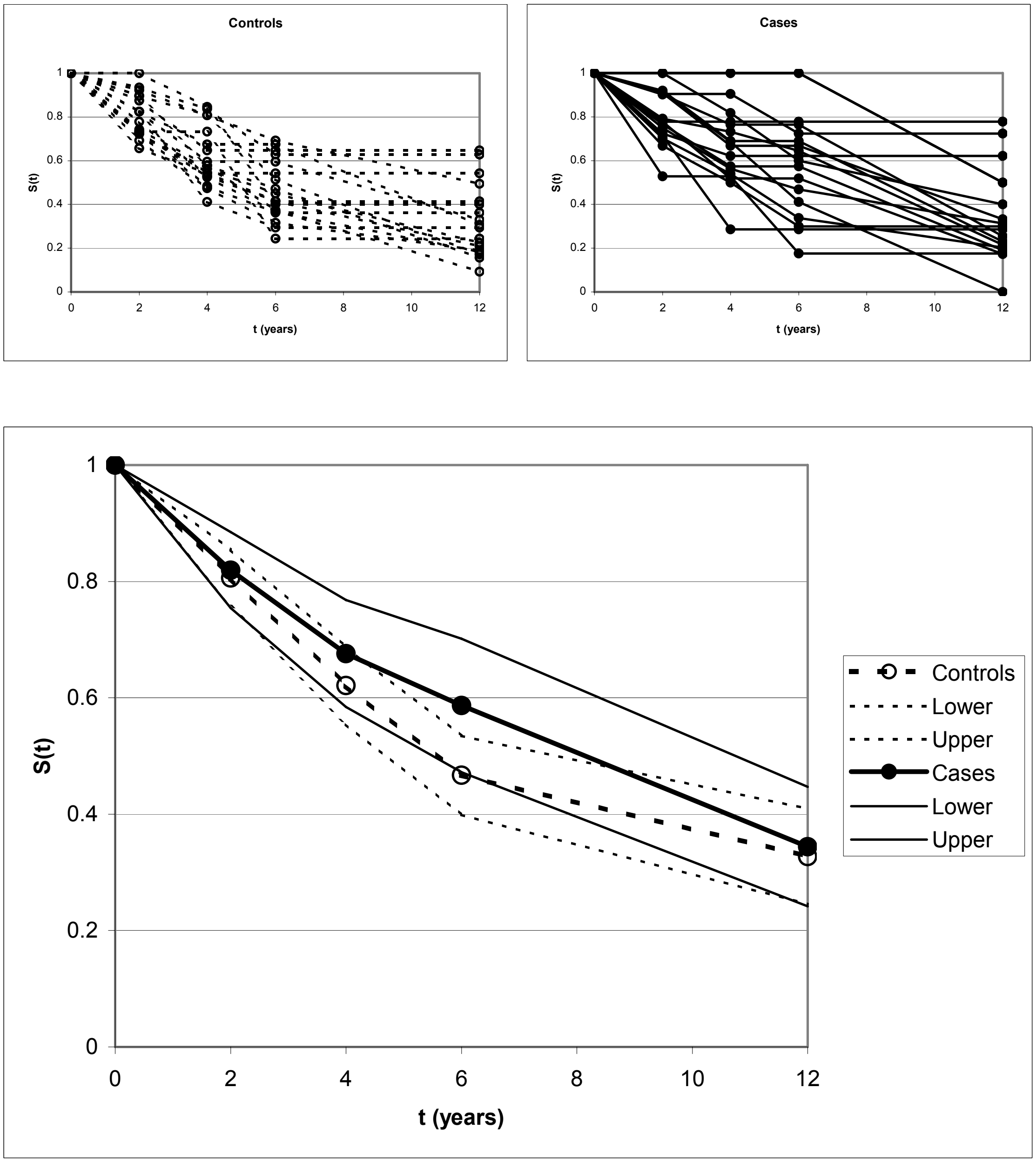
Figure 3.3 Life table survivor functions for time to first screening DRE stratified by case/control status and collapsed matching variables age group and region, with 95% confidence limits



3.2.5 Life table fully stratified by case/control status, the collapsed matching variables age group and region, and 3-level propensity score

Finally, we fully stratify the life table by case/control status, the collapsed matching variables age group and region, and the 3-level categorical propensity score.

Figure 3.4 Life table survivor functions for time to first screening DRE fully stratified by case/control status, the collapsed matching variables age group and region, and 3-level propensity score; the upper two graphs show the individual survivor functions; the lower graph shows the mean survivor functions with t-based (17 d.f.) 95% confidence limits



There are 18 strata in this life table besides case/control status. The survivor functions for the individual strata are plotted along with mean survivor functions and t-based (17 d.f.) 95% confidence limits for controls and cases in Figure 3.4. As seen in the previous two life tables, cases show a lower rate of first screening DRE than controls. The confidence limits indicate possible statistical significance, but this is informal. Formal tests are performed in the next section.

3.3 Stratified Mantel-Haenszel odds ratio

The next step in the DRE study is to reduce the stratified life table into a single overall measure of effect. This is accomplished with the Mantel-Haenszel odds ratio.

3.3.1 Background

Suppose there are data that have been stratified into k strata, with a 2 by 2 table for each stratum cross-classifying a dichotomous outcome (e.g., a disease) with a dichotomous explanatory variable (e.g., a prophylactic).

Table 3.4 Example 2 by 2 table for stratum j

Disease / Prophylactic	Prophylactic + (X=1)	Prophylactic – (X=0)	Total
Disease + (Y=1)	n_{11j}	n_{12j}	n_{1+j}
Disease – (Y=0)	n_{21j}	n_{22j}	n_{2+j}
Total	n_{+1j}	n_{+2j}	n_{++j}

Table 3.4 illustrates this example for stratum j. The odds ratio for the relative odds of disease (Y=1) between prophylactic users (X=1) and nonusers (X=0) is

$$\theta = \frac{P(Y=1 \mid X=1)/(1-P(Y=1 \mid X=1))}{P(Y=1 \mid X=0)/(1-P(Y=1 \mid X=0))}$$

The odds ratio estimator for the stratum j 2 by 2 table is developed as

$$\hat{\theta}=\frac{\hat{P}(Y=1\mid X=1)/\left(1-\hat{P}(Y=1\mid X=1)\right)}{\hat{P}(Y=1\mid X=0)/\left(1-\hat{P}(Y=1\mid X=0)\right)}=\frac{\left(n_{11j}/n_{+1j}\right)/\left(n_{21j}/n_{+1j}\right)}{\left(n_{12j}/n_{+2j}\right)/\left(n_{22j}/n_{+2j}\right)}=\frac{\left(n_{11j}n_{22j}\right)}{\left(n_{12j}n_{21j}\right)}$$

An important property of the odds ratio is symmetry, that is, the relative odds of the higher level of one variable between the two levels of the other is the same regardless of which variable is viewed as the outcome. To see this, consider:

$$\begin{aligned}
\theta_{X\to Y} &= \frac{P(Y=1\mid X=1)/\left(1-P(Y=1\mid X=1)\right)}{P(Y=1\mid X=0)/\left(1-P(Y=1\mid X=0)\right)}\\
&= \frac{P(Y=1\mid X=1)/P(Y=0\mid X=1)}{P(Y=1\mid X=0)/P(Y=0\mid X=0)}\\
&= \frac{\left[P(X=1\mid Y=1)\dfrac{P(Y=1)}{P(X=1)}\right]\Big/\left[P(X=1\mid Y=0)\dfrac{P(Y=0)}{P(X=1)}\right]}{\left[P(X=0\mid Y=1)\dfrac{P(Y=1)}{P(X=0)}\right]\Big/\left[P(X=0\mid Y=0)\dfrac{P(Y=0)}{P(X=0)}\right]}\\
&= \frac{P(X=1\mid Y=1)/P(X=0\mid Y=1)}{P(X=1\mid Y=0)/P(X=0\mid Y=0)}\\
&= \frac{P(X=1\mid Y=1)/\left(1-P(X=1\mid Y=1)\right)}{P(X=1\mid Y=0)/\left(1-P(X=1\mid Y=0)\right)}\\
&= \theta_{Y\to X}
\end{aligned}$$

This property will be important to the later stages of the analysis.

The asymptotic variance of the log odds ratio is derived in Agresti [12] using the delta method to be:

$$V\hat{a}r\left(\log\left(\hat{\theta}\right)\right)=\left(\frac{1}{n_{11}}+\frac{1}{n_{12}}+\frac{1}{n_{21}}+\frac{1}{n_{22}}\right)$$

This can be used to construct a 95% normal-based confidence interval for the log odds ratio, which translates when the limits are exponentiated into a 95% confidence interval for the odds ratio:

$$\exp\left[\log\left(\hat{\theta}\right)\pm Z_{.025}\sqrt{V\hat{a}r\left(\log\left(\hat{\theta}\right)\right)}\right]$$

The Mantel-Haenszel odds ratio [18] is an estimator for a single common odds ratio across k strata:

$$\hat{\theta}_{MH} = \frac{\sum_{j=1}^{k} \left(n_{11j} n_{22j} / n_{++j} \right)}{\sum_{j=1}^{k} \left(n_{12j} n_{21j} / n_{++j} \right)}$$

This estimator weights larger strata more heavily. The asymptotic variance of the log Mantel-Haenszel odds ratio was developed by Robins et al [22] and given in Agresti [18]:

$$V\hat{a}r\left(\log\left(\hat{\theta}_{MH}\right)\right) = \frac{1}{2R^2} \sum_{j=1}^{k} \frac{1}{n_{++j}} \left(n_{11j} + n_{22j}\right) R_j + \frac{1}{2S^2} \sum_{j=1}^{k} \frac{1}{n_{++j}} \left(n_{12j} + n_{21j}\right) S_j +$$
$$\frac{1}{2RS} \sum_{j=1}^{k} \frac{1}{n_{++j}} \left[\left(n_{11j} + n_{22j}\right) S_j + \left(n_{12j} + n_{21j}\right) R_j \right]$$

where $R_j = n_{11k} n_{22k} / n_{++k}$, $S_j = n_{12k} n_{21k} / n_{++k}$, $R = \sum_{j=1}^{k} R_j$, $S = \sum_{j=1}^{k} S_j$, and $\hat{\theta} = R/S$. The estimator is considered valid under standard as well as sparse asymptotics in which k increases with n. The result depends in part on the assumption of independence between probabilities of exposure in cases and controls of different 2 by 2 tables.

An important question to consider when calculating the Mantel-Haenszel odds ratio is whether there really is one common odds ratio across strata. The first way one might approach this question is by looking at a graph plotting the probabilities in the two treatment groups across strata. If the lines cross or their relationship otherwise differs substantially between strata then homogeneity of the odds ratio may not be a reasonable assumption. The Breslow-Day test [13] for homogeneity of odds ratios across strata can answer the question more formally. The test statistic for the null hypothesis of a single common odds ratio across k strata is:

$$X_{k-1}^{2}=\sum_{j=1}^{k}\frac{\left(n_{j11}-E\left(n_{j11}\mid\hat{\theta}_{MH}\right)\right)^{2}}{Var\left(n_{j11}\mid\hat{\theta}_{MH}\right)}$$

which is referred to a chi-square distribution with k-1 degrees of freedom. This test requires a large enough sample size in each stratum that at least 80% of the expected cell counts are 5 or greater. Note that the existence of a common odds ratio is not a prerequisite for the Mantel-Haenszel odds ratio. If no common odds ratio exists across strata, the Mantel-Haenszel odds ratio still provides an estimate of overall effect; in that case it simply should be understood that the effect is not meant to uniformly represent all members of the population.

3.3.2 Calculating an overall measure of effect from a stratified life table

The method of combining intervals from a stratified (two treatment groups) life table into a single odds ratio to estimate a treatment effect is explained by Kahn and Sempos [17]. The general approach treats the intervals in the life table as different strata, despite the violation of independence presented by the fact that all the subjects in any interval certainly occupied all previous intervals. Mantel [15] provided the theoretical justification for treating life table intervals as separate strata in order to calculate an overall test of significance for the difference between survivor functions. He used the construction of orthogonal (independent) variables to show that the Mantel-Haenszel chi-square (1 d.f.) test statistic for testing the overall significance of a set of 2 by 2 tables could also be used to compare life table curves:

$$\chi_{1}^{2}=\frac{\left[\left|\sum_{j=1}^{k}D_{j}^{(1)}-E\left(\sum_{j=1}^{k}D_{j}^{(1)}\right)\right|-\frac{1}{2}\right]^{2}}{\sum_{j=1}^{k}Var\left(D_{j}^{(1)}\right)}$$

where the k strata are the life table intervals and

$$E\left(D_j^{(1)}\right) = \frac{N_j^{(1)}\left(D_j^{(0)} + D_j^{(1)}\right)}{N_j^{(0)} + N_j^{(1)}} \text{ and}$$

$$Var\left(D_j^{(1)}\right) = \left(D_j^{(0)} + D_j^{(1)}\right)\frac{N_j^{(1)}N_j^{(0)}}{\left(N_j^{(0)} + N_j^{(1)}\right)^2}\frac{N_j^{(0)} + N_j^{(1)} - D_j^{(0)} - D_j^{(1)}}{N_j^{(0)} + N_j^{(1)} - 1}$$

are the expectation and variance of a hypergeometric random variable obtained under the null hypothesis and fixed table margins.

The asymptotic variance of the log Mantel-Haenszel odds ratio is no longer valid because of the dependence between tables. Instead, confidence limits for the Mantel-Haenszel odds ratio can be obtained from Miettinen's test-based limits as explained by Kahn and Sempos [11]. The idea is to consider that the Mantel-Haenszel chi-square statistic above could alternatively have been obtained by:

$$\chi_1^2 = \frac{\left[\log\left(\hat{\theta}_{MH}\right) - E\left(\log\left(\hat{\theta}_{MH}\right)\right)\right]^2}{Var\left(\log\left(\hat{\theta}_{MH}\right)\right)} \underset{H_0}{=} \frac{\left[\log\left(\hat{\theta}_{MH}\right) - 0\right]^2}{Var\left(\log\left(\hat{\theta}_{MH}\right)\right)}$$

This equation can be solved for the variance of the log Mantel-Haenszel odds ratio. This variance estimate can be used to obtain confidence limits using the usual standard normal quantiles. Kahn and Sempos point out that although the estimate is strictly valid only under the null hypothesis, they quote a study by Greenland that indicates reasonableness of the estimate except under the most extreme departures from the null (odds ratio>10.0 or <0.10).

The first step in the procedure is to produce a 2 by 2 table from each life table interval, cross-classifying treatment group with failure indicator. Subjects who fail within an interval contribute to the failure cell for their treatment group in that interval's 2 by 2 table, while subjects surviving beyond an interval contribute to the survival cell for their treatment group. Subjects censored in an interval contribute 0.5 to the survival cell, consistent with the effective sample size calculation. Only subjects still at risk at the start of an interval contribute to that interval's 2 by 2 table. Each interval is treated as a separate stratum and the Mantel-Haenszel odds ratio is calculated. If there are additional variables (besides the treatment group) on which the life table was stratified, the combination of those variables with interval number forms the strata. The Mantel-Haenszel odds ratio estimates the overall effect of treatment on the odds of the outcome whose event time was modeled in the life table.

In the following sections, the stratified life tables obtained earlier will be analyzed with the Mantel-Haenszel odds ratio. This will estimate the effect of being a case on the odds of having a screening DRE. However, since odds ratios are symmetric, it also estimates the effect of having a screening DRE on the odds of caseness; this is the study objective. The statistical significance of overall differences will be assessed with the Mantel-Haenszel chi-square test and test-based confidence limits will be produced.

3.3.3 Calculating an overall effect from the life table stratified only by case/control status

First we calculate an overall effect from the life table stratified only by case/control status. This life table was shown in Table 3.3. There are 4 intervals and therefore 4 2 by 2 tables to be constructed from the life table, each cross-classifying case/control status with screening DRE. Table 3.5 shows the 2 by 2 table created for the interval [4,6 years) from the stratified life table shown in Table 3.3. The individual odds ratio for this 2 by 2 table is 0.663 with 95% confidence interval (0.318, 1.379).

Table 3.5 The 2 by 2 table for the interval [4,6 years) from the stratified life table shown in Table 3.3

Case/control status	Screening DRE in the interval	No screening DRE by the end of the interval	Total at risk (effective sample size)
Controls	27	91.5	118.5
Cases	13	66.5	79.5
Total	40	158	198

Before calculating the Mantel-Haenszel odds ratio, we might consider whether there is evidence against the existence of a common odds ratio. Figure 3.2 shows that the survivor functions do not cross, and the gap does not vary much except out at 12 years where the curves meet. The sample size in this life table is sufficient to use the Breslow-Day test for homogeneity of odds ratios, so we can test more formally. The chi-square test statistic is 2.93 on 3 degrees of freedom (p-value=0.403); there is no evidence against the existence of a common odds ratio.

The Mantel-Haenszel odds ratio is 0.840 with 95% test-based confidence interval (0.617, 1.143). The Mantel-Haenszel chi-square test statistic is 1.232 (p-value=0.267). Unadjusted for matching variables or confounders, there is little statistical evidence of a mild protective effect of having a screening DRE on the odds of developing MPC, and only in terms of the confidence interval leaning towards the protective side of null. The evidence is nowhere near strong enough to reject the null hypothesis at any reasonable alpha level.

3.3.4 Calculating an overall effect from the life table stratified by case/control status and collapsed matching variables age group and region

Next we calculate the Mantel-Haenszel odds ratio from the life table that was adjusted for the matching variables. This stratification produced 6 strata besides case/control status, and with 4 intervals in each stratum, the Mantel-Haenszel odds ratio is based on 24 2 by 2 tables. Table 3.6 shows the distribution of 2 by 2 table sizes.

Table 3.6 Distribution of 2 by 2 table sizes from the matched life table analysis

Effective sample size	Number of tables	Effective sample size	Number of tables	Effective sample size	Number of tables	Effective sample size	Number of tables
8.5	1	31.5	1	53.5	1	79.5	1
14.5	1	34	1	56.5	1	102	1
15	2	35	1	61	1	128	1
15.5	1	37	2	66.5	1	135	1
17	1	38	1	67.5	1		
22.5	1	46	1	69.5	1		

The sample sizes appear to be sufficient to apply the Breslow-Day test for homogeneity of odds ratios. The chi-square test statistic is 24.51 on 23 degrees of freedom (p-value=0.376); there is no evidence against the existence of a common odds ratio.

The Mantel-Haenszel odds ratio is 0.848 with 95% test-based confidence interval (0.621, 1.156). The Mantel-Haenszel chi-square test statistic is 1.088 (p-value=0.297). After adjusting for the matching variables, there is even less evidence of a mild protective effect of having a screening DRE on the odds of developing MPC.

3.3.5 Calculating an overall effect from the life table fully stratified by case/control status, the collapsed matching variables age group and region, and 3-level propensity score

Finally, we calculate the Mantel-Haenszel odds ratio from the life table fully stratified by case/control status, the collapsed matching variables age group and region, and 3-level propensity score to adjust for confounders. This stratification produced 18 strata besides case/control status, and with 4 intervals in each stratum, the Mantel-Haenszel odds ratio is based on 72 2 by 2 tables. Table 3.7 shows the distribution of 2 by 2 table sizes.

Table 3.7 Distribution of 2 by 2 table sizes from the fully stratified life table analysis

Effective sample size	Number of tables	Effective sample size	Number of tables	Effective sample size	Number of tables	Effective sample size	Number of tables
2	1	10.5	1	18	1	26	1
2.5	3	11.5	2	20	1	30	1
3.5	2	12	3	21	1	30.5	1
4.5	2	12.5	3	21.5	1	35.5	1
5	3	13	3	22	2	36	2
5.5	2	13.5	1	22.5	2	40.5	1
6	3	14	3	23	1	42.5	1
7	2	14.5	2	23.5	2	43.5	1
7.5	4	16	3	24	2	48.5	1
9.5	2	17	1	24.5	2	52	1

Table sizes range from 2 to 52 (effective sample size). This degree of stratification has created sparseness: many tables have empty columns, and overall table sizes are generally small.

Fortunately, the Mantel-Haenszel odds ratio is designed to combine sparse data such as these into a stable estimate of a common effect.

Is there a common odds ratio? Considering the effect across intervals, the mean survivor function curves in Figure 3.4 tell a similar story to when the life table was stratified only by case/control status; recall that in that situation no evidence was found against a common effect. However, in this case, the stratification of the Mantel-Haenszel odds ratio is across 72 2 by 2 tables, with strata formed not only on interval, but also on the collapsed matching variables and the 3-level propensity score. Visually, there is no realistic way to ascertain homogeneity of odds ratios, and the Breslow-Day test cannot be applied due to the small table sizes. Instead, we perform another stratified life table analysis, only on 3-level propensity score without stratifying by matching variables, solely for the purpose of helping us assess homogeneity of odds ratios in the fully stratified analysis. The simpler analysis does produce sufficient table sizes to apply the Breslow-Day test. The chi-square test statistic is 6.98 on 11 degrees of freedom (p-value=0.801); there is no evidence against the existence of a common odds ratio across propensity score levels. We proceed under the assumption that stratifying on the matching variables and propensity score together will not alter the homogeneity of odds ratios found when stratifying on either alone, and calculate the Mantel-Haenszel odds ratio.

The Mantel-Haenszel odds ratio is 0.861 with 95% test-based confidence interval (0.611, 1.215). The Mantel-Haenszel chi-square test statistic is 0.722 (p-value=0.396). After adjusting for the matching variables and confounders together, the already weak evidence of a mild protective effect of having a screening DRE on the odds of developing MPC is even weaker.

3.4 Results

The Mantel-Haenszel odds ratio describing the overall effect of having a screening DRE on the odds of developing MPC, after adjusting for the matched case-control design as well as several potentially confounding variables, was found to be 0.861 with 95% confidence interval (0.611, 1.215). This is based on a Mantel-Haenszel chi-square test p-value of 0.396. Like most of the other studies on DRE efficacy before this one (with the exception of the Jacobsen et al paper that was biased) the estimate is mildly protective but not statistically significant.

This study finds only weak evidence of a mild protective effect of screening with digital rectal exam in the prevention of metastatic prostate cancer.

4 CHAPTER FOUR: SIMULATION

The methodology outlined in this report is novel and therefore demands scrutiny. In the previous chapter, we endeavored to justify each step from a theoretical standpoint. In this chapter, we assess the validity of the method using Monte Carlo simulation programmed in SAS 9.1.3.

4.1 Building the population

The first step in studying this method by simulation is to construct an artificial population from which a case-control sample can be drawn, matching the DRE study population as closely as possible.

To determine the appropriate size of the simulated population, we must estimate the proportion of the actual population in-scope for the DRE study as those data were collected (between August 1, 1999 and May 31, 2002, centered at the beginning of 2001). This is defined as male, aged 35-74 in 1990, which describes 19.8% of the 2000/2001 Ontario population. (Note that age is defined in 1990. This is to provide a common frame of reference; since the data were collected over 2 years, age at time of data collection would not define a coherent population.) The total population in the 6 study counties as of the 2001 Canadian census was 5.19 million, or about 1.0 million in-scope men. The simulated population therefore is set at 1 million.

235 new cases of MPC were found in the DRE study, and this is the number of cases dispersed in the simulated population. Although this number is known to be less than the actual number of people with MPC during the DRE data collection period, the simulation will analyze the estimation of the effect of DRE on applicable MPC according to the criteria for caseness in that study; there is the assumption of generalizability to more general MPC.

We must also consider the distribution of 1990 age. Although the DRE study sample was also matched on region, for simplicity we will match simulated samples only according to age and assume that the results extend to situations of more numerous strata.

Table 4.1 Distribution of 5-year age group (in 1990) in the 2000/2001 DRE study population and sampled cases

1990 Age Group	35-39	40-44	45-49	50-54	55-59	60-64	65-69	70+
% Population	22.41	19.71	15.33	11.01	10.83	9.35	6.08	5.28
% Sample cases	1.28	2.13	7.23	10.64	21.7	22.13	26.81	8.08

Figure 4.1 Distribution of 5-year age group (in 1990) in the 2000/2001 DRE study population and sampled cases

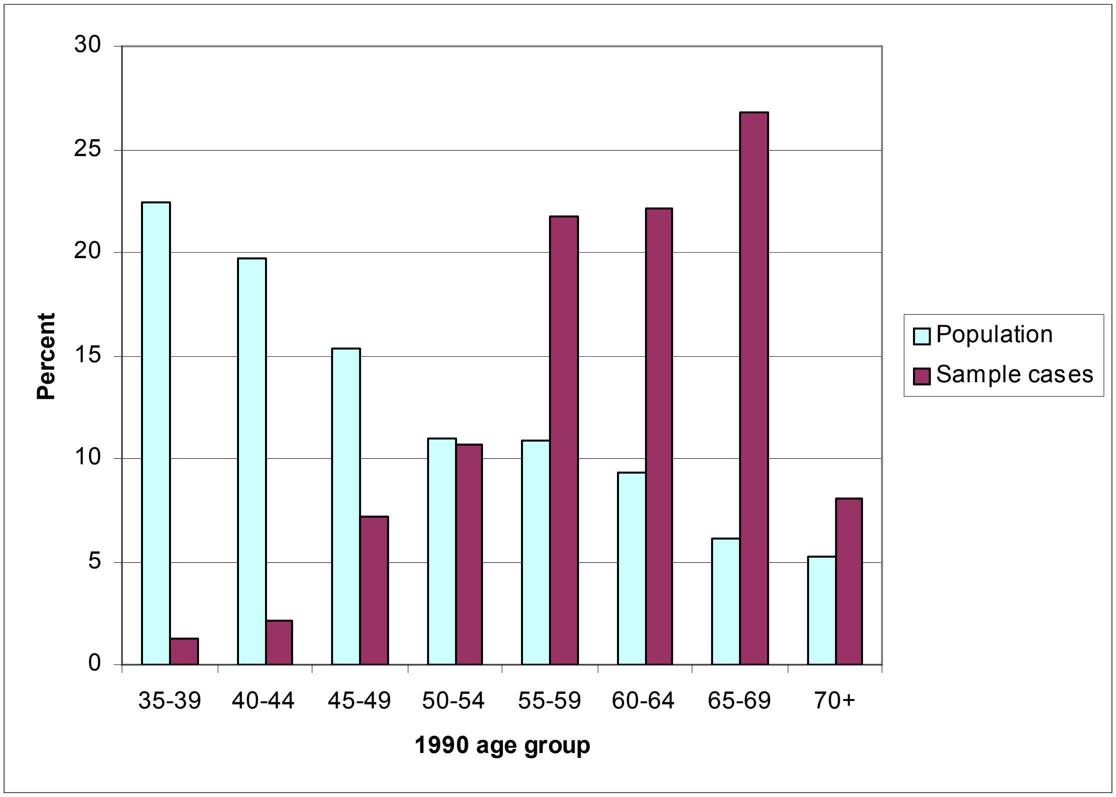


Table 4.1 and Figure 4.1 show the distribution of 5-year age group in 1990 in the 2000/2001 DRE study population and amongst the 235 sampled cases. Together these show a strong association between the cross-sectional risk of MPC and age; prevalence of new cases by age group increases monotonically until the 70+ group when it drops off, possibly because those

who would get MPC are likely to have it and have died before 80 years of age (70 in 1990). The distribution in the DRE study population was estimated using the Ontario portion of the 2000/2001 Canadian Community Health Survey (CCHS), a cross-sectional Statistics Canada survey.

These numbers are sufficient to estimate the appropriate age-specific prevalence of new cases for the simulated population. The number of cases and controls in the simulated population according to age group are obtained as follows for each age group i:

$n_i = p_i N$

$a_i = ap_i^{(1)}$ (cases)

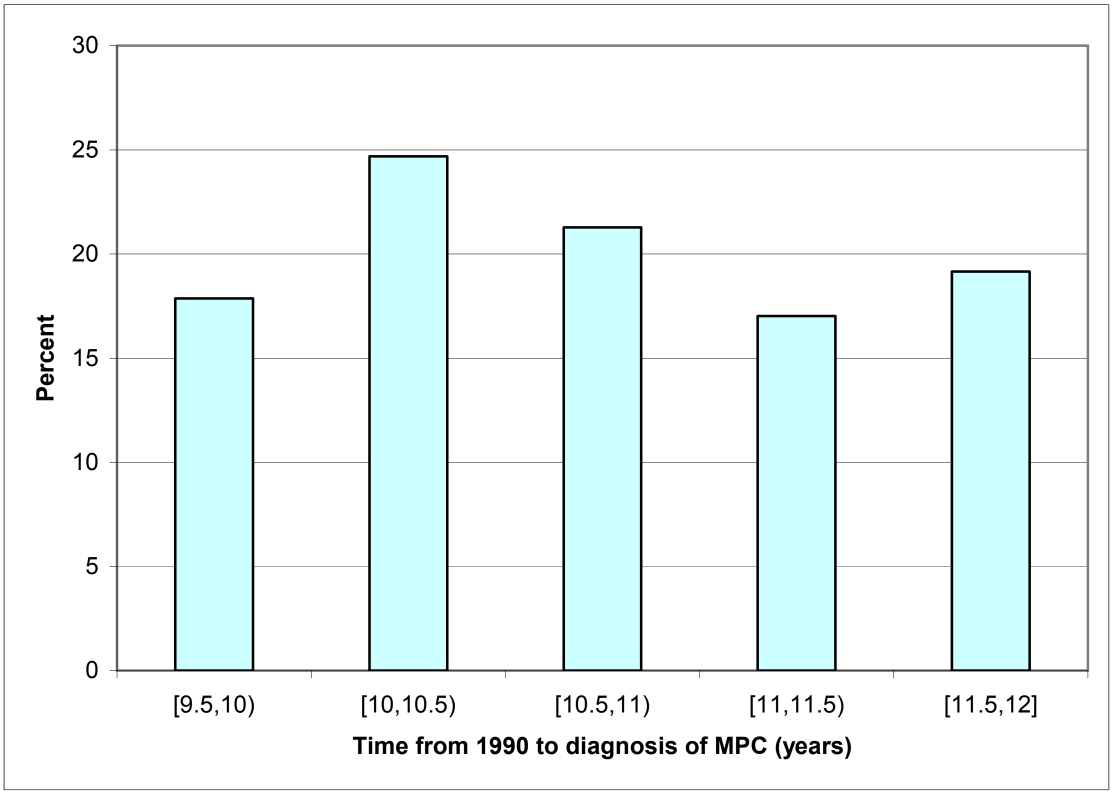
$b_i = n_i - a_i$ (controls)

$r_i = a_i / n_i$ (age-specific prevalence)

where p_i is the proportion of the DRE study population in age group i, $p_i^{(1)}$ is the proportion of sampled cases in age group i, and a=235 cases.

Cases in the simulated population were assigned a date of MPC diagnosis according to a uniform distribution between 10 and 12 years. Figure 4.2 shows that this approximately matches the distribution of MPC diagnosis in the DRE study data.

Figure 4.2 Distribution of time from 1990 to diagnosis of MPC in the DRE study cases (10 subjects within 2 months of 12 years are collapsed into bin [11.5,12])



Event times can take on any distribution with a non-negative support. The survivor function S(t) is the probability of not having the event up to the time t. It equals the integral of the density from 0 to t, or 1 minus the CDF. Two distributions commonly used to describe event times are the exponential and Weibull. The exponential was the first widely recognized lifetime distribution, while the Weibull is thought by some to be the most widely used today [20]. The exponential density is:

$$f(t) = \lambda e^{-\lambda t}$$

while the Weibull density is:

$$f(t) = \lambda\beta(\lambda t)^{\beta-1} e^{\left\{-(\lambda t)^{\beta}\right\}}$$

The exponential is a special case of the Weibull where $\beta = 1$. λ is the hazard rate in the exponential density and the rate parameter in the Weibull. β is the shape parameter in the Weibull. It can be shown that if the exponential distribution fits the data then log(S(t)) should be linearly related to t, and if the Weibull distribution fits the data then log(-log(S(t)) should be linearly related to t.

The exponential regression model is:

$$\ln(\mathrm{T}) = \mathbf{X}\beta + \varepsilon$$

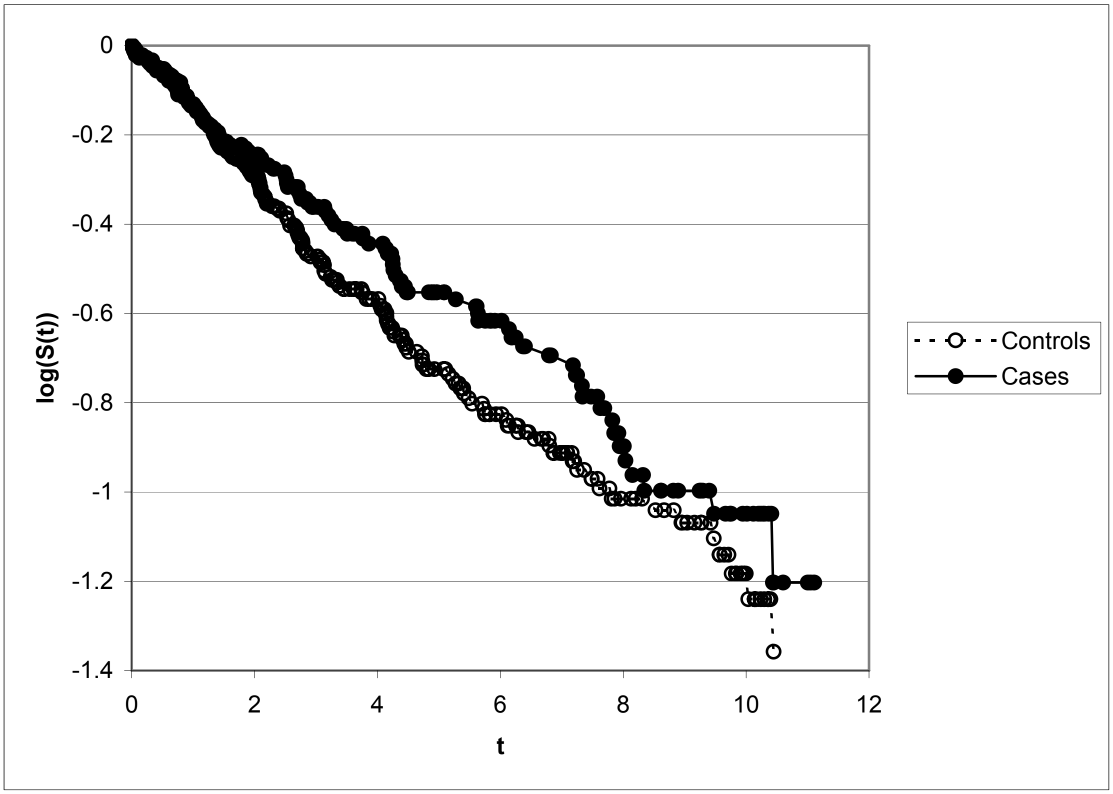
where T is the event time and ε is a random error term distributed according to the standard extreme value distribution [21]. λ is estimated as $\hat{\lambda} = e^{-\mathbf{X}\beta}$. The Weibull regression model is:

$$\ln(\mathrm{T}) = \mathbf{X}\gamma + \sigma\varepsilon$$

where $\beta = 1/\sigma$, and again $\hat{\lambda} = e^{-\mathbf{X}\beta}$ with ε a random error term distributed according to the standard extreme value distribution [21].

Rate of DRE screening in the DRE study data was modeled with both the exponential distribution and the more general Weibull and the former was found to fit the DRE study data equally well (LRT p-value=0.29). Further, Figure 4.3 shows that the plot of the log of the non-parametric Kaplan-Meier survival curve [5] (stratified by case/control status) against time is reasonably straight. This indicates that the exponential distribution fits the time to first screening DRE reasonably well; we will use the exponential distribution to model DRE event times.

Figure 4.3 log(S(t)) versus t from the Kaplan-Meier survivor function for first screening DRE stratified by case/control status



The estimated hazard rate of DRE screening for controls in the DRE study data is 0.135 per year. For the exponential distribution, the rate parameter is related to the probability of DRE over time according to the relationship:

$$P(\text{DRE between 1990 and 1999}) = p = 1 - S(10) = 1 - e^{-10\lambda} \Rightarrow \lambda = \frac{-1}{10}\ln(1-p)$$

Solving this yields a probability of DRE in controls of 0.741 over ten years (the approximate duration of the data collection). Recall that the adjusted odds ratio in the study was 0.861. Letting this describe a ten-year odds ratio, and since odds ratios are symmetric, this implies a probability of DRE in cases of 0.711 between 1990 and 1999 according to the formula:

$$\theta = \frac{p_1/(1-p_1)}{p_0/(1-p_0)} \Rightarrow p_1 = \frac{p_0\theta}{p_0\theta + 1 - p_0}$$

This in turn requires a rate of DRE in cases of 0.124 per year. Each case and control in the simulated population is assigned a random time to first DRE measured from 1990 according to the exponential distribution with the appropriate rate.

Next we consider prostate-related symptoms. 26.2% of controls and 21.3% of cases in the DRE study data had symptoms at baseline. Subjects were assigned symptoms at baseline according to these probabilities. For the remaining subjects, exponential regression was used to estimate rates of symptoms in each third of the 1990 decade. These models fit as well as the Weibull (LRT p-value=0.16, 0.09, 0.78), and plots of log(S(t)) against t were reasonably straight. This produced the following rates of first symptoms onset for controls versus cases in each third of the decade: 0.0593 vs. 0.0588, 0.0815 vs. 0.0797, and 0.102 vs. 0.152. The steady increase from third to third reflects the effect of aging on the rate of symptoms. The more dramatic increase in symptoms in the third third amongst cases was expected; these are largely symptoms of PC. Time to first onset of symptoms was generated randomly for each simulated baseline-asymptomatic subject according to these rates. If the first symptom did not appear in the first third then an event time was generated from 3 and 1/3 years using the rates for the second third. If that first symptom did not appear in the second third then an event time was generated from 6 and 2/3 years using the rates for the third third, and this would be the time of first symptoms onset.

Finally, we generate the reference date. In the DRE study, about 10% of cases who had a first DRE before their reference date were screen-detected by that DRE, which leaves 90% who were detected by other means. Thus for the simulation we must generate a competing diagnostic event. It is reasonable to suppose that time to diagnostic event is related to symptoms. Modeling time to competing diagnostic event as a function of symptoms in the DRE study data, we find that the Weibull model provides a better fit to the data than the exponential (LRT p-value<0.001), and the plot of log(-log(S(t)) against t is reasonably straight. The event parameters are λ=0.0905 and β=3.292 for asymptomatic time giving an expected time of 9.9 years, and λ=0.1433 and

β =1.642 for symptomatic time giving a (shorter) expected time of 6.2 years. Competing diagnostic event times were generated for each simulated case according to the appropriate distribution. If the asymptomatic event time exceeded the time of first symptoms onset, a new event time was generated from first symptoms onset with the symptomatic distribution. Diagnostic event times exceeding the date of MPC diagnosis were set equal to that date. Diagnostic event times were then compared to the first DRE for the simulated case. If the DRE happened before the diagnostic event, then with probability 0.1 the DRE detects the PC and the reference date is set to the earliest of the diagnostic event time and one month after the DRE (to allow time for biopsy). Otherwise, the reference date is set to the diagnostic event time.

4.2 Drawing and analyzing a case-control sample

The next step is to take a case-control sample from the simulated population. First, all 235 cases are selected. Two controls are then randomly selected to match each case on the basis of age group. This sample is analyzed according to the method followed in the DRE study. For simplicity, age group is the sole matching variable. Also, no confounding variables are included in this simulation so the propensity score step is omitted. Thus, life tables are stratified only on age group, instead of age group, region and propensity score. Each analysis produces a Mantel-Haenszel odds ratio and test-based confidence limits.

4.3 Varying populations

To properly study the performance of the method, it is necessary to simulate populations and analyses under a variety of conditions, not just those which match the parameters in the DRE study. Beginning with the population described in Section 4.1, several parameters were varied: the probability of screen-detection in cases given they had a screening test is set to 0.1, 0.5 and 0.9; censoring versus no censoring (no symptoms); small (235 cases) versus larger (470 cases) case groups; and moderate true protective effect (ten-year OR=0.861) versus a strong effect

(0.500). In addition, more life table intervals (3 intervals with endpoints at 0, 3, 6 and 12 years) in the analysis was compared to fewer intervals (2 intervals with endpoints at 0, 6 and 12 years). This describes 24 populations in total, each analyzed 2 different ways.

Each combination of population parameters describes "true" probabilities and rates. The actual effect observed in a given simulated population may appear quite different, however. Even if the true effect of DRE within the decade is 0.861 for example, in a given population one might observe an odds ratio of 1.1 or something else due to chance. The "true" effect actually describes a comparison of hazard rates and event probabilities between cases and controls. Any given group of cases and controls can exhibit observed DRE counts quite contrary to the expected counts; in one population cases may have more DREs, while in another controls will have more DREs. It is only through replication that the law of large numbers begins to ensure that "true" effects are seen in practice. Thus to assess the methodology, each combination of parameters should be used to generate many simulated populations and samples. From each combination of population parameters, 100 simulated populations are generated, and from each simulated population, 1 case-control sample is drawn and analyzed with more (3) and fewer (2) life table intervals. This totals 2,400 populations and case-control samples, and 4,800 analyses.

4.4 Results

The population odds ratio for each simulated population is calculated as the stratified Mantel-Haenszel case-control odds ratio on the full simulated population (PMHOR). In general, since observation time begins in 1990 and ends at censoring times that are not the same between cases and controls, PMHOR is not expected to equal the true odds ratio. It is desirable to confirm that in the absence of censoring the odds ratio would be right and therefore the relative driving force behind DRE processes for cases versus controls correctly warrants being described with the "true" ten-year odds ratio. The uncensored population odds ratio is calculated as the stratified Mantel-Haenszel case-control odds ratio on the full simulated population (UCPMHOR), counting all DREs between 1990 and 1999 regardless of the censoring processes, symptoms and competing diagnostic events. UCPMHOR is expected to equal the true odds ratio. This is confirmed in each parameter combination by taking and back-transforming normal-based confidence limits about log-UCPMHOR using the 100 replicates, and checking that the limits for UCPMHOR contain the true odds ratio.

The probability of a given screening test detecting the disease in a case (P(SD|DRE)) does not affect the odds ratios at the population level. This is because: a) cases and controls are not matched in the general population, so controls are unaffected by the earlier ending of observation time in cases; and b) the cases' DRE status is unaffected since all subjects who are possibly observed for a shorter time due to screen-detection will still have their DRE recorded just prior to detection. There are 8 combinations of the remaining population parameters that produce different population odds ratios. These are summarized inTable 4.2.

Table 4.2 Mean simulated population odds ratios for the 8 population groups with distinct odds ratios, based on 300 replicates each, with 95% confidence intervals

Pop #s	True OR	Cases	Censor-ing	PMHOR	95% CI PMHOR	UC-PMHOR	95% CI UC-PMHOR
1-3	0.861	235	Yes	0.901	(0.881,0.906)	0.858	(0.834,0.863)
4-6	0.861	235	No	0.639	(0.623,0.643)	0.871	(0.846,0.876)
7-9	0.861	470	Yes	0.908	(0.896,0.914)	0.871	(0.857,0.876)
10-12	0.861	470	No	0.643	(0.634,0.647)	0.865	(0.852,0.871)
13-15	0.500	235	Yes	0.642	(0.626,0.646)	0.501	(0.488,0.504)
16-18	0.500	235	No	0.396	(0.387,0.398)	0.508	(0.496,0.511)
19-21	0.500	470	Yes	0.643	(0.632,0.647)	0.499	(0.492,0.502)
22-24	0.500	470	No	0.391	(0.385,0.393)	0.504	(0.497,0.508)

As expected, PMHOR does not equal (the 95% confidence intervals do not contain) the true odds ratio. Uncensored (asymptomatic) populations tend to underestimate it, while censored populations tend to overestimate it. The negative bias in PMHOR in the uncensored combinations is expected; controls are observed for the full ten years, while cases are only observed until the earliest of ten years and their competing diagnostic event, less than ten years on average. The positive bias in PMHOR in the censored combinations is because cases are observed for a longer time; controls are censored at higher rates in the first 2/3 of the decade, and 26.2% (compared to 21.3% of cases) are already symptomatic at baseline and have 0 person-time. The confidence limits around UCPMHOR confirm that in the absence of censoring, the true ten-year odds ratio is accurate in each population; the true odds ratio falls inside the confidence limits in all 8 factor combinations.

A single case-control sample is drawn from each population and analyzed twice (with more and fewer intervals), each analysis producing a life table Mantel-Haenszel odds ratio (LTMHOR). For comparison, the usual (age) stratified Mantel-Haenszel case-control odds ratio is also produced (MHOR), which as explained in the introduction is expected to be biased when there are screen-detected cases. But censoring will also (unfairly, perhaps) bias the MHOR, due to the fact that controls and cases are censored at different rates and therefore have unequal opportunities to be screened from 1990 forward. Therefore another comparison odds ratio is also calculated, the even-time stratified Mantel-Haenszel case-control odds ratio (ETMHOR). This is obtained by taking the earliest censoring time in each matched set of three subjects and censoring all three at that time. There is some loss of person-time in this method. Test-based confidence limits are calculated for LTMHOR, and asymptotic limits are calculated for MHOR and ETMHOR.

Conditional logistic regression (conditioning on the un-collapsed matching variable age group) was also used to analyze each sample as an alternative to MHOR as well as an even-time alternative to ETMHOR. The mean results were extremely close in all combinations to MHOR and ETMHOR (point estimates and confidence intervals equal to within 0.005 in all but one combination, which was equal to within 0.006). Therefore for brevity the conditional logistic regression results are omitted, and MHOR and ETMHOR are understood to represent both themselves and their conditional logistic regression counterparts.

4.4.1 Point estimates

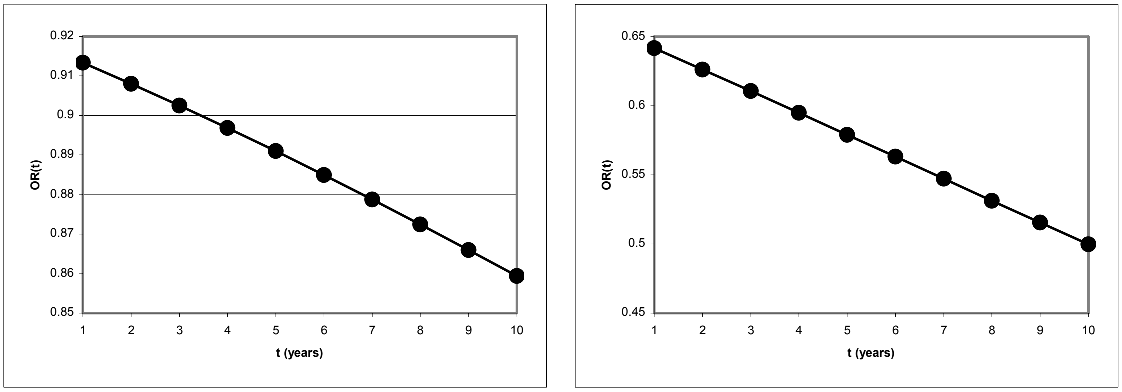
The mean sample odds ratios (with LTMHOR based on 3 life table intervals) are shown in Table 4.3 for all 24 factor combinations along with 95% confidence limits for the LTMHOR, calculated by taking and back-transforming normal-based confidence limits about log-LTMHOR using the 100 replicates. Between MHOR, ETMHOR and LTMHOR, the best estimator (closest mean estimate to the true odds ratio) is in bold in each row.

Table 4.3 Mean sample odds ratios based on 100 replicates in each combination of parameters, life table odds ratio calculated with more (3) intervals; the closest mean estimate to the true odds ratio is in bold

Pop #	True OR	Obs Time	Corr True OR	Cases	Censoring	P(SD|DRE)	MHOR	ET-MHOR	LT-MHOR	95% CI LT-MHOR
1	0.861	4.53	0.895	235	Yes	0.1	1.191	0.964	**0.960**	(0.919,0.977)
2	0.861	3.99	0.898	235	Yes	0.5	1.328	1.088	**0.966**	(0.925,0.984)
3	0.861	3.44	0.901	235	Yes	0.9	1.543	1.242	**0.992**	(0.948,1.011)
4	0.861	8.65	0.870	235	No	0.1	0.984	0.984	**0.923**	(0.895,0.939)
5	0.861	7.17	0.879	235	No	0.5	1.383	1.383	**0.924**	(0.894,0.939)
6	0.861	5.75	0.888	235	No	0.9	2.009	2.009	**0.976**	(0.940,0.994)
7	0.861	4.55	0.895	470	Yes	0.1	1.174	0.963	**0.952**	(0.930,0.966)
8	0.861	3.98	0.898	470	Yes	0.5	1.348	1.090	**0.973**	(0.948,0.988)
9	0.861	3.44	0.901	470	Yes	0.9	1.548	1.218	**0.993**	(0.967,1.008)
10	0.861	8.67	0.870	470	No	0.1	0.977	0.977	**0.915**	(0.900,0.926)
11	0.861	7.20	0.879	470	No	0.5	1.409	1.409	**0.942**	(0.924,0.954)
12	0.861	5.75	0.888	470	No	0.9	1.956	1.956	**0.962**	(0.941,0.976)
13	0.500	4.58	0.586	235	Yes	0.1	0.830	0.711	**0.660**	(0.630,0.673)
14	0.500	4.14	0.593	235	Yes	0.5	0.892	0.761	**0.649**	(0.615,0.662)
15	0.500	3.72	0.600	235	Yes	0.9	1.044	0.901	**0.698**	(0.668,0.711)
16	0.500	8.73	0.520	235	No	0.1	**0.579**	**0.579**	0.618	(0.599,0.629)
17	0.500	7.58	0.538	235	No	0.5	0.792	0.792	**0.635**	(0.613,0.646)
18	0.500	6.44	0.556	235	No	0.9	1.017	1.017	**0.645**	(0.624,0.656)
19	0.500	4.56	0.586	470	Yes	0.1	0.839	0.713	**0.665**	(0.647,0.675)
20	0.500	4.13	0.593	470	Yes	0.5	0.913	0.787	**0.664**	(0.645,0.675)
21	0.500	3.73	0.599	470	Yes	0.9	1.001	0.855	**0.668**	(0.649,0.679)
22	0.500	8.72	0.520	470	No	0.1	**0.584**	**0.584**	0.619	(0.607,0.627)
23	0.500	7.59	0.538	470	No	0.5	0.777	0.777	**0.632**	(0.618,0.640)
24	0.500	6.45	0.556	470	No	0.9	1.011	1.011	**0.643**	(0.630,0.652)

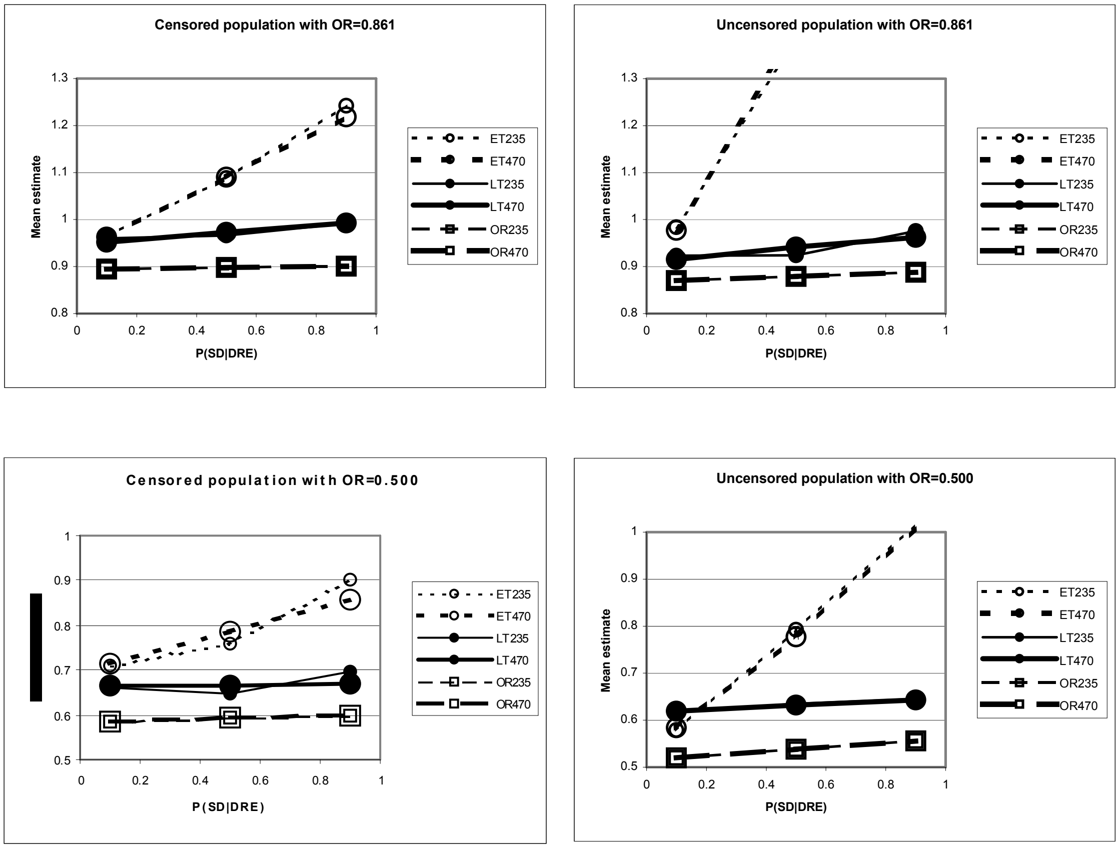
Besides the factors, also shown are mean observation time and a time-corrected true odds ratio. Mean observation time is the average time from 1990 to the earliest of the reference date and the censoring date in the samples drawn from each combination. The time-corrected true odds ratio reflects the fact that when subjects are observed for less than ten years the odds ratio for an event changes, assuming rates of DRE remain constant in cases and controls. This is because the relative event probability between two groups with exponential event times and different hazard rates varies with observation time. Figure 4.4 illustrates the phenomenon; it shows that odds ratio is nearly linearly related to observation time between 1 and 10 years.

Figure 4.4 Odds ratio as a function of observation time, when true ten-year odds ratio is 0.861 based on a hazard rate for DRE screening of 0.135 per year in controls and 0.124 per year in cases (left); and when true ten-year odds ratio is 0.500 based on a hazard rate for DRE screening of 0.135 per year in controls and 0.0887 per year in cases (right)



In 22 of 24 combinations (92%), Table 4.3 shows LTMHOR to have a closer mean estimate to the true odds ratio (less biased) than MHOR or ETMHOR. Populations 16 and 22, both with true ten-year odds ratios 0.500 (strong) and P(SD|DRE)=0.1, have mean MHOR and ETMHOR tied (since in uncensored populations samples matched on observation time already have "even time") and superior to LTMHOR. But these factor combinations may not even be realistic, since if a screening test is effective enough for so strong an odds ratio, it would most likely need to detect the disease in cases with greater than 10% probability.

Figure 4.5 Mean sample odds ratios based on 100 replicates in each combination of parameters, life table odds ratio calculated with more (3) intervals; left figures are from censored populations while right are uncensored (asymptomatic); top figures have true ten-year OR=0.861 while lower ones have true ten-year OR=0.500

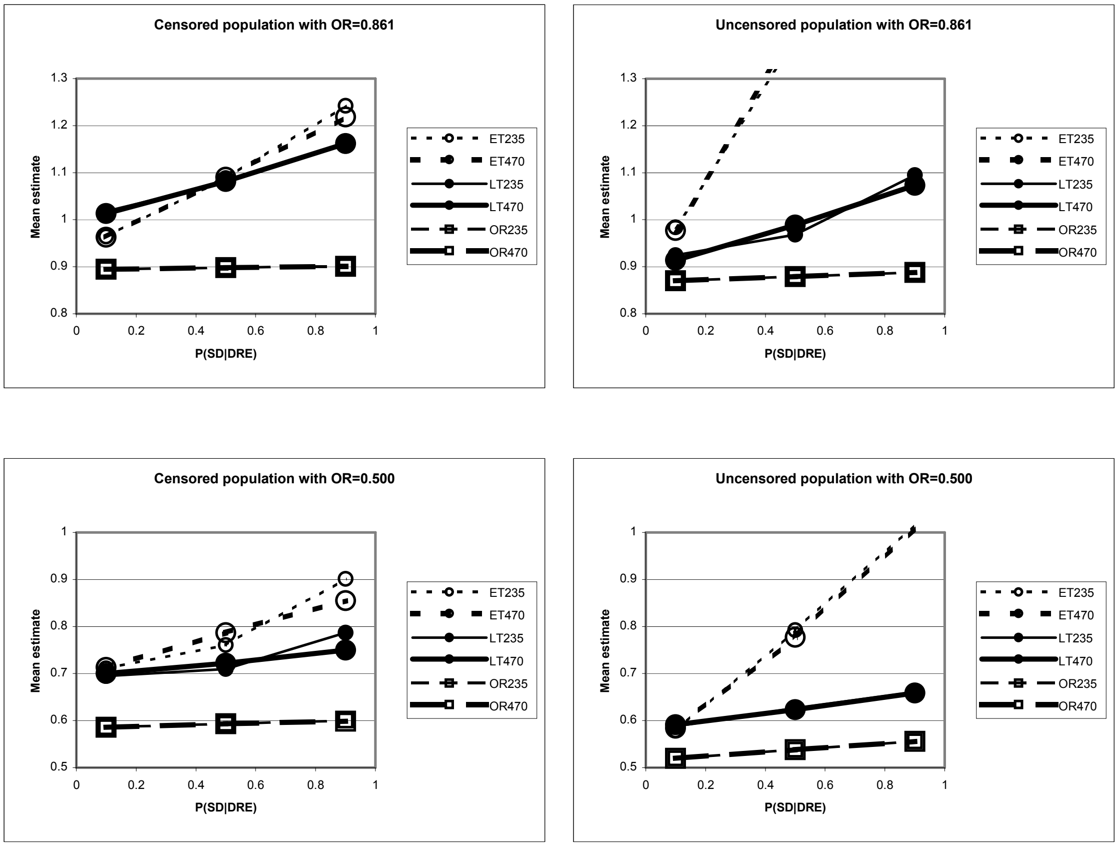


The results are summarized graphically in Figure 4.5. ET235 and ET470 are the mean ETMHOR estimates from populations with 235 and 470 cases, respectively. LT235 and LT470 are the mean LTMHOR estimates. OR235 and OR470 are the time-adjusted true odds ratios based on the average observation time in the samples. MHOR is not compared here because it is always equal to or worse than ETMHOR. The graph shows that LTMHOR is generally superior to ETMHOR. Only for a strong effect (OR=0.500) with P(SD|DRE)=0.1 is LTMHOR worse (more biased) than ETMHOR, which as mentioned before is an unrealistic combination. The plots show that as P(SD|DRE) increases, the bias in ETMHOR increases steeply. LTMHOR is also positively biased, but its bias's dependence on P(SD|DRE) is minimal. For analyzing powerful screening tests (represented on the plot as P(SD|DRE)=0.9) LTMHOR is highly superior to ETMHOR. Sample size (number of cases) appears to have no systematic effect on the expected value of LTMHOR or ETMHOR. Censoring processes slightly increase the bias in LTMHOR, uniformly across P(SD|DRE). For ETMHOR, censoring has a mitigating effect, worsening relatively good estimates at the low end of P(SD|DRE) (0.1) and improving terribly biased estimates at the high end (0.9). The improvement in ETMHOR for higher P(SD|DRE) results from censoring reducing the apparent dependence between the observation time for matched sets and the DRE events in their cases, by randomly omitting a subset of the DREs in screen-detected cases. This is not enough to eliminate the greater bias in ETMHOR over that of LTMHOR, however. As explained earlier, the "true" odds ratio is based on the ten-year event probabilities of DRE screening. That the time-adjusted true odds ratios increases with P(SD|DRE) is no surprise; controls are matched to cases on observation time, and detecting more cases with screening results in less observation time for everyone. Comparing the time-adjusted odds ratios to LTMHOR, most of the increase in the LTMHOR's bias with increasing P(SD|DRE) is eliminated. Unfortunately there remains some baseline residual bias even above the time-adjusted odds ratio, but this is small compared with ETMHOR, and conservative.

Table 4.4 Mean sample odds ratios based on 100 replicates in each combination of parameters, life table odds ratio calculated with fewer (2) intervals; the closest mean estimate to the true odds ratio is in bold

Pop #	True OR	Obs Time	Corr True OR	Cases	Censoring	P(SD|DRE)	MHOR	ET-MHOR	LT-MHOR	95% CI LT-MHOR
1	0.861	4.53	0.895	235	Yes	0.1	1.191	**0.964**	1.018	(0.973,1.038)
2	0.861	3.99	0.898	235	Yes	0.5	1.328	1.088	**1.079**	(1.032,1.099)
3	0.861	3.44	0.901	235	Yes	0.9	1.543	1.242	**1.163**	(1.109,1.186)
4	0.861	8.65	0.870	235	No	0.1	0.984	0.984	**0.925**	(0.890,0.942)
5	0.861	7.17	0.879	235	No	0.5	1.383	1.383	**0.968**	(0.930,0.986)
6	0.861	5.75	0.888	235	No	0.9	2.009	2.009	**1.095**	(1.044,1.116)
7	0.861	4.55	0.895	470	Yes	0.1	1.174	**0.963**	1.013	(0.988,1.029)
8	0.861	3.98	0.898	470	Yes	0.5	1.348	1.090	**1.081**	(1.051,1.098)
9	0.861	3.44	0.901	470	Yes	0.9	1.548	1.218	**1.162**	(1.128,1.181)
10	0.861	8.67	0.870	470	No	0.1	0.977	0.977	**0.914**	(0.896,0.926)
11	0.861	7.20	0.879	470	No	0.5	1.409	1.409	**0.989**	(0.966,1.004)
12	0.861	5.75	0.888	470	No	0.9	1.956	1.956	**1.073**	(1.044,1.090)
13	0.500	4.58	0.586	235	Yes	0.1	0.830	0.711	**0.695**	(0.661,0.708)
14	0.500	4.14	0.593	235	Yes	0.5	0.892	0.761	**0.710**	(0.670,0.724)
15	0.500	3.72	0.600	235	Yes	0.9	1.044	0.901	**0.787**	(0.750,0.803)
16	0.500	8.73	0.520	235	No	0.1	**0.579**	**0.579**	0.588	(0.567,0.599)
17	0.500	7.58	0.538	235	No	0.5	0.792	0.792	**0.627**	(0.602,0.639)
18	0.500	6.44	0.556	235	No	0.9	1.017	1.017	**0.661**	(0.635,0.673)
19	0.500	4.56	0.586	470	Yes	0.1	0.839	0.713	**0.700**	(0.680,0.711)
20	0.500	4.13	0.593	470	Yes	0.5	0.913	0.787	**0.722**	(0.700,0.735)
21	0.500	3.73	0.599	470	Yes	0.9	1.001	0.855	**0.750**	(0.726,0.764)
22	0.500	8.72	0.520	470	No	0.1	**0.584**	**0.584**	0.591	(0.578,0.599)
23	0.500	7.59	0.538	470	No	0.5	0.777	0.777	**0.623**	(0.608,0.632)
24	0.500	6.45	0.556	470	No	0.9	1.011	1.011	**0.659**	(0.643,0.668)

Figure 4.6 Mean sample odds ratios based on 100 replicates in each combination of parameters, life table odds ratio calculated with fewer (2) intervals; left figures are from censored populations while right are uncensored (asymptomatic); top figures have true ten-year OR=0.861 while lower ones have true ten-year OR=0.500



Next we consider analyses based on 2 life table intervals. Table 4.4 and Figure 4.6 summarize the results. Censoring has a greater impact on the bias than it did with 3 intervals. In the (unrealistic) combination with true odds ratio 0.500 and P(SD|DRE)=0.1, fewer life table intervals actually improved LTMHOR under no censoring. However, in general LTMHOR is more biased when based on fewer intervals. But even so, LTMHOR is generally superior to ETMHOR, with the closer mean estimate in 20/24 combinations (83%).

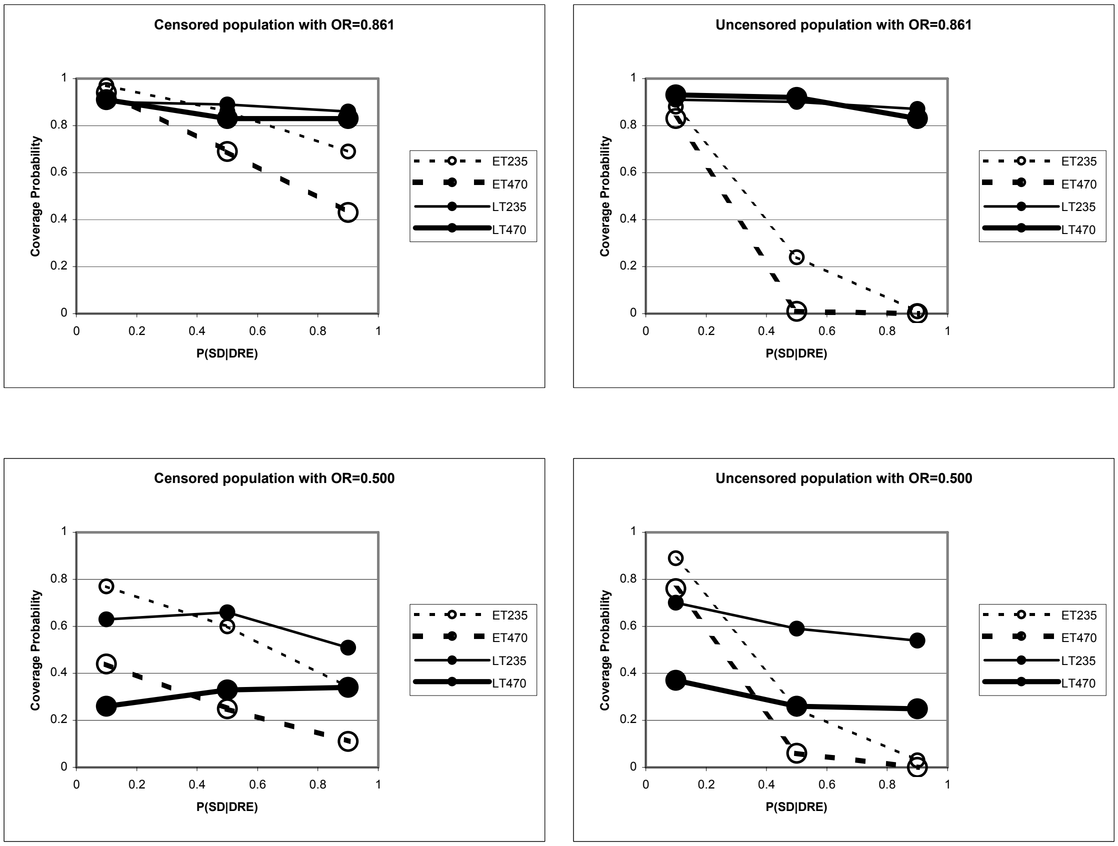
4.4.2 Coverage probabilities

Finally we consider coverage probabilities (CPs) of the 95% confidence intervals for sample odds ratios. CP is estimated as the average of an indicator variable for the confidence limits containing the true odds ratio.

Table 4.5 Coverage probabilities of 95% confidence intervals for sample odds ratios based on 100 replicates in each combination of parameters, life table odds ratio calculated with more (3) intervals; the closest CP to 0.95 is in bold

Pop #	True OR	Obs Time	Corr True OR	Cases	Cen-sor-ing	P(SD|DRE)	Covg Prob MHOR	Covg Prob ET-MHOR	Covg Prob LT-MHOR	95% CI Covg Prob LT-MHOR
1	0.861	4.53	0.895	235	Yes	0.1	0.51	**0.97**	0.90	(0.824,0.936)
2	0.861	3.99	0.898	235	Yes	0.5	0.31	0.86	**0.89**	(0.824,0.936)
3	0.861	3.44	0.901	235	Yes	0.9	0.08	0.69	**0.86**	(0.800,0.921)
4	0.861	8.65	0.870	235	No	0.1	0.88	0.88	**0.91**	(0.848,0.951)
5	0.861	7.17	0.879	235	No	0.5	0.24	0.24	**0.90**	(0.836,0.944)
6	0.861	5.75	0.888	235	No	0.9	0.01	0.01	**0.87**	(0.800,0.921)
7	0.861	4.55	0.895	470	Yes	0.1	0.24	**0.94**	0.91	(0.848,0.951)
8	0.861	3.98	0.898	470	Yes	0.5	0.02	0.69	**0.83**	(0.753,0.890)
9	0.861	3.44	0.901	470	Yes	0.9	0.00	0.43	**0.83**	(0.753,0.890)
10	0.861	8.67	0.870	470	No	0.1	0.83	0.83	**0.93**	(0.874,0.965)
11	0.861	7.20	0.879	470	No	0.5	0.01	0.01	**0.92**	(0.861,0.958)
12	0.861	5.75	0.888	470	No	0.9	0.00	0.00	**0.83**	(0.753,0.890)
13	0.500	4.58	0.586	235	Yes	0.1	0.16	**0.77**	0.63	(0.477,0.659)
14	0.500	4.14	0.593	235	Yes	0.5	0.11	0.60	**0.66**	(0.497,0.678)
15	0.500	3.72	0.600	235	Yes	0.9	0.03	0.34	**0.51**	(0.350,0.533)
16	0.500	8.73	0.520	235	No	0.1	**0.89**	**0.89**	0.70	(0.517,0.697)
17	0.500	7.58	0.538	235	No	0.5	0.25	0.25	**0.59**	(0.447,0.631)
18	0.500	6.44	0.556	235	No	0.9	0.03	0.03	**0.54**	(0.398,0.582)
19	0.500	4.56	0.586	470	Yes	0.1	0.02	**0.44**	0.26	(0.169,0.325)
20	0.500	4.13	0.593	470	Yes	0.5	0.00	0.25	**0.33**	(0.204,0.368)
21	0.500	3.73	0.599	470	Yes	0.9	0.01	0.11	**0.34**	(0.212,0.379)
22	0.500	8.72	0.520	470	No	0.1	**0.76**	**0.76**	0.37	(0.195,0.357)
23	0.500	7.59	0.538	470	No	0.5	0.06	0.06	**0.26**	(0.160,0.314)
24	0.500	6.45	0.556	470	No	0.9	0.00	0.00	**0.25**	(0.110,0.247)

Figure 4.7 Coverage probabilities of 95% confidence intervals for sample odds ratios based on 100 replicates in each combination of parameters, life table odds ratio calculated with more (3) intervals; left figures are from censored populations while right are uncensored (asymptomatic); top figures have true ten-year OR=0.861 while lower ones have true ten-year OR=0.500



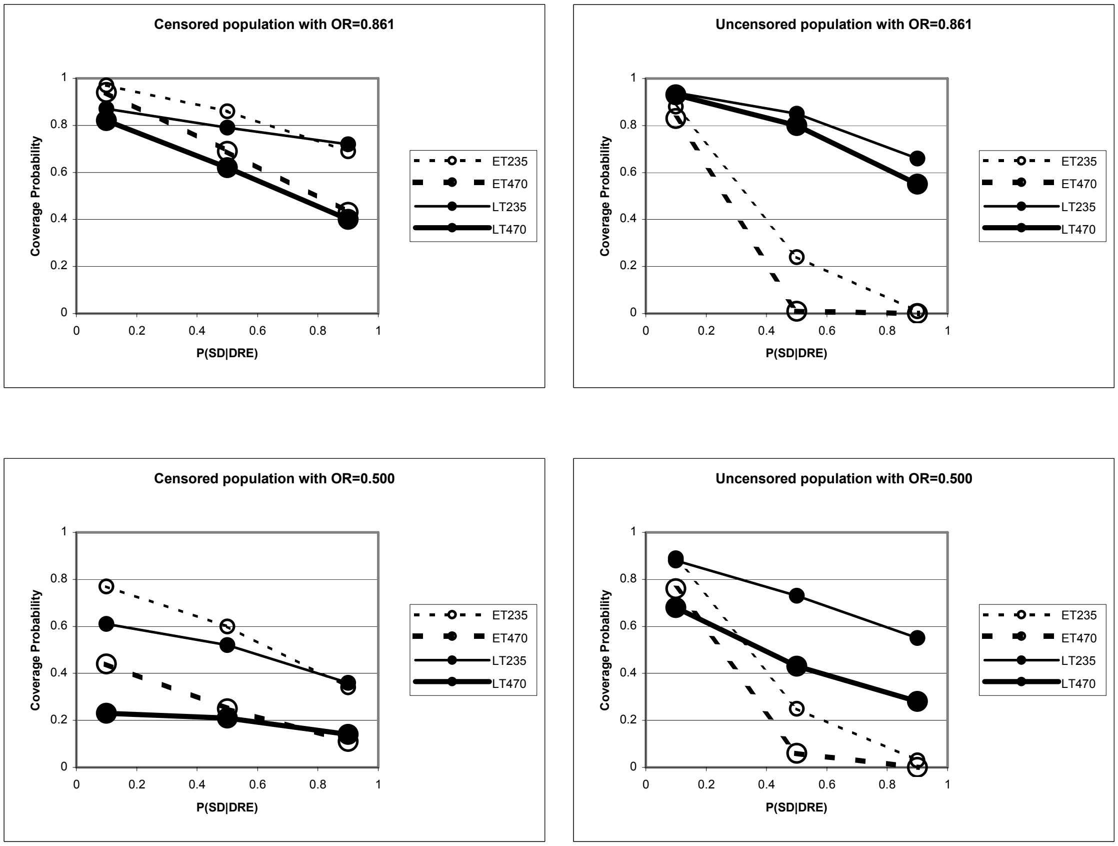
Coverage probabilities from the analyses based on 3 life table intervals are summarized in Table 4.5 and Figure 4.7. In 18/24 combinations (75%), LTMHOR has a CP closer to 95% than ETMHOR. CP tends to be worse as the point estimates become more biased, which makes obvious sense. For a more extreme effect (odds ratio=0.500), CP suffers compared with a milder odds ratio (0.861); but in terms of relative odds ratios this is also in line with what was observed in the point estimates. Censoring appears to have no systematic effect on the CPs for LTMHOR, but once again for ETMHOR it is a mitigating influence; higher CPs are lowered and lower ones are raised. This is again in line with the bias of the point estimates. Note that smaller case groups (235 cases) improve the coverage probability for both LTMHOR and ETMHOR especially under a stronger effect, even though the point estimates were generally close. This is because if two point estimates are equally biased, a wider confidence interval about the estimate will have better coverage probability for the true value than a narrower one.

It can be noted that CP for LTMHOR is high in both the censored and uncensored groups when the true odds ratio is 0.861. In those groups it is about 90% except when P(SD|DRE) is most extreme (0.9). This reflects well on the DRE study. In the absence of the small positive bias in LTMHOR the coverage would probably be close to 95%, indicating that Miettinen's test-based limits are performing well.

Table 4.6 Coverage probabilities of 95% confidence intervals for sample odds ratios based on 100 replicates in each combination of parameters, life table odds ratio calculated with fewer (2) intervals; the closest CP to 0.95 is in bold

Pop #	True OR	Obs Time	Corr True OR	Cases	Cen-sor-ing	P(SD|DRE)	Covg Prob MHOR	Covg Prob ET-MHOR	Covg Prob LT-MHOR	95% CI Covg Prob LT-MHOR
1	0.861	4.53	0.895	235	Yes	0.1	0.51	**0.97**	0.87	(0.800,0.921)
2	0.861	3.99	0.898	235	Yes	0.5	0.31	**0.86**	0.79	(0.708,0.857)
3	0.861	3.44	0.901	235	Yes	0.9	0.08	0.69	**0.72**	(0.632,0.796)
4	0.861	8.65	0.870	235	No	0.1	0.88	0.88	**0.94**	(0.861,0.958)
5	0.861	7.17	0.879	235	No	0.5	0.24	0.24	**0.85**	(0.776,0.906)
6	0.861	5.75	0.888	235	No	0.9	0.01	0.01	**0.66**	(0.569,0.743)
7	0.861	4.55	0.895	470	Yes	0.1	0.24	**0.94**	0.82	(0.753,0.890)
8	0.861	3.98	0.898	470	Yes	0.5	0.02	**0.69**	0.62	(0.528,0.706)
9	0.861	3.44	0.901	470	Yes	0.9	0.00	**0.43**	0.40	(0.313,0.493)
10	0.861	8.67	0.870	470	No	0.1	0.83	0.83	**0.93**	(0.874,0.965)
11	0.861	7.20	0.879	470	No	0.5	0.01	0.01	**0.80**	(0.719,0.865)
12	0.861	5.75	0.888	470	No	0.9	0.00	0.00	**0.55**	(0.457,0.640)
13	0.500	4.58	0.586	235	Yes	0.1	0.16	**0.77**	0.61	(0.487,0.669)
14	0.500	4.14	0.593	235	Yes	0.5	0.11	**0.60**	0.52	(0.408,0.592)
15	0.500	3.72	0.600	235	Yes	0.9	0.03	0.34	**0.36**	(0.248,0.421)
16	0.500	8.73	0.520	235	No	0.1	**0.89**	**0.89**	0.88	(0.812,0.929)
17	0.500	7.58	0.538	235	No	0.5	0.25	0.25	**0.73**	(0.632,0.796)
18	0.500	6.44	0.556	235	No	0.9	0.03	0.03	**0.55**	(0.457,0.640)
19	0.500	4.56	0.586	470	Yes	0.1	0.02	**0.44**	0.23	(0.152,0.303)
20	0.500	4.13	0.593	470	Yes	0.5	0.00	**0.25**	0.21	(0.135,0.281)
21	0.500	3.73	0.599	470	Yes	0.9	0.01	0.11	**0.14**	(0.079,0.200)
22	0.500	8.72	0.520	470	No	0.1	**0.76**	**0.76**	0.68	(0.579,0.752)
23	0.500	7.59	0.538	470	No	0.5	0.06	0.06	**0.43**	(0.341,0.523)
24	0.500	6.45	0.556	470	No	0.9	0.00	0.00	**0.28**	(0.204,0.368)

Figure 4.8 Coverage probabilities of 95% confidence intervals for sample odds ratios based on 100 replicates in each combination of parameters, life table odds ratio calculated with fewer (2) intervals; left figures are from censored populations while right are uncensored (asymptomatic); top figures have true ten-year OR=0.861 while lower ones have true ten-year OR=0.500



Coverage probabilities from the analyses based on 2 life table intervals are summarized in Table 4.6 and Figure 4.8. In only 13/24 combinations (54%), LTMHOR has a CP closer to 95% than ETMHOR. However, the graph shows that controlling for sample size, LTMHOR tends to win big when it wins, compared with an only moderately better ETMHOR when that performs best. Again, CP sensibly tends to be worse as the point estimates become more biased. Censoring has a greater impact on CP for LTMHOR than when LTMHOR was based on 3 intervals, which is also what was observed in the point estimates.

In the context of the DRE study, it can be noted that the CP for LTMHOR remains high in both the censored and uncensored groups when the true odds ratio is 0.861 and P(SD|DRE) is moderate (0.1). The DRE study did even better, based on more (3) life table intervals.

5 CHAPTER FIVE: DISCUSSION AND CONCLUSION

The life table Mantel-Haenszel odds ratio (LTMHOR) developed in this study provides a better estimate of the odds ratio from case-control screening data than previous methods used in this area. Although there is some residual bias, it is relatively small and in the conservative direction. On the other hand, the time-adjusted Mantel-Haenszel odds ratio (ETMHOR) alternative is extremely biased, and the bias increases steeply with the probability that a screened case is detected by the test. Unfortunately, the more efficacious a screening test is, the higher this proportion will be in general, and the more biased the results. Conditional logistic regression as an alternative to MHOR does not improve matters. The fundamental problem with using these estimators on case-control screening data is the dependence between screening tests in screen-detected cases and the end of the observation time in their matched controls. Stopping the observation time in controls immediately after a screening test is counted in the case is a perfect design for recording as many screening tests in cases as possible while minimizing those counted in controls.

LTMHOR solves this problem with its approach to censoring. Estimates are based on life table event probabilities in cases versus controls, and the life table is constructed to handle censoring appropriately (as long as that censoring is non-informative). To assess whether this is reasonable, consider each censoring process. Firstly, for controls, the reference date (date of diagnosis in the matched case) forms an independent censoring process from screening tests in the control because the case and control are unconnected and unaware of each other. Symptomatic progression in controls, the other censoring process, also runs independently of screening tests, because in the absence of symptoms, a test marked as screening probably is—but what about as a control nears the onset of symptoms? If the first symptoms are about to appear in a few months, can that affect the rate of screening DREs? Since the symptoms have not yet appeared, the answer should be no. Next we consider censoring processes for cases. It is assumed for cases that the competing diagnosis of PC (non-DRE-detected PC) is a process that runs independently of screening DREs. In non-screen-detected cases who do not develop symptoms, the reasoning is that there are no outward indications that could influence frequency of screening DREs. Symptomatic progression in cases is a separate censoring process, and a non-informative one by an argument analogous to that given for controls.

There are alternative survival analysis methods that might have been considered instead of the life table. One might try continuous time approaches rather than the discetized life table, for example fitting parametric regression models such as piecewise Weibull or exponential, then calculating an odds ratio to describe the difference in event probabilities over different lengths of time. One of the reasons for using the life table in this study was that there is some scientific literature using this approach at least to describe the effect of an exposure in stratified prospective data. While this project went further by applying this to the retrospective, case-control situation and then reversed the interpretation of the odds ratio, this is less of a leap than might otherwise have been the case had we fit continuous time models and proceeded from there. In trying to publish scientific results, smaller leaps are more readily accepted than giant ones, in part because smaller leaps tend to lead to fewer mistakes. Still another approach might have been the log rank test to assess if two curves differ. While this would have provided a p-value, it does not naturally lead to an estimate of the odds ratio, which was desired in this study.

There are also alternative general approaches other than survival analysis that might have been developed for this problem. During this study some were considered. One approach would have been to begin with a 2 by 2 table crossing DRE with case/control status, then adjust the cell sizes to reflect the estimated effects of dependence bias. The adjustment would incorporate an estimate of when screen-detected cases would have been diagnosed had they not been detected by a DRE, and how that additional observation time would have altered the DRE counts in matched controls. This is a very novel approach that would have suffered from weaker scientific background (i.e., no backing in literature). Still another approach might have involved estimating a DRE rate ratio, which would directly estimate and compare rates of DRE between cases and controls rather than modeling time to first DRE. While rate of DRE might actually be the more appropriate measure considering that one single DRE in a lifetime will probably not help you very much, again it does not naturally reverse itself into an estimate for the effect of DRE on the odds of MPC. Further, time to first DRE perhaps serves well as a proxy for rate of DRE, if it can be assumed that shorter time to first DRE indicates a higher rate of DRE throughout a lifetime. In any case, these alternatives and others may be explored more fully in the future and better solutions may be found, such is the case with all science.

LTMHOR performs better than ETMHOR, but considering how poorly ETMHOR performs this is perhaps nothing to brag about. Considering instead how LTMHOR performs in its own right, it does provide good estimates (not too biased) under various combinations of true ten-year odds ratio, screen-detection probability for cases, censoring or not beyond that due to the competing diagnostic event (and/or the reference date in controls), and smaller or larger case groups. Miettinen's test-based limits provide good coverage probability when the true odds ratio is moderate and the case group is not too large. But that said, there is room for improvement. That the life table Mantel-Haenszel odds ratio remains slightly biased might form the basis of additional work in this area. In the meantime, that the bias is small and (importantly) in the conservative direction bodes well for both the results of the current DRE study and others like it in the future that might make use of this estimator.

In this study, the life table Mantel-Haenszel odds ratio describing the effect of having a screening DRE on the odds of MPC, after adjusting for the matched case-control design as well as several potentially confounding variables, is 0.861 with 95% confidence interval (0.611, 1.215). This constitutes only weak evidence of a mild protective effect of screening with digital rectal exam in the prevention of metastatic prostate cancer. However, considering that the simulation indicates a small positive bias in LTMHOR, and the fact that several previous studies also found mildly protective (if statistically insignificant) results, continued screening with DRE is advised at least until a more definitive study is undertaken.

REFERENCES

[1] Weir HK, Thun MJ, Hankey BF, Ries LA, Howe HL, Wingo PA, et al. Annual report to the nation on the status of cancer, 1975-2000, featuring the uses of surveillance data for cancer prevention and control. J Natl Cancer Inst. 2003;95:1276-99.

[2] Johansson JE, Holmberg L, Johansson S, Bergstrom R, Adami HO. Fifteen-year survival in prostate cancer. A prospective, population-based study in Sweden. JAMA 1997;277:467-71.

[3] Weiss NS. Application of the case-control method in the evaluation of screening. Epidemiol Rev 1994;16:102-8.

[4] Jacobsen SJ, Bergstralh EJ, Katusic SK, Guess HA, Darby CH, Silverstein MD et al. Screening digital rectal examination and prostate cancer mortality: a population based case control study. Urology 1998;52:173 9.

[5] Lawless JF. Statistical Models and Methods for Lifetime Data. John Wiley & Sons, Inc. Hoboken, NJ, USA. 2003. p. 80.

[6] Richert-Boe KE, Humphrey LL, Glass AG, Weiss NS. Screening digital rectal examination and prostate cancer mortality: a case-control study. J Med Screen 1998;5:99 103.

[7] Friedman GD, Hiatt RA, Quesenberry CP Jr, Selby JV. Case control study of screening for prostatic cancer by digital rectal examinations. Lancet 1991;337:1526 9.

[8] Weinmann S, Richert-Boe K, Glass AG, Weiss NS. Prostate cancer screening and mortality: a case-control study (United States). Cancer Causes Control. 2004 Mar;15(2):133-8.

[9] Cronin KA, Weed DL, Connor RJ, Prorok PC. Case-control studies of cancer screening: theory and practice. J Natl Cancer Inst 1998; 90: 498-504.

[10] Kahn HA, Sempos CT. Statistical Methods in Epidemiology. Oxford University Press, New York, New York, USA. p. 186.

[11] Kahn HA, Sempos CT. Statistical Methods in Epidemiology. Oxford University Press, Inc., New York, NY, USA. 1989. p. 119.

[12] Agresti A. Categorical Data Analysis. John Wiley & Sons, Inc. Hoboken, NJ, USA. 2002. pp. 75-76.

[13] Agresti A. Categorical Data Analysis. John Wiley & Sons, Inc. Hoboken, NJ, USA. 2002. p. 258.

[14] Joffe MM, Rosenbaum PR. Invited commentary: propensity scores. Am J Epidemiol. 1999;150:327-33.

[15] Mantel N. Evaluation of survival data and two new rank order statistics arising in its consideration. Cancer Chemother Rep. 1966 Mar;50(3):163-70.

[16] Lawless JF. Statistical Models and Methods for Lifetime Data. John Wiley & Sons, Inc. Hoboken, NJ, USA. 2003. pp. 128-136.

[17] Kahn HA, Sempos CT. Statistical Methods in Epidemiology. Oxford University Press, Inc., New York, NY, USA. 1989. pp. 186-187.

[18] Agresti A. Categorical Data Analysis. John Wiley & Sons, Inc. Hoboken, NJ, USA. 2002. p. 234.

[19] Hosmer DW, Lemeshow S. Applied Logistic Regression. John Wiley & Sons, Inc. USA. 1989. pp. 187-189.

[20] Lawless JF. Statistical Models and Methods for Lifetime Data. John Wiley & Sons, Inc. Hoboken, NJ, USA. 2003. pp. 17-18.

[21] Lawless JF. Statistical Models and Methods for Lifetime Data. John Wiley & Sons, Inc. Hoboken, NJ, USA. 2003. pp. 270, 296.

[22] Robins J., Breslow N., Greenland, S. Estimators of the Mantel-Haenszel variance consistent in both sparse data and large-strata limiting models. Biometrics. 1986 Jun;42(2):311-23.