

INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

UMI

**A Bell & Howell Information Company
300 North Zeeb Road, Ann Arbor MI 48106-1346 USA
313/761-4700 800/521-0600**

UNIVERSITY OF ALBERTA

Cholecystectomy as a Risk for Colon Cancer

by

Donna Cheryl Ranneris Turner



**A thesis submitted to the Faculty of Graduate Studies and Research in partial
fulfillment of the requirements for the degree of Doctor of Philosophy.**

IN

MEDICAL SCIENCES - ONCOLOGY

EDMONTON, ALBERTA

FALL 1997



**National Library
of Canada**

**Acquisitions and
Bibliographic Services**

**395 Wellington Street
Ottawa ON K1A 0N4
Canada**

**Bibliothèque nationale
du Canada**

**Acquisitions et
services bibliographiques**

**395, rue Wellington
Ottawa ON K1A 0N4
Canada**

Your file Votre référence

Our file Notre référence

The author has granted a non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.

The author retains ownership of the copyright in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de cette thèse sous la forme de microfiche/film, de reproduction sur papier ou sur format électronique.

L'auteur conserve la propriété du droit d'auteur qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

0-612-23080-5

University of Alberta

Library Release Form

Name of Author: Donna Cheryl Ranneris Turner
Title of Thesis: Cholecystectomy as a Risk for Colon Cancer
Degree: Doctor of Philosophy
Year this Degree Granted: 1997

Permission is hereby granted to the University of Alberta Library to reproduce single copies of this thesis and to lend or sell such copies for private, scholarly, or scientific research purposes only.

The author reserves all other publication and other rights in association with the copyright in the thesis, and except as hereinbefore provided, neither the thesis nor any substantial portion thereof may be printed or otherwise reproduced in any material form whatever without the author's prior written permission.

Donna Ranneris Turner

11127 - 73 Avenue

Edmonton, Alberta

Canada

T6G 0C5

June 23, 1997

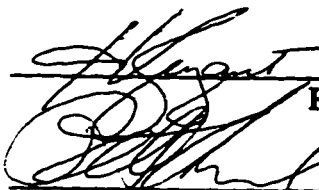
University of Alberta

Faculty of Graduate Studies and Research

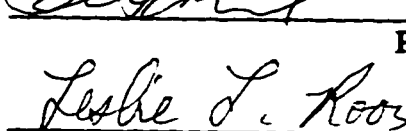
The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies and Research for acceptance, a thesis entitled *Cholecystectomy as a Risk for Colon Cancer* submitted by Donna Cheryl Ranneris Turner in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Medical Sciences - Oncology.



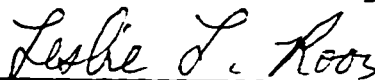
Anthony L.A. Fields, MD



Heather E. Bryant, MD, PhD



Patrick A. Hessel, PhD



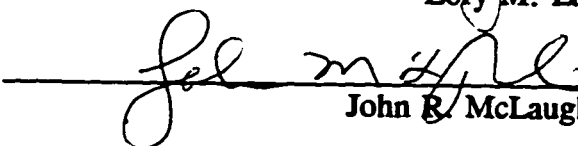
Leslie L. Roos, PhD



Garth L. Warnock, MD



Lory M. Laing, PhD



John R. McLaughlin, PhD

ABSTRACT

The purpose of this epidemiologic study was to investigate the risk of colon cancer following cholecystectomy in Alberta. Others have proposed that there is a risk of colon cancer because removal of the gallbladder results in a continuous flow of bile through the digestive tract, constantly exposing the colon to carcinogenic secondary bile acids.

Record linkage of files from the Alberta Health Care Insurance Plan (AHCIP) and the Alberta Cancer Registry was used to determine the frequency of colon cancer in over 90,000 Albertans who received cholecystectomy between 1973 and 1993. Different record linkage approaches were used to achieve the best possible results. Estimates of risk in the population were expressed using standardized incidence ratios (SIRs). Statistical significance was determined using 95% confidence intervals (CIs), based on tables that provided limits for Poisson data. An alternative comparison group was selected from the AHCIP, those with stripping and ligation of varicose veins, to control for potential overestimation of the expected number of cases which would result in underestimated SIRs.

A total of 606 colon cancer cases were found in the cholecystectomy cohort. With the general population as the comparison group, the risk was statistically significant under the assumption of no induction ($SIR=1.09$, 95% $CI=1.01-1.19$). More biologically plausible induction periods of 10 and 15 years showed no increase in risk, with SIRs of 0.92 (95% $CI: 0.78 - 1.07$) and 0.93 (95% $CI: 0.69 - 1.21$) respectively. Males were found to have some significant increase in risk for the first

5 years of induction, while females had non-significantly low SIRs. With the alternative comparison group, there was no significant difference in the SIRs overall (SIR=1.10 with 95% CI=0.94-1.29, and SIR=1.15 with 95% CI=0.86-1.51 for 10 and 15 years' induction, respectively). The patterns of the sex-specific risk estimates were similar to those determined using the general population, but the magnitude of the SIRs was exaggerated. However, the sex-specific colon cancer risk estimates using the varicose vein cohort were opposite to those observed for the cholecystectomy cohort, suggesting limited usefulness of the varicose vein cohort as an alternative comparison group.

Concordant with previous large-sample studies, this study cannot conclude that colon cancer risk is increased following cholecystectomy at plausible induction periods.

ACKNOWLEDGEMENTS

Many people contributed to the success of this project and I would like to acknowledge certain individuals specifically.

I would like to thank my thesis committee, Dr. Patrick Hessel, Dr. Leslie Roos and Dr. Garth Warnock, and the external examiners, Dr. Lory Laing and Dr. John McLaughlin, for their support. I am especially grateful for the patient guidance of my supervisors, Dr. Anthony Fields and Dr. Heather Bryant.

Members of the Division of Epidemiology, Prevention and Screening provided ongoing support. In particular, I would like to thank Ms. Mary-Lynn Gantefoer and Ms. Herta Gaedke for their assistance with the chart reviews, Mr. Voon Siaw and Mr. Andrew MacMillan for programming suggestions, Dr. Penny Brasher and Dr. Christine Friedenreich for methodologic advice, and Dr. Shirley Fincham for her thorough editing. Mr. Andre Wajda of the Manitoba Centre for Health Policy and Evaluation and Mr. Larry Svenson of Alberta Health provided technical advice on record linkage using administrative data.

I gratefully acknowledge the financial support received from the Alberta Heritage Foundation for Medical Research, in the form of a graduate studentship, and the Alberta Cancer Board.

Finally, I would like to thank my family, including the Ranneris, Love and Turner clans, for their continued support and belief in my abilities. I am especially grateful for all the contributions of my own personal data manager and spouse, Ken, and for the patience and good humour of our daughter, Tegan.

TABLE OF CONTENTS

Signature Page

Abstract

Acknowledgements

Table of Contents

List of Tables

List of Figures

Chapter

I.	Introduction	1
II.	Literature Review	5
A.	The Biological Perspective: Cholecystectomy and Colon Cancer ...	5
1.	Cholecystectomy and the Epidemiology of Gallstone Disease	5
2.	The Epidemiology of Colon Cancer	7
3.	Evidence for the Association between Cholecystectomy and Colon Cancer	9
4.	Gastric Procedures as Confounders in the Cholecystectomy-Colon Cancer Relationship	17
5.	Stripping and Ligation of Varicose Veins as a Comparison Procedure	17
B.	Methods: Linking Administrative Data for Health Research	18
1.	Basic Steps	20

a.	Searching	22
b.	Matching	26
c.	Separating Links from Nonlinks	29
2.	Errors Influencing the Linkage Process	31
3.	Summary	35
III.	Methods	36
A.	Assessing Data Quality	36
1.	Data from the Alberta Health Care Insurance Plan	36
a.	Registrant File	37
b.	Claims File	38
c.	Combined Information	38
2.	Data from the Alberta Cancer Registry	39
B.	Record Linkage	40
C.	Comparing Linkage to Chart Review	42
D.	Analysis of Risk	43
1.	Exclusion Criteria	43
2.	Standardized Incidence Ratios	43
a.	Person-years at Risk	44
b.	Rates	45
c.	Calculation of Expected Numbers	46
d.	Adjustment for Induction	46
e.	Adjustment for Gastric Procedures	47
f.	95 % Confidence Intervals	48

3.	Sensitivity Analysis	48
4.	Proportional Hazards Modelling	49
5.	Nested Case-Control: Fat as a Confounder	50
E.	Stratified Analysis by History of Colon Cancer and Cholecystectomy	51
IV.	Results	52
A.	Cohort Preparation	52
1.	Alberta Health Care Insurance Plan Cohorts	52
2.	Colon Cancer Cohort	57
B.	Data Quality	58
1.	Alberta Health Care Insurance Plan Cohorts	58
2.	Colon Cancer Cohort	61
C.	Record Linkage	62
1.	Searching: Determining the Linkage Strategy	62
2.	Matching: Deterministic Linkage	63
3.	Matching: Probabilistic Linkage	63
4.	Separation: Identifying Probable Pairs	64
5.	Comparison of Linkage Strategies: Deterministic versus Probabilistic	66
6.	Variable Agreement in Linked Pairs	69
7.	Chart Review: Confirming Record Linkage	71
8.	Secondary Endpoints: Linkage to Other Relevant Cancers ..	74
D.	Estimates of Risk	78

1.	Colon Cancer Risk in the Cholecystectomy Cohort	
	Compared to the General Population	78
a.	Determination of the Final Cohort	79
b.	Overall Risk by Type of Linkage	79
c.	Risk by Subsite	87
2.	Colon Cancer Risk in the Cholecystectomy Cohort	
	Compared to the Varicose Vein Cohort	87
3.	Colon Cancer Risk in the Cholecystectomy Cohort	
	Adjusted for Confounding by Gastric Procedures	108
E.	Sensitivity Analysis: Influence of Completeness of AHCIP	
	Effective Dates	118
F.	Stratified Analysis: Characteristics of Individuals with	
	Colon Cancer versus the Cohort	124
G.	Summary	129
V.	Discussion	131
A.	Overview	131
B.	Methodology	131
1.	Data Quality	131
a.	General Comments	131
	i AHCIP Data	132
	ii Alberta Cancer Registry Data	138
b.	Influence of Alternative Comparison Groups	138
2.	Record Linkage	139

3.	Comparison of Chart Review and Linkage Results	141
C.	Assessment of Risk	142
1.	Cholecystectomy as a Risk for Colon Cancer: Compared to the General Population	142
a.	All Colon Subsites Combined	142
b.	Risk by Colon Subsite	147
2.	Cholecystectomy as a Risk for Colon Cancer: Adjusting for Gastric Procedures	150
3.	Cholecystectomy as a Risk for Colon Cancer: Compared to the Varicose Vein Cohort	150
a.	All Colon Subsites Combined	150
b.	Risk by Colon Subsite	156
4.	Influence of AHCIP Effective Date Quality on Risk Estimates	156
D.	Strengths and Limitations of the Current Study	158
1.	Strengths	158
2.	Limitations	160
E.	Conclusions and Recommendations	164
	References	166

Appendix

A.	Alberta Health Care Insurance Plan Fee Codes for Procedures	179
B.	Measures of Information, AHCIP and Colon Cancer Cohorts	183
C.	Deterministic Linkage Results	184
D.	Probabilistic Results: LinkPro Output	188
E.	Probabilistic Linkage Thresholds	192
F.	Person-years at Risk for the Cholecystectomy Cohort (5 Years Induction): Individuals Identified by Either Type of Linkage	200
G.	Colon Cancer Rates in the Alberta Population, 1969-1993	203

LIST OF TABLES

Table 1.	Bile Acid Composition in Normal Patients and in Gallstone Patients Before and After Cholecystectomy	16
Table 2.	Threshold Example: Distribution of Weights for Links and Nonlinks	32
Table 3.	Proportion of Procedures by Fiscal Year, Cohort and Procedures ...	56
Table 4.	Proportion of Coverage Records Missing Date Elements, All Records and Individual-specific	59
Table 5.	Agreement Between Deterministic and Probabilistic Linkage Strategies: Colon Cancer as Outcome	68
Table 6.	Variable Agreement in Linked Pairs: AHCIP and Cancer Registry Linking Variables (Colon Cancer Cohort Only)	70
Table 7.	Record Linkage versus Chart Review: Cancer Charts Only	72
Table 8.	Record Linkage versus Chart Review: Cancer Charts and Physician Records	73
Table 9.	Agreement Between Deterministic and Probabilistic Linkage Strategies: Other Relevant Cancers as Outcome	77
Table 10.	Summary of Final File Preparation, by Linkage Approach	80
Table 11.	Comparison of Standardized Incidence Ratios (SIRs) and 95 % Confidence Intervals (CIs) for Colon Cancer in the Cholecystectomy Cohort versus the General Alberta Population, Using Deterministic Linkage Only	81

Table 12.	Comparison of Standardized Incidence Ratios (SIRs) and 95 % Confidence Intervals (CIs) for Colon Cancer in the Cholecystectomy Cohort versus the General Alberta Population, Using Deterministic Linkage Only, by Sex	82
Table 13.	Comparison of Standardized Incidence Ratios (SIRs) and 95 % Confidence Intervals (CIs) for Colon Cancer in the Cholecystectomy Cohort versus the General Alberta Population, Using Probabilistic Linkage Only	83
Table 14.	Comparison of Standardized Incidence Ratios (SIRs) and 95 % Confidence Intervals (CIs) for Colon Cancer in the Cholecystectomy Cohort versus the General Alberta Population, Using Probabilistic Linkage Only, by Sex	84
Table 15.	Comparison of Standardized Incidence Ratios (SIRs) and 95 % Confidence Intervals (CIs) for Colon Cancer in the Cholecystectomy Cohort versus the General Alberta Population, Using Both Deterministic and Probabilistic Linkage	85
Table 16.	Comparison of Standardized Incidence Ratios (SIRs) and 95 % Confidence Intervals (CIs) for Colon Cancer in the Cholecystectomy Cohort versus the General Alberta Population, Using Both Deterministic and Probabilistic Linkage, by Sex	86
Table 17.	Comparison of Standardized Incidence Ratios (SIRs) and 95 % Confidence Intervals (CIs) for Right Colon Cancer in the Cholecystectomy Cohort versus the General Alberta	

	Population	90
Table 18.	Comparison of Standardized Incidence Ratios (SIRs) and 95 % Confidence Intervals (CIs) for Right Colon Cancer in the Cholecystectomy Cohort versus the General Alberta Population, by Sex	91
Table 19.	Comparison of Standardized Incidence Ratios (SIRs) and 95 % Confidence Intervals (CIs) for Mid-Colon Cancer in the Cholecystectomy Cohort versus the General Alberta Population	92
Table 20.	Comparison of Standardized Incidence Ratios (SIRs) and 95 % Confidence Intervals (CIs) for Mid-Colon Cancer in the Cholecystectomy Cohort versus the General Alberta Population, by Sex	93
Table 21.	Comparison of Standardized Incidence Ratios (SIRs) and 95 % Confidence Intervals (CIs) for Left Colon Cancer in the Cholecystectomy Cohort versus the General Alberta Population	94
Table 22.	Comparison of Standardized Incidence Ratios (SIRs) and 95 % Confidence Intervals (CIs) for Left Colon Cancer in the Cholecystectomy Cohort versus the General Alberta Population, by Sex	95
Table 23.	Comparison of Standardized Incidence Ratios (SIRs) and 95 % Confidence Intervals (CIs) for Colon Cancer in the	

	Cholecystectomy Cohort versus the Varicose Vein Cohort	98
Table 24.	Comparison of Standardized Incidence Ratios (SIRs) and 95 % Confidence Intervals (CIs) for Cancer in the Cholecystectomy Cohort versus the Varicose Vein Cohort, by Sex	99
Table 25.	Comparison of Standardized Incidence Ratios (SIRs) and 95 % Confidence Intervals (CIs) for Right Colon Cancer in the Cholecystectomy Cohort versus the Varicose Vein Cohort ...	100
Table 26.	Comparison of Standardized Incidence Ratios (SIRs) and 95 % Confidence Intervals (CIs) for Right Colon Cancer in the Cholecystectomy Cohort versus the Varicose Vein Cohort, by Sex	101
Table 27.	Comparison of Standardized Incidence Ratios (SIRs) and 95 % Confidence Intervals (CIs) for Mid-Colon Cancer in the Cholecystectomy Cohort versus the Varicose Vein Cohort	102
Table 28.	Comparison of Standardized Incidence Ratios (SIRs) and 95 % Confidence Intervals (CIs) for Mid-Colon Cancer in the Cholecystectomy Cohort versus the Varicose Vein Cohort, by Sex	103
Table 29.	Comparison of Standardized Incidence Ratios (SIRs) and 95 % Confidence Intervals (CIs) for Left Colon Cancer in the Cholecystectomy Cohort versus the Varicose Vein Cohort	104
Table 30.	Comparison of Standardized Incidence Ratios (SIRs) and	

	95 % Confidence Intervals (CIs) for Left Colon Cancer in the Cholecystectomy Cohort versus the Varicose Vein Cohort, by Sex	105
Table 31.	Comparison of Standardized Incidence Ratios (SIRs) and 95 % Confidence Intervals (CIs) for Colon Cancer in the Varicose Vein Cohort versus the General Alberta Population	109
Table 32.	Comparison of Standardized Incidence Ratios (SIRs) and 95 % Confidence Intervals (CIs) for Colon Cancer in the Varicose Vein Cohort versus the General Alberta Population, by Sex	110
Table 33.	Comparison of Standardized Incidence Ratios (SIRs) and 95 % Confidence Intervals (CIs) for Right Colon Cancer in the Varicose Vein Cohort versus the General Alberta Population	112
Table 34.	Comparison of Standardized Incidence Ratios (SIRs) and 95 % Confidence Intervals (CIs) for Right Colon Cancer in the Varicose Vein Cohort versus the General Alberta Population, by Sex	113
Table 35.	Comparison of Standardized Incidence Ratios (SIRs) and 95 % Confidence Intervals (CIs) for Mid-Colon Cancer in the Varicose Vein Cohort versus the General Alberta Population ...	114
Table 36.	Comparison of Standardized Incidence Ratios (SIRs) and 95 % Confidence Intervals (CIs) for Mid-Colon Cancer in the Varicose Vein Cohort versus the General Alberta Population,	

	by Sex	115
Table 37.	Comparison of Standardized Incidence Ratios (SIRs) and 95 % Confidence Intervals (CIs) for Left Colon Cancer in the Varicose Vein Cohort versus the General Alberta Population ...	116
Table 38.	Comparison of Standardized Incidence Ratios (SIRs) and 95 % Confidence Intervals (CIs) for Left Colon Cancer in the Varicose Vein Cohort versus the General Alberta Population, by Sex	117
Table 39.	Comparison of Standardized Incidence Ratios (SIRs) and 95 % Confidence Intervals (CIs) for Colon Cancer in the Cholecystectomy Cohort versus the General Alberta Population, Adjusting for Presence of Gastric Procedures	119
Table 40.	Comparison of Standardized Incidence Ratios (SIRs) and 95 % Confidence Intervals (CIs) for Colon Cancer in the Cholecystectomy Cohort versus the General Alberta Population, Adjusting for Presence of Gastric Procedures, by Sex	120
Table 41.	Comparison of Standardized Incidence Ratios (SIRs) and 95 % Confidence Intervals (CIs) for Colon Cancer in the Cholecystectomy Cohort versus the General Alberta Population, Individuals with Complete AHCIP Effective Dates Only	121
Table 42.	Comparison of Standardized Incidence Ratios (SIRs) and 95 % Confidence Intervals (CIs) for Colon Cancer in the Cholecystectomy Cohort versus the General Alberta Population,	

	Individuals with Complete AHCIP Effective Dates Only, by Sex	122
Table 43.	Comparison of Standardized Incidence Ratios (SIRs) and 95 % Confidence Intervals (CIs) for Colon Cancer in the Cholecystectomy Cohort versus the Varicose Vein Cohort, Individuals with Complete AHCIP Effective Dates Only	125
Table 44.	Comparison of Standardized Incidence Ratios (SIRs) and 95 % Confidence Intervals (CIs) for Colon Cancer in the Cholecystectomy Cohort versus the Varicose Vein Cohort, Individuals with Complete AHCIP Effective Dates Only, by Sex	126
Table 45.	Point Estimates of Standardized Incidence Ratios and Range of Values, by Induction.....	130
Table 46.	Comparison of Variables Available for Linkage: Netherlands (based on data from Van den Brandt, 1990) and Alberta	134
Table 47.	Average Number of Registration Numbers per Unique Identifier, by First Service Year	137

LIST OF FIGURES

Figure 1.	Summary of Processing: Alberta Health Care Insurance Plan Files	53
Figure 2.	Comparison of Standardized Incidence Ratios for Colon Cancer Following Cholecystectomy: Different Linkage Approaches, by Induction	88
Figure 3.	Comparison of Standardized Incidence Ratios for Colon Cancer Following Cholecystectomy: Using Different Linkage Approaches, by Induction and Sex	89
Figure 4.	Comparison of Standardized Incidence Ratios for Colon Cancer Following Cholecystectomy: by Colon Subsite, Using Rates of Colon Cancer in the General Alberta Population	96
Figure 5.	Analysis of Standardized Incidence Ratios for Colon Cancer Following Cholecystectomy Using Different Comparison Groups: the General Alberta Population versus the Varicose Vein Cohort	106
Figure 6.	Comparison of Standardized Incidence Ratios for Colon Cancer Following Cholecystectomy: by Colon Subsite, Using Rates of Colon Cancer in the Varicose Vein Cohort	107
Figure 7.	Standardized Incidence Ratios for Colon Cancer Following Stripping and Ligation of Varicose Veins, Using Rates of Colon Cancer in the General Alberta Population	111
Figure 8.	Sensitivity Analysis: Standardized Incidence Ratios for	

Colon Cancer Following Cholecystectomy: All Records versus
Records With Complete Effective Dates, Using Rates of Colon
Cancer in the General Alberta Population 123

Figure 9. Sensitivity Analysis: Standardized Incidence Ratios for
Colon Cancer Following Cholecystectomy: All Records versus
Records With Complete Effective Dates, Using Rates of Colon
Cancer in the Varicose Vein Cohort 127

CHAPTER I. INTRODUCTION

Both colon cancer and gallstones are common diseases in Western populations. Recently, medical scientists have been interested in determining if cholecystectomy, the usual treatment for gallstones, puts individuals at risk for developing colon cancer. The issue remains controversial despite numerous other studies.

Proponents of an association between cholecystectomy and colon cancer suggest that the link between these two factors involves altered bile acid profiles and flow.¹⁻⁵ The tumour-promoting nature of bile, especially that of secondary bile acids, is well established.^{1,6} Additionally, the development of gallstones (particularly the more common cholesterol-based stone) has also been hypothesized to be the result of altered bile composition.⁷⁻⁹ In individuals with intact gallbladders, the colon is buffered from the damaging influence of the bile acids, as the latter are released from the gallbladder in response to fat in the duodenum and are diluted by other food in the intestine. This protective mechanism is lost with the removal of the gallbladder, because bile is released continuously. Aside from the usual caustic effect, the composition of the bile changes to become increasingly potent. These physiologic changes form the basis for the study hypothesis:

Individuals who have undergone cholecystectomy for benign gallbladder disease have a significantly increased risk for subsequent development of colon cancer.

The specific objectives of the study were:

1. To examine the risk of colon cancer following cholecystectomy, for both genders combined, for both genders separately and for each tumour site (left, right and mid-colon).
2. To evaluate the effectiveness of two types of record linkage as methods for cohort follow-up using Alberta data.
3. To provide a statistical model for determining the profile of individuals most at risk for developing colon cancer following cholecystectomy.

A non-concurrent cohort design was used to investigate the hypothesized association, by following a population-based cohort of patients who had cholecystectomy in Alberta between 1973 and 1992, in two population registries, the Alberta Cancer Registry and the Alberta Health Care Insurance Plan registrant database. The cohort's incidence of colon cancer was determined using automated record linkage, a technique shown to have great potential in cancer epidemiology. The observed incidence of colon cancer in the cholecystectomy cohort was compared to the expected incidence in the general population and in another insurance plan-derived comparison group (individuals undergoing stripping and ligation of varicose veins), using standardized incidence ratios for induction periods of 0, 1, 2, 5, 10 and 15 years.

The study design incorporated the additional comparison group (the "varicose vein" cohort) to address concerns about bias and confounding in the disease-exposure relationship. With the general population as the comparison group, potential bias existed because of differing data collection methods for the numerator and the

denominator of the standardized incidence ratio. That is, record linkage was used to determine the observed number of individuals with colon cancer, but the expected number of individuals with cancer were based on the rate of cancer in the general population, which did not require file linkage. By using linkage to ascertain the outcomes for both the cholecystectomy cohort (which became the observed cases in the numerator) and the varicose vein cohort (which became the rate base for determining the expected number of cases in the denominator), any effect of bias that might be associated with the determination of outcomes was minimized.

A main confounder in a study of cholecystectomy and colon cancer is dietary fat. In this study, the role of fat was partially addressed with the varicose vein cohort, which included individuals with an age and sex distribution similar to that of the population with cholecystectomy. In addition, obesity, gender, and reproductive factors (especially number of pregnancies), could also be risk factors for varicose veins and potential confounding factors in the cholecystectomy-colon cancer association. By comparing the risk estimates for the varicose vein cohort, the cholecystectomy cohort, and the general population, cholecystectomy's role in colon cancer was assessed with some control of these other risk factors. Further investigation of the confounding effect of fat was originally planned, based on data collected for an unpublished colon cancer case-control study, which included information about diet, gallbladder disease and cholecystectomy. However, methodological problems (notably lack of a probability model to connect the case-control sample to the cohort) led to the decision to abandon further analysis of these

data.

In summary, this study was designed to evaluate the effect of a common treatment on a specific health outcome, using the largest population-based cohort reported to date in the cholecystectomy-colon cancer literature. Benefits of the study included long follow-up and the use of two comparison groups to address issues of confounding and generalizability. In addition, the study investigated the efficacy of record linkage in cancer research using Alberta's administrative databases. These features enabled the study to enhance knowledge of the effect of cholecystectomy on colon cancer incidence and the use of administrative data in health research.

CHAPTER II. LITERATURE REVIEW

A. The Biological Perspective: Cholecystectomy and Colon Cancer

1. *Cholecystectomy and the Epidemiology of Gallstone Disease*

Cholecystectomy is a common procedure in the western world. In Alberta between 1973 and 1992, over 90,000 patients underwent this operation, averaging approximately 4,500 procedures per year. In that period, the rate of cholecystectomy performed in Alberta varied somewhat: the age-standardized surgical rates per 100,000 population in Alberta were 282.1 in 1976, 188.7 in 1981-82, and 209.0 in 1985-86.¹⁰

Cholecystectomy is generally performed in response to cholelithiasis, that is, the presence of gallstones in the gallbladder. Data from the Framingham study suggests the prevalence of gallstones in 55-65 year-olds in the United States is approximately 10% in men and 20% in women.¹¹ Certain other sub-groups are at higher risk for gallstones, including North American aboriginals,^{12,13} diabetics,¹⁴ and patients with other digestive disorders such as cirrhosis of the liver¹⁵ and regional enteritis.^{16,17} Additionally, there is evidence that cholelithiasis is associated with obesity,^{11,18,25} but the relationship is not necessarily linear,¹⁸ nor is it always strongest for the highest categories of body mass.^{11,19} Some studies of women have reported that the strongest association with obesity is in the younger age groups.^{18,20,25}

Since a higher prevalence of cholelithiasis is found in women, the role of

hormones has also been investigated. The use of supplemental estrogens has been associated with increased rates of cholecystectomy^{26,27} and there has been concern that oral contraceptives may increase risk for gallbladder disease.²⁸⁻³⁰ Natural exposure to high hormone levels (as measured by multiparity), is also considered a risk factor.²⁵

In short, epidemiologic studies have led to the "5 Fs" profile of the typical Western patient with gallstone disease: fair, fat, forty, fertile and female. In North American and European patients, gallstones are formed mostly of cholesterol. The other type of gallstone is composed primarily of bilirubinate, but is relatively rare outside Asia. Contrary to the profile of risks for the formation of cholesterol stones, risks for developing bilirubinate stones include exposure to *Escherichia coli* and the presence of calcium carbonate, calcium chloride and high molecular weight organic compounds in the bile.³¹

Three mechanisms are thought to play a role in cholesterol gallstone formation. First, bile's ability to dissolve cholesterol may be exceeded so that microcrystals of cholesterol precipitate from the solution. Such action produces bile which is supersaturated due to increased cholesterol and/or decreased concentrations of bile salts or lecithin.⁷ It has been suggested that people who form cholesterol gallstones have abnormal bile, with a significantly decreased total bile salt pool and lower daily bile salt excretions.⁸ This would lead to lower excretion of lecithin from the liver,⁹ which would lead to a relative excess of cholesterol in the bile. Johnson and Kaplan noted that cholesterol stones do not always form in patients with supersaturated bile and a second step is required for gallstone formation.³² A

nucleation catalyst is needed to combine the cholesterol microcrystals into a macroscopic stone. Although the actual gallstone growth process is not well understood,⁷ candidates for the essential pronucleating agent include mucous glycoproteins (mucin), calcium, bilirubin and small molecular weight proteins.³³ Regardless of which agent is at work, patients who form gallstones have bile with a shorter nucleation time.³⁴ Part of the reason for decreased nucleation time involves gallbladder hypomotility, the last mechanism associated with gallstone development. Gallbladder hypomotility has been shown to precede stone formation,^{35,36} and, although it is not the primary cause of gallstones, it is likely to promote gallstone formation.³⁷

The factors associated with increased risk of gallstones in epidemiologic studies can be attributed to changes in bile that are conducive to gallstone formation. In women, the lithogenicity of bile fluctuates during the menstrual cycle and increases during pregnancy and high-dose exogenous estrogen therapy. Female hormones enhance the liver's uptake of lipoproteins, supported by animal studies showing an increased expression of low-density lipoprotein receptors following estrogen treatment.^{38,39} Further, Kern and Everson suggested that contraceptive steroids exert their effect on biliary cholesterol by regulating cholesterol flow into and out of the hepatic cholesterol pool.⁴⁰

2. *The Epidemiology of Colon Cancer*

Colon cancer is a major public health concern in the western world. In

Canada in 1997, estimates show that 12.5% of all new cases of cancer (excluding non-melanoma skin cancer) were colorectal cancer, for an incidence rate of 50 per 100,000 and a mortality rate of approximately 18.5 per 100,000 (rates are age-standardized to the 1991 Canadian population).⁴¹ In the last decade, estimates of the lifetime probability of developing colorectal cancer have been over 6%, while the probability of dying of colorectal cancer is almost 3%.^{41,42}

Various factors have been associated with colon cancer incidence, including increasing age⁴³ or earlier birth cohort^{44,45} and increased body mass.⁴⁶⁻⁴⁸ Risk factors have been shown to vary by colon subsite.⁴⁹⁻⁵² However, the major focus of colon cancer epidemiology has been on diet, with increased risk associated with high consumption of fat (especially of animal origin)^{53,54} and low consumption of fibre.^{55,56}

The mechanisms for the effect of diet have not been completely resolved. For dietary fat, the consensus is that a high intake of animal fat and cholesterol changes the composition of bile acids and neutral sterols in the large bowel, thus modifying the bacteria in the colon. These compounds are thought to be transformed into carcinogenic secondary bile acids and cholesterol metabolites.^{57,58} Weisburger noted in his review that omega-6-fatty acids stimulate the enzymes producing bile acids from cholesterol and that bile acids have a cytotoxic effect that results in tumour promotion in the colon.⁵⁹

The protection offered by high fibre diets was hypothesized by Burkitt.^{60,61} Fibre decreases intestinal transit time, thereby reducing the contact time between tumour carcinogens and promoters and the mucosa and increasing the water content of

the intestinal lumen, which allows dilution of harmful substances and affects the absorption and excretion of carcinogens and tumour promoters in the colon.^{62,63}

However, Dwyer and Ausman pointed out that attempts to separate the role of dietary fibre from that of dietary fat have not been completely successful; fat is still a potentially strong confounder in the fibre-colon cancer association.⁶⁴ These authors also suggested that high-fibre diets may be low in a variety of other carcinogenic substances, or high in protective substances.⁶⁴

The connection between various dietary elements and colon cancer may involve bile composition as the common factor. Aries and colleagues hypothesized that colorectal cancer is caused by substances produced as a result of colonic bacterial flora's metabolism of a benign substrate.⁶⁵ Based on the diet consumed, concentration of the substrate, composition of the flora and metabolic activity of the flora changes and effectively influences the amount of carcinogenic metabolite. Subsequent work focused on bile acids as the substrates which are metabolized by the flora, producing carcinogenic secondary bile acids (deoxycholic acid and lithocholic acid) and a host of other carcinogenic substances including keto bile acids, sulphate esters, unsaturated bile acids and allo bile acids.⁶

3. *Evidence for the Association between Cholecystectomy and Colon Cancer*

As noted, the growth of malignant tumours in the colon has been linked to factors associated with composition and transit of bile acids in the bowel. Compared to residents of developing nations, people in industrialized nations have higher

concentrations of secondary bile acids in their feces and higher incidence of colon cancer.⁶⁶ Fecal concentrations of secondary bile acids are higher in people with colon cancer than in people without and the association may be strongest for women (80% of women and 65% of men with colon cancer have been shown to have high fecal bile acid levels).⁶⁷ Hill suggested that the carcinogenic effect of bile acids is due to stimulation of the growth of small benign adenomas, with a corresponding increased risk of malignant change.⁶ In a summary of other research, McMichael and Potter concluded that, in the colon, higher levels of bile acids enhance epithelial cell proliferation and tumour yield, while the concentration of secondary bile acids influences the rate of progression of carcinogenesis.¹

Cholecystectomy alters the profile and circulation of the bile acids, producing a carcinogenic environment. When the gallbladder is removed, regulation of the transit of bile acids through the intestinal tract is lost, leading to more continuous secretion.¹ A greater amount of bile enters the intestine unaccompanied by food, and the composition of total bile acids changes because the bile acid pool circulates faster, as there is no gallbladder to slow the process, so that bile acids are exposed to intestinal bacteria for longer periods. This exposure leads to increased levels of secondary bile acids that return to the liver in the enterohepatic circulation, resulting in decreased synthesis of primary bile acids due to feedback inhibition.^{2,3} Of note, the excretion of carcinogenic deoxycholic acid has been shown to increase significantly.^{4,5}

These effects have led to interest in evaluating cholecystectomy as it relates to the development of colon cancer. Werner et al documented a 70% incidence of colon

carcinomas in rodents with cholecystectomy, compared to a 16% incidence in rodents without.⁶⁸ Hickman et al found that cholecystectomy induced pre-neoplastic changes in the murine colonic crypt.⁶⁹ Other animal-based studies suggested that cholecystectomy enhances tumorigenesis in the presence of other carcinogens.^{70,71} In humans, the mitotic index of colonic crypt epithelium is higher following cholecystectomy, demonstrating that cholecystectomy is associated with enhanced proliferative activity of the colonic mucosa, which is associated with cancer promotion.⁷²

Several epidemiologic approaches have been used to study the effect of cholecystectomy on colon cancer in human populations, but there is no consensus in the published literature. A number of cross-sectional studies have been published,⁷³⁻⁷⁷ but lack of a control group and the inability to ensure that the exposure precedes the outcome severely limit the conclusions that can be drawn from this body of evidence. However, it is interesting to note that two of these studies reported a gradient of colorectal cancer risk associated with cholecystectomy, with a predilection for tumours in the proximal colon.^{76,77} Seven autopsy studies have been published with similarly mixed findings.⁷⁸⁻⁸⁴ As with correlational studies, autopsy studies are unable to permit causal inferences and are also frequently influenced by selection bias. However, two of the reviewed studies reported an association between cholecystectomy and proximal colon cancer in women.^{82,83}

More than 30 case-control studies have been published on the topic of cholecystectomy and colon cancer. Nineteen were hospital-based studies; of these,

nine showed a significant association,⁸⁵⁻⁹³ with three observing an increased risk of proximal colon cancer in women.^{85,90,92} There are several possible explanations for the negative findings of the remaining ten hospital-based studies,⁹⁴⁻¹⁰³ including the over-matching problems inherent in hospital controls and insufficient sample size, leading to inadequate statistical power. In fact, the majority of the studies reporting nonsignificant findings (7 out of 10), used fewer than 600 subjects, which may be insufficient to detect a moderate risk, while the majority reporting significant findings (7 out of 9), had more than 600 subjects. Of eight other case-control studies using population controls,¹⁰⁴⁻¹¹² only one found an increased risk for colon cancer, confined to proximal cancer only.¹⁰⁴ Most of the studies finding no increased risk used colorectal cancers and not just colon cancer alone, but lower rectal cancer incidence may obscure higher colon cancer risk in the colorectal cancer rate. The only Canadian work published to date observed a significant reduction in colorectal cancer risk following cholecystectomy in females.¹¹² There is no obvious reason for this inconsistent finding.

Under the circumstances, the cohort study is the strongest feasible epidemiologic design. However, the ten published cohort studies did not achieve clear consensus, with four studies finding a significant association that was particularly evident in women and for proximal colon cancer.¹¹³⁻¹¹⁶ The remaining six studies showed no association.¹¹⁷⁻¹²² Null findings cannot be dismissed solely on the basis of insufficient power due to inadequate sample size, because two of the studies are based on a Swedish cohort of more than 16,000 subjects.^{117,118} However, it is

interesting to note that the study by Ekblom and colleagues is an expansion of this cohort and with over 60,000 subjects, these authors showed a significant association for proximal colon cancer in women 15 years following cholecystectomy.¹¹³ The previous studies may have had insufficient power to detect this subgroup's risk because there were relatively few women with sufficient years at risk.

A recent Dutch study used a case-cohort approach to investigate the association between colon cancer and cholecystectomy.¹²³ In this approach, all cases are combined with a subset of the remainder of the cohort; the relative risk is estimated based on the maximum likelihood function. From an original cohort of over 100,000 subjects, almost 4,000 were included for the analysis, which demonstrated a significant increase in risk for both men and women. In women, risk was particularly high for proximal colon cancer.

A recent meta-analysis summarized much of the epidemiologic research.¹²⁴ Risk estimates from 38 studies (5 cohort studies and 33 case-control studies), were pooled, with a resulting significant colorectal cancer risk of 1.21 for males (95 % confidence interval: 1.04-1.40) and 1.24 for females (95 % confidence interval: 1.10-1.40). In case-control studies providing information about colon cancer subsite following cholecystectomy, risk for proximal colon cancer was 1.88 (95 % confidence interval: 1.54-2.30), but the risk for distal colorectal cancer was not increased. In addition, the meta-analysis reviewed several studies of cholecystectomy and colon adenomas, concluding that there was evidence for an increased risk in colon polyp growth (especially for women), approximately 10 years post-cholecystectomy. Given

that polyps are generally accepted to be precursors of carcinoma in the colon, this evidence suggests that the carcinogenic process is well established at 10 years following surgery. It also suggests that the appropriate time to investigate the association between cholecystectomy and colon cancer is at least 10 years post-cholecystectomy, to allow time for tumour formation.

Several design issues have contributed to the inconsistency of the reviewed findings. As noted, sample size is inadequate in many of the studies, particularly when the study group is subdivided by several factors of interest (e.g., by gender, age, and colon subsite). Insufficient numbers lead to inadequate statistical power, so that real differences may be obscured by chance.

The interval time between cholecystectomy and diagnosis of colon cancer is also an issue that requires further attention in subsequent research. Short interval times do not allow for a suitable induction period and it is likely that some tumours are diagnosed shortly after cholecystectomy because of medical attention (ascertainment bias). Additionally, a risk may not be observed even in studies with relatively long intervals, such as that of Adami and colleagues,^{117,118} because of temporary changes that delay the beginning of the induction period. For example, if patients decrease the amount of dietary fat consumed for a period post-cholecystectomy, their colon cancer risk may decrease even in the presence of altered bile acids resulting from cholecystectomy. Returning to a high fat diet could then exacerbate the bile acid imbalance and increase risk after a delay.

The spectrum of variables investigated may have played a role in the

inconsistency of the results. Most studies examined differences in risk associated with colon subsite, patient age and gender, but tumour histology and stage at diagnosis were often not considered, especially with sufficient cases in each stratum to detect significant risk.

There has also been some evidence that gallstones themselves, form the integral step in the carcinogenic process, not cholecystectomy. The bile acid profile has been shown to change dramatically following the development of gallstones as well as following cholecystectomy, as is shown in Table 1. A number of epidemiologic studies have investigated the association between gallstones and colon cancer, with the data summarized in the meta-analysis described previously.¹²⁴ A significantly increased risk for colorectal cancer was reported following detection of gallstones (relative risk = 1.24), with a higher risk for proximal colon cancer (relative risk = 1.55).

In summary, a reasonable mechanism (altered bile acid composition and circulation), has been proposed to explain an association between cholecystectomy and colon cancer. However, the epidemiologic evidence is not consistent, largely because of differences in study design, populations, attention to bias and confounding, sample size and power. There is some evidence suggesting that cholecystectomy does increase colon cancer risk, especially in the proximal colon and in women. The risk may be difficult to detect since it is only modestly elevated, with estimates of an increase of 20% for all colorectal cancer and almost 90% for proximal colon tumours.

Table 1. Bile Acid Composition in Normal Patients and in Gallstone Patients Before and After Cholecystectomy (based on data in Bouchier¹²⁵ and Almond et al⁴)

Bile Acid	Bile Acid Composition		
	Normal Patients ¹²⁵	Gallstone Patients ⁴	
		Before Cholecystectomy	After Cholecystectomy
Cholic	45 %	34 %	28.5 %
Chenodeoxycholic	35 %	43 %	38.5 %
Deoxycholic	15 %	20 %	30 %
Lithocholic	5 %	3 %	3 %

4. *Gastric Procedures as Confounders in the Cholecystectomy-Colon Cancer Relationship*

Under the altered bile acid hypothesis, the cholecystectomy-colon cancer relationship could be confounded by certain gastric procedures. These procedures may be performed at the same time as cholecystectomy and may independently alter bile composition and transit. Gastric ulcer procedures involving vagotomy may alter bile habits, potentially changing the bile acid profile^{126,127} and biliary kinetics.¹²⁸ Conversely, hiatus hernia repair (fundoplication), can involve reduced gastric motility¹²⁹ and without allowing drainage, biliary transit time may be increased. The extent to which these procedures are performed simultaneously with cholecystectomy is unknown.

5. *Stripping and Ligation of Varicose Veins as a Comparison Procedure*

An additional comparison group is useful where there are concerns about bias or confounding in the exposure-disease relationship. Given that linkage is used to address cholecystectomy as a risk for colon cancer, the effects of one possible source of bias and one potential confounder may be alleviated by comparing the colon cancer experience of the cholecystectomy cohort to that of another cohort of patients undergoing surgery. A suitable comparison group could include individuals who have had ligation and stripping of varicose veins. This group is more like the cholecystectomy cohort than is the general population, as indicated by a similar age and sex distribution.¹⁰ Further, it has been suggested that obesity (and consumption

of dietary fat), gender and reproductive factors (especially childbirth) are risk factors for varicose veins,^{130,131} and may also be confounding factors in the cholecystectomy-colon cancer association. By comparing the risk estimates for the varicose vein cohort, the cholecystectomy cohort and the general population, it should be more feasible to assess the role of cholecystectomy in colon cancer while minimizing the influence of these other risk factors. If the stated hypothesis is true, colon cancer will still be more prevalent in the cholecystectomy cohort.

Using another procedure as a comparison group has the additional advantage of avoiding bias by collecting data for the risk estimate's numerator and denominator in a similar way. Any bias that might be associated with record linkage can be minimized if linkage is used to ascertain outcomes for both the cholecystectomy cohort (contributing to the numerator as the observed number of individuals with colon cancer) and the varicose vein cohort (contributing to the denominator as the rates that determine the expected number of individuals with colon cancer).

B. Methods: Linking Administrative Data for Health Research

Stated most simply, record linkage involves matching records from each of two files such that the union represents the experience of one individual. The two files are merged such that records referring to the same individual are connected, while records without corresponding mates remain separate, with minimal misclassification. Each file must be arranged in individual-specific rows (records),

with fields (variables), containing the potential identifying attributes of the individual.

The use of record linkage in the medical environment is not a novel phenomenon, although its popularity has increased significantly in the past few years. The term "record linkage" was initially used by Dunn 50 years ago, in a paper describing the creation of individual "books of life" to be used for administrative and statistical purposes.¹³²

Record linkage has been used in Canadian health research largely because of the existence of large administrative databases that support the government-run, population-based health care insurance plans. These databases contain potential exposure and outcome information, especially for discrete events, such as surgery.¹³³ Provincial and federal databases, including vital statistics registers and cancer registries, provide additional outcome information. Recent publications have focused on the clinical effectiveness of the Papanicolaou smear for cervical cancer,^{134,135} the influenza vaccine¹³⁶ and breast cancer screening.¹³⁷ Health risks following tubal sterilization¹³⁸ and tonsillectomy¹³⁹ have also been investigated. Mortality in certain occupational cohorts has been ascertained using linkage, notably synthetic textile workers,¹⁴⁰ petroleum industry employees¹⁴¹ and uranium miners.¹⁴²

In times of economic restraint, the main advantage of record linkage is its cost-effectiveness. The greatest efficiency gains occur when large, routinely collected administrative databases are used as data sources. These files contain data collected for managerial purposes, such as physician and hospital billing under Canadian provincial health insurance plans.¹³⁴ Discrete events (e.g., surgery or other health

care service) and longitudinal analyses have been shown to be particularly amenable to processing by record linkage, with follow-up rates as good as or better than those from primary data collection.¹⁴³ For chronic diseases, where the elapsed time between exposure and disease is many years, record linkage is attractive because a study can be completed in less time with fewer resources than are required by traditional epidemiologic methods. Researchers are able to study large numbers of people at relatively low cost because less time is required for the identification of eligible subjects and the acquisition of the necessary information about exposure and subsequent disease.

The main disadvantages of using administrative data are that the investigator is not involved with data collection and therefore cannot be sure of data quality and data on confounders may not be available.¹⁴⁴ In addition, linkage techniques do not protect the chronic disease epidemiologist from the risks associated with losing subjects because identifiers may change during the interval between initial exposure and the development of the outcome (e.g., changes in marital status and surname). Such variations may interfere with the ability of record linkage to connect records pertaining to the same individual.

1. Basic Steps

In general, two approaches to record linkage have been identified, deterministic and probabilistic linkage. Deterministic linkage involves a comparison of the variables in each candidate record pair and generates matched pairs based on

the number of agreements between identifiers in the two files. This approach provides a simple categorical assessment of the likelihood of a true match (i.e., yes, no, possible). Probabilistic linkage carries the process further, using more of the information in the data to provide a numeric estimate of the likelihood that the records are a true match. Each method has advantages. Deterministic linkage is relatively simple and is used primarily when data are known to be complete and to have low levels of coding errors. Since probabilistic linkage involves generating weights (probabilities) for each potential link, this method is most advantageous when few variables are linked, data are incomplete or coding errors are common.¹⁴⁵ However, probabilistic linkage is much more complex than deterministic linkage and so accuracy versus simplicity often becomes an important trade-off.¹⁴⁵⁻¹⁴⁷

Regardless of the method used, there are basic steps that guide the linkage process. In an early work on record linkage, Newcombe suggested that there are two primary steps required in any linkage: searching and matching.¹⁴⁸ In the searching step, the aim is to limit the number of times potentially linkable records are not compared, while limiting the number of comparisons (i.e., optimizing the search). In the matching step, rules are applied to determine whether or not a pair of records refers to the same individual, given that some of the personal information agrees and some disagrees. There is also a third step in record linkage, in which linked pairs are separated from unlinked pairs. In probabilistic linkage, explicit thresholds are determined to identify candidate pairs as linked or unlinked.

a. **Searching**

The searching step involves the use of blocks, effectively sorting the files on one or more variables to make the search more efficient. In the usual application, since the discovery of a true pair is a reasonably rare event and the variables are arranged in 2^n configurations (for n fields), the sample size required to ensure that all truly linked pairs are identified approaches all possible pairs.¹⁴⁹ For example, if Files A and B both have 10 records and 4 linkage variables there would be a total of 100 candidate pairs, with 16 possible configurations. Following the implementation of blocks, the discovery of linked pairs becomes a more frequent event in the delimited subset of $A \times B$.

To increase the searching efficiency, the blocking variables must be information-rich. In addition, the variables chosen as blocking variables must be of high quality, as the determination of a link is dependent on an exact match on the blocking variable. For example, given the candidate pairs

	<i>Birthdate</i>	<i>Name</i>
A.	1975.12.01	Honeydew, Bunsen
	1975.05.31	Beaker
B.	1981.03.15	Frog, Kermit The
	1982.03.15	Frog, Kermit The

blocking on birthyear would allow the comparison of the records in pair A because the birthyear agrees, but there would be no comparison of the records in pair B because birthyear does not agree, even though all other information is identical.

Several measures can be calculated to assess the usefulness of potential blocking variables, including the average number of cases per variable level (pocket size), the discriminating power, the Shannon entropy statistic and the merit ratio.

Average pocket size is determined by

$$P_s = n/i$$

where n is the total file size and i is the number of levels in a given variable.¹⁴⁵ For example, in a file with 20 observations and 12 possible values for birthmonth (i.e., January through December), then the average pocket size is 1.67. In the most ideal case, the pocket size is 1: i.e., the average pocket has only one observation for each level of the variable.

The main limitation to the pocket size is that it does not reflect the distribution of a variable's values. For example, in 100 observations, there may be 99 males and 1 female, 75 males and 25 females, or 10 males and 90 females, but the average pocket size is always 50. Other measures compensate for this lack of information on the frequency distribution.

Discriminating power provides an indication of how well a particular variable distinguishes between records representing different individuals. The discriminating power (D_p) is defined as

$$D_p = \log_2(1/C_s)$$

where C_s is the coefficient of specificity, which represents the extent to which a file will be divided by a particular variable.¹⁴⁸ The coefficient of specificity is defined as

$$C_s = \sum P_x^2$$

where P_x is the proportion of the file in the x th block.¹⁴⁸ Newcombe notes that the coefficient of specificity is simply a weighted average of the pocket sizes.¹⁴⁸ The larger the discriminating power (or the smaller the coefficient of specificity), the more discriminating the variable.

Returning to birthmonth as an example, if each month were represented equally in the file, then $P_x = 1/12$ for each month and $P_x^2 = 1/144$. The coefficient of specificity is $\Sigma P_x^2 = (12)*(1/144)$, or 0.083. Thus the discriminating power is $\log_2(1/0.083)$, or 3.58.

The Shannon entropy statistic can also be used to determine the amount of information available in any file.¹⁴⁵ The Shannon entropy is defined as

$$S_E = -\Sigma P_x * \log_2 P_x$$

where P_x is the proportion of the file in any given pocket. As with the discriminating power, the larger the Shannon entropy the more informative the variable. The Shannon entropy statistic has an upper bound determined by $\log_2 n$, so that all values of the statistic will be no greater than that value.¹⁵¹

In the birthmonth example used for calculating the discriminating power, the Shannon entropy is $-12*[(1/12)*\log_2(1/12)]$, or 3.58. If n is 12 then the upper bound is 3.58, and if n is 24 then the maximum value is 4.58.

Finally, the benefit of using any particular variable for blocking can be determined by the merit ratio. The ideal blocking variables will be those with the highest merit ratio, as they will be the most reliable, with the fewest discrepancies in correctly matched pairs and considerable discriminating ability. The merit ratio is

calculated by

$$M_t = D_p/I$$

where I is the likelihood of inconsistency (or discrepancy) of the variable in linkable pairs of records.¹⁴⁸ The likelihood of inconsistency is simply the frequency of discrepancies expressed as a percent of all linkable pairs.¹⁴⁸ Shannon entropy may also be used in place of discriminating power.

In the birthmonth example, the merit ratio can be calculated given information indicating the reliability of the data in linkable pairs. From a sample or previous study, it may be known that birthmonth disagrees in truly linked pairs 5 % of the time. Therefore, the merit ratio is $3.58/0.05$, or 71.6.

Determining a minimum set of primary variables maximizes computing efficiency, but the variables should have considerable discriminating ability, as determined by calculating the discriminating power, the Shannon entropy or the merit ratio. Roos and Wajda note that when determining the set of variables to be used for linkage, it is advisable to start with variables that individually have lower values of discriminating power (although in combination they may have considerable power), so that variables with higher levels of discriminating power can be used to resolve ties generated by the linkage.¹⁴⁵ Using this approach, information content would be determined for the primary variables in combination and the remaining variables individually. Since the data will not be error-free, the statistics on the additional variables estimate their ability to deal with the effects of measurement error in the primary variable set.

b. Matching

The matching step requires that the computer is provided sufficient information to imitate the human decision-making process, thereby determining the records most likely to represent linked pairs. If all the variables in the candidate pair agree, then the likelihood that the records relate to the same individual is high; conversely, complete disagreement suggests that the record pair refers to different individuals.¹⁵² Difficulty arises when some of the variables agree but some disagree.

Although determining the best approach for searching is the same for either type of linkage, it is in the matching phase that deterministic and probabilistic linkage differ. In deterministic linkage, if the majority of the variables in the candidate pair agree or if subjectively-determined essential variables agree, then the pair is considered linked. However, in probabilistic linkage, the likelihood that a candidate pair represents a true link is quantified as a weight or probability.

In probabilistic linkage, the question is, "How typical is that comparison outcome among linked pairs of records, as compared with unlinkable pairs brought together at random?"¹⁵⁰ The answer involves calculating the frequency ratio (similar to betting odds), which is defined as

$$\text{Frequency Ratio} = \frac{\text{frequency of outcome (x,y) among linked pairs}}{\text{frequency of outcome (x,y) among unlinkable pairs}}$$

where x is the variable (and its value), for the record from file A, and y is the variable (and its value), for the record from file B.¹⁵⁰ The outcome of interest

involves the occurrence of any specified event and can be agreement or disagreement on a variable or combination of variables.

For example, out of 15 record pairs, one may have 10 potentially linked pairs, while the remaining 5 are unlinkable pairs generated by random merging of records. The outcome of interest is agreement on birthmonth; in the linkable pairs, birthmonth agrees 8 out of 10 times, while in the unlinkable pairs birthmonth agrees only in only 1 out of the 5 pairs. The frequency ratio is $(8/10)/(1/5)$, or 4. Thus, in the situation where birthmonth agrees, the chance that the pair represents a true link is 4:1.

Frequency ratios can also be calculated for configurations involving more than one variable. Extending the previous example, if the outcome of interest involves not only agreement on birthmonth but on birthyear as well, one must look at the frequency of both variables in the linkable and unlinkable pairs. Therefore, if the frequency of agreement on birthyear in the linkable pairs is 9 out of 10, but the agreement in the unlinkable pairs is 2 out of 5, then the frequency ratio for agreement on both birthyear and birthmonth is $(9/10)(8/10)/(2/5)(1/5)$, or 9. The likelihood that any record pair agreeing on both variables is a true link is increased to 9:1.

In practice, both the numerator and the denominator of the frequency ratio are estimated by linkage software programs based on estimates from previous studies or by assuming the frequency in the linkable pairs is similar to one or other of the files and modifying it by the probability of error in that variable (determined empirically or iteratively).^{149,153}

The frequency of the outcome in the unlinkable pairs can be calculated either

by determining the number of disagreements in a file of unlinkable pairs or by estimation. To create a file of unlinkables, Newcombe suggests assigning numbers to the two files of interest at random, sorting the files by these numbers and re-numbering according to their rank in the sorted files.¹⁵⁰ The files can then be merged and any pairs suspected to be true matches removed (e.g., those with similar sounding surnames, as determined by the phonetic soundex code and the same birthdate). On the other hand, estimating the frequency in the unlinkables involves some knowledge of the characteristics of the variables in question (i.e., numeric or character type, range and distribution), in order to generate the probability that a variable will disagree in any candidate pair of records. Since the probability of finding a link is a relatively rare event, estimation often assumes a sample size of all possible record pair combinations (i.e., $A \times B$ for Files A and B).

Frequency ratios are considered to be global when value-specific differences in discriminating power are ignored. However, it is often useful to create value-specific frequency ratios, since certain values are more common (such as the name "Smith" as opposed to "Schwarzenegger"), with the result that the associated discriminating powers provide different degrees of information. In addition, a certain degree of flexibility can be incorporated into the calculation of frequency ratios, so that instead of a simple binary structure (agree/disagree), the frequency ratios are stratified (fully agree/partially agree/disagree).¹⁵⁰

When the logarithm of the frequency ratio is calculated, the result is called a weight. The weights are usually defined as

$$\text{weight} = \log_2 \frac{(\text{"outcome" frequency in linked pairs})}{(\text{"outcome" frequency in unlinked pairs})}$$

One of the benefits of taking the logarithm is that the likelihood that two records refer to one individual can be expressed by summing the variable-specific weights to determine the overall weight. This is equivalent to multiplying all the frequency ratios associated with a pair of records, assuming that the various agreements and disagreements are independent of each other.¹⁵⁰ Recalling the frequency ratio examples, the weight for agreement on birthmonth is $\log_2(4)$, or 2.0, and on both birthyear and birthmonth is $\log_2(9)$, or 3.17. A weight of zero is equivalent to odds of 1:1 that linkage is correct, while a weight of +1 suggests odds of 2:1 and +2 is odds of 4:1. Negative odds halve the probability of a true link (e.g., -1 = 1:2).¹⁴² The overall weight is a relative measure of the probability that a pair of records refer to one individual, rather than an absolute measure.

c. Separating Links from Nonlinks

Regardless of the type of linkage, the process of separating linked pairs from unlinked pairs involves comparing the respective variables within each record pair to determine whether the records are linked or not. In deterministic linkage overall agreement is based on simple inspection of the variables. Following the calculation of odds in probabilistic linkage, a level of odds (the threshold) is set, to separate the linked pairs from the unlinked pairs. Newcombe notes that,

"It is not so much a matter of picking needles out of a haystack, as of progressively getting rid of the haystack without losing the needles."¹⁵⁰

There are no valid shortcuts in determining true links from false links.

Newcombe cautions that decisions must be made with a frequency ratio or odds, but not with the ratio's components alone.¹⁵⁰ In theory, the threshold is set where the absolute odds is 50:50, but in practice the threshold may have to be set above or below that point in order to reach a desired ratio of false positive and false negative links.¹⁵⁰ The final threshold is often determined empirically, based on simple inspection of the set of potentially linked pairs.¹⁵³

It may be more practical to assign two thresholds instead of one: candidate pairs with weights greater than the upper threshold are considered to be linked, while pairs with weights below the lower threshold are not linked. The pairs with weights between the thresholds exist in a grey zone and require verification. Based on work by Fellegi and Sunter,¹⁵⁴ Jaro proposed an algorithm which allows the calculation of threshold weights.¹⁴⁹ M is the set of all truly linked pairs, while U is the set of truly unlinked pairs. There are 2^n possible combinations (agreements and disagreements), of n components (variables), for which the composite weights can be determined.

The maximum weight for an unlinkable pair is the weight of the configuration where

$$\sum \Pr(\bullet \mid M) < \text{ideal probability of classifying a true link as unlinked} \\ \text{(false negative rate);}$$

the minimum weight for a linkable pair is the weight of the configuration where

$$1 - \sum \Pr(\bullet \mid U) < \text{ideal probability of classifying a truly unlinked pair as}$$

linked (false positive rate).

Weights between these two thresholds are the undecided cases. One could argue that the ideal probability of any misclassification is 0, but the number of candidate pairs with indeterminate weights may be too large to be manually verified efficiently.

For example, following linkage it may be observed that the candidate pairs have weights between +4 and -4. The truly linked pairs and the truly unlinked pairs occur with the frequency shown in Table 2. Without additional information, it would be difficult to assign a candidate pair with a weight between -2 and +1 to either the linked or unlinked categories with certainty. If 20% is the acceptable level of false negatives, then the weight for the lower threshold is $\Sigma \Pr(\bullet | M) < 20\%$, the cumulative frequency in the linked pairs that is still less than 20%. From the table, this point is -2. If the acceptable level of false positives is also 20%, then the upper threshold is $(1 - \Sigma \Pr(\bullet | U)) < 20\%$. The point (1 - the cumulative frequency of the unlinked pairs) at which the weight is still less than 20% is at -1. Therefore, all candidate pairs with weights between -2 and -1 are in the grey zone and require verification.

2. *Errors Influencing the Linkage Process*

Random errors arise as a result of incorrect data entry or lack of entry of available information. Systematic errors occur where the original source of the information does not reflect the true experience of the subject. While random error affects reliability, systematic error (bias) is considered to be more serious, since it can

Table 2. Threshold Example: Distribution of Weights for Links and Nonlinks

Weight	True Links	Cumulative Frequency of True Links (%)	True Nonlinks	Cumulative Frequency of True Nonlinks (%)
-4	0	0	2	20
-2	1	10	5	70
-1	0	10	2	90
+1	0	10	1	100
+2	3	40	0	100
+3	4	80	0	100
+4	2	100	0	100

affect validity. It has been suggested that error is a particular concern in medical research using computer data where a disease is difficult to diagnose because the errors tend to be systematic, which threaten the study's validity.¹⁴⁴

Numbers are prone to several common random errors: transcription or substitution errors, where digits are incorrectly recorded because of mishearing, misreading or miskeying; transposition errors, in which correct digits are entered in the wrong order; and shift errors, which occur as a result of addition or omission of zeros.¹⁵⁵ When the numbers are part of an identifying system, the error rates can be minimized by the incorporation of a check-digit routine. However, similar errors could occur in any numeric or character variables where such checks would be ineffective.

In the actual process of linkage, other errors can occur. The three main problems are issues associated with blocking, correlated identifiers and thresholds.

Blocking is a potential source of error if the blocking variables of a truly linked pair disagree so that the records are never brought together for comparison. To minimize the effect of the trade-off inherent in blocking (loss of true links versus resource consumption), it is important to estimate the reliability and discriminating power of candidate linkage variables, perhaps using merit ratios.¹⁵⁰ Losses due to blocking can be estimated by comparing every search record with every record in the file being searched; although this is ideal, it is often impossible with finite computing resources. Alternatives include the use of a different file (e.g., tax records to determine vital status) and the use of an extended search (e.g., just one variable)

combined with an alternative search (e.g., different search criteria) so that the linkage net is cast widely. In practice, Newcombe notes that the most useful blocking variables are likely to be a personal number, a second surname, or the first given name plus the date of birth.¹⁵⁰

Correlation between variables can lead to problems in determining weights because they interfere with the assumption of independence. Correlated discrepancies are those where the informant is confused or deliberately gives incorrect answers, resulting in a downward bias of the odds. Conversely, correlated agreement occurs among variables where identifying variables are likely to occur together. For example, in minority geographical and ethnic groups, names and places of birth are likely to be common within the group, even if these values are rare in the overall study group. If variables are known to be highly correlated, then a multiple agreement should not be weighted heavily, particularly when assessing links with only moderate weight.

Following linkage, errors can occur if the threshold denoting the acceptance/rejection cutoff is too high so that there may be an increase in false negatives, or too low so that there are increased numbers of false positives. The potential errors associated with the threshold are directly related to errors in discriminating power.¹⁵⁰ Insufficient discriminating power can lead to higher numbers of false positive and false negative links. High levels of discriminating power leave little doubt as to the best point for the threshold, since the number of true links and truly unlinked records being misclassified is likely to be small.

Because of linkage errors, there may be an imbalance between the number of false positive links and false negative links. Therefore, simplified linkage procedures may be adequate if the investigator is only interested in statistical outcomes because there is an assumed balance between false positives and false negatives. This assumption can be verified by using three different thresholds in analysis, one moderate and two extremes. As in other sensitivity analysis, if the same statistical associations are maintained, the result is likely to be real and not due to an imbalance of false negatives and false positives.

3. *Summary*

Record linkage is a viable methodology that is appealing to many health researchers because of its cost-effectiveness. However, researchers must consider the errors in administrative data that can threaten the validity of the study. These can usually be resolved by careful attention to details of the information provided. Record linkage might actually improve data quality overall by highlighting areas for improvement in routinely-collected data.¹⁴⁵

In many cases, the limitations of record linkage are most evident in situations where the method is inappropriately applied. Research involving high quality data and routinely collected administrative data will be well-served by linkage, while studies based on poor, error-prone data or qualitative outcomes will not.

CHAPTER III. METHODS

In this study, the risk of colon cancer following cholecystectomy was assessed by linking records from the Alberta Health Care Insurance Plan (AHCIP) and records from the Alberta Cancer Registry. Data quality was assessed for both data sources before linkage was initiated. Effectiveness of record linkage was evaluated using a chart review.

Risk estimates were calculated using standardized incidence ratios based on person-years at risk. The risk estimates were subjected to sensitivity analyses to investigate the ability of different record linkage approaches to identify colon cancer in the cholecystectomy cohort and to examine the effect of assumptions about incomplete dates from AHCIP. Modelling of the risk was attempted using proportional hazards regression.

Unless otherwise noted, the statistical software package SAS¹⁵⁶ was used for analysis. Statistical significance was determined using two-tailed p-values with $\alpha=0.05$. No adjustments were made to the significance level to compensate for multiple comparisons.

A. Assessing Data Quality

1. *Data from the Alberta Health Care Insurance Plan*

Two files were provided by AHCIP; the registrant or cross-reference file

described registrants and their health care insurance history in Alberta and the claims file described the procedures performed on these individuals. This reflects the approach to data capture used by AHCIP, in which data pertaining to an individual's eligibility for registration is stored and maintained separately from data concerning payment of service claims submitted by physicians.

a. Registrant File

The original registrant file contained 185,103 records, representing 143,647 individuals. The variables in this file included identifying information, such as the patient's initials, date of birth, sex, and AHCIP registration number, the head registrant's last name and a family status indicator (1 signifying the head registrant and 2 or greater indicating dependents). For each individual, a health care insurance history was provided, including the date that registration was effective, the date it was terminated (termed "cancellation" if the head registrant's information was ended and "deletion" if the termination only applied to a dependent) and the reason for termination. Records relating to the same person were linked by AHCIP and were identified using a unique lifetime identifier. AHCIP has indicated that this internal linkage was more reliable for registrants with active coverage since 1983. Data quality was assessed by examining simple frequencies of the variables, with particular attention to identification of missing or outlying data.

b. Claims File

The original claims file had 164,600 records, representing 143,647 individuals who had undergone specified biliary, gastric and varicose veins procedures in the time period of interest. The variables included the patient's sex, date of birth, registration number and unique lifetime identifier, the date and fiscal year that the procedure was performed, the patient's age at the time of the procedure, the procedure's fee code and the amount paid to the physician by AHCIP. As with the registrant file, the data quality investigation in the claims file focused on simple frequencies of these variables to identify missing data and outliers.

The fee codes, representing all possible codes used between April, 1973 and September, 1993, are shown in Appendix A. The original dataset included procedures such as choledochostomy, transduodenal sphincteroplasty and choledochenterostomy because the descriptions noted they could be performed with or without cholecystectomy. Cholecystostomy, a procedure where gallstones are removed but the gallbladder remains intact, was also included. However, analysis of the cancer risk for this project was limited to the most common situation involving cholecystectomy alone.

c. Combined Information

Simple frequencies of the data in each of the AHCIP files alone illuminated only some of the data quality issues. The files were combined to allow analysis of the service information in relation to registrant history, particularly the frequency of

services that appeared to occur outside coverage and the reasons for termination of coverage in these cases.

2. *Data from the Alberta Cancer Registry*

The Alberta Cancer Registry has existed for over 50 years and, for more than 20 years, reporting of malignancies to the Registry has been required by law in Alberta. Patient information is abstracted and coded by health record technicians and entered into a computer. For this study, all Albertans diagnosed with cancers of the colon, biliary tract or pancreas between July, 1969 and December, 1993 inclusive, were identified through the Alberta Cancer Registry. Colon cancer patients included those diagnosed with cancer of the colon and the rectosigmoid junction but not cancer of the rectum, as rectal cancer is thought to have different risk factors. Patients diagnosed with gallbladder cancer were identified so that those whose diagnosis occurred within 6 months of cholecystectomy could be removed (i.e., the reason for cholecystectomy was malignant and not benign gallbladder disease). Patients diagnosed with pancreatic cancer or cancer of the biliary ducts at any time following cholecystectomy were identified and follow-up was truncated at the date of diagnosis; these patients sometimes undergo surgical treatment that involves removing the gallbladder (e.g., duodenopancreatectomy or Whipple procedure).

Although this population-based registry is relatively complete and of solid quality, frequencies were determined to illustrate the extent of missing patient identifiers and diagnostic data. Method of diagnosis was of particular interest, as the

numbers of histologically verified cases and of cases identified only through death certificates are indicative of registry quality.

B. Record Linkage

As described previously, record linkage has three main steps: searching, matching and separating links from non-links. The two approaches to record linkage, deterministic and probabilistic, benefit from a common strategy for determining the most efficient searching patterns but differ in the matching and separation steps.

For the searching step, the pocket size, discriminating power and Shannon entropy were calculated for variables in the colon cancer file and the entire AHCIP dataset, which included the cholecystectomy cohort as well as the confounding gastric procedures cohort and the comparison varicose vein cohort. Based on the results of these calculations, a blocking strategy was developed to maximize searching efficiency.

In the matching step, both deterministic and probabilistic strategies were applied. Deterministic linkage used the SQL procedure in SAS¹⁵⁶ and probabilistic linkage used the LinkPro macro.¹⁵⁷ Agreement of the results from these two approaches was assessed using the kappa statistic.¹⁵⁸

The approach for separating links and non-links from the pool of candidate pairs differed for deterministic and probabilistic linkage. In deterministic linkage, records agreeing on AHCIP number or all variables except AHCIP number were

presumed to be linked, while records agreeing on fewer than half of the identifying variables (including disagreement on last name soundex and at both initials) were considered to be unlinked. Candidate pairs where a majority of variables agreed, including last name soundex or one initial, were subjected to manual review.

The approach described by Jaro¹⁴⁹ was used to assist with threshold determination for probabilistic linkage. All possible variable configurations for a pair of records were determined (e.g., sex matched, last name matched, and birthdate mismatched), and the frequencies observed by each configuration were calculated for the candidate linked pairs and a sample of randomly linked pairs. The frequencies were then examined by probabilistic weight.

In general, Jaro's approach uses these frequencies to establish an acceptable level of false negatives and an acceptable level of false positives, which determines the maximum threshold weight for the unlinked pairs and the minimum threshold weight from the linked pairs. Manual resolution occurs for all candidate pairs in the matched set with weights between the two thresholds. In this study, the matched set (M), was very large and covered a wide range of weights as a result of the lenient parameters used to ensure that all possible pairs were identified. This led to a substantial number of pairs in M that were unlikely to be true links. Therefore, when the random set of unlinked pairs (U), was generated, the most extreme cases were used to determine the lower threshold so that the probability of false positives was very small ($<0.001\%$). Using this point, the probability of false negatives appeared to be very large (around 70%), but this compensated for the overly-relaxed criteria of

the initial linkage.

In both cases, there were some records that required manual review to establish their status. Two reviewers examined all candidate pairs; concordance between the reviewers was assessed by the kappa statistic. Subsequently, disagreements were resolved by consensus.

Individuals identified with colon cancer in the Alberta Cancer Registry were linked to the AHCIP files first. The experience acquired from this linkage was then applied to linkage between the other relevant cancer cases (biliary and pancreatic) and the AHCIP files.

C. Comparing Linkage to Manual Review

The reliability of record linkage was assessed by examining the charts of a subset of the individuals with colon cancer for a history of cholecystectomy. For convenience, only those individuals with charts at the Cross Cancer Institute in Edmonton, Alberta were considered to be eligible.

All eligible individuals who were found to have cholecystectomy followed by colon cancer in record linkage and who had at least 10 years' induction were included in the review. A similar number of individuals diagnosed with colon cancer but without a history of cholecystectomy (assuming a 10-year induction period), were sampled. Each chart was examined by two reviewers. Concordance between the reviewers was determined using the kappa statistic, and disagreements were resolved

by consensus.

Following the review, individuals without a history of cholecystectomy in the chart (regardless of linkage status) were identified; the attending physician for each of these individuals was contacted to determine the presence or absence of previous cholecystectomy. The kappa statistic was used to quantify agreement between record linkage and the cancer chart review as well as record linkage and the "extended" review involving physician contact.

D. Analysis of Risk

1. Exclusion Criteria

Patients diagnosed with colon cancer before cholecystectomy were excluded from the analysis. In addition, patients diagnosed with gallbladder cancer within 6 months of cholecystectomy were removed from the file of linked records for analysis, since cancer and not gallstones would be the reason for cholecystectomy. Patients diagnosed with pancreatic cancer or cancer of the biliary ducts at any time following cholecystectomy did not contribute to follow-up from the date of diagnosis, since these patients sometimes undergo surgical treatment that involves removing the gallbladder (for example, duodenopancreatectomy or Whipple procedure).

2. Standardized Incidence Ratios

Standardized incidence ratios (SIRs) compare the observed number of

individuals with cancer to the number expected in a comparison population. In this study, the observed number was provided by record linkage. The strategy to determine the expected number of cases involved multiplying the person-years at risk accumulated by the cholecystectomy cohort by the colon cancer rates in a comparison population.

Separate SIRs were calculated for colon subsite (left, mid, and right colon) and each sex. SIRs based on observed values less than 10 were not reported because of concerns regarding the stability of the risk estimate and expected values of 0 were set to a small value (0.1), to allow calculation of an approximate SIR and confidence limits.

a. Person-years at Risk

Person-years at risk began accruing on the date that cholecystectomy occurred and ended at the date of diagnosis of cancer, death or migration from the province. Patients who left the province but returned after less than one year accumulated person-years throughout their absence, since it was assumed that any cancer diagnosed would have been reported to the Alberta Cancer Registry. If the absence was more than one year, person-years were not accumulated during the absence, but resumed on the patient's return to Alberta. Person-years at risk were stored in a matrix that specified the accrual by calendar year, sex and 5-year agegroups.

b. Rates

Two comparison groups were used for assessing risk of colon cancer following cholecystectomy. The general population was the primary comparison group, as is standard practice. Concerns about bias introduced by record linkage led to the identification of a second comparison group, the varicose vein cohort. Colon cancer rates were calculated slightly differently for each of these comparison groups.

All colon cancer cases identified in the Alberta Cancer Registry with diagnosis dates between July, 1969 and December, 1993 were used to calculate cancer rates when the general population was used as the comparison group. These were the same individuals identified in the colon cancer cohort, described previously. Population figures for the rate denominators were from Statistics Canada, received and stored electronically by the Alberta Cancer Board. Rates were determined by dividing the number of cases of cancer by the population for each calendar year, sex and 5-year agegroup category.

To minimize bias introduced by differential methods of collecting data for the numerator and the rates in the denominator, the cancer rates in the varicose vein group was used in a second SIR calculation. By using linkage to ascertain the outcomes for both the cholecystectomy cohort (which become the observed cases in the numerator) and the varicose vein cohort (which become the rate base for determining the expected number of cases in the denominator), any effect of bias that might be associated with the determination of outcomes was minimized. Patients having both cholecystectomy and varicose vein procedures were excluded from this

analysis.

When the varicose vein cohort was used as the comparison group, the number of individuals with colon cancer in the varicose vein cohort formed the numerator of the rate. The denominator was the number of person-years at risk accrued by the individuals in the varicose vein cohort. The rates were determined by dividing the number of colon cancer cases by the total number of person-years at risk accumulated by the varicose vein cohort in calendar year, sex and 5-year agegroup categories.

c. Calculation of Expected Numbers

The expected number of cases was determined by multiplying the calendar year, sex and 5-year agegroup-specific matrix for the cholecystectomy cohort's person-years at risk by the matrix of rates of colon cancer in the comparison population. The matrices' year, sex and age presentation allowed for appropriate year, sex and age-adjustment of the resulting expected numbers and SIRs.

d. Adjustment for Induction

SIRs were determined for 0, 1, 2, 5, 10 and 15 years of induction, meaning that observed cases and person-years at risk were accrued beginning 0, 1, 2, 5, 10 or 15 years respectively following cholecystectomy. Inductions less than 5 years were included for completeness, but were not considered biologically important.

e. Adjustment for Gastric Procedures

Since gastric procedures may be performed at the same time as cholecystectomy and have been shown to alter bile transit and to increase cancer risk, they may confound the association between cholecystectomy and colon cancer. The confounding effect of these procedures was addressed by calculating SIRs adjusted for history of gastric surgery. This was accomplished by stratifying the cholecystectomy cohort's person-years matrix and the rate matrix by history of gastric procedure.

Individuals with a history of both cholecystectomy and gastric surgery were identified from the AHCIP claims file. Individuals accumulated person-years at risk in the unexposed stratum until the date they underwent a gastric procedure, if applicable. Those individuals who had gastric surgery then accumulated person-years at risk in the exposed stratum beginning on the date of surgery.

The rate matrix was also adjusted to account for differing rates of colon cancer in the population according to presence or absence of gastric procedure. As described, the rates of cancer in the general population were calculated by dividing the observed number of colon cancer cases in the population by annual population figures, which are equivalent to person-years at risk. Since record linkage had identified the number of individuals with gastric procedures who subsequently developed colon cancer, the number of individuals who developed colon cancer *without* a history of previous gastric surgery was determined to be the difference between the total number of cases of colon cancer in the population and the number of colon cancer cases in the gastric procedures cohort. This calculation provided the

numerator for the cancer rates in the general population by history of gastric surgery. The denominator was calculated in a similar fashion, with the difference between the total population (i.e., the person-years at risk contributed by all Albertans) and the person-years at risk accrued by the gastric cohort representing the person-years at risk for individuals *without* gastric surgery. Rates were determined by dividing the number of colon cancer cases with and without a history of gastric surgery by the appropriate person-years at risk.

The expected numbers were determined by multiplying the cholecystectomy cohort's stratified person-years at risk matrix by the stratified rate matrix, with both matrices now stratified by calendar year, sex, 5-year agegroup and gastric surgery exposure. This adjustment was only performed using the general population as the comparison group.

f. 95 % Confidence Intervals

The ratio of the observed to expected numbers showed the risk of cancer in the cholecystectomy cohort relative to a comparison population. 95 % confidence intervals (CIs) were determined as described by Bailer and Ederer.¹⁵⁹ Linear interpolation was used to determine limits where the table did not present exact data.

3. *Sensitivity Analysis*

Sensitivity analysis was used to investigate the magnitude and direction of potential bias resulting from assumptions about the data.

First, assumptions were made about individuals missing AHCIP effective dates. It was assumed that those missing these dates were most likely to have coverage beginning at July 1, 1969, the first date of comprehensive health care insurance in Alberta. This decision ensured that the date of service occurred during coverage for most of the affected cohort, but it raised questions about data quality. Therefore, the SIRs were re-calculated using only records with complete effective dates.

The second set of assumptions were made when links were identified. There were cases that were identified by either deterministic or probabilistic linkage but not by both. SIRs were calculated using the individuals identified by deterministic linkage alone, probabilistic linkage alone, or a combination of both linkage methods to determine the extent of the effect of type of linkage.

4. *Proportional Hazards Modelling*

SIRs provide a composite measure of risk. To provide a profile of the cholecystectomy patient at risk for colon cancer, proportional hazards regression was applied to investigate the effect of independent variables, such as age at surgery, year of surgery and sex on the hazard function over time. However, the relative rarity of the outcome and the paucity of explanatory variables precluded meaningful analysis. Thus, the third objective (outlined on page 2) was not pursued further in this study.

5. *Nested Case-control: Fat as a Confounder*

Dietary fat was identified as a significant confounder in the relationship between cholecystectomy and colon cancer. Initially, an adjustment for dietary fat was planned based on data from an unpublished Alberta-based case-control study. As with many published studies of diet and colon cancer, there were concerns about the case-control study's assessments of diet and disease status, but there was also a significant statistical barrier to the effective use of the case-control study to adjust for dietary fat in the cholecystectomy cohort. Most studies that use a nested case-control study to acquire data on confounders use all the cases and randomly select from the controls. This ensures that the case-control study is representative of the larger cohort. Even if a point estimate could be calculated for the current study, the lack of a probability model connecting the case-control and cohort studies means that the corresponding standard error would be a non-trivial calculation.

Therefore, an adjustment based on data collected for the case-control study would be meaningless because the statistical premise for adjusting the cohort risk estimates with the case-control results is weak at best. Dietary confounding would be appropriately assessed using a truly nested case-control study, in which cases and controls would be sampled randomly from the cohort to provide an appropriate probability model for determining the connection between the case-control and cohort risks.

E. Stratified Analysis by History of Colon Cancer and Cholecystectomy

The datasets included additional variables that could provide an indication of which individuals may be at higher risk for developing colon cancer following cholecystectomy. The SIR analysis only adjusted for age, sex and calendar year, with a special calculation performed to investigate the influence of gastric surgery. Therefore, the differences between individuals with cholecystectomy and colon cancer and those without both factors were addressed using variables available in the cohorts. In the cholecystectomy cohort, the average age at surgery, the year of service and sex distribution for individuals who were found to have developed colon cancer were compared to those who were not found to have colon cancer. In the colon cancer cohort, the average age at diagnosis, year of diagnosis, sex distribution, and tumour stage, grade and histology were compared for those with previous cholecystectomy, relative to those without the procedure. Differences by age were assessed by t-tests while differences by other variables were assessed using X^2 tests.

Characteristics were compared for individuals who had the opportunity to be identified with both cholecystectomy and colon cancer for each of the assumed induction periods. For example, at 15 years of induction, the set of individuals in the cholecystectomy cohort without colon cancer was restricted to those with service dates before 1979; similarly, the set of individuals in the cancer cohort without a history of cholecystectomy was restricted to those diagnosed since 1988.

CHAPTER IV. RESULTS

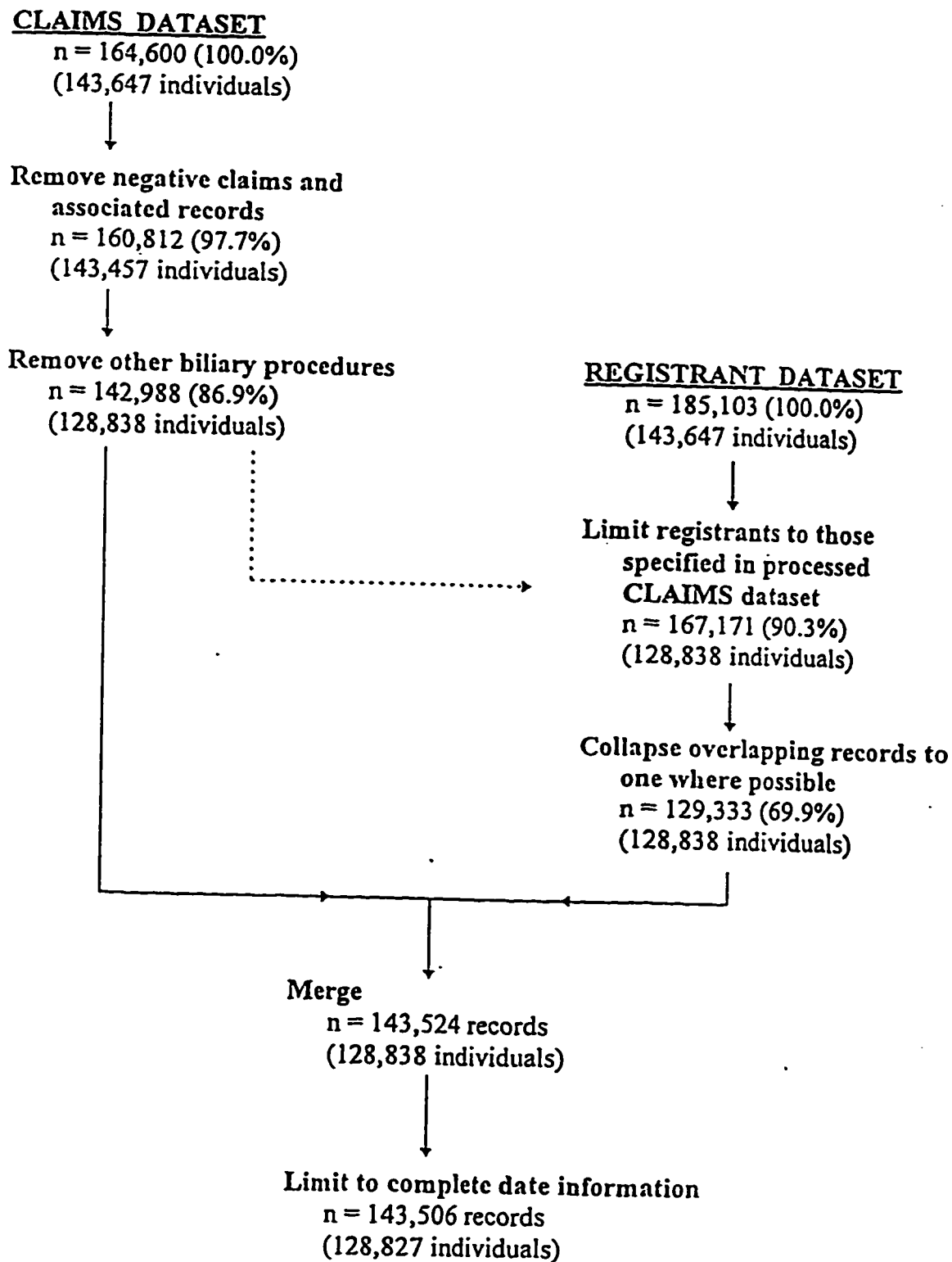
A. Cohort Preparation

1. *Alberta Health Care Insurance Plan Cohorts*

In the original claims dataset, there were 164,600 records, representing 143,647 individuals who received at least one of the specified biliary, varicose vein or gastric procedures between April, 1973 and March, 1993 (inclusive). Each individual identified in the claims dataset had records defining their AHCIP registration history in the registrant dataset, which had 185,103 records. Figure 1 summarizes the processing that occurred for this study.

Initial analysis showed a substantial number of records with a negative amount paid to the physician. Subsequent investigation revealed the accounting format used in the AHCIP claims file, where an incorrect entry would be revised by duplicating the record, adding a negative sign to the amount paid and re-entering the correct record. An algorithm was developed to remove the incorrect entries and their associated negative claims, which involved sorting by the absolute value of the amount paid and cancelling duplicate records. The claims dataset was found to have 1,900 records with negative claims, or 1.2% of the total, resulting from corrections to previously entered information. The algorithm to remove these claims and their associated records resulted in the removal of 3,788 records (2.3% of the total) and 190 patients (0.1% of all individuals).

Figure 1. Summary of Processing: Alberta Health Care Insurance Plan Files 53



An additional 17,824 (10.8% of the total) records were removed because the study excluded the more complicated biliary procedures, representing 14,619 patients (10.2% of all individuals).

After this processing, there were 142,988 (86.9% of the total) records, describing 128,838 patients (89.7% of all individuals). The exclusion criteria applied to the claims dataset reduced the number of records in the registrant dataset to 167,171 (90.3% of the original total). The records in the registrant dataset were collapsed where coverage periods overlapped, resulting in a further reduction to 129,333 (69.9% of the total) records to describe the 128,838 individuals.

The two files were joined using PROC SQL in SAS, which uses a many-to-many merge and resulted in 143,524 records to describe the procedures received and the registrant history simultaneously. Preliminary results showed that 17 records (for 10 individuals) had no date of birth and one record had a date of birth in the early 1800s. These records were removed because the date of birth information is essential in the calculation of person-years at risk, leaving 143,506 records for 128,827 individuals.

The final cohort was associated with 142,970 claims. The majority (92,537, or 64.7%) were cholecystectomy claims, followed by claims for varicose vein procedures (25,424, or 17.8%) and gastric procedures (25,009, or 17.5%). Most individuals (117,358, or 91.1%) had only one claim in the dataset, with 9,402 (7.3%) having two claims and 2,067 (1.6%) having three or more claims. The maximum number of claims per person was eight.

The median age of the entire AHCIP cohort was 46 years; 67% were females (n=95,857). The median age of individuals having cholecystectomy was 45 years and 72.2% were females. The demographics of the comparison group (the varicose veins cohort) had similar characteristics, with a median age of 46 years and 72.8% females, while the gastric procedure group had a median age of 51 years and fewer females (42.1%). Table 3 shows the relative proportion of procedures by fiscal year, with fairly uniform distributions across the years, except for the varicose vein group.

Chi-square tests were used to compare the distributions of certain variables for the cholecystectomy cohort and the comparison group, the varicose vein cohort. The analysis of the distribution of fiscal year of service was highly significant ($X^2 = 2,460.52$, $p < 0.001$), indicating a dissimilar distribution across the years of service for the cohorts. As shown in Table 3, varicose veins procedures were less frequent in the earliest years. A similar result was found when age distribution was compared for these cohorts ($X^2 = 3,295.42$, $p < 0.001$), and inspection of the data shows that cholecystectomy was less frequent between the ages of 35 years and 64 years, compared to varicose vein procedures. Only a marginal difference was observed in an analysis of distribution by sex ($X^2 = 4.02$, $p = 0.05$). These differences suggest that the varicose vein cohort may not be the ideal comparison group in the study of cholecystectomy and colon cancer, but the use of age-, sex- and year-specific person-years at risk tables ensured adjustment for the dissimilar distributions in these strata.

Table 3. Proportion of Procedures by Fiscal Year, Cohort and Procedures

	Cholecystectomy Cohort (n=92,537)	Varicose Veins Cohort (n=25,424)	Gastric Procedures Cohort (n=25,009)	Total (n=142,970)
1973/74 to 1977/78	25,054 (27.1%)	3,152 (12.4%)	7,175 (28.7%)	35,381 (24.7%)
1978/79 to 1982/83	19,908 (21.5%)	6,760 (26.6%)	5,669 (22.7%)	32,337 (22.6%)
1983/84 to 1987/88	22,970 (24.8%)	6,831 (26.9%)	5,904 (23.6%)	35,705 (25.0%)
1988/89 to 1992/93	24,605 (26.6%)	8,681 (34.1%)	6,261 (25.0%)	39,547 (27.7%)

2. *Colon Cancer Cohort*

At the beginning of November, 1995, there were 20,940 records for colon and rectosigmoid junction tumours in the Alberta Cancer Registry, representing 20,296 individuals. Only the first diagnosis was retained for patients with more than one cancer occurrence. The group was restricted to residents of Alberta who were diagnosed between July 1, 1969 and December 31, 1993 whose registration was complete (or archived). Although 12,861 individuals with colon cancer met these criteria, four were excluded because of missing sex and birthdate information leaving 12,857 records for analysis.

The proportion of cases diagnosed in each 5-year interval increased from 10.7% in 1969-1973 to 27.6% in 1989-1993. This is expected in an aging and expanding population. The median age at diagnosis was 70 years, with almost equal sex distribution (6,476 males, or 50.4%; 6,381 females, or 49.6%). Most of the patients (10,908, or 84.8%) were diagnosed with colon cancer, with the remainder having cancer of the rectosigmoid junction. Almost all were invasive tumours (12,576, or 97.8%) with a few in situ (188, or 1.5%) and borderline (93, or 0.7%) tumours. The majority of the tumours (9,364, or 72.8%) were unspecified adenocarcinomas, followed by mucin-producing adenocarcinomas (716, or 5.6%), unspecified carcinomas (710, or 5.5%) and mucinous adenocarcinomas (539, or 4.2%). Morphology was unknown for 183 (1.4%) patients.

B. Data Quality

1. Alberta Health Care Insurance Plan Cohorts

Date and type of service was always complete in the AHCIP cohorts, although personal identifiers were not. Of the 167,157 original AHCIP coverage records associated with the final cohort, only 436 (0.3%) were missing last name and two (<0.1%) had a missing AHCIP number. None of the cohort was missing both initials, but 45,195 (27.0%) had only one initial. With one exception, the first initial was always present. Information on sex (male or female) was always complete. Table 4 shows the proportion of records missing date information. A small proportion (< 0.1%) of the AHCIP records were missing year, month and day of birth and were excluded from the cohort during initial processing because these variables were required for calculation of age in assessment of person-years at risk.

Records missing effective dates were assigned July 1, 1969, as the best estimate of the first date of coverage. Records missing last dates of coverage were assigned March 31, 1995, as it was assumed that these people were still residents of the province. The most common reason for termination of AHCIP coverage (i.e., for individuals whose last date was complete), was death (17,006, or 50.9%) followed by migration from the province (6,502, or 19.4%) and migration from the province under unknown circumstances (4,365, or 13.1%).

AHCIP service (procedure) dates were then examined to determine the proportion of individuals recorded as undergoing surgery in Alberta during a lapse in

Table 4. Proportion of Coverage Records Missing Date Elements, All Records and Individual-specific

		All Records (n=167,157)	Individual- specific Records (n=128,827)
Date of birth	Complete	157,366 (94.1%)	122,632 (95.2%)
	Missing month	2 (<0.1%)	2 (<0.1%)
	Missing day	9,677 (5.8%)	6,081 (4.7%)
	Missing month and day	112 (0.1%)	112 (0.1%)
Effective date of coverage	Complete	127,775 (76.4%)	92,125 (71.5%)
	Missing year, month and day	39,382 (23.6%)	36,702 (28.5%)
Final date of coverage	Complete	71,725 (42.9%)	33,433 (26.0%)
	Missing year, month and day	95,432 (57.1%)	95,394 (74.0%)

coverage. Of the 136,311 unique individual-specific service dates, only 252 (0.2%) (representing 249 individuals) occurred during a lapse in coverage. Most (180, or 71.4%) were cholecystectomies, followed by gastric procedures (55, or 21.8%) and varicose vein procedures (17, or 6.7%).

The majority of the procedures occurring during a lapse in AHCIP coverage occurred after the last date of coverage (242, or 96.0%). Most of the individuals receiving these procedures were female (176, or 72.7%) and/or aged 15 to 34 years at date of service (135, or 55.8%). The most common reason for termination of coverage for these individuals was the apparent detection of a duplicate record, where a dependent's effective date and cancellation date coincided (92, or 38.0%), followed by death (48, or 19.8%), unconfirmed residence (42, or 17.4%) and "other", which was often a child leaving parents' coverage (38, or 15.7%). If the end of coverage had been extended by one year, 116 (47.9%) of these procedures would have occurred during coverage; if coverage had been extended by two years, 159 (65.7%) would have occurred during coverage. Very few individuals would have benefited from adjusting the effective year back one or two years, with two and three procedures covered respectively.

One potential problem was demonstrated by the cholecystectomy cohort. Although any individual has only one gallbladder to be removed, 235 individuals (0.3%) had more than one cholecystectomy. It was suspected that the additional procedures were incorrectly entered or assigned to the wrong individual in the AHCIP claims file, with the most likely explanation being incorrect entry of a family

member's number. The latter situation would pose a problem for accurate record linkage, if 235 individuals with cholecystectomy were not represented because correct identifying data were unavailable. Additionally, the start and end of follow-up may be incorrect for the 235 affected individuals, since the first cholecystectomy was retained although there is no way of knowing if this was the correct record to keep. In practical terms, however, the error introduced would be negligible given the low frequency of these apparent duplicate procedures. The extent of this problem in the other cohorts is unknown because individuals can have multiple varicose vein procedures or gastric procedures; however, a similar proportion of duplicates could be expected with comparable influence on the follow-up data.

2. *Colon Cancer Cohort*

Most of the identifying information from the Alberta Cancer Registry was complete. First and last names were always present, but 6,293 individuals (48.9%) were missing middle names. AHCIP numbers were missing for 1,097 (8.5%) of the cohort. Birthdate was complete for 12,709 (98.8%) individuals with colon cancer, with 34 (0.3%) missing day and 114 (0.9%) missing month and day. Note that records were deleted for three individuals missing all birthdate elements and one with unknown sex.

Diagnosis date was complete for 11,912 (92.6%) individuals with colon cancer, but day was unknown for 158 patients (1.5%) and month and day were unknown for 790 (6.1%) individuals. In 1969-1973 only 57.7% of all colon cancer

patients had complete diagnosis date information (only the day was missing for an additional 40.8%), but these data were complete for over 95% of the individuals diagnosed after that time.

Most individuals (12,198 or 94.9%) had their colon cancer diagnosed using histology or cytology (pathology reports), with 215 (1.7%) diagnosed by radiology and 160 (1.2%) by clinical means. Only 128 (1.0%) were diagnosed exclusively by death certificate. One individual was missing method of diagnosis.

C. Record Linkage

1. Searching: Determining the Linkage Strategy

Measures of the informativeness of the variables were calculated in this step. Appendix B shows the pocket size (cases/level), Shannon entropy and discriminating power for several combinations of variables and for individual variables separately.

These results led to the linkage strategy for matching. The deterministic linkage strategy was based on the combinations of variables providing the highest discriminating power, starting with AHCIP number and all variables matching. Many combinations of variables were used to detect possible matches, down to very loose criteria, such as agreement on birthyear and sex. The probabilistic linkage strategy relied less on discriminating power and made more allowance for poor data quality, dividing the process into two phases. First, birthyear alone was used as the blocking variable; records not contributing to candidate pairs using birthyear were subjected to

a second procedure using first initial and sex together for blocking. In both phases, four variables were required to match for a pair to be considered as a candidate link. This approach left variables with higher discriminating power to resolve ties and mid-weight pairs.

2. *Matching: Deterministic Linkage*

Using the SQL procedure in SAS, a total of 18,552 candidate pairs were found. However, many were not truly likely pairs. The strategy for separating the most likely candidate pairs from the unlikely candidates is described below.

As shown in Appendix C, one-to-one matches were found when the AHCIP number, alone or in combination with other variables, was used to merge the files and when all variables excluding AHCIP number were used in the merge. Most other variable combinations resulted in many-to-many merges, such that each record had more than one candidate match in the other file. This situation made the identification of cases difficult since each record had several candidate partners. The manual resolution of these pairs is described in the separation step (below).

3. *Matching: Probabilistic Linkage*

In the first phase of probabilistic linkage, files were blocked on birthyear only and agreement on a minimum of 3 additional variables was required for candidate pairs. Pairs meeting these criteria were removed before the second phase, in which the files were blocked by sex and first initial and agreement on 3 additional variables

was required for candidate pairs. Alternative last names and first initials were included as additional records and LinkPro ascertained the best match for each individual. The output from LinkPro is in Appendix D.

In the first phase, 6,613 candidate pairs were identified with an additional 51,687 unresolved pairs (ties). The unresolved pairs were equivalent to the many-to-many merges described for deterministic linkage. Many pairs with extreme weights were easily classified as links or non-links. Both candidate and unresolved pairs between the threshold weights guided by Jaro's approach¹⁴⁹ were printed for manual resolution (see below).

The candidate pairs were removed from the files before the second phase of probabilistic linkage. The second phase identified 2,134 candidate pairs and 10,482 unresolved pairs. The unresolved and candidate pairs between the threshold weights were resolved manually (see below).

4. Separation: Identifying Probable Pairs

The pairs most likely to represent links were separated from all the candidate pairs based on inspection of the output and manual review of some of the candidate pairs in deterministic linkage. AHCIP number was thought to be the most robust variable, so that pairs agreeing on AHCIP number (1,593 or 8.6%), were considered to be probable links. Pairs that did not agree on at least four out of the remaining seven variables, and those that did not agree on last name soundex or at least one initial, were considered to be unlinked (15,324 or 82.6%). Only 1,635 (8.8%)

original candidate pairs required manual resolution.

The probabilistic linkage results were assessed empirically, guided by a variation of Jaro's approach¹⁴⁹ (Appendix E). The upper threshold weights were established based on logic. Since agreement on the AHCIP number or agreement on all other variables was considered a match in the deterministic linkage, the upper threshold was set at the point where the AHCIP number and at least one other variable disagreed. This occurred for records with weights less than 32.36 in the first phase of probabilistic linkage and for records with weights less than 32.31 in the second phase of probabilistic linkage. The lower thresholds were set where no further elements of the randomly generated set of unlinked pairs (*U*) occurred, which was for records with weights greater than 16.60 in the first phase and for records with weights greater than 17.30 in the second phase. Based on these thresholds, 184 candidate records required manual resolution from the first phase with 684 candidate records from the second phase.

Unresolved pairs are generated by LinkPro where the program cannot determine the best pair because the weights are the same for two or more candidate pairs. Therefore, all unresolved pairs with weights greater than the lower threshold must also be reviewed manually. Based on this criterion, 21 pairs from the first phase and 399 pairs from the second phase of linkage qualified for review.

Reviewer agreement was very satisfactory. For the 1,635 record pairs reviewed from deterministic linkage, an additional 64 (3.9% of those reviewed), were identified by both reviewers for inclusion and 1,552 (94.9%), were identified for

exclusion by both reviewers. Agreement was 98.8%, with a kappa of 0.87. The remaining 19 cases were discussed because the reviewers did not agree about whether they should be included or not, or because at least one reviewer could not decide either way. Two of the re-reviewed cases (10.5%) were included by consensus.

Of the 868 record pairs reviewed from probabilistic linkage, 30 (3.5% of those reviewed) were included by both reviewers, 802 (92.4%) were excluded by both reviewers and two cases (0.2%) could not be classified by either reviewer. This led to inter-reviewer agreement of 96.1%, with kappa at 0.65. After reviewing the 36 cases where the reviewers did not agree or where the reviewers could not classify the case, eight of the re-reviewed cases (22.2%) were included by consensus.

Unresolved cases (ties) generated by LinkPro were also rated by the reviewers; ties appearing in Phase I linkage were not removed for Phase II linkage and so could appear as ties for Phase II as well as Phase I. Of the nine candidates from Phase I, five (55.6%) were included by both reviewers and three (33.3%) were excluded; of the 55 candidates identified from Phase II linkage, the same five cases (9.1%) were included by both reviewers and 48 (87.3%) were excluded. Agreement was 88.9% for Phase I and 96.4% for Phase II, with kappa at 0.80 and 0.82 respectively. The three unclassifiable cases were re-reviewed and rejected by consensus.

5. *Comparison of Linkage Strategies: Deterministic versus Probabilistic*

Before manual resolution, 1,593 pairs were considered to be clear matches from deterministic linkage, with an additional 1,635 pairs requiring manual review.

Probabilistic linkage resulted in 1,595 probable pairs (1,558 from the first phase, 37 from the second phase), with 868 pairs (184 from the first phase, 684 from the second phase) requiring manual review, plus the unresolved cases generated as ties by LinkPro. Following the review of unresolved pairs, 66 more pairs (4.0% of the reviewed pairs) were identified in the deterministic linkage and 38 more pairs (4.4% of the reviewed pairs) were identified from probabilistic linkage (with five cases from the ties). There was a total of 1,659 cases ascertained by deterministic linkage and 1,638 cases identified by probabilistic linkage.

In total, 1,670 individuals were identified by at least one linkage strategy. Table 5 shows the agreement pattern for the two strategies. The two strategies agreed for the majority of the cases (99.7%) and level of agreement beyond chance, as assessed by the kappa statistic, was excellent ($\kappa = 0.99$).

Most of the cases (8 out of 11, or 72.7%), undetected by deterministic linkage but identified in probabilistic linkage, were missed because the appropriate record had been removed in an earlier deterministic step. Inspection of the competing candidate records showed that the removal was inappropriate, usually resulting from overemphasis on soundex code matches. In three cases (27.3%), AHCIP number matched but more appropriate matches were found in probabilistic linkage. Each of these three deterministic matches was based on AHCIP number and other variables, such as sex, first initial and birthdate, mismatched considerably. The corresponding probabilistic linkages were based on agreement of most variables except for AHCIP number. Analysis using the combined results of both linkage used the information

Table 5. Agreement Between Deterministic and Probabilistic Linkage**Strategies: Colon Cancer as Outcome**

Deterministic Assessment	Probabilistic Assessment		Total
	Matched Pair	Not Matched Pair	
Matched Pair	1,627	32	1,659
Not Matched Pair	11	11,187	11,198
Total	1,638	11,219	12,857

 $\text{kappa} = 0.99$

from the probabilistic linkage in these cases for a total of 1,667 individuals.

Three of the 32 individuals (9.4%), undetected by probabilistic linkage were, therefore, incorrect matches in deterministic linkage, based on over-reliance on the AHCIP number. The remaining 29 individuals (90.6%) were missed by probabilistic linkage because a poor match was found in the first phase, with blocking variable birthyear, and the record was removed from contention for the second phase of linkage, where the appropriate match would have been found.

6. *Variable Agreement in Linked Pairs*

The frequency of agreement between the linking variables in the matched pairs was assessed to give an indication of the higher quality variables that may be more useful in the future (Table 6). In particular, these values could be used to calculate the merit ratio, adjusting the discriminating power by the frequency of disagreement in linkable pairs in subsequent studies using data from the Alberta Cancer Registry and the provincial health care insurance plan. Variables from the AHCIP dataset and the cancer registry agreed over 95% in most linked pairs, except for day of birth (about 85%) and middle initial (about 72%). AHCIP number was useful in both linkage strategies, although the higher emphasis placed on it in deterministic linkage resulted in higher proportions of records agreeing on AHCIP number using that method, as opposed to the probabilistic method.

Table 6. Variable Agreement in Linked Pairs: AHCIP and Cancer Registry**Linking Variables (Colon Cancer Cohort Only)**

Variable	Agreement in Deterministic Linkage (n = 1,659)	Agreement in Deterministic Linkage: AHCIP Matches Only (n = 1,593)	Agreement in Probabilistic Linkage (n = 1,638)	Agreement in Either Linkage* (n = 1,667)
Birthyear	1,590 (95.8%)	1,524 (95.7%)	1,596 (97.4%)	1,593 (95.6%)
Birthmonth	1,613 (97.2%)	1,548 (97.2%)	1,595 (97.4%)	1,620 (97.2%)
Birthday	1,421 (85.7%)	1,365 (85.7%)	1,417 (86.5%)	1,428 (85.7%)
First Initial	1,602 (96.6%)	1,536 (96.4%)	1,598 (97.6%)	1,607 (96.4%)
Middle Initial	1,205 (72.6%)	1,154 (72.4%)	1,192 (72.8%)	1,205 (72.3%)
Last Name	1,616 (97.4%)	1,551 (97.4%)	1,596 (97.4%)	1,624 (97.4%)
Last Name Soundex	1,645 (99.2%)	1,579 (99.1%)	1,625 (99.2%)	1,653 (99.2%)
Sex	1,646 (99.2%)	1,580 (99.2%)	1,630 (99.5%)	1,654 (99.2%)
AHCIP Number	1,593 (96.0%)	1,593 (100.0%)	1,561 (95.3%)	1,590 (95.4%)

Recall 3 AHCIP matches in deterministic linkage were corrected by competing candidates from probabilistic linkage in combined analysis.

7. *Chart Review: Confirming Record Linkage*

The individuals with colon cancer identified by record linkage were considered diseased for the calculation of risk in the study. However, a chart review was used to provide an external perspective and assessment of the effectiveness of record linkage using Alberta data sources.

Of the 162 individuals diagnosed with colon cancer at least 10 years after cholecystectomy, 102 (63.0%) had charts at the Cross Cancer Institute, but 10 were non-reporting patients with incomplete charts. Of the remaining 92 patients, 2 (2.2%) were identified by deterministic linkage only, with the remainder being identified by both types of linkage. Unlinked records were frequency matched by diagnosis year. Of the 3,510 eligible cases, 92 were selected for review.

Four charts (2.2%) were missing, but the remaining 180 charts were examined by two reviewers. Agreement between the reviewers was 0.79, as assessed by the kappa statistic. Agreement with record linkage was 0.56, with specifics shown in Table 7 below.

The physicians of the 90 individuals whose chart review result did not indicate history of cholecystectomy were contacted. The records for many of this group (36, or 40.0%) were unavailable because the physicians could not be contacted or the charts had been destroyed. The outcome of the manual review was updated using the information acquired about the remaining 54 individuals (60.0%). Twelve more individuals were found to have a history of cholecystectomy. The final assessment of agreement between manual review and record linkage was 0.62 (Table 8).

Table 7. Record Linkage versus Chart Review: Cancer Charts Only

Chart Review Assessment	Linkage Assessment		Total
	Matched Pair	Not Matched Pair	
Matched Pair	76	14	90
Not Matched Pair	26	64	90
Total	102	78	180

$\text{kappa} = 0.56$

Table 8. Record Linkage versus Chart Review: Cancer Charts and Physician Records

Chart Review Assessment	Linkage Assessment		Total
	Matched Pair	Not Matched Pair	
Matched Pair	85	17	102
Not Matched Pair	17	61	78
Total	102	78	180

$\text{kappa} = 0.62$

Date of cholecystectomy was not recorded for many of the colon cancer patients who had previous cholecystectomy recorded in their cancer charts (75, or 83.3%). If an approximate date was provided, most charts (14, or 93.3%) included only the year of the procedure. In the follow-up with physicians, no complete dates of cholecystectomy were provided; year alone was available for 7 patients (58.3%).

8. *Secondary Endpoints: Linkage to Other Relevant Cancers*

As discussed, patients diagnosed with malignant gallbladder disease were removed from the cohort. In addition, patients diagnosed with pancreatic cancer or cancer of the biliary ducts at any time following cholecystectomy had follow-up truncated at the date of diagnosis, since their treatment often involved removing the gallbladder. To identify these individuals, records of patients diagnosed with gallbladder, biliary and pancreatic tumours were submitted for linkage to the AHCIP cohort following the colon cancer cohort linkage.

Data quality for the records of these individuals was comparable to that of the colon cancer cohort. Using the same exclusion criteria previously described, 659 Albertans were diagnosed with gallbladder cancer between July, 1969 and December, 1993. The majority (602, or 91.4%) were diagnosed by pathology and only 10 (1.5%) individuals were reported to the cancer registry based on the death certificate only. The majority of the patients were female (456, or 69.2%). Most of the identifying information was complete, although middle initial was missing for 380 (57.7%) patients and AHCIP number was missing for 94 (14.3%).

Biliary cancer was diagnosed in 484 Albertans between July, 1969 and December, 1993; again, most individuals were diagnosed by pathology reports (360, or 74.4%) with only a few notifications by death certificate only (16, or 3.3%). The sex distribution for individuals diagnosed with biliary cancer was approximately equal (249 males, or 51.4%, and 235 females, or 48.6%). Middle initials and AHCIP number were again most likely to be missing (272 cases, or 56.2%, and 88 cases, or 18.2%, respectively).

Pancreatic cancer was diagnosed in 4,124 individuals between July, 1969 and December, 1993. Most of these patients (2,788, or 67.6%) were diagnosed by pathology report and only 160 (3.9%) were reported by death certificate only. As with biliary cancer, slightly more than half (2,317, or 56.2%) were men. Approximately half (2,126, or 51.6%) of the patients' records were missing middle initial and almost a quarter (961, or 23.3%) were missing AHCIP number.

The magnitude of the discriminating power was slightly lower in the cohort of individuals with gallbladder, biliary or pancreatic cancer than in the colon cancer cohort. This was expected, given there were fewer individuals in this cohort compared to the colon cancer cohort. However, the relative magnitude was similar and the linkage approaches described for the colon cancer cohort were applied.

Deterministic linkage yielded 1,137 probable (AHCIP number) matches and an additional 655 matches which were submitted for manual review. An additional 5,647 matches fulfilled less important deterministic criteria, so were not considered to be potential links. Probabilistic linkage was again performed in two phases, first

blocking on birthyear and then blocking on first initial and sex. The agreement constructs developed in the colon cancer cohort linkage were used to determine the cutoffs for probabilistic linkage. In the first phase, 1,143 probable matches were identified (weights > 29.51), with 99 requiring manual review (weights between 16.95 and 29.51). Weights less than 16.95 were assigned to 1,996 matches and were not considered to be potential links. In the second phase, 22 probable matches were identified (weights > 29.37), with 189 other matches identified for manual review (weights between 17.83 and 29.37). Weights less than 17.83 were assigned to 456 matches and these were excluded.

Reviewer agreement was again very satisfactory. The deterministic linkage review resulted in an observed level of agreement of 98.5%, with kappa=0.95; 105 (16.0% of all reviewed pairs) were confirmed as linked pairs by both reviewers, and 5 more were added by consensus. The probabilistic review, excluding ties, resulted in an observed agreement of 94.8%, with kappa=0.83; 44 (15.3% of all reviewed pairs) were confirmed as links by both reviewers and 9 more were added by consensus. Six probabilistic ties were also considered to be linked pairs by both reviewers.

Table 9 shows the agreement between the two types of linkage. Observed agreement between the approaches was 99.3%, with kappa=0.98. Most individuals (28 out of 29, or 96.6%) identified through deterministic but not probabilistic linkage disagreed on birthyear; the remaining pair identified in deterministic linkage matched on birthyear but few other variables, leading to a low probabilistic weight.

Table 9. Agreement Between Deterministic and Probabilistic Linkage
Strategies: Other Relevant Cancers as Outcome

Deterministic Assessment	Probabilistic Assessment		Total
	Matched Pair	Not Matched Pair	
Matched Pair	1,218	29	1,247
Not Matched Pair	6	4,009	4,015
Total	1,224	4,038	5,262

$\kappa = 0.98$

Individuals identified through probabilistic but not deterministic linkage were usually identified in later steps of the deterministic linkage (4 out of 6, or 66.7%).

Additionally, one pair was excluded in deterministic linkage as the result of a mismatch in an early deterministic step and another was excluded in the manual review in deterministic, but not probabilistic, linkage.

The linkage between AHCIP records and the secondary cancer endpoints were examined for agreement between linking variables. As with the colon cancer cohort, agreement for most variables from the AHCIP dataset and the cancer registry was over 95 % in most linked pairs, except for AHCIP number (approximately 90 %), day of birth (approximately 80 %) and middle initial (approximately 73 %).

D. Estimates of Risk

1. Colon Cancer Risk in the Cholecystectomy Cohort Compared to the General Population

The risk for colon cancer following cholecystectomy was based on the results of linkage, but different linkage approaches did not identify the same individuals. Therefore, risks were calculated for individuals identified (a) by deterministic linkage, (b) by probabilistic linkage, and (c) by either deterministic or probabilistic linkage, providing a sensitivity analysis for linkage.

a. Determination of the Final Cohort

The files were prepared using a standard protocol to avoid bias and to allow comparisons between the various linkage approaches. The final determination of the number of individuals with cholecystectomy who subsequently developed colon cancer is summarized in Table 10. Clearly, there was little difference in the final cohort based on the linkage approach used.

b. Overall Risk by Type of Linkage

As discussed, age-, sex- and calendar year-specific colon cancer rates in the Alberta population were applied to the corresponding person-years at risk to determine the risk of colon cancer for the cholecystectomy cohort. Induction periods of 0, 1, 2, 5, 10 and 15 years were used. Appendix F provides an example of the person-years at risk matrix, using the cohort defined by either type of linkage and a 5-year induction period. Appendix G shows the age-, sex- and year-specific colon cancer rates for Alberta (1969-1993).

The SIRs are shown by induction period in tables. First, the person-years at risk and the number of individuals in the cohort are displayed, followed by the number of individuals observed to have colon cancer and the number expected, based on the person-years at risk in the cohort and the rates in the comparison group. The final columns provide the risk estimate and 95% CIs. The results of linkage are presented in Tables 11 and 12 for deterministic linkage, Tables 13 and 14 for probabilistic linkage, and Tables 15 and 16 for both linkage types combined. As

Table 10. Summary of Final File Preparation, by Linkage Approach

	Deterministic Linkage	Probabilistic Linkage	Either Linkage
Individuals with cholecystectomy = 92,301			
Removed from cohort:			
Gallbladder cancer diagnosed within 6 months of cholecystectomy	176	174	177
Biliary/pancreatic cancer diagnosed on or before cholecystectomy	114	111	115
Colon cancer diagnosed before cholecystectomy	376	374	381
Colon cancer diagnosed at same time as cholecystectomy	264	261	264
Cholecystectomy after last date of follow-up	172	171	172
First and last dates of follow-up equal	26	26	26
Total removed	1,128	1,117	1,135
(% of total cohort)	(1.2%)	(1.2%)	(1.2%)
Follow-up truncated:			
Gallbladder cancer diagnosed more than 6 months from cholecystectomy	3	2	3
Biliary/pancreatic cancer diagnosed after cholecystectomy	218	216	219
Colon cancer diagnosed after cholecystectomy	607	596	609
Total truncated	828	814	831
(% of total cohort)	(0.9%)	(0.9%)	(0.9%)
Individuals with multiple cholecystectomies	231	231	231
(% of total cohort)	(0.3%)	(0.3%)	(0.3%)
Colon cancer diagnosed after follow-up	2	3	3
(% of individuals with cancer)	(0.3%)	(0.5%)	(0.5%)
Final cohort	91,173	91,184	91,166
(% of total cohort)	(98.8%)	(98.8%)	(98.8%)
Individuals with colon cancer	605	593	606
(% of final cohort)	(0.7%)	(0.7%)	(0.7%)

Table 11. Comparison of Standardized Incidence Ratios (SIRs) and 95 % Confidence Intervals (CIs) for Colon Cancer in the Cholecystectomy Cohort versus the General Alberta Population, Using Deterministic Linkage Only

Induction (Years)	Cohort		Individuals with Colon Cancer		SIR	95 % CI
	Person-years at Risk	Individuals	Observed	Expected		
0	826,102.0	91,173	605	553.8	1.09	(1.01-1.18)
1	736,832.8	86,884	526	510.4	1.03	(0.94-1.12)
2	653,384.6	80,336	475	467.6	1.02	(0.93-1.11)
5	441,363.6	61,515	348	345.8	1.01	(0.90-1.12)
10	198,106.3	37,022	161	176.2	0.91	(0.78-1.07)
15	58,407.2	19,609	52	57.3	0.91	(0.68-1.19)

Table 12. Comparison of Standardized Incidence Ratios (SIRs) and 95 % Confidence Intervals (CIs) for Colon Cancer in the Cholecystectomy Cohort versus the General Alberta Population, Using Deterministic Linkage Only, by Sex

Induction (Years)	Cohort		Individuals with Colon Cancer		SIR	95% CI
	Person-years at Risk	Individuals	Observed	Expected		
Females						
0	610,395.5	65,979	337	330.1	1.02	(0.91-1.14)
1	545,666.4	63,062	284	305.9	0.93	(0.82-1.04)
2	485,065.9	58,400	257	281.9	0.91	(0.80-1.03)
5	330,433.2	45,088	186	212.2	0.88	(0.76-1.01)
10	150,501.9	27,732	89	111.5	0.80	(0.64-0.98)
15	45,078.6	14,966	26	37.5	0.69	(0.45-1.02)
Males						
0	215,706.4	25,194	268	223.7	1.20	(1.06-1.35)
1	191,166.4	23,822	242	204.5	1.18	(1.04-1.34)
2	168,318.7	21,936	218	185.8	1.17	(1.02-1.34)
5	110,930.4	16,427	162	133.6	1.21	(1.03-1.41)
10	47,604.4	9,290	72	64.7	1.11	(0.87-1.40)
15	13,328.6	4,643	26	19.8	1.31	(0.86-1.93)

Table 13. Comparison of Standardized Incidence Ratios (SIRs) and 95 % Confidence Intervals (CIs) for Colon Cancer in the Cholecystectomy Cohort versus the General Alberta Population, Using Probabilistic Linkage Only

Induction (Years)	Cohort		Individuals with Colon Cancer		SIR	95 % CI
	Person-years at Risk	Individuals	Observed	Expected		
0	826,168.9	91,184	593	554.0	1.07	(0.99-1.16)
1	736,892.0	86,890	515	510.5	1.01	(0.92-1.10)
2	653,439.9	80,338	464	467.7	0.99	(0.90-1.09)
5	441,414.0	61,516	342	345.9	0.99	(0.89-1.10)
10	198,136.7	37,027	158	176.3	0.90	(0.76-1.05)
15	58,417.7	19,611	52	57.3	0.91	(0.68-1.19)

Table 14. Comparison of Standardized Incidence Ratios (SIRs) and 95 % Confidence Intervals (CIs) for Colon Cancer in the Cholecystectomy Cohort versus the General Alberta Population, Using Probabilistic Linkage Only, by Sex

Induction (Years)	Cohort		Individuals with Colon Cancer	SIR	95 % CI
	Person-years at Risk	Individuals			
Females					
0	610,411.9	65,987	332	1.01	(0.90-1.12)
1	545,678.4	63,065	279	0.91	(0.81-1.03)
2	485,077.0	58,399	252	0.89	(0.79-1.01)
5	330,447.4	45,087	183	0.86	(0.74-1.00)
10	150,507.4	27,735	88	0.79	(0.63-0.97)
15	45,078.6	14,966	26	0.69	(0.45-1.02)
Males					
0	215,757.0	25,197	261	1.17	(1.03-1.32)
1	191,213.6	23,825	236	1.15	(1.01-1.31)
2	168,362.9	21,939	212	1.14	(0.99-1.31)
5	110,966.6	16,429	159	1.19	(1.01-1.39)
10	47,629.3	9,292	70	1.08	(0.84-1.37)
15	13,339.0	4,645	26	1.31	(0.86-1.92)

Table 15. Comparison of Standardized Incidence Ratios (SIRs) and 95 % Confidence Intervals (CIs) for Colon Cancer in the Cholecystectomy Cohort versus the General Alberta Population, Using Both Deterministic and Probabilistic Linkage

Induction (Years)	Cohort		Individuals with Colon Cancer		SIR	95 % CI
	Person-years at Risk	Individuals	Observed	Expected		
0	826,052.4	91,166	606	553.7	1.09	(1.01-1.19)
1	736,791.0	86,876	527	510.3	1.03	(0.95-1.13)
2	653,350.7	80,328	476	467.6	1.02	(0.93-1.11)
5	441,350.5	61,509	348	345.8	1.01	(0.90-1.12)
10	198,105.6	37,021	162	176.2	0.92	(0.78-1.07)
15	58,406.8	19,609	53	57.3	0.93	(0.69-1.21)

Table 16.

Comparison of Standardized Incidence Ratios (SIRs) and 95 % Confidence Intervals (CIs) for Colon Cancer in the Cholecystectomy Cohort versus the General Alberta Population, Using Both Deterministic and Probabilistic Linkage, by Sex

Induction (Years)	Cohort		Individuals with Colon Cancer		SIR	95 % CI
	Person-years at Risk	Individuals	Observed	Expected		
Females						
0	610,347.1	65,973	339	330.0	1.03	(0.92-1.14)
1	545,624.7	63,055	286	305.8	0.94	(0.83-1.05)
2	485,031.1	58,393	259	281.8	0.92	(0.81-1.04)
5	330,417.3	45,082	186	212.2	0.88	(0.76-1.01)
10	150,499.8	27,731	89	111.5	0.80	(0.64-0.98)
15	45,078.6	14,966	26	37.5	0.69	(0.45-1.02)
Males						
0	215,705.3	25,193	267	223.7	1.19	(1.05-1.35)
1	191,166.3	23,821	241	204.5	1.18	(1.03-1.34)
2	168,319.6	21,935	217	185.8	1.17	(1.02-1.33)
5	110,933.2	16,427	162	133.6	1.21	(1.03-1.41)
10	47,605.9	9,290	73	64.7	1.13	(0.89-1.42)
15	13,328.2	4,643	27	19.8	1.36	(0.90-1.99)

shown in Figures 2 and 3, there is no appreciable difference in risk based on deterministic, probabilistic or the combined linkage results. Therefore, subsequent calculations of risk used the combined linkage findings.

c. Risk by Subsite

SIRs were calculated by subsite. Tables 17 and 18 present the results for right colon cancer, which included tumours of the cecum, appendix and ascending colon (ICDO codes C18.0-C18.2). Tables 19 and 20 present data for mid-colon cancer, which included tumours of the hepatic flexure, transverse colon and splenic flexure (ICDO codes C18.3-C18.5). Tables 21 and 22 show results for left colon cancer, which included tumours of the descending colon, sigmoid colon and rectosigmoid junction (ICDO codes C18.6, C18.7 and C19.9). Data are not presented for excluded and "other" colon, which included tumours classified as overlapping lesions of the colon (C18.8) and "Colon, not otherwise specified" (C18.9). As shown in Figure 4, there were no statistically significant risk estimates for any site, sex and induction combination, and there was no apparent trend of increasing or decreasing risk from the right to left colon.

2. *Colon Cancer Risk in the Cholecystectomy Cohort Compared to the Varicose Vein Cohort*

As discussed, colon cancer rates in another cohort of individuals identified by AHCIP, those with stripping and ligation of varicose veins, were determined to

Figure 2. Comparison of Standardized Incidence Ratios for Colon Cancer Following Cholecystectomy: Different Linkage Approaches, by Induction

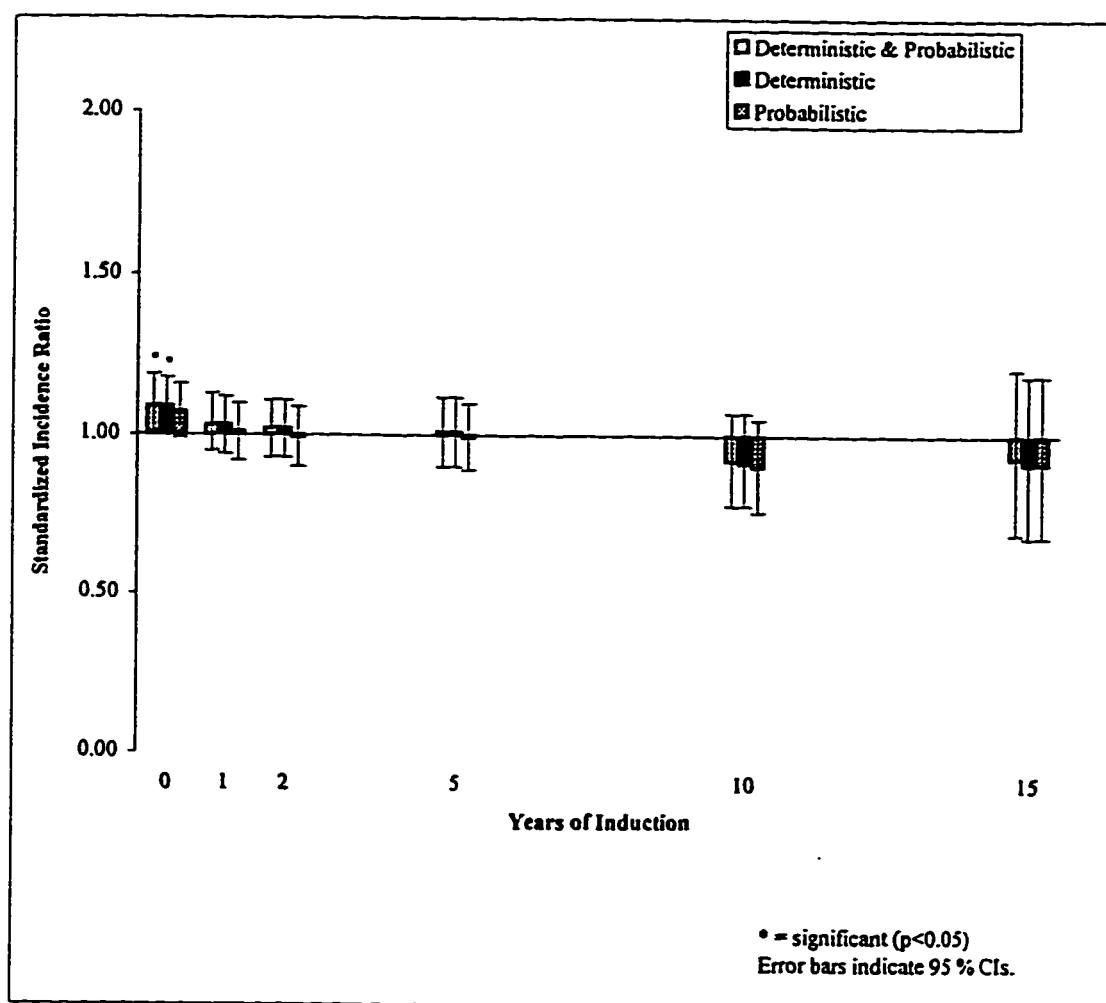
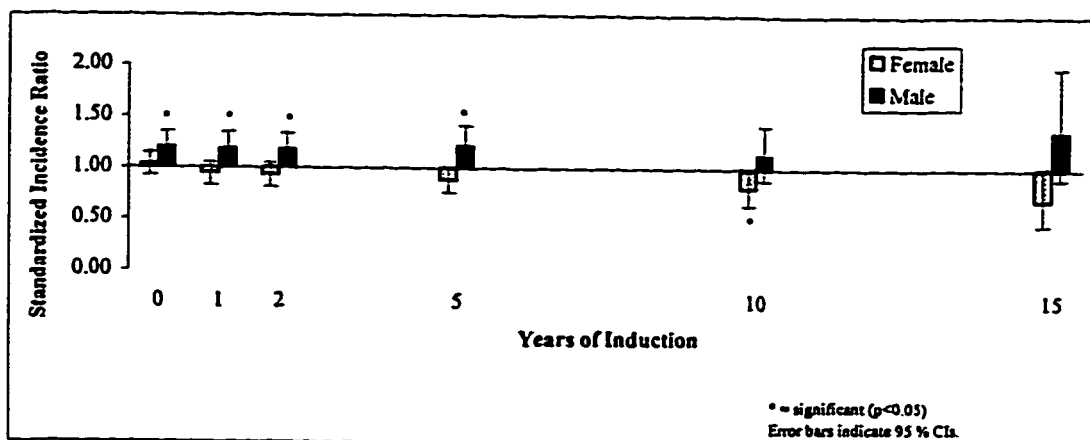
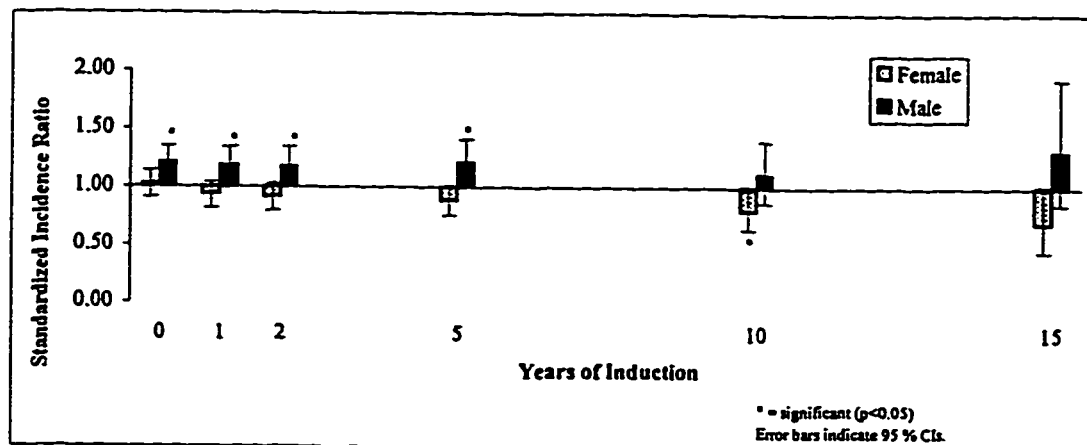


Figure 3. Comparison of Standardized Incidence Ratios for Colon Cancer Following Cholecystectomy: Using Different Linkage Approaches, by Induction and Sex

Deterministic + Probabilistic



Deterministic



Probabilistic

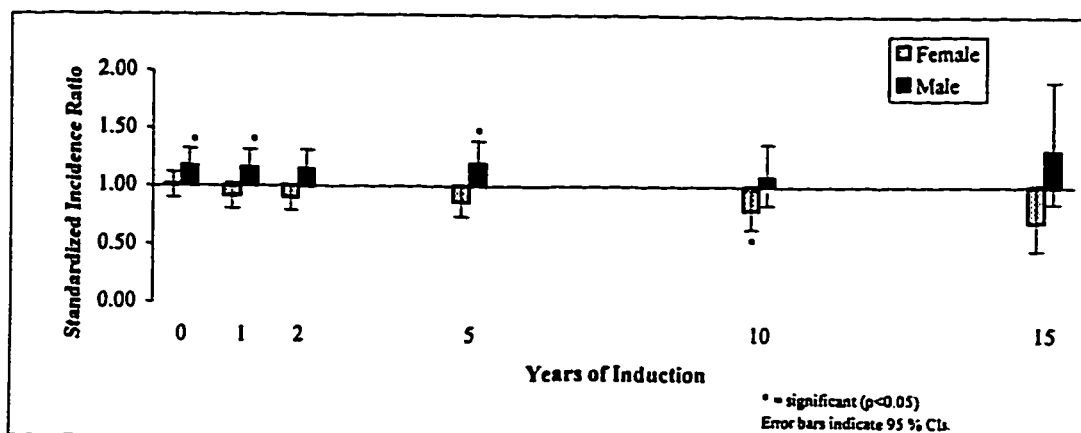


Table 17. Comparison of Standardized Incidence Ratios (SIRs) and 95 % Confidence Intervals (CIs) for Right Colon Cancer in the Cholecystectomy Cohort versus the General Alberta Population

Induction (Years)	Cohort		Individuals with Colon Cancer		SIR	95 % CI
	Person-years at Risk	Individuals	Observed	Expected		
0	826,052.4	91,166	186	165.9	1.12	(0.97-1.29)
1	736,791.0	86,876	153	153.8	0.99	(0.84-1.17)
2	653,350.7	80,328	138	141.6	0.97	(0.82-1.15)
5	441,350.5	61,509	104	106.4	0.98	(0.80-1.19)
10	198,105.6	37,021	49	55.7	0.88	(0.65-1.16)
15	58,406.8	19,609	23	18.7	1.23	(0.78-1.85)

Table 18. Comparison of Standardized Incidence Ratios (SIRs) and 95 % Confidence Intervals (CIs) for Right Colon Cancer in the Cholecystectomy Cohort versus the General Alberta Population, by Sex

Induction (Years)	Cohort		Individuals with Colon Cancer		SIR	95 % CI
	Person-years at Risk	Individuals	Observed	Expected		
Females						
0	610,347.1	65,973	110	104.0	1.06	(0.87-1.28)
1	545,624.7	63,055	88	96.9	0.91	(0.73-1.12)
2	485,031.1	58,393	79	89.8	0.88	(0.70-1.10)
5	330,417.3	45,082	59	68.9	0.86	(0.65-1.10)
10	150,499.8	27,731	32	37.3	0.86	(0.59-1.21)
15	45,078.6	14,966	13	13.0	1.00	(0.53-1.71)
Males						
0	215,705.3	25,193	76	62.0	1.23	(0.97-1.54)
1	191,166.3	23,821	65	56.9	1.14	(0.88-1.46)
2	168,319.6	21,935	59	51.8	1.14	(0.87-1.47)
5	110,933.2	16,427	45	37.5	1.20	(0.88-1.61)
10	47,605.9	9,290	17	18.4	0.92	(0.54-1.48)
15	13,328.2	4,643	10	5.7	1.75	(0.84-3.21)

Table 19. Comparison of Standardized Incidence Ratios (SIRs) and 95 % Confidence Intervals (CIs) for Mid-Colon Cancer in the Cholecystectomy Cohort versus the General Alberta Population

Induction (Years)	Cohort		Individuals with Colon Cancer	SIR	95 % CI
	Person-years at Risk	Individuals			
0	826,052.4	91,166	Observed 107 Expected 87.0	1.23	(1.01-1.49)
1	736,791.0	86,876	88 80.3	1.10	(0.88-1.35)
2	653,350.7	80,328	82 73.6	1.11	(0.89-1.38)
5	441,350.5	61,509	60 54.3	1.10	(0.84-1.42)
10	198,105.6	37,021	27 27.7	0.98	(0.64-1.42)
15	58,406.8	19,609	* *	*	*

* Insufficient number of cases for calculation of SIR and CI.

Table 20. Comparison of Standardized Incidence Ratios (SIRs) and 95 % Confidence Intervals (CIs) for Mid-Colon Cancer in the Cholecystectomy Cohort versus the General Alberta Population, by Sex

Induction (Years)	Cohort		Individuals	Individuals with Colon Cancer		SIR	95% CI
	Person-years at Risk	Observed		Expected			
Females							
0	610,347.1	58	65,973	51.6	1.12	(0.85-1.45)	
1	545,624.7	46	63,055	47.9	0.96	(0.70-1.28)	
2	485,031.1	44	58,393	44.1	1.00	(0.73-1.34)	
5	330,417.3	31	45,082	33.1	0.94	(0.64-1.33)	
10	150,499.8	15	27,731	17.3	0.87	(0.48-1.43)	
15	45,078.6	*	14,966	*	*	*	
Males							
0	215,705.3	49	25,193	35.4	1.38	(1.02-1.83)	
1	191,166.3	42	23,821	32.4	1.29	(0.93-1.75)	
2	168,319.6	38	21,935	29.5	1.29	(0.91-1.77)	
5	110,933.2	29	16,427	21.2	1.37	(0.92-1.96)	
10	47,605.9	12	9,290	10.3	1.16	(0.60-2.02)	
15	13,328.2	*	4,643	*	*	*	

* Insufficient number of cases for calculation of SIR and CI.

Table 21. Comparison of Standardized Incidence Ratios (SIRs) and 95 % Confidence Intervals (CIs) for Left Colon Cancer in the Cholecystectomy Cohort versus the General Alberta Population

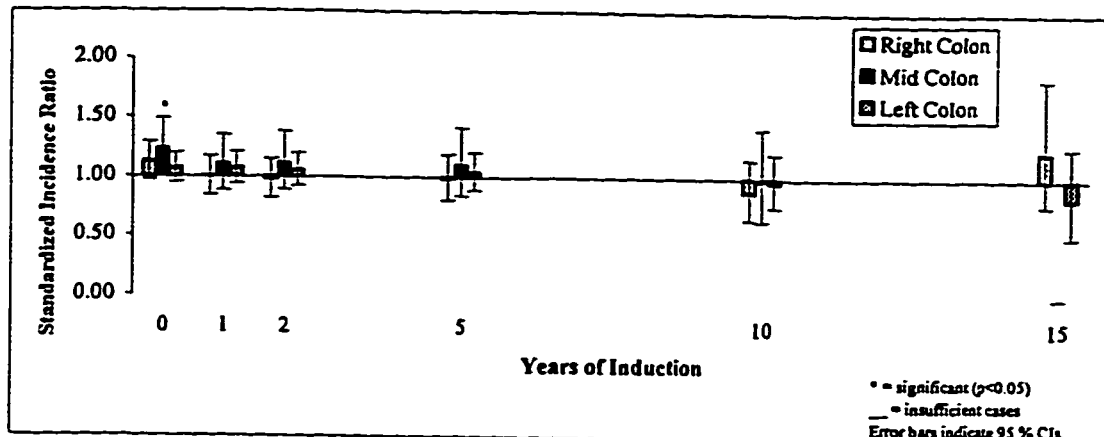
Induction (Years)	Cohort		Individuals with Colon Cancer		SIR	95 % CI
	Person-years at Risk	Individuals	Observed	Expected		
0	826,052.4	91,166	276	258.0	1.07	(0.95-1.20)
1	736,791.0	86,876	254	236.9	1.07	(0.94-1.21)
2	653,350.7	80,328	227	216.2	1.05	(0.92-1.20)
5	441,350.5	61,509	164	157.9	1.04	(0.89-1.21)
10	198,105.6	37,021	76	79.0	0.96	(0.76-1.21)
15	58,406.8	19,609	21	25.3	0.83	(0.51-1.27)

Table 22. Comparison of Standardized Incidence Ratios (SIRs) and 95 % Confidence Intervals (CIs) for Left Colon Cancer in the Cholecystectomy Cohort versus the General Alberta Population, by Sex

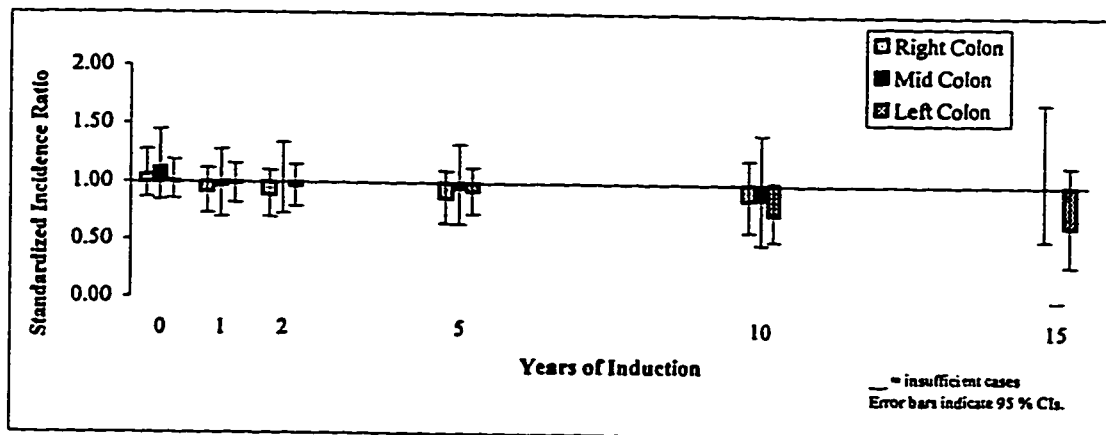
Induction (Years)	Cohort		Individuals with Colon Cancer		SIR	95% CI
	Person-years at Risk	Individuals	Observed	Expected		
Females						
0	610,347.1	65,973	150	148.1	1.01	(0.86-1.19)
1	545,624.7	63,055	134	136.6	0.98	(0.82-1.16)
2	485,031.1	58,393	120	125.3	0.96	(0.79-1.15)
5	330,417.3	45,082	85	92.9	0.92	(0.73-1.13)
10	150,499.8	27,731	35	47.7	0.73	(0.51-1.02)
15	45,078.6	14,966	10	15.7	0.64	(0.31-1.17)
Males						
0	215,705.3	25,193	126	109.9	1.15	(0.96-1.37)
1	191,166.3	23,821	120	100.3	1.20	(0.99-1.43)
2	168,319.6	21,935	107	90.9	1.18	(0.96-1.42)
5	110,933.2	16,427	79	65.1	1.21	(0.96-1.51)
10	47,605.9	9,290	41	31.3	1.31	(0.94-1.78)
15	13,328.2	4,643	11	9.6	1.15	(0.57-2.06)

Figure 4. Comparison of Standardized Incidence Ratios for Colon Cancer Following Cholecystectomy: by Colon Subsite, Using Rates of Colon Cancer in the General Alberta Population

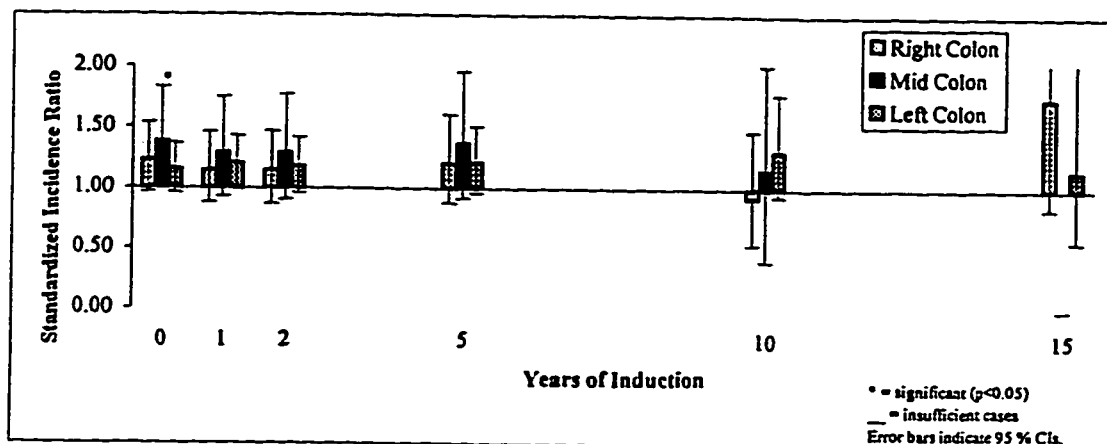
Male & Female



Female



Male



minimize bias associated with record linkage and to examine colon cancer risk in a group that may have been more similar to the cholecystectomy cohort than was the general population. The varicose vein cohort was prepared using a protocol similar to that described for the cholecystectomy cohort. There were 19,746 individuals in the unedited varicose vein cohort. Those with gallbladder, biliary or pancreatic cancer had follow-up terminated at date of diagnosis; out of 32 individuals identified with these cancers, only two were diagnosed before undergoing a varicose vein procedure and were removed from the cohort, leaving 19,744 individuals for analysis.

Only 127 individuals in the varicose vein cohort (0.6% of the group) were diagnosed with colon cancer. When individuals with cancer diagnosis before varicose vein surgery were removed (26, or 20.5% of the cohort), the cohort was reduced to 19,718. An additional 16 individuals were removed from the cohort because their dates of service occurred after their follow-up; one person had the same date for the beginning and end of follow-up (resulting in no person-years at risk), so was deleted. After this processing, 19,701 individuals remained in the cohort (99.8% of the original cohort), including data for 100 individuals with colon cancer.

SIRs were based on the number of patients with only cholecystectomy and the rate of colon cancer in patients with only varicose veins. Data for 6 individuals who had undergone both procedures were removed before analysis. Tables 23 and 24 show the overall risk estimates, and Tables 25 - 30 show the risks by colon subsite. The risk for males was significantly elevated when the varicose vein group was used as the comparison group, except for cancer of the left colon (Figures 5 and 6).

Table 23. Comparison of Standardized Incidence Ratios (SIRs) and 95 % Confidence Intervals (CIs) for Colon Cancer in the Cholecystectomy Cohort versus the Varicose Vein Cohort

Induction (Years)	Cohort		Individuals with Colon Cancer		SIR	95 % CI
	Person-years at Risk	Individuals	Observed	Expected		
0	809,974.8	89,667	600	522.7	1.15	(1.06-1.24)
1	722,203.8	85,409	522	483.5	1.08	(0.99-1.18)
2	640,195.1	78,927	471	443.3	1.06	(0.97-1.16)
5	432,082.0	60,331	343	317.1	1.08	(0.97-1.20)
10	193,721.7	36,240	161	145.9	1.10	(0.94-1.29)
15	57,067.3	19,168	53	46.0	1.15	(0.86-1.51)

Table 24. Comparison of Standardized Incidence Ratios (SIRs) and 95 % Confidence Intervals (CIs) for Cancer in the Cholecystectomy Cohort versus the Varicose Vein Cohort, by Sex

Induction (Years)	Cohort			Individuals with Colon Cancer		SIR	95 % CI
	Person-years at Risk	Individuals	Observed	Expected			
Females							
0	596,795.1	64,734	334	353.8	0.94	(0.85-1.05)	
1	533,305.3	61,841	282	329.8	0.86	(0.76-0.96)	
2	473,900.9	57,227	255	305.2	0.84	(0.74-0.95)	
5	322,525.7	44,090	182	220.9	0.82	(0.71-0.95)	
10	146,774.8	27,064	88	101.8	0.86	(0.70-1.07)	
15	43,924.8	14,595	26	35.9	0.72	(0.47-1.06)	
Males							
0	213,179.7	24,933	266	168.9	1.57	(1.39-1.78)	
1	188,898.5	23,568	240	153.8	1.56	(1.37-1.77)	
2	166,294.2	21,700	216	138.1	1.56	(1.36-1.79)	
5	109,556.3	16,241	161	96.2	1.67	(1.43-1.95)	
10	46,946.9	9,176	73	44.1	1.65	(1.30-2.08)	
15	13,142.5	4,573	27	10.1	2.67	(1.76-3.89)	

Table 25. Comparison of Standardized Incidence Ratios (SIRs) and 95 % Confidence Intervals (CIs) for Right Colon Cancer in the Cholecystectomy Cohort versus the Varicose Vein Cohort

Induction (Years)	Cohort		Individuals with Colon Cancer		SIR	95 % CI
	Person-years at Risk	Individuals	Observed	Expected		
0	809,974.8	89,667	185	163.6	1.13	(0.97-1.31)
1	722,203.8	85,409	152	151.7	1.00	(0.85-1.18)
2	640,195.1	78,927	137	139.3	0.98	(0.83-1.16)
5	432,082.0	60,331	103	98.4	1.05	(0.85-1.27)
10	193,721.7	36,240	49	43.5	1.13	(0.83-1.49)
15	57,067.3	19,168	23	16.5	1.39	(0.88-2.09)

Table 26. Comparison of Standardized Incidence Ratios (SIRs) and 95 % Confidence Intervals (CIs) for Right Colon Cancer in the Cholecystectomy Cohort versus the Varicose Vein Cohort, by Sex

Induction (Years)	Cohort		Individuals with Colon Cancer	SIR	95 % CI
	Person-years at Risk	Individuals	Observed	Expected	
Females					
0	596,795.1	64,734	109	140.1	0.78 (0.64-0.94)
1	533,305.3	61,841	87	130.6	0.67 (0.53-0.82)
2	473,900.9	57,227	78	120.8	0.65 (0.51-0.81)
5	322,525.7	44,090	58	87.2	0.66 (0.50-0.86)
10	146,774.8	27,064	32	41.9	0.76 (0.52-1.08)
15	43,924.8	14,595	13	16.5	0.79 (0.42-1.35)
Males					
0	213,179.7	24,933	76	23.5	3.23 (2.55-4.05)
1	188,898.5	23,568	65	21.0	3.09 (2.39-3.94)
2	166,294.2	21,700	59	18.5	3.19 (2.43-4.12)
5	109,556.3	16,241	45	11.1	4.05 (2.95-5.42)
10	46,946.9	9,176	17	1.6	10.70 (6.19-17.05)
15	13,142.5	4,573	10	0*	100 (48.08-183.82)

*0.1 used for calculation of SIR and CI.

Table 27. Comparison of Standardized Incidence Ratios (SIRs) and 95 % Confidence Intervals (CIs) for Mid-Colon Cancer in the Cholecystectomy Cohort versus the Varicose Vein Cohort

Induction (Years)	Cohort		Individuals with Colon Cancer		SIR	95 % CI
	Person-years at Risk	Individuals	Observed	Expected		
0	809,974.8	89,667	107	56.1	1.91	(1.56-2.31)
1	722,203.8	85,409	88	51.9	1.70	(1.36-2.09)
2	640,195.1	78,927	82	47.7	1.72	(1.37-2.14)
5	432,082.0	60,331	60	32.6	1.84	(1.40-2.37)
10	193,721.7	36,240	27	10.1	2.67	(1.76-3.89)
15	57,067.3	19,168	*	*	*	*

* Insufficient number of cases for calculation of SIR and CI.

Table 28. Comparison of Standardized Incidence Ratios (SIRs) and 95 % Confidence Intervals (CIs) for Mid-Colon Cancer in the Cholecystectomy Cohort versus the Varicose Vein Cohort, by Sex

Induction (Years)	Cohort		Individuals with Colon Cancer		SIR	95 % CI
	Person-years at Risk	Individuals	Observed	Expected		
Females						
0	596,795.1	64,734	58	41.0	1.41	(1.07-1.83)
1	533,305.3	61,841	46	38.2	1.20	(0.88-1.61)
2	473,900.9	57,227	44	35.3	1.25	(0.91-1.67)
5	322,525.7	44,090	31	24.6	1.26	(0.86-1.79)
10	146,774.8	27,064	15	6.7	2.23	(1.24-3.67)
15	43,924.8	14,595	*	*	*	*
Males						
0	213,179.7	24,933	49	15.0	3.26	(2.41-4.31)
1	188,898.5	23,568	42	13.7	3.06	(2.21-4.14)
2	166,294.2	21,700	38	12.3	3.08	(2.18-4.23)
5	109,556.3	16,241	29	8.1	3.60	(2.42-5.17)
10	46,946.9	9,176	12	3.4	3.57	(1.84-6.24)
15	13,142.5	4,573	*	*	*	*

*Insufficient number of cases for calculation of SIR and CI.

Table 29. Comparison of Standardized Incidence Ratios (SIRs) and 95 % Confidence Intervals (CIs) for Left Colon Cancer in the Cholecystectomy Cohort versus the Varicose Vein Cohort

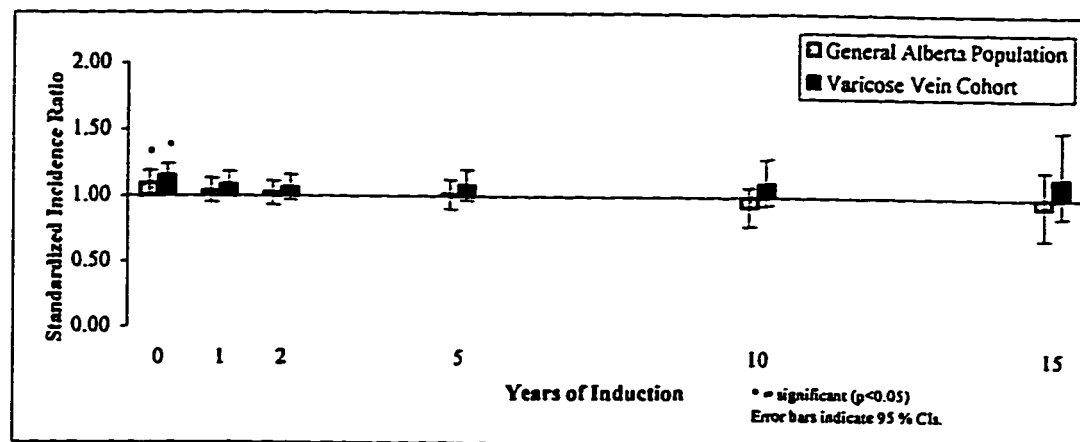
Induction (Years)	Cohort		Individuals with Colon Cancer		SIR	95 % CI
	Person-years at Risk	Individuals	Observed	Expected		
0	809,974.8	89,667	271	253.1	1.07	(0.95-1.21)
1	722,203.8	85,409	250	232.9	1.07	(0.94-1.22)
2	640,195.1	78,927	223	212.3	1.05	(0.92-1.20)
5	432,082.0	60,331	160	151.6	1.06	(0.90-1.23)
10	193,721.7	36,240	75	75.9	0.99	(0.78-1.24)
15	57,067.3	19,168	21	22.2	0.95	(0.59-1.45)

Table 30. Comparison of Standardized Incidence Ratios (SIRs) and 95 % Confidence Intervals (CIs) for Left Colon Cancer in the Cholecystectomy Cohort versus the Varicose Vein Cohort, by Sex

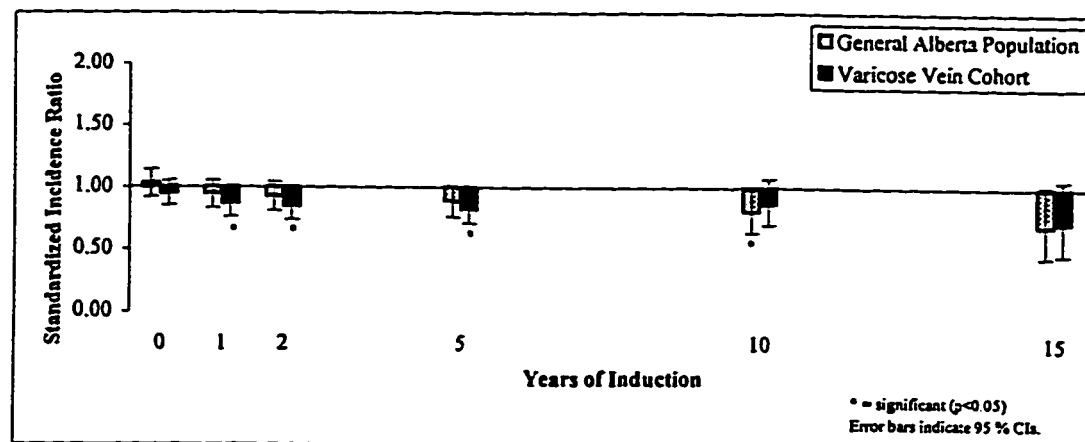
Induction (Years)	Cohort		Individuals with Colon Cancer		SIR	95 % CI
	Person-years at Risk	Individuals	Observed	Expected		
Females						
0	596,795.1	64,734	146	136.7	1.07	(0.90-1.26)
1	533,305.3	61,841	131	126.9	1.03	(0.86-1.23)
2	473,900.9	57,227	117	116.9	1.00	(0.83-1.20)
5	322,525.7	44,090	82	83.6	0.98	(0.78-1.22)
10	146,774.8	27,064	34	40.9	0.83	(0.57-1.16)
15	43,924.8	14,595	10	14.6	0.68	(0.33-1.26)
Males						
0	213,179.7	24,933	125	116.4	1.07	(0.89-1.28)
1	188,898.5	23,568	119	106.0	1.12	(0.93-1.34)
2	166,294.2	21,700	106	95.3	1.11	(0.91-1.35)
5	109,556.3	16,241	78	68.1	1.15	(0.91-1.43)
10	46,946.9	9,176	41	35.0	1.17	(0.84-1.59)
15	13,142.5	4,573	11	7.5	1.46	(0.73-2.62)

Figure 5. Analysis of Standardized Incidence Ratios for Colon Cancer Following Cholecystectomy Using Different Comparison Groups: the General Alberta Population versus the Varicose Vein Cohort

Male & Female



Female



Male

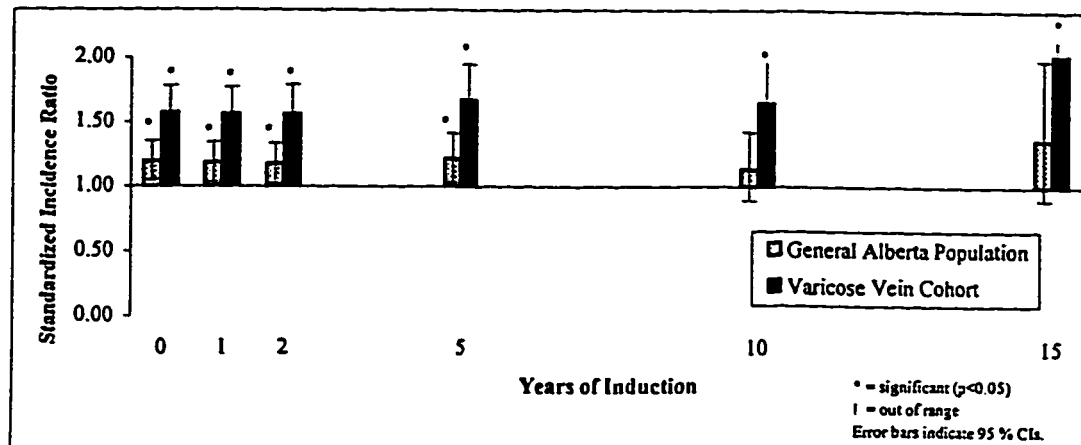
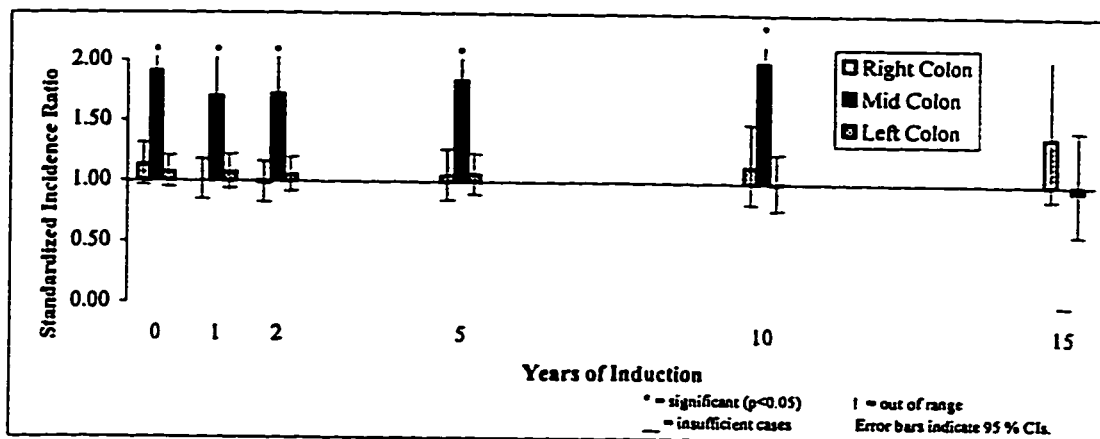
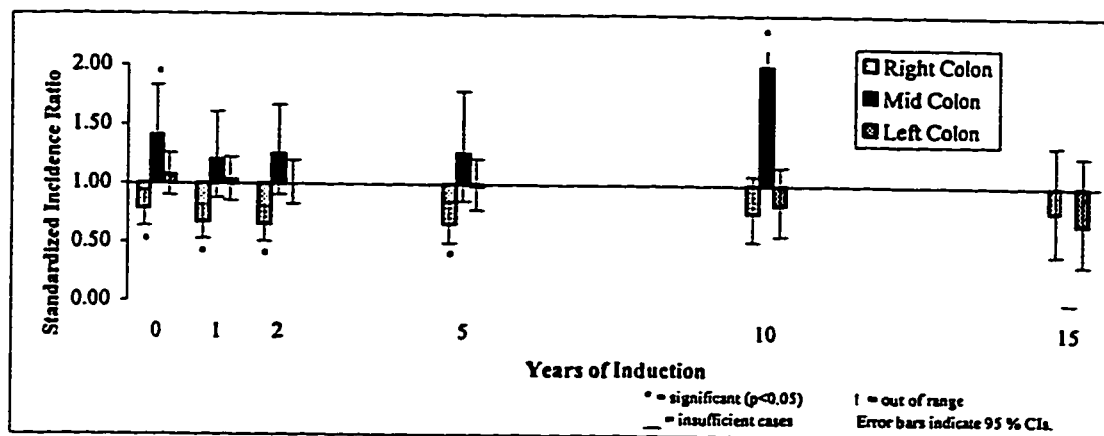


Figure 6. Comparison of Standardized Incidence Ratios for Colon Cancer Following Cholecystectomy: by Colon Subsite, Using Rates of Colon Cancer in the Varicose Vein Cohort

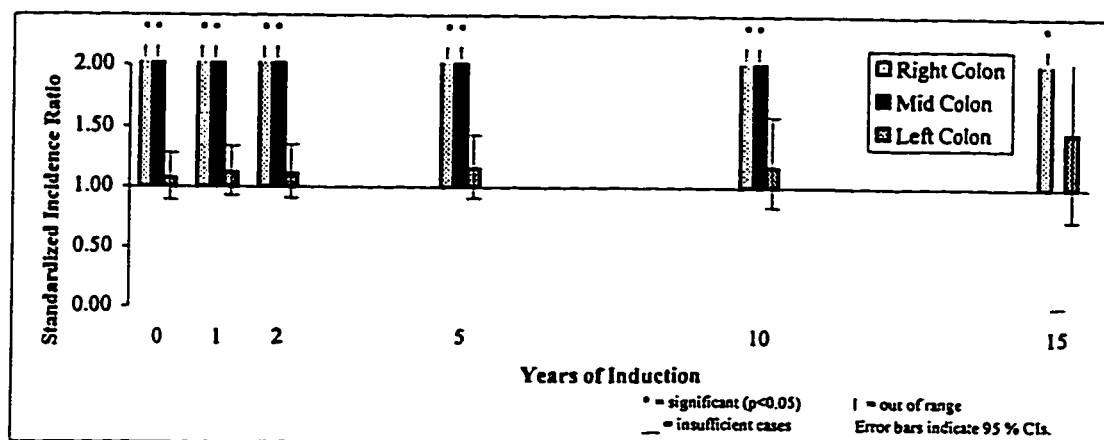
Male & Female



Female



Male



SIRs were also calculated for colon cancer risk in the varicose vein cohort compared to the general population to aid in the interpretation of the results (Tables 31 and 32). Figure 7 shows that, when compared to the general population, the direction of the risk was different for the cholecystectomy cohort compared to the varicose vein cohort. In particular, the risk was somewhat increased for males in the cholecystectomy cohort and somewhat decreased for males in the varicose vein cohort. The effect was magnified in subsite analysis (Tables 33 - 38).

3. *Colon Cancer Risk in the Cholecystectomy Cohort Adjusted for Confounding by Gastric Procedures*

The potential for confounding in the cholecystectomy-colon cancer relationship was partly addressed by evaluating the risk adjusted for the presence of gastric procedures.

As with the other AHCIP cohorts, the gastric procedures cohort underwent processing before it was submitted for analysis. Initially, there were records for 21,950 individuals in the cohort. A total of 827 individuals were diagnosed with gallbladder, biliary or pancreatic cancer, and most (685) were removed from the cohort because the diagnosis preceded the service.

Only 363 individuals (1.7% of the cohort) were diagnosed with colon cancer, and many of these (207) were removed from the cohort because the gastric procedures were performed at or after diagnosis of colon cancer. This left 21,058 individuals in the cohort. However, some (52) individuals had a service date after the last date of

Table 31. Comparison of Standardized Incidence Ratios (SIRs) and 95 % Confidence Intervals (CIs) for Colon Cancer in the Varicose Vein Cohort versus the General Alberta Population

Induction (Years)	Cohort		Individuals with Colon Cancer		SIR	95 % CI
	Person-years at Risk	Individuals	Observed	Expected		
0	159,743.3	19,701	100	94.7	1.06	(0.86-1.28)
1	140,297.3	18,942	92	87.5	1.05	(0.85-1.29)
2	122,247.6	17,152	82	80.1	1.02	(0.81-1.27)
5	77,578.1	12,625	59	58.6	1.01	(0.77-1.30)
10	28,826.3	7,155	28	26.9	1.04	(0.69-1.50)
15	4,617.4	2,608	*	*	*	*

* Insufficient number of cases for calculation of SIR and CI.

Table 32. Comparison of Standardized Incidence Ratios (SIRs) and 95 % Confidence Intervals (CIs) for Colon Cancer in the Varicose Vein Cohort versus the General Alberta Population, by Sex

Induction (Years)	Cohort		Individuals	Individuals with Colon Cancer		SIR	95 % CI
	Person-years at Risk			Observed	Expected		
Females							
0	119,001.4		14,522	72	62.2	1.16	(0.91-1.46)
1	104,655.5		13,991	68	57.4	1.18	(0.92-1.50)
2	91,309.5		12,693	61	52.6	1.16	(0.89-1.49)
5	58,258.2		9,354	46	38.6	1.19	(0.87-1.59)
10	21,855.7		5,375	23	17.9	1.29	(0.81-1.93)
15	3,492.8		1,996	*	*	*	*
Males							
0	40,741.8		5,179	28	32.6	0.86	(0.57-1.24)
1	35,641.8		4,951	24	30.1	0.80	(0.51-1.19)
2	30,938.1		4,459	21	27.5	0.76	(0.47-1.17)
5	19,319.9		3,271	13	20.0	0.65	(0.35-1.11)
10	6,970.6		1,780	*	*	*	*
15	1,124.5		612	*	*	*	*

* Insufficient number of cases for calculation of SIR and CI.

Figure 7. Standardized Incidence Ratios for Colon Cancer Following Stripping and Ligation of Varicose Veins, Using Rates of Colon Cancer in the General Alberta Population

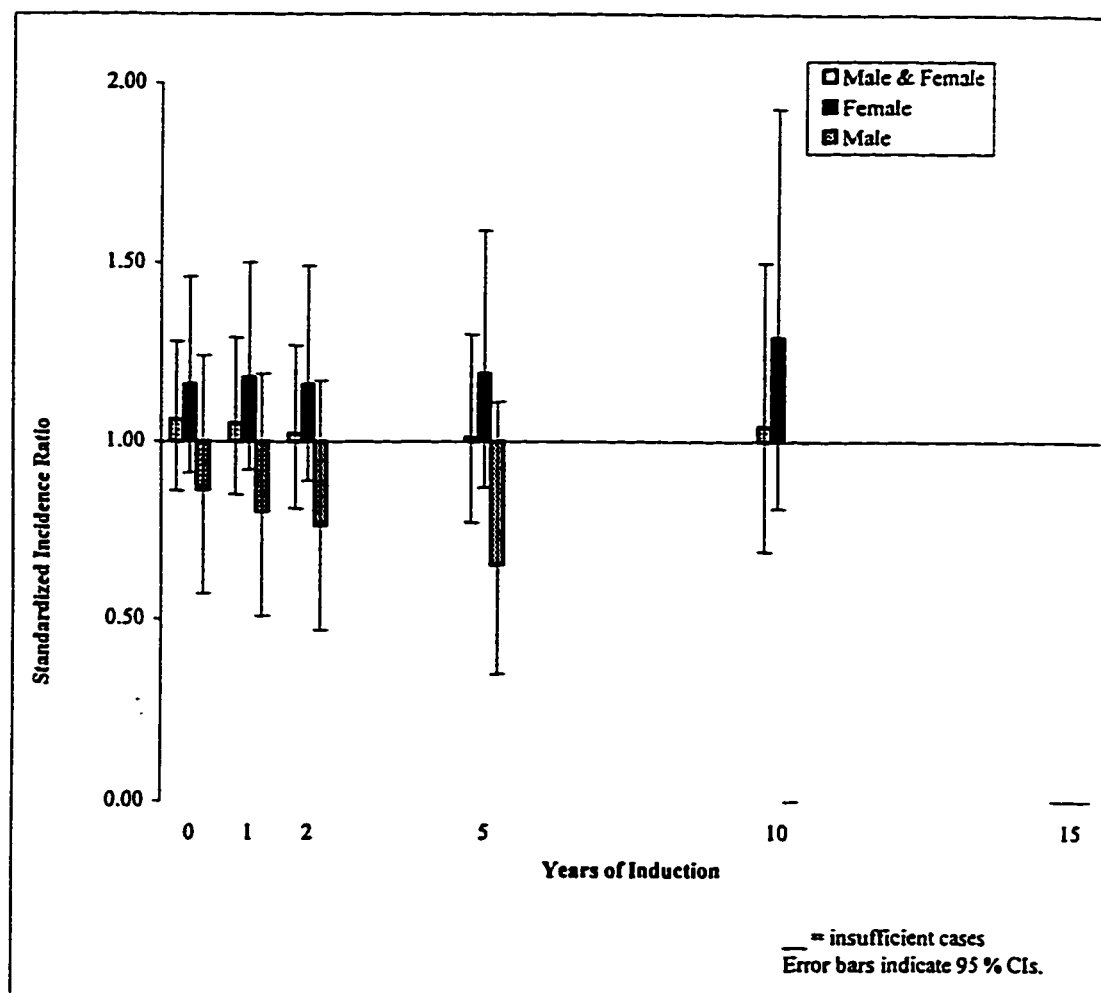


Table 33. Comparison of Standardized Incidence Ratios (SIRs) and 95 % Confidence Intervals (CIs) for Right Colon Cancer in the Varicose Vein Cohort versus the General Alberta Population

Induction (Years)	Cohort		Individuals with Colon Cancer		SIR	95 % CI
	Person-years at Risk	Individuals	Observed	Expected		
0	159,743.3	19,701	31	27.5	1.13	(0.77-1.60)
1	140,297.3	18,942	29	25.5	1.14	(0.76-1.63)
2	122,247.6	17,152	25	23.6	1.06	(0.68-1.57)
5	77,578.1	12,625	19	17.6	1.08	(0.65-1.69)
10	28,826.3	7,155	*	*	*	*
15	4,617.4	2,608	*	*	*	*

* Insufficient number of cases for calculation of SIR and CI.

Table 34. Comparison of Standardized Incidence Ratios (SIRs) and 95 % Confidence Intervals (CIs) for Right Colon Cancer in the Varicose Vein Cohort versus the General Alberta Population, by Sex

Induction (Years)	Cohort		Individuals	Individuals with Colon Cancer		SIR	95 % CI
	Person-years at Risk			Observed	Expected		
Females							
0	119,001.4		14,522	27	18.6	1.45	(0.95-2.11)
1	104,655.5		13,991	25	17.3	1.44	(0.93-2.13)
2	91,309.5		12,693	23	16.0	1.44	(0.91-2.15)
5	58,258.2		9,354	17	12.1	1.41	(0.82-2.25)
10	21,855.7		5,375	*	*	*	*
15	3,492.8		1,996	*	*	*	*
Males							
0	40,741.8		5,179	*	*	*	*
1	35,641.8		4,951	*	*	*	*
2	30,938.1		4,459	*	*	*	*
5	19,319.9		3,271	*	*	*	*
10	6,970.6		1,780	*	*	*	*
15	1,124.5		612	*	*	*	*

* Insufficient number of cases for calculation of SIR and CI.

Table 35. Comparison of Standardized Incidence Ratios (SIRs) and 95 % Confidence Intervals (CIs) for Mid-Colon Cancer in the Varicose Vein Cohort versus the General Alberta Population

Induction (Years)	Cohort		Individuals with Colon Cancer		SIR	95 % CI
	Person-years at Risk	Individuals	Observed	Expected		
0	159,743.3	19,701	*	*	*	*
1	140,297.3	18,942	*	*	*	*
2	122,247.6	17,152	*	*	*	*
5	77,578.1	12,625	*	*	*	*
10	28,826.3	7,155	*	*	*	*
15	4,617.4	2,608	*	*	*	*

* Insufficient number of cases for calculation of SIR and CI.

Table 36. Comparison of Standardized Incidence Ratios (SIRs) and 95 % Confidence Intervals (CIs) for Mid-Colon Cancer in the Varicose Vein Cohort versus the General Alberta Population, by Sex

Induction (Years)	Cohort		Individuals	Individuals with Colon Cancer		SIR	95 % CI
	Person-years at Risk	Individuals		Observed	Expected		
Females							
0	119,001.4	14,522		*	*	*	*
1	104,655.5	13,991		*	*	*	*
2	91,309.5	12,693		*	*	*	*
5	58,258.2	9,354		*	*	*	*
10	21,855.7	5,375		*	*	*	*
15	3,492.8	1,996		*	*	*	*
Males							
0	40,741.8	5,179		*	*	*	*
1	35,641.8	4,951		*	*	*	*
2	30,938.1	4,459		*	*	*	*
5	19,319.9	3,271		*	*	*	*
10	6,970.6	1,780		*	*	*	*
15	1,124.5	612		*	*	*	*

* Insufficient number of cases for calculation of SIR and CI.

Table 37. Comparison of Standardized Incidence Ratios (SIRs) and 95 % Confidence Intervals (CIs) for Left Colon Cancer in the Varicose Vein Cohort versus the General Alberta Population

Induction (Years)	Cohort		Individuals with Colon Cancer		SIR	95 % CI
	Person-years at Risk	Individuals	Observed	Expected		
0	159,743.3	19,701	52	45.6	1.14	(0.85-1.50)
1	140,297.3	18,942	48	42.0	1.14	(0.84-1.52)
2	122,247.6	17,152	43	38.2	1.12	(0.81-1.52)
5	77,578.1	12,625	31	27.6	1.12	(0.76-1.60)
10	28,826.3	7,155	15	12.4	1.21	(0.68-2.00)
15	4,617.4	2,608	*	*	*	*

* Insufficient number of cases for calculation of SIR and CI.

Table 38. Comparison of Standardized Incidence Ratios (SIRs) and 95 % Confidence Intervals (CIs) for Left Colon Cancer in the Varicose Vein Cohort versus the General Alberta Population, by Sex

Induction (Years)	Cohort		Individuals with Colon Cancer	SIR	95 % CI
	Person-years at Risk	Individuals	Observed	Expected	
Females					
0	119,001.4	14,522	32	29.3	(0.75-1.54)
1	104,655.5	13,991	31	26.9	(0.78- 1.64)
2	91,309.5	12,693	27	24.5	(0.73-1.61)
5	58,258.2	9,354	21	17.6	(0.73-1.82)
10	21,855.7	5,375	11	8.0	(0.69-2.47)
15	3,492.8	1,996	*	*	*
Males					
0	40,741.8	5,179	20	16.3	(0.75-1.89)
1	35,641.8	4,951	17	15.1	(0.66-1.81)
2	30,938.1	4,459	16	13.8	(0.66-1.89)
5	19,319.9	3,271	10	9.9	(0.48-1.85)
10	6,970.6	1,780	*	*	*
15	1,124.5	612	*	*	*

* Insufficient number of cases for calculation of SIR and CIs.

follow-up while others (18) had no follow-up (first and last dates of follow-up were the same) resulting in their removal from the cohort. This left records for 20,988 individuals with a history of gastric surgery (95.6% of the original cohort), including data for 156 individuals with colon cancer. As shown in Tables 39 and 40, there was no change in risk estimates following adjustment.

E. Sensitivity Analysis: Influence of Completeness of AHCIP Effective Dates

As noted, some individuals (36,702, or 28.5%) identified in the AHCIP datasets were missing the effective date of AHCIP coverage. A sensitivity analysis was performed to examine the effect of assuming that these missing dates were actually July 1, 1969, the first date of public health insurance in Alberta.

SIRs were re-calculated for the 65,086 individuals in the cholecystectomy cohort whose records had complete date information. A higher proportion of females (21,662, or 32.8%) were lost as a result compared to males (4,418, or 17.5%); records for 81 individuals with colon cancer (13.4% of the original group with cancer) were also removed, leaving 525 for analysis.

Tables 41 and 42 display the SIRs using the Alberta population as the comparison group. Figure 8 shows that the analysis of records with complete date information produced SIRs slightly less than the initial estimates and colon cancer risk for males became statistically nonsignificant. The SIRs were also calculated using the varicose vein cohort as the comparison group. For this analysis, both cohorts

Table 39. Comparison of Standardized Incidence Ratios (SIRs) and 95 % Confidence Intervals (CIs) for Colon Cancer in the Cholecystectomy Cohort versus the General Alberta Population, Adjusting for Presence of Gastric Procedures

Induction (Years)	Cohort		Individuals with Colon Cancer		SIR	95 % CI
	Person-years at Risk	Individuals	Observed	Expected		
0	826,052.4	91,166	606	554.5	1.09	(1.01-1.18)
1	736,791.0	86,876	527	510.7	1.03	(0.95-1.12)
2	653,350.7	80,328	476	467.8	1.02	(0.93-1.11)
5	441,350.5	61,509	348	345.5	1.01	(0.90-1.12)
10	198,105.6	37,021	162	175.9	0.92	(0.78-1.07)
15	58,406.8	19,609	53	57.1	0.93	(0.69-1.21)

Table 40. Comparison of Standardized Incidence Ratios (SIRs) and 95 % Confidence Intervals (CIs) for Colon Cancer in the Cholecystectomy Cohort versus the General Alberta Population, Adjusting for Presence of Gastric Procedures, by Sex

Induction (Years)	Cohort		Individuals with Colon Cancer	SIR	95 % CI
	Person-years at Risk	Individuals	Observed		
Females					
0	610,347.1	65,973	339	1.03	(0.92-1.14)
1	545,624.7	63,055	286	0.94	(0.83-1.05)
2	485,031.1	58,393	259	0.92	(0.81-1.04)
5	330,417.3	45,082	186	0.88	(0.76-1.01)
10	150,499.8	27,731	89	0.80	(0.65-0.99)
15	45,078.6	14,966	26	0.70	(0.46-1.02)
Males					
0	215,705.3	25,193	267	1.19	(1.05-1.34)
1	191,166.3	23,821	241	1.17	(1.03-1.33)
2	168,319.6	21,935	217	1.16	(1.01-1.33)
5	110,933.2	16,427	162	1.21	(1.03-1.41)
10	47,605.9	9,290	73	1.13	(0.88-1.42)
15	13,328.2	4,643	27	1.36	(0.90-1.98)

Table 41. Comparison of Standardized Incidence Ratios (SIRs) and 95 % Confidence Intervals (CIs) for Colon Cancer in the
Cholecystectomy Cohort versus the General Alberta Population, Individuals with Complete AHCIP Effective Dates
Only

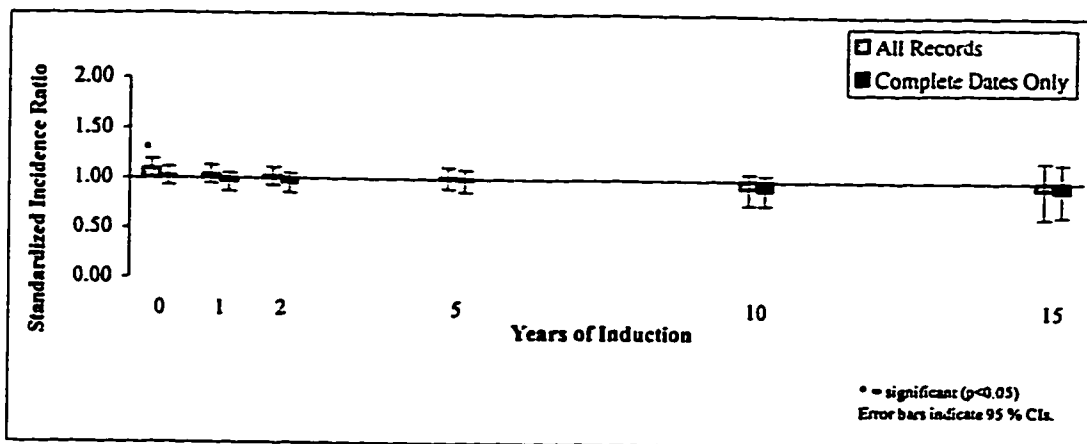
Induction (Years)	Cohort		Individuals with Colon Cancer		SIR	95 % CI
	Person-years at Risk	Individuals	Observed	Expected		
0	637,122.7	65,086	525	515.8	1.02	(0.93-1.11)
1	572,813.5	63,093	458	477.4	0.96	(0.87-1.05)
2	511,851.8	59,109	417	439.1	0.95	(0.86-1.05)
5	353,379.1	46,827	320	328.1	0.98	(0.87-1.09)
10	163,483.2	29,706	152	168.6	0.90	(0.76-1.06)
15	49,189.4	16,398	49	54.7	0.90	(0.66-1.19)

Table 42. Comparison of Standardized Incidence Ratios (SIRs) and 95 % Confidence Intervals (CIs) for Colon Cancer in the Cholecystectomy Cohort versus the General Alberta Population, Individuals with Complete AHCIP Effective Dates Only, by Sex

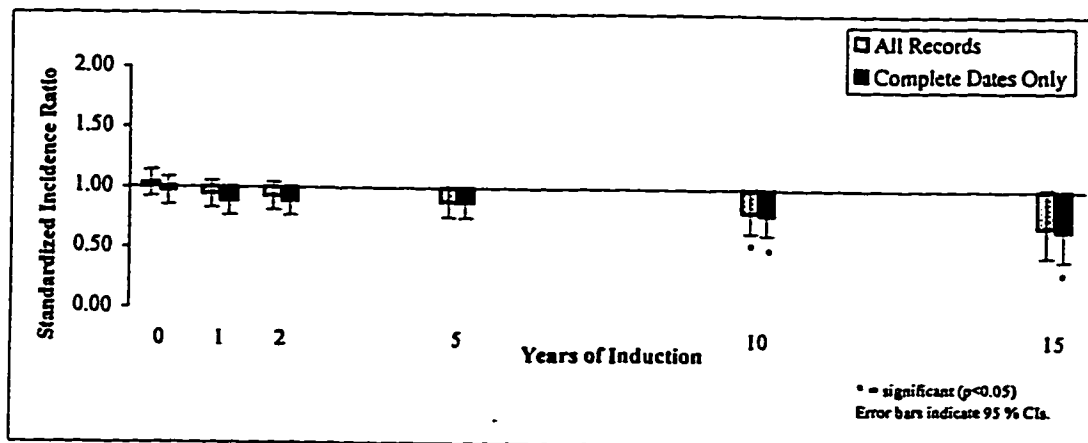
Induction (Years)	Cohort		Individuals	Individuals with Colon Cancer		SIR	95 % CI
	Person-years at Risk			Observed	Expected		
Females							
0	444,274.9		44,311	291	302.6	0.96	(0.85-1.08)
1	400,460.7		43,027	247	281.4	0.88	(0.77-1.00)
2	358,872.8		40,368	229	260.1	0.88	(0.77-1.00)
5	250,302.5		32,254	172	197.6	0.87	(0.75-1.01)
10	118,081.8		21,002	82	104.5	0.78	(0.62-0.97)
15	36,314.0		11,941	23	35.1	0.66	(0.42-0.98)
Males							
0	192,847.8		20,775	234	213.2	1.10	(0.96-1.25)
1	172,352.8		20,066	211	196.0	1.08	(0.94-1.23)
2	152,979.0		18,741	188	179.0	1.05	(0.90-1.21)
5	103,076.7		14,573	148	130.5	1.13	(0.96-1.33)
10	45,401.4		8,704	70	64.1	1.09	(0.85-1.38)
15	12,875.4		4,457	26	19.6	1.32	(0.87-1.94)

Figure 8. Sensitivity Analysis: Standardized Incidence Ratios for Colon Cancer Following Cholecystectomy: All Records versus Records With Complete Effective Dates, Using Rates of Colon Cancer in the General Alberta Population

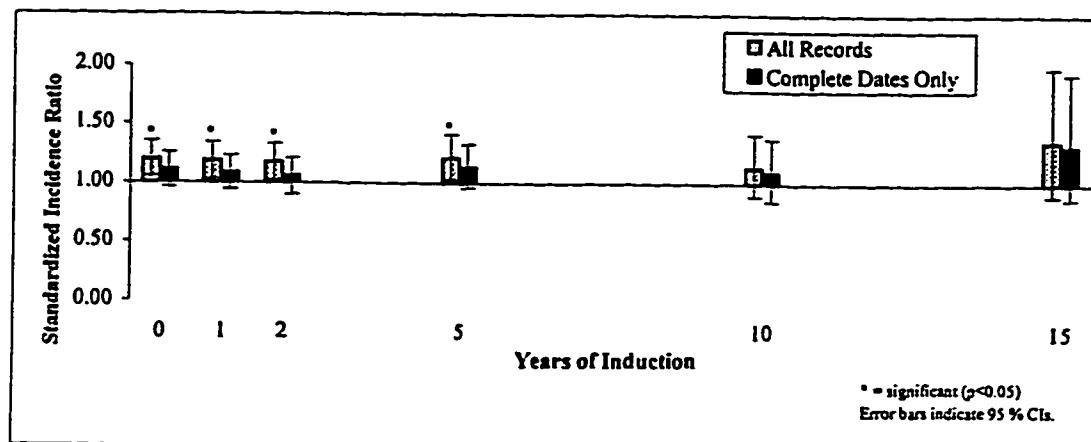
Male & Female



Female



Male



incorporated only individuals with complete AHCIP effective dates (Tables 43 and 44). As shown in Figure 9, the overall SIRs for the early (0- to 5-year) induction periods were significantly increased, and low rates in females became statistically nonsignificant.

F. Stratified Analysis: Characteristics of Individuals with Colon Cancer versus the Cohort

In total, there were 606 individuals with colon cancer that occurred after cholecystectomy, representing only 0.7% of the total cholecystectomy cohort. Comparing these individuals to the remainder of the cholecystectomy cohort, they were older at the time of service (average age, 62.6 years versus 46.0 years for members of the cohort without colon cancer, $p < 0.001$) and were also more likely to have had their surgery in earlier years (median, 1980 versus 1983 for non-cases; $X^2=171.9$ with 20 degrees of freedom, $p=0.001$). These values make intuitive sense, as younger individuals and those with more recent procedures would not have had adequate time to develop colon cancer. Interestingly, individuals with colon cancer were divided almost evenly between the sexes, with 339 (55.9%) females and 267 (44.1%) males, while there was a considerably higher proportion of females in the remainder of the cholecystectomy cohort (65,634 females, or 72.5%; $X^2=82.3$ with 1 degree of freedom, $p=0.001$).

In the colon cancer cohort, 11,718 individuals were diagnosed since April,

Table 43. Comparison of Standardized Incidence Ratios (SIRs) and 95 % Confidence Intervals (CIs) for Colon Cancer in the Cholecystectomy Cohort versus the Varicose Vein Cohort, Individuals with Complete AHCIP Effective Dates Only

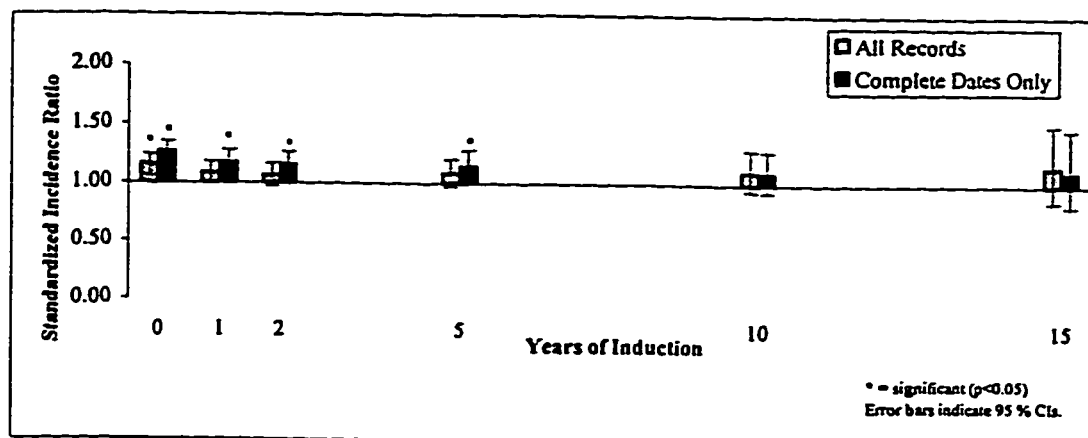
Induction (Years)	Cohort		Individuals with Colon Cancer		SIR	95 % CI
	Person-years at Risk	Individuals	Observed	Expected		
0	623,819.7	63,876	519	413.5	1.26	(1.15-1.37)
1	560,714.8	61,908	453	386.7	1.17	(1.07-1.28)
2	500,907.4	57,978	412	359.5	1.15	(1.04-1.26)
5	345,587.7	45,868	315	275.7	1.14	(1.02-1.28)
10	159,713.7	29,056	151	138.0	1.09	(0.93-1.28)
15	48,009.6	16,013	49	44.2	1.11	(0.82-1.47)

Table 44. Comparison of Standardized Incidence Ratios (SIRs) and 95 % Confidence Intervals (CIs) for Colon Cancer in the Cholecystectomy Cohort versus the Varicose Vein Cohort, Individuals with Complete AHCIP Effective Dates Only, by Sex

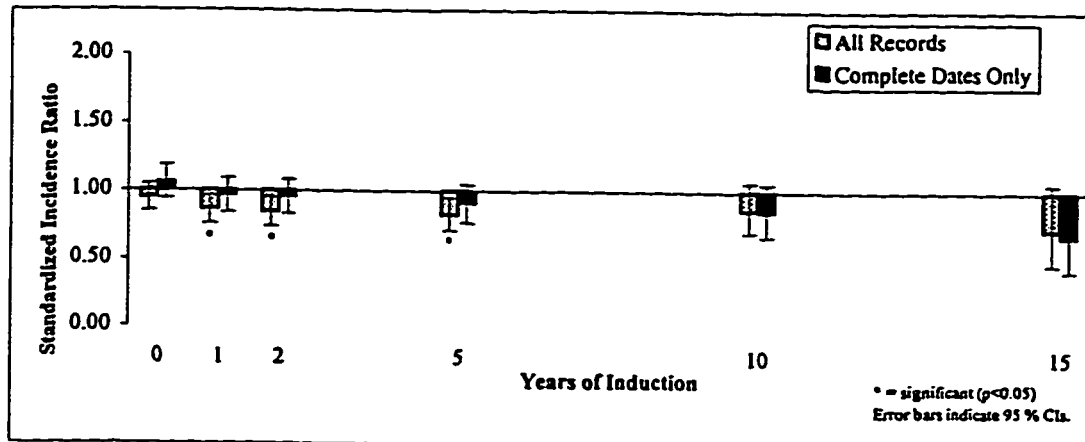
Induction (Years)	Cohort		Individuals	Individuals with Colon Cancer		SIR	95 % CI
	Person-years at Risk			Observed	Expected		
Females							
0	433,297.9		43,334	286	269.0	1.06	(0.94-1.19)
1	390,456.3		42,069	243	253.4	0.96	(0.84-1.09)
2	349,806.1		39,447	225	237.5	0.95	(0.83-1.08)
5	243,808.5		31,462	168	185.3	0.91	(0.77-1.05)
10	114,943.1		20,460	81	95.5	0.85	(0.67-1.06)
15	35,313.5		11,623	23	34.4	0.67	(0.42-1.00)
Males							
0	190,521.8		20,542	233	144.5	1.61	(1.41-1.83)
1	170,258.6		19,839	210	133.4	1.57	(1.37-1.80)
2	151,101.4		18,531	187	122.0	1.53	(1.32-1.77)
5	101,779.2		14,406	147	90.5	1.62	(1.37-1.91)
10	44,770.6		8,596	70	42.5	1.65	(1.29-2.08)
15	12,696.1		4,390	26	9.8	2.66	(1.74-3.90)

Figure 9. Sensitivity Analysis: Standardized Incidence Ratios for Colon Cancer Following Cholecystectomy: All Records versus Records With Complete Effective Dates, Using Rates of Colon Cancer in the Varicose Vein Cohort

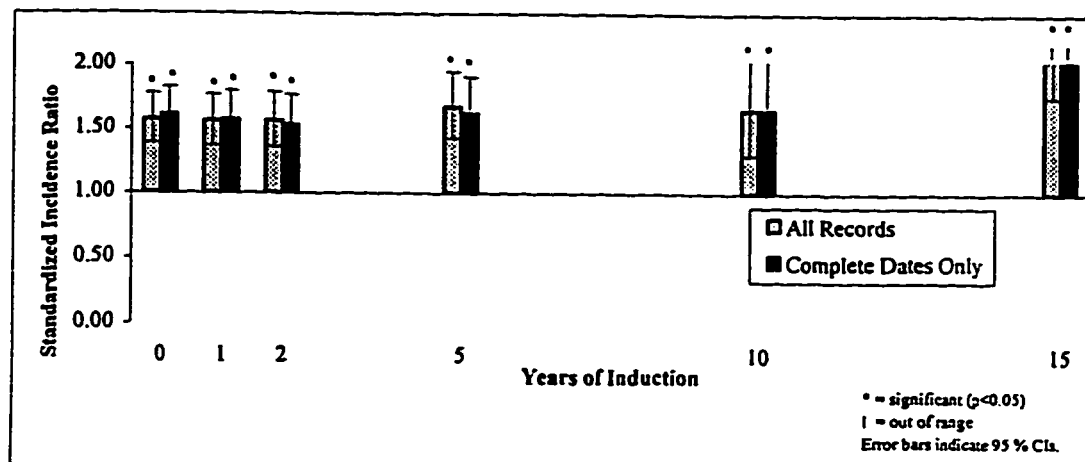
Male & Female



Female



Male



1973 (the first date of service in the AHCIP cohorts). The 606 individuals with cholecystectomy before diagnosis represented only 5.2% of the total cancer cohort. The average age at diagnosis was similar for cases regardless of cholecystectomy status (average age, 69.5 years for those with cholecystectomy versus 68.6 years for those without cholecystectomy; not significant) but the median year of diagnosis was later for patients with cholecystectomy (median diagnosis year, 1988 for cholecystectomy group versus 1985 for others; $X^2=136.8$ with 20 degrees of freedom, $p=0.001$). Individuals with cholecystectomy were more likely to be female (339, or 55.9% in the cholecystectomy group versus 5,663, or 51.0% for others; $X^2=11.0$ with 1 degree of freedom, $p=0.001$).

Initially, analysis by stage and histology of the tumour was proposed. However, stage information was missing for the majority of the cancer cohort (8,647 missing, or 73.8%). In addition, most tumours were adenocarcinomas regardless of the individual's cholecystectomy history (497, or 82.0% for the cholecystectomy group versus 9,023, or 81.2% for others; not significant). A slightly higher proportion of patients without a history of cholecystectomy had tumours with grade 0 (50, or 8.3% for cholecystectomy group versus 1,669, or 15.0% for others; $X^2=26.9$ with 5 degrees of freedom, $p=0.001$); however, there was little difference for other grades suggesting limited clinical significance.

These characteristics were compared for individuals with colon cancer following cholecystectomy and the remainder of the cohorts for each induction period used to calculate SIRs. The only differences by induction appeared in the colon

cancer cohort, where differences in sex and grade distribution became nonsignificant at 5 years. Under assumptions of 5 or 10 years of induction, individuals with a history of cholecystectomy were slightly older when they were diagnosed with colon cancer, but the difference was not clinically significant. Average age at diagnosis for individuals with cholecystectomy was 70.3 years compared to 69.0 years for others, assuming 5 years of induction ($p = 0.02$) and 68.9 years compared to 71.5 years, assuming 10 years of induction ($p = 0.003$).

G. Summary

Many estimates of colon cancer risk following cholecystectomy were calculated in this study. Table 45 presents a summary of the data, with a sensitivity analysis combining the SIRs for the different types of linkage, comparison groups, adjustments and levels of completeness of date information. Subsite-specific SIRs were not included.

The point estimates are those obtained by both types of linkage when the general population was used as the comparison group, since they were based on the most stable rates and were comparable with the approaches used in other cohort studies published in the literature. The range of SIRs is the highest and lowest point estimates obtained.

Table 45. Point Estimates of Standardized Incidence Ratios and Range of Values, by Induction

	Point Estimate	Range of Point Estimates	
		Lowest	Highest
Males and Females			
0	1.09	1.02	1.26
1	1.03	0.96	1.17
2	1.02	0.95	1.15
5	1.01	0.98	1.14
10	0.92	0.90	1.10
15	0.93	0.90	1.15
Females			
0	1.03	0.94	1.06
1	0.94	0.86	0.96
2	0.92	0.84	0.95
5	0.88	0.82	0.91
10	0.80	0.78	0.86
15	0.69	0.66	0.72
Males			
0	1.19	1.10	1.61
1	1.18	1.08	1.57
2	1.17	1.05	1.53
5	1.21	1.13	1.62
10	1.13	1.08	1.65
15	1.36	1.31	2.67

V. DISCUSSION

A. Overview

The nature of the risk for colon cancer imposed by cholecystectomy cannot readily be determined in simple qualitative or quantitative terms. The methodology used to study the association requires careful consideration and the putative etiologic association may exist only for subgroups. The study methods will be discussed first to provide the appropriate context for the discussion of the etiologic conclusions.

B. Methodology

1. *Data Quality*

a. General Comments

The effectiveness of record linkage is highly contingent on adequate data quality. The data from AHCIP were of particular concern, as the data were collected for business, not research, purposes. The most notable influences on AHCIP data quality were the presence of negative claims records in the original dataset, non-registrants who apparently received services, rates of missing identifiers, the agreement of birthdate and sex in the two AHCIP datasets and the concern for ineffective linkage in the AHCIP registrant database, resulting in multiple registrant history records for each individual, particularly before 1983. Several potential

problems were also identified in the data from the Alberta Cancer Registry, most notably the rate of missing identifiers.

In general, poor data quality increases error. Where the varicose vein cohort was used as the comparison group, the resulting SIRs may have been nonsignificant because of increased random error associated with that cohort's relatively small size and data quality issues that would affect both AHCIP cohorts similarly. However, where the general population was used as the comparison group, error associated with data quality in the exposed cohort would lead to bias. That is, increased measurement error would be expected to result in underlinkage and therefore, a relative decrease in the observed number of cases, without a concomitant decrease in the expected number of cases (i.e., SIRs below the true value).

i AHCIP Data

The discovery of the negative claims records was fortuitous, since including these records would have incorrectly inflated the number of claims per individual. Further, 190 patients (0.1 % of the cohort) without cholecystectomy would have been included in the exposed cohort. Under the bile acid hypothesis, these individuals would be at lower risk for colon cancer than the truly exposed individuals, so the error would be expected to dilute the magnitude of the effect. The affected proportion of the cohort was small, suggesting that the magnitude of the error effect was negligible.

Another problem with the AHCIP data arose with the identification of a number of individuals who did not have coverage at the time of procedure. The

influence of this problem was relatively small since only 0.2% of the cohort was affected, but the majority of the effect was probably among females in younger agegroups. Since young women are more likely to undergo last name and registration number changes than men or older women, it is likely that these individuals were Alberta residents at the time of the procedure, but use of the incorrect registrant number and underlinkage by AHCIP led to the apparent lapse in coverage. The data showing that extending coverage by one or two years would have resulted in valid coverage support this suggestion. In this study, any of these individuals who were found to have colon cancer were not included in the observed portion of the SIR, but they also had their person-years at risk truncated. This would lead to lower observed and lower expected values, possibly increasing error in the SIRs, but minimizing bias.

Very few individuals in the cohort had missing identifiers. Of the cohort, less than 0.1% were missing AHCIP numbers or date of birth, 0.3% were missing last name and none were missing sex. A larger proportion of the cohort (27.0%) were missing one initial, with serious implications for the discriminating power of the initials from the AHCIP datasets and rendering them only moderately useful in linkage. Interestingly, the case-cohort study by Goldbohm et al¹²³ referenced work by Van den Brandt et al¹⁶⁰, showing very high data availability in the Netherlands' data sources. As shown in Table 46, birthdate, gender, family name (equivalent to maiden name) and postal codes were always available in both the cohort file and the cancer registry, while others were less prevalent. Alberta data were very comparable, suggesting that data availability may be similar to that of other places. However, the

Table 46. Comparison of Variables Available for Linkage: Netherlands
(based on data from Van den Brandt¹⁶⁰) and Alberta

Variable	Cohort		Cancer Registry	
	Netherlands ¹⁶⁰	Alberta	Netherlands ¹⁶⁰	Alberta
Birthdate	100 %	95.2 %	100 %	98.8 %
Gender	100 %	100 %	100 %	100 %
Family Name	100 %	99.7 % *	100 %	100 % *
Prefix of family name	13 %	NA	12 %	NA
Married Name	87 %	99.7 % *	85 %	100 % *
First Initial	100 %	100 %	99 %	100 %
Place of Birth	100 %	NA	42 %	NA
Residential postal code	100 %	NA	100 %	NA

NA: Not Applicable

*: Current last name

Alberta information on last names may not be as reliable as the Dutch information, since the European data sources collected family name (meaning maiden name), which an individual retains for a lifetime, while the family name provided by the Alberta data sources was the current last name. Further, the family name provided by the AHCIP referred to the head registrant's last name and not the patient's last name. Thus the level of completeness of last name in Alberta's data sources cannot be used to infer data quality.

Validity was also assessed by examining the agreement within AHCIP files and between AHCIP files and the Alberta Cancer Registry. The only variables common to the AHCIP files were the unique personal identifiers (unique lifetime identifier, linked to the AHCIP registrant numbers), birthdate and sex. For the same unique identifiers, sex agreed consistently, but there was some disagreement for birthdate between the claims dataset and the registrant dataset. When birthdates were compared, 673 (0.5%) of the individual and service-specific records were missing different birthdate elements. Further information on validity was assessed by examining the agreement of variables in the pairs obtained by linking the cholecystectomy cohort with the colon cancer cases from the Alberta Cancer Registry. As shown in Table 6 (*Results*, page 70), the highest level of agreement was found for sex and last name soundex (over 99%) and the lowest agreement was for day of birth (85.7%) and middle initial (72.3%).

The level of variable disagreement among linked pairs suggests that some pairs may have been missed in record linkage and that there were some inaccuracies in

person-years at risk, but with only minor implications for the risk estimates. These data suggest priorities for future linkages between AHCIP files and the Alberta Cancer Registry. However, accurate assessment of data quality requires a follow-up study of individuals identified in the AHCIP files, to ascertain data accuracy in static variables (such as date of birth), and history in dynamic variables (such as names). This strategy would come under the auspices of Alberta Health and would be cumbersome to implement.

Previous studies using AHCIP data have suggested that underlinkage exists in the registrant dataset.¹⁶¹ A similar situation was observed in the current study, since the number of registrant numbers for each individual was lower for individuals receiving services in earlier years of the cohort (Table 47). Bias could result when the general population was used for comparison since person-years at risk and therefore the expected number of cases, may have been incorrectly inflated. Assuming that data from later years were more accurate, the person-years from the earlier periods (pre-1983) may have been overestimated. However, this bias should have been alleviated by the use of the alternative comparison group, which was subjected to the same conditions as the cholecystectomy cohort.

A final consideration with the AHCIP data is the extent to which exposure was captured. True cholecystectomies were likely always reported because physicians were not paid unless they were reported. Data entry errors, leading to the recording of cholecystectomy in a patient who did not receive the procedure, were more difficult to assess, requiring a data quality study as described previously.

**Table 47. Average Number of Registration Numbers per Unique Identifier, by
First Service Year**

First Service Year	Average Number of Registration Numbers per Unique Identifier
1973 - 1975	1.16
1976 - 1978	1.20
1979 - 1981	1.25
1982 - 1984	1.29
1985 - 1987	1.34
1988 - 1990	1.38
1991 - 1993	1.41

ii Alberta Cancer Registry Data

Most identifying information was complete in the Alberta Cancer Registry, except for a large proportion of missing middle names (48.9%). AHCIP number was missing for almost 10% of the individuals in the cancer cohort. Complete AHCIP number history was not available on the Registry, limiting the usefulness of AHCIP registration number in record linkage. These limitations reflect the usual use of the data, which is registration, not research involving linkage with AHCIP files. However, as shown in Table 46, Alberta's cancer registry compared favourably with previously published work from a European cancer registry.

The Alberta Cancer Registry was likely to have recorded most colon cancer cases occurring in Albertans. A common way of assessing the quality of case ascertainment in cancer registries is through the proportion of cases notified by death certificate only. This was low overall and over time. Few cases would be included as colon cancer cases that were not true diagnoses; this number would be difficult to estimate.

b. Influence of Alternative Comparison Groups

As noted, bias to lower SIRs was possible when the general population was used as the comparison group, since person-years at risk could be inflated, leading to increases in expected values relative to observed. Another AHCIP-identified group was used as the comparison to adjust for this bias.

A cohort of individuals who had undergone stripping and ligation of varicose

veins was used as the comparison group, because preliminary investigation based on minimal summary information showed that the average age and the sex distribution were similar to the cholecystectomy cohort. Although there may be some common risk factors for cholecystectomy and varicose veins, it is unlikely that varicose veins are a risk for colon cancer. These characteristics suggested that the varicose vein cohort would be a suitable unbiased unexposed group for this study. However, analysis of the complete dataset indicated that this group was not as well matched to the cholecystectomy cohort as first thought. Notably, there were significant differences in service year and age at service for the two cohorts.

The low rate of varicose vein procedures from 1973 through 1978 remains unexplained, but may have resulted from the introduction of new fee schedules in 1978, incorporating more varicose vein procedures. Thus the age- and sex-specific rates of colon cancer in the varicose vein group may be underestimated for the longer induction periods and may have contributed to unstable SIRs at 15 years of induction.

The varicose vein cohort was also considerably younger than the cholecystectomy cohort. Although much of this difference was taken into account with the age- and sex-specific stratification of the SIRs, there may have been insufficient outcome events, because this relatively young group had simply not reached the age at which colon cancer becomes more common.

2. *Record Linkage*

One of the main goals of the study was to examine the differences between the

different types of record linkage, for future application in studies using AHCIP and Alberta Cancer Registry data. Almost the same number of individuals were identified by deterministic and probabilistic linkage before manual review, although the probabilistic approach limited the number of candidate pairs for manual review to approximately half of the number identified in the deterministic approach. The expanded manual review in deterministic linkage resulted in the identification of only 21 (1.3%) more linked individuals for deterministic linkage compared to probabilistic linkage following manual resolution. However, the extremely high level of agreement for case ascertainment between the two types of linkage suggests that the probabilistic approach was more efficient, since the identification of the additional individuals by the deterministic approach required considerably greater resources.

The conclusion that there was little difference between the two approaches was further emphasized by the resulting SIRs. Probabilistic linkage resulted in slightly lower risk estimates, but statistical significance or non-significance, was the same for both linkage methods. The only exception was for males and females combined under the assumption of no induction. However, the statistical significance of the deterministic estimate was only borderline (95% CI=1.01-1.18).

The similarity of results between the linkage types may reflect the available data. There were few variables for matching, so that the few variable combinations could be easily programmed for the deterministic linkage and were essentially the same as the decision rules used by the probabilistic program. The probabilistic program would likely have been more efficient and less error-prone had there been

more variables available. Therefore, the data appear to have been of sufficient quality that either linkage approach could have been used.

Both linkage types may have a role in other environments, but in the current setting, probabilistic linkage would be recommended for future projects using AHCIP and Alberta Cancer Registry data, because of its efficiency gains. Additionally, the availability of a consistent, pre-tested computer program gives the added advantage of methodologic reliability.

3. *Comparison of Chart Review and Linkage Results*

A chart review was undertaken to define some parameters of the error rate in linkage. The agreement was moderate, with some individuals identified by linkage but not chart review and the reverse. This suggests that linkage may perform better than chart review in terms of answering certain questions, especially those requiring the actual date of a procedure. Of the 90 individuals identified with cholecystectomy in the chart review, over 80% were missing date of cholecystectomy. Of those that provided a date, over 90% only indicated the year of the procedure. When physicians were contacted to resolve the presence or absence of cholecystectomy, charts could not be found for 40% of the patients, and dates of procedures were no more readily available, with 50% of the available charts having only some indication of the time that cholecystectomy occurred.

These results suggest that a traditional retrospective chart review would not be suitable to address the question of cholecystectomy as a risk for colon cancer.

Indeed, charts from both the attending physicians and the Alberta Cancer Registry were missing relevant information. This supports the assertion that linkage is of particular value when data involve discrete exposures, such as surgeries.¹⁴⁴

C. Assessment of Risk

1. *Cholecystectomy as a Risk for Colon Cancer: Compared to the General Population*

a. All Colon Subsites Combined

The overall risk for colon cancer following cholecystectomy was close to unity for all induction periods, ranging from a significantly high risk of 1.09 (95 % CI=1.01-1.19) when no induction was assumed to low, nonsignificant values of 0.92 (95 % CI=0.78-1.07) and 0.93 (95 % CI=0.69-1.21) for inductions of 10 and 15 years, respectively. A similar lack of an association was observed in the majority of the previous cohort studies investigating this relationship.¹¹⁷⁻¹²² The significantly increased risk when no induction was used is consistent with medical attention bias, which may occur when cholecystectomy does not resolve abdominal complaints and further investigation in the immediate post-surgical period reveals colon cancer. Although risk estimates were not consistently statistically significant, the general trend was for lower SIRs at later induction periods. This may result from long-term dietary changes, such as decreased intake of dietary fat, or from continued medical follow-up, including colonic screening and polypectomy, in the cholecystectomy cohort, such that

the exposed group was actually protected from developing colon cancer. Early detection of cancer resulting from increased post-surgical medical attention in the cholecystectomy cohort would also lead to a decreased SIR in the short term.

However, it is unlikely that these practices occur frequently after surgical follow-up is complete. Therefore, the SIRs associated with longer induction periods (especially 15 years), would be unaffected by such bias.

In addition to the biological influences, data quality may play a significant role in the SIRs. Instability of the data in the earliest years of the cohort may have contributed to overestimation of person-years at risk, resulting in inflated expected values and risk estimates biased to sub-unity values.

The risk for colon cancer among the female members of the cholecystectomy cohort was generally below the null value. The risk was highest with no induction (SIR=1.03, not significant) and declined to a minimum at 15 years of induction (SIR=0.69, not significant). With 10 years of induction, the risk was significantly low (SIR=0.80, 95% CI=0.64-0.98). The lack of significance for the 15-year point, despite a continued drop in the point estimate, was likely the result of insufficient power to detect a risk of this magnitude, as only 26 cases remained at that time. As with the overall risk, the reduced risk for females could be the result of either a protective etiologic factor or poor data quality. Etiologic explanations include changes in health care utilization and lifestyle factors, such as decreasing dietary fat and increasing physical activity, which might be different for women with cholecystectomy compared to women in the general population, resulting in lower

colon cancer incidence in the cholecystectomy cohort. Unfortunately, there are no data to test these hypotheses. Interestingly, the only other Canadian study of cholecystectomy and colon cancer published to date reported a statistically significant reduction in colon cancer risk for females.¹¹² Although the biological explanation remains unresolved, the previous study's authors suggested that the results were consistent with shorter fecal transit time associated with the increased cycling of the enterohepatic circulation following cholecystectomy.

On the other hand, data quality would be expected to affect the risk estimate for females more than the estimate for males, since females change key identifiers as their marital status changes. This could contribute to underlinkage of AHCIP registration number history, resulting in the same individual being represented several times in the cholecystectomy cohort and in inflated person-years at risk.

The increased risk for males was unexpected, as the consensus from previously published work is for increased risk for females but not for males. In this study, the risk remained around 1.20, with all CIs excluding 1, for all induction periods up to and including 5 years. The risk declined slightly to 1.13 at 10 years of induction and rose to 1.36 at 15 years, but the CIs were not significant and became very wide, suggesting lack of statistical power resulting from few observed cases of colon cancer: the number of individuals with cancer were 73 and 27 for 10 and 15 years of induction, respectively. Moorehead et al⁹⁰ noted that power is typically too low to assess risk of colon cancer following cholecystectomy for males, because relatively few males undergo the procedure.

These data may indicate lifestyle or medical care differences among males who have had cholecystectomy, compared to males in the general population, but there are no data to support or refute this assertion. As proposed in the study's hypothesis, the observed risk may result from exposure to unbuffered bile acids following cholecystectomy, but this cannot be ascertained for individuals in this study.

Data quality issues were less likely to affect the risks for males compared to females, because key identifiers that change for females with marital status likely remained constant for males. Bias due to underlinkage, with concomitant overestimation of person-years at risk and expected numbers of cases, is unlikely, given the magnitude of the observed effects.

In summary, the overall SIRs represented a weighted average of the sex-specific SIRs. However, the directions of the sex-specific risks were dissimilar. The null values observed for males and females combined were the result of a decreased risk for females being offset by an increased risk for males.

Although most studies reported a higher colon cancer risk for females compared to males following cholecystectomy, a few observed a higher risk for males. Three were small studies that may not have had sufficient power to detect a difference: Kune et al's¹⁰⁸ Australian case-control study (odds ratio for males: 1.41, 95% CI=0.7-2.9; odds ratio for females: 1.07, 95% CI=0.2-7.0), Gudmundsson et al's¹¹⁹ Swedish cohort study (relative risk for males: 2.86, $p = 0.09$; relative risk for females: 0.49, $p = 0.43$), and Vobecky et al's¹¹² Canadian case-control study. Only an assessment of statistical significance was provided in the last study: not

significant for males and $p = 0.05$ for females. Goldbohm et al's¹²³ case-cohort study in a Dutch population only had 5 years of follow-up at publication, but showed elevated cancer risk for both sexes, with males at greater risk than females (relative risk for males: 1.81, $p = 0.02$; relative risk for females: 1.47, $p = 0.05$ for females). A study of an Icelandic population showed the greatest concordance with the current study's results, finding a significant increase in risk for males starting at 11 years post-cholecystectomy (relative risk for males: 2.73, 95 % CI=1.25-5.19; relative risk for females not provided). However, none of these studies offered any rationale for the difference by sex. One would expect most dietary and other lifestyle factors to be similar amongst all individuals with cholecystectomy, regardless of sex. However, it is possible that, while all individuals make dietary changes, the protective effect is more evident for women. This would fit with secular trends of colorectal cancer that show larger decreases in incidence and mortality for women compared to men,^{41,162} in an era of lower consumption of dietary fat.¹⁶³ There are no data available for the current study's population, but Goldbohm et al¹²³ showed comparable risk estimates even following adjustment for parity, obesity, alcohol intake and various dietary factors.

Selection bias is also a possible explanation for the higher risk of colon cancer observed for males alone. Because the profile of individuals at risk for gallstones, the "five F's", focuses on females, women presenting with abdominal pain may be more likely to undergo cholecystectomy than men presenting with similar symptoms. Therefore, men who undergo cholecystectomy may be at extreme risk for both

gallstones and colon cancer (in particular, they may be obese), while women who undergo cholecystectomy may be more similar to their counterparts in the general population. Thus, there would be bias associated with which individuals are selected for gallbladder surgery based on gender.

The difference in the direction of the risk for males and females suggests that there may be competing explanatory factors contributing unequally to the sex-specific SIRs. It is possible that the true magnitude of the risk is reflected in the SIRs for males, and that data quality problems or an underlying biological factors specific to females result in their lower SIRs. While data on potential protective biological characteristics are not available for individuals in this cohort, it is possible that females who underwent cholecystectomy adopted lower fat, higher fibre diets and undertook more physical activity. Hormonal effects may also contribute to an observed protective effect, since parity, supplemental estrogen use and oral contraceptives have been associated with increased gallstone disease,²⁵⁻²⁹ but decreased risk of colonic cancer.^{164,165}

b. Risk by Colon Subsite

The trends observed for the colon persisted in the analysis of risk by colon subsite. The explanations proposed for the patterns also apply to the subsite-specific results. However, discussion of patterns by subsite are generally limited by decreasing sample size and insufficient power.

In previous studies, the colonic subsite most at-risk for cancer following

cholecystectomy was the right (ascending) colon. This was thought to result from greater exposure to secondary bile acids. In the current study, no association was seen when the general population was used as the comparison group. The pattern of risk for females was similar to the risk for males and females combined, with a slight but not significantly elevated risk at 0 years of induction (overall SIR: 1.12; SIR for females: 1.06) and sub-unity risk for all other induction periods except for a nonsignificant rise at 15 years of induction (overall SIR: 1.23; SIR for females: 1.00). The trend, although not significant, was different for males, however, with elevated risk throughout, except for an unexplained decrease at 10 years of induction. Interestingly, the magnitude of the risk at 15 years increased to 1.75; the 95% CI was wide and lack of significance may have been simply due to inadequate sample size and power, since only 10 males in the cholecystectomy cohort were diagnosed with right colon cancer following a 15-year interval.

The results for cancer of the mid-colon were unusual. Some of the previously published studies divided the colon into only two subsites, proximal and distal, so could not address cancer risk of the mid-colon. Overall, the risk estimates for the mid-colon were increased for the 0-, 1-, 2- and 5-year induction periods, but only significantly under the assumption of no induction (SIR: 1.23; 95% CI=1.01-1.49). The overall SIR achieved significance largely due to the significant risk estimate for males at 0 years of induction. The latter situation could result from medical attention bias, but the significance of the mid-colon risk and lack of significance for either right or left colon is difficult to explain and therefore may be attributed to chance. As with

colon cancer risk overall, the direction of the risk for cancer of the mid-colon was different for males and females. Bias due to differential data quality or biological variations resulting in a relative increase in cancer risk for males and/or decreased cancer risk for females could explain this variation, as discussed previously.

However, lack of significant risk due to diminishing sample size and power precludes meaningful discussion of the sex-specific differences in risk for this colon subsite.

Overall, the risk for left colon cancer was not significant at any induction, although the risk declined from 1.07 to 0.83 over time. A similar pattern was observed for females, with a more substantial, but still nonsignificant, decrease in risk (SIRs of 1.01 to 0.64 for 0 and 15 years of induction, respectively). A nonsignificant elevation in risk (near 1.20) was observed for males for all induction periods. Again, data quality and/or biological differences may contribute to the patterns of risk, but are impossible to quantify in this study.

The general pattern by subsite was not expected under the hypothesis that exposure to carcinogenic bile acids leads to increased rates of colon cancer. This hypothesis would be supported by data showing a risk gradient with highest risk for cancer of the right colon and lowest risk for cancer of the left colon. In fact, risk appeared to be relatively constant at each induction point for all subsites. If anything, the risk was slightly elevated for mid-colon cancer relative to right and left colon cancer; possible explanations for this include an inappropriate choice of subsite definitions, a chance occurrence or no relationship, under the proposed carcinogenic mechanism. Since the subsite definitions were consistent with other studies where a

left to right pattern was found, chance is a likely explanation. As noted, small sample sizes limited the power of the subsite-specific analysis and constrained discussion, beyond extrapolating arguments proposed for cancer risk for the colon as a whole.

2. *Cholecystectomy as a Risk for Colon Cancer: Adjusting for Gastric Procedures*

Data on patients who had undergone gastric procedures were acquired to compensate for potential confounding. Confounding was possible since many patients were thought to undergo cholecystectomy and gastric procedures at the same time, and gastric procedures have been associated with colon cancer risk. However, the magnitudes of the adjusted and unadjusted risk estimates were very similar, suggesting that gastric procedures are not an important confounder in the relationship between cholecystectomy and colon cancer. In fact, less than 4% of the cholecystectomy cohort had gastric procedures.

3. *Cholecystectomy as a Risk for Colon Cancer: Compared to the Varicose Vein Cohort*

a. All Colon Subsites Combined

The cohort of individuals with varicose vein procedures was used as a comparison group to compensate for any bias incurred by using record linkage to determine the observed but not the expected values of the SIRs. The rates of colon cancer in the varicose vein cohort incorporated any linkage effects, resulting in error

but not bias in the SIR. Further, the varicose vein cohort was thought to be a more suitable comparison group than the general population, because there are some common risk factors for varicose veins and cholecystectomy that could be partially controlled using the additional comparison group.

The risk for colon cancer for males and females combined was not statistically significant except for a marginally significant increase under the assumption of no induction (SIR: 1.15; 95 % CI=1.06-1.24). This conclusion was also drawn when the general population was used as the comparison group. However, the risk estimates remained above unity even at 15 years of induction, suggesting that the relative inflation of expected values may be corrected when a comparison group is selected from the same source as the cohort of interest.

If females were more susceptible to underlinkage because of changing identifiers, using an AHCIP-derived comparison cohort would decrease bias but increase error. That is, underestimation of the observed and the expected numbers would be anticipated to be at the same rates, leading to SIRs near unity. However, risk remained below unity and even became significantly low for the shorter induction periods (SIR for 1 year of induction: 0.86; 95 % CI=0.76-0.96; SIR for 5 years of induction: 0.82; 95 % CI=0.71-0.95). The nonsignificant SIRs at 10 and 15 years of induction may be attributed to inadequate power ($n=88$ and $n=26$ for 10 and 15 years of induction, respectively). The risk estimates for the more biologically plausible induction periods (5, 10 and 15 years) ranged from 0.82 at 5 years (95 % CI=0.71-0.95) to 0.72 at 15 years (95 % CI=0.47-1.06). The direction of the

apparent protective effect is consistent with the findings of the only previous Canadian study,¹¹² although the magnitude of the effect was considerably greater (odds ratio=0.27, $p = 0.05$) in the previous study.

The risks for males were significantly increased for all induction periods. The magnitude of the risk was consistently near 1.60 for 0 through 10 years of induction, increasing to 2.67 for 15 years of induction. While this increase may reflect random fluctuation, the pattern is consistent with an induction effect occurring at some point after 10 years. Of previous publications, only Nielsen et al¹²¹ showed a significant increase in risk (risk estimate=2.73; 95 % CI=1.25-5.19) for males with at least 11 years follow-up post-cholecystectomy.

The sex-specific patterns of risk observed using the rates of colon cancer in an alternative, AHCIP-derived comparison group were comparable with the previous results, based on cancer rates in the general population. This suggests that data quality issues did not bias the results substantially and that other factors were responsible for the increased colon cancer risk following cholecystectomy for males and the decreased risk for females. Plausible biological mechanisms for these findings were described previously. Notably, females with cholecystectomy may make more significant lifestyle changes that protect them from developing colon cancer and/or males may be at increased risk because they are not protected from the carcinogenic effect of bile acids by female hormones.

On the other hand, using the varicose vein cohort instead of the general population as the comparison group may have replaced one set of biases with another.

That is, the observed risks could result if the varicose vein cohort was an unsuitable comparison group with biases and confounding leading to overestimated colon cancer rates for females and underestimated colon cancer rates for males.

To investigate these possibilities, the risk of colon cancer in the varicose vein cohort was calculated using the general population as the comparison group. Unfortunately, the conclusions based on this analysis were limited because low numbers of cases led to unreportable SIRs for 15 years of induction for all risk estimates and for 10 years of induction for male estimates. For the induction periods where stable SIRs could be calculated, the overall risk estimate for colon cancer in the varicose vein cohort was very close to unity. Although non-significant, the risk estimates were slightly over the null value. Lack of significance suggests a chance finding, but the patterns in the risk estimates deserve some consideration.

One interpretation of these results centres around data quality. Critics of record linkage, particularly when administrative data are used in cancer studies, may suggest that low SIRs occur simply because of the methodology. AHCIP data for females are thought to be particularly error-prone because of identifier changes associated with changes in marital status. However, the slight excesses shown in the varicose vein cohort's SIRs suggested that this was not the case. The risk for colon cancer in the varicose vein cohort showed an increased, though statistically nonsignificant, risk for females, with SIRs near 1.20. Conversely, the risk for colon cancer in the males of the varicose vein cohort was lower, though statistically nonsignificant, compared to the general population, with a marked decline in risk

estimates over time. These trends persisted in the subsite analysis, although many SIRs were unreportable because of insufficient numbers of individuals with colon cancer.

On the other hand, these data may be interpreted by examining varicose veins' association with colon cancer. Although there is no evidence that varicose veins or their treatment are directly associated with colon cancer, there are common risk factors for the two conditions. Varicose veins are more prevalent in obese individuals, with multiparous females at particular risk.^{130,131} Since increased dietary fat intake, a correlate of obesity, is associated with increased risk of colon cancer, the varicose vein cohort may have higher rates of colon cancer than the general population. Conversely, if the females in the varicose vein cohort had higher rates of supplemental estrogen use and were more likely to be multiparous than females in the general population, there would be a decreased risk of colon cancer in the varicose vein cohort. Therefore, the observed patterns could be explained if females with varicose veins were more overweight as a result of consuming high levels of dietary fat, offsetting any protection from hormone exposure. However, since obesity is also a risk factor for cholecystectomy, SIRs for colon cancer risk in the cholecystectomy cohort using the cancer rates in the varicose vein cohort would be expected to moderate to near unity. The observed exaggeration of the protective effect shown with the alternative comparison group argues against an explanation involving common co-factors. This assumes, however, that the postulated common risk factors for cholecystectomy, varicose veins and colon cancer are in effect in this population.

For example, it is possible that decreased dietary fat is associated with the protective effect observed for the cholecystectomy cohort while fat is not actually associated with varicose veins. This would lead to a systematic difference between the two AHCIP cohorts that could explain the observed patterns of risk.

The trend to reduced colon cancer risk for males in the varicose vein cohort could not be explained by the obesity-dietary fat co-factor, since high intake of dietary fat would result in an increased risk of colon cancer. Another unknown factor may lead to the protective effect. Socioeconomic status (SES) may be a useful factor to examine. Lower rates of colon cancer may be found in men who have undergone varicose veins procedures if higher SES men are more likely to undergo varicose veins procedures and to have lower-fat diets and higher levels of physical activity. This effect may not occur for women since stripping and ligation of varicose veins could be more universal for all social classes. Further, if men who undergo cholecystectomy are more obese and of lower SES, then the varicose vein cohort would be an unsuitable comparison group due to systematic differences associated with SES.

The explanations for the risk patterns remain speculative because of the lack of individual-level information regarding confounders. However, the SIR patterns were similar regardless of the comparison group used. In summary, using the varicose vein cohort as the comparison group allowed for estimation of colon cancer risk following cholecystectomy with minimal methodologic bias, while using the general population allowed for stability of rates and generalizability, given the larger

population.

b. Risk by Colon Subsite

The patterns of risk by colon subsite were similar to those observed when the general population was used as the comparison group, although the excess right and mid-colon cancer risk for males and the excess mid-colon cancer risk overall became significant. The general explanation for these trends is analogous to the discussion for the trends described for the colon overall. Additional discussion of the findings is limited by the paucity of individuals with cancer at each subsite, contributing to low power. In addition, since the varicose vein group had few or no individuals with cancer at certain subsites with the specified induction periods, the risk estimates were often unstable because the rates for the expected numbers were missing for some age, sex, year and subsite categories. For example, there were no males in the varicose vein cohort contributing to the expected rates at 15 years of induction, leading to an infinitely large SIR because the rates in the denominator were 0. Therefore, it is assumed that the risk of right colon cancer was large for males at 15 years of induction, as was the case where the general population was used as the comparison group, but the precise magnitude of the risk was impossible to ascertain.

4. *Influence of AHCIP Effective Date Quality on Risk Estimates*

Approximately one-third of the study group did not have AHCIP effective dates. For the majority of the analysis, it was assumed that these registrants were

identified in the provincial insurance plan at its inception (July 1, 1969). A sensitivity analysis, where only records with complete effective dates were included, showed little difference in the general pattern of risk. As with the analysis using all records, the overall risk for cancer using the records with complete dates alone remained close to unity when the general population was used as the comparison group. Risk declined over time for females and increased over time for males. However, the risk estimates for earlier (0- to 5-year) induction periods for males became statistically nonsignificant in the analysis of the records with complete dates alone, presumably due to loss of power. The direction of risk also remained consistent with the original analysis when the varicose vein cohort was used in the sensitivity analysis. However, the statistically significantly low SIRs observed for females became nonsignificant and the overall SIR became significant for the early (0- to 5-year) induction periods.

These results followed the predicted direction of bias. When SIRs for records with complete dates were compared to all records and the general population was used as the comparison group, the expected numbers were only slightly deflated, because the rates of cancer stayed constant even though the observed and expected numbers were reduced. However, the comparison of SIRs using the rates of cancer in the varicose vein cohort were based on a larger reduction of expected values because the rates were also affected by the complete date restriction. Thus, the sensitivity analysis using the varicose vein cohort as the comparison group shows the effect of date completeness more explicitly and suggests that there was an inflation of person-years at risk and therefore, expected numbers, leading to underestimated SIRs.

However, at biologically relevant induction periods, the conclusions did not differ by clinically meaningful amounts.

D. Strengths and Limitations of the Current Study

1. Strengths

The design of this study incorporated several features that contributed to its credibility. First, the study was a nonconcurrent cohort study, which allowed for follow-up of exposed, cholecystectomized individuals with long induction, using administrative data to minimize selection bias. Previous studies investigating the association between cholecystectomy and colon cancer have been restricted because of small sample sizes leading to low power, bias and lack of compensation for an adequate induction period.

This study was the largest one to investigate the association between cholecystectomy and colon cancer to date. Past case-control studies may have been unable to find an association because cholecystectomy was a relatively rare exposure, leading to few exposed cases in the study population. Some of the smaller cohort studies may also have had limited power, because colon cancer was a relatively rare outcome, and, given the moderate degree of risk, few cases would be detected in the exposed group.

Power issues were exacerbated where insufficient induction was allowed; that is, the interval between cholecystectomy and colon cancer would have to be fairly

long to achieve the number of endpoints necessary. In two of the previously published cohort studies,^{122,123} follow-up was very short and would likely lead to no observable increase in risk because of insufficient induction. Power may have been a particular problem for determining risks by age- and subsite-specific subgroups in other studies, as well as in this study. For example, other studies' observations that cancer risk was increased for females with cholecystectomy but not for males could result because cholecystectomy was considerably more prevalent in females, leading to adequate power to find a difference in this group only.

Risk estimates calculated using various induction periods allowed investigation of bias as well as possible etiology. Of note, Hyvärinen and Partinen showed peak levels of colorectal cancer at 15 and 23 years.⁷⁵ However, some previous studies did not consider the interval between cholecystectomy and colon cancer. For example, many of the retrospective studies, including formal case-control studies and simple chart reviews, could have documented an increased risk because they did not consider induction. This possibility was supported by the current study, in which the SIRs were significant when no induction was assumed, but often became nonsignificant over time. Adami et al also observed steady declines in risk from significant levels (risk estimate of 2.86; 95% CI=1.56-4.79) in the first post-operative year to nonsignificantly decreased levels (risk estimate of 0.80) at 12 to 15 years post-operatively.¹¹⁸ The phenomenon was demonstrated again in a study of diabetes as a risk for pancreatic cancer, where the odds ratio dropped from 3.04 (95% CI=2.21-4.17) to 1.43 (95% CI=0.98-2.07), when only patients with diabetes for more than

three years' duration were considered.¹⁶⁶ Further, the risk estimates declined as induction increased, from 3.04 at baseline (0 years of induction) to 0.73 when at least 15 years of induction was assumed. Therefore, increased risk estimates with very short induction periods are suggestive of increased medical attention post-cholecystectomy as symptoms are investigated more closely after surgery.

Biases inherent in many of the retrospective medical record reviews were avoided by using health care insurance plan data. In their meta-analysis, Giovannucci et al¹²⁴ pointed out that many of the record reviews may not have been blind and that information on cholecystectomy may be more complete in the records of cancer patients, which could lead to an apparent, but false, increase in risk. In this study, reviewers of both the linkage results and the medical charts were blind to exposure status, ensuring that their conclusions were not influenced by knowledge of an individual's surgical history. The use of AHCIP data also minimized selection bias, compared to studies that used convenience samples, or direct contact with subjects.

Generalizability was limited in previous studies investigating cholecystectomy and colon cancer because relative incidence by subsite was used instead of external population controls. In contrast, this study used the Alberta population as its main comparison group, which maximized the generalizability of the results, particularly to residents of the province.

2. *Limitations*

Giovannucci et al¹²⁴ suggested that there may not be any increase in colon

cancer risk until 10 to 15 years post-cholecystectomy. Although this study allowed for calculation of SIRs using these induction periods, there were fewer subjects available for the later induction periods, leading to lower statistical power. In addition, the greatest data quality concerns were for AHCIP data before 1983, which coincides with the group available for 10 or more years' induction. Therefore, nonsignificant SIRs at the later induction periods may not be indicative of a true lack of risk, but of small numbers and poor data quality, leading to underestimated SIRs with wide confidence intervals.

Two comparison groups were used in this study. As discussed, there are limitations to the conclusions derived from either of them: the varicose vein cohort had few individuals contributing to the age-, sex-, and year-specific rates, while the general population potentially introduced methodologic bias associated with record linkage. Regardless of the comparison group used, some individuals with cholecystectomy would have been included; for example, some individuals would have had cholecystectomy elsewhere, or before the inception of the provincial health care insurance plan in 1969. Although the effect of this contamination was probably minimal, it could result in SIRs biased toward unity.

The influence of potential confounders, such as dietary fat, obesity, hormones and presence of gallstones, were not addressed in this study because of the unavailability of suitable information to apply to the cohort. Assuming that dietary fat and obesity are positively associated with both cholecystectomy and colon cancer, adjusted SIRs would likely decrease. Conversely, assuming that female hormones

protect against colon cancer, the adjusted SIRs would likely increase, except for males' SIRs, which would remain constant.

However, there is little consensus about the relevance of these factors in the literature and studies that have investigated the influence of both cholecystectomy and dietary factors on colon cancer risk suggest that they may be relatively unimportant. Initially, the largest concern for this study was the confounding effects of dietary fat, but the protective effects of cruciferous vegetables and fibre were more important in altering colon cancer risk than the risk imposed by dietary fat in several studies.^{88,104} In the case-cohort study by Goldbohm et al¹²³, adjustment for parity, obesity, alcohol intake and other dietary factors relevant to colon cancer made very little difference in risk; adjustment decreased the risk between cholecystectomy and colon cancer from 1.81 to 1.78 for males and increased the risk from 1.47 to 1.51 in females.

An issue that remains unresolved by this study is whether cholecystectomy or simply the presence of gallstones might be the more important risk factor. If cholecystectomy patients are already at risk for colon cancer due to gallstone disease, adjusting for gallstones would decrease the risk estimates. However, since some individuals may have asymptomatic gallstones, the expected numbers may also be slightly inflated. Therefore, true adjustment for gallstones would require screening of both the exposed cohort and the comparison groups, as attempted by Mannes et al.¹⁶⁷ These authors found an association between adenomas and cholecystectomy but not adenomas and gallstones.

Data quality was a potentially limiting factor in this study. Even though

agreement between variables was found to be comparable with other published work, data availability does not imply validity and the magnitude of the validity issue remained largely unresolved in this study. Further, variables such as tumour stage that were potentially useful in discriminating between individuals found to have both cholecystectomy and colon cancer and those with only one factor were not available for the majority of the individuals in the study.

Although the results of this study may be applied to the Alberta population, they cannot be extended further. As noted by Moorehead and McKelvey,¹⁶⁸ the difference in conclusions between this and other studies may be partly due to the underlying rates of colon cancer in populations. Recent statistics for 1983 to 1987 showed that Sweden's incidence rate for colon cancer (adjusted to the world standard population) was 17.5 per 100,000 while Alberta's incidence rate was 21.8 per 100,000.¹⁶⁹ The similarity of rates suggest that an association may not be documented in either population because of relatively low rates compared to areas where an increased risk following cholecystectomy was reported, and therefore insufficient numbers of individuals with colon cancer in the study group contribute to low statistical power. Further, different populations have different risk factor prevalences, which may influence the rates of both colon cancer and gallstones and may act as confounders.

There were many analyses incorporated into this study, with risk estimates calculated for the sexes separately and combined, for colon subsites separately and combined and for six induction periods. The level of statistical significance was not

adjusted to compensate for multiple testing. This adjustment would be difficult to carry out given the interdependence of a large number of the analyses performed in the course of this study.

E. Conclusions and Recommendations

The primary objectives of this study were to examine the risk of colon cancer following cholecystectomy, for both genders combined, for both genders separately and for each tumour site (left, right and mid-colon), and to evaluate the effectiveness of two types of record linkage as methods for cohort follow-up using Alberta data.

The study did not establish an overall increased risk for colon cancer following cholecystectomy for Albertans. A possible increase in the risk for males may have been obscured by decreasing numbers of study subjects and diminishing power at biologically relevant induction periods. No right-left gradient was observed in the analysis by subsite, as might be expected from previous research, although there was a trend to a significantly increased mid-colon cancer risk, which remains unexplained.

These results do not support the hypothesized mechanism of increased risk resulting from altered bile acids in the colon. However, the results could be explained if bile acids were carcinogenic in the colon, but the presence of protective factors, such as exposure to female hormones through exogenous estrogen use or pregnancy, masked the effect in females.

Record linkage, using either the deterministic or the probabilistic approach,

was found to be suitable for addressing the question of colon cancer risk following cholecystectomy using AHCIP and Alberta Cancer Registry data. Data deficiencies were identified, which may have limited the ability to link individuals, but overall record linkage compared favourably with traditional chart review.

The recommendations from this study are:

1. The data should be analyzed again in 5 years and in 10 years, when a higher proportion of the cohort has achieved suitable induction and the power of the study to detect a true difference in risk will be higher. A truly nested case-control study could also be undertaken to investigate the role of diet and other confounders on the relationship between cholecystectomy and colon cancer, although obtaining a sufficient number of exposed cases to allow detection of this moderate effect may be problematic. Statistical modelling using Poisson regression should be investigated in future analysis.
2. Future linkage studies using AHCIP data and the Alberta Cancer Registry should use probabilistic linkage. Although this method tended to underestimate risk slightly, the differences were not clinically significant and the gains made using deterministic linkage required substantially more human resources.
3. Data from AHCIP, particularly before 1983, remain questionable. Further review and specification of data validity is required before other nonconcurrent cohort studies using this data source are initiated.

REFERENCES

1. McMichael AJ, Potter JD. Host factors in carcinogenesis: certain bile-acid metabolic profiles that selectively increase the risk of proximal colon cancer. *J Natl Cancer Inst* 1985;75(2):185-91.
2. Pomare EW, Heaton KW. The effect of cholecystectomy on bile salt metabolism. *Gut* 1973;14:753-62.
3. Malagelada JR, Go VLW, Summerskill WHJ, Gamble WS. Bile acid secretion and biliary bile acid composition altered by cholecystectomy. *Am J Digest Dis* 1973;18:455-9.
4. Almond HR, Vlahcevic ZR, Bell CC, Gregory DH, Swell L. Bile acid pools, kinetics and biliary lipid composition before and after cholecystectomy. *N Engl J Med* 1973;289:1213-6.
5. Breuer NF, Jaekel S, Dommes P, Goebell H. Fecal bile acid excretion pattern in cholecystectomized patients. *Dig Dis Sci* 1986;31(9):953-60.
6. Hill MJ. Bile, bacteria and bowel cancer. *Gut* 1983;24:871-5.
7. Admirand W. The pathogenesis of gallstones. In: Sleisenger MH, Fordtran JS, editors. *Gastrointestinal disease: pathophysiology, diagnosis, management*. Philadelphia:Saunders, 1973: 1110-5.
8. Vlahcevic ZR, Bell CC Jr, Buhac I, Farrar JT, Swell L. Diminished bile and pool size in patients with gallstones. *Gastroenterology* 1970;59:165-73.
9. Nilsson S, Schersten T. Importance of bile acids for phospholipid secretion into human bile. *Gastroenterology* 1969;57:525-32.
10. Alberta Health. Rates of selected surgical procedures in Alberta and Canada 1976 to 1985-86, selected years. Edmonton:Alberta Health, 1988.
11. Friedman DK, Kannel WB, Dawber TR. Epidemiology of gallbladder disease: Observations in the Framingham study. *J Chronic Dis* 1966;19:273-92.
12. Boss LP, Lanier AP, Dohan PH, Bender TR. Cancers of the gallbladder and biliary tract in Alaskan natives: 1970-1979. *J Natl Cancer Inst* 1982;69:1005-7.
13. Lowenfels AB, Lindstrom CG, Conway MJ, Hastings PR. Gallstones and risk of gallbladder cancer. *J Natl Cancer Inst* 1985;75:77-80.

14. Diehl AK, Haffner SM, Hazuda HP, Stern MP. Coronary risk factors and clinical gallbladder disease: approach to the prevention of gallstones? *Am J Public Health* 1987;77:841-5.
15. Bouchier IAD. Postmortem study of the frequency of gallstones in patients with cirrhosis of the liver. *Gut* 1969;10:705-10.
16. Heaton KW, Read AE. Gall stones in patients with disorders of the terminal ileum and disturbed bile salt metabolism. *Brit Med J* 1969;3:494-6.
17. Cohen S, Kaplan M, Gottlieb L, Paterson J. Liver disease and gallstones in regional enteritis. *Gastroenterology* 1971;60:237-45.
18. Petitti DB, Sidney S. Obesity and cholecystectomy among women: implications for prevention. *Am J Prev Med* 1988;4:327-30.
19. Pixley F, Wilson D, McPherson K, Mann J. Effect of vegetarianism on development of gallstones in women. *Br Med J* 1985;291:11-2.
20. Scragg RKR, McMichael AJ, Baghurst PA. Diet, alcohol, and relative weight in gallstone disease: a case control study. *Br Med J* 1984;288:1113-9.
21. Rome Group for the Epidemiology and Prevention of Cholelithiasis (GREPCO). Prevalence of gallstone disease in an Italian adult female population. *Am J Epidemiol* 1984;119:796-805.
22. Layde PM, Vessey MP, Yeates D. Risk factors for gallbladder disease: a cohort study of young women attending family planning clinics. *J Epidemiol Community Health* 1982;36:274-8.
23. Honore LH. The lack of positive association between symptomatic cholesterol cholelithiasis and clinical diabetes mellitus: a retrospective study. *J Chronic Dis* 1980;33:465-9.
24. Smith DA, Gee MI. Dietary survey to determine the relationship between diet and cholelithiasis. *Am J Clin Nutr* 1979;32:1519-26.
25. Bernstein RA, Werner LH, Rimm AA. Relationship of gallbladder disease to parity, obesity, and age. *Health Serv Rep* 1973;88:925-36.
26. Petitti DB, Sidney S, Perlman JA. Increased risk of cholecystectomy in users of supplemental estrogen. *Gastroenterology* 1988;94:91-5.
27. Weinstein MC. Estrogen use in postmenopausal women - costs, risks, and

- benefits. *N Engl J Med* 1980;303:308-16.
28. Fortney JA, Harper JM, Potts M. Oral contraceptives and life expectancy. *Stud Fam Plann* 1986;17:117-25.
 29. Strom BL, Tamragouri RN, Morse ML, Lazar EL, West SL, Stolley PD et al. Oral contraceptives and other risk factors for gallbladder disease. *Clin Pharmacol Ther* 1986;39:335-41.
 30. Rosenfield A. Oral and intrauterine contraception: a 1978 risk assessment. *Am J Obstet Gynecol* 1978;132:92-106.
 31. Maki T. Pathogenesis of calcium bilirubinate gallstone: role of *E. coli*, β -glucuronidase and coagulation by inorganic ions, polyelectrolytes and agitation. *Ann Surg* 1966;164:90-100.
 32. Johnston DE, Kaplan MM. Pathogenesis and treatment of gallstones. *N Engl J Med* 1993;328:412-21.
 33. Strasberg SM, Harvey PRC. Biliary cholesterol transport and precipitation: introduction and overview of conference. *Hepatology* 1990;12:1S-5S.
 34. Holzbach RT, Barnhart RL, Nader JM. Pathogenesis of cholesterol gallstone disease: the physico-chemical defect. In: Northfield T, Jazrawi R, Zentler-Munro, editors. *Bile acids in health and disease*, Norwell:Kluwer Academic Publishers, 1988: 117-33.
 35. Baxter JN. Gall-bladder emptying. *J Gastroenterol Hepatol* 1989; 4:353-72.
 36. Behar J, Lee KY, Thompson WR, Biancani P. Gallbladder contraction in patients with pigment and cholesterol stones. *Gastroenterology* 1989; 97:1479-84.
 37. Everson GT. Gallbladder function in gallstone disease. *Gastroenterol Clin North Am* 1991; 20:85-110.
 38. Windler EET, Kovanen PT, Chao Y-S, Brown MS, Havel RJ, Goldstein JL. The estradiol-stimulated lipoprotein receptor of rat liver. *J Biol Chem* 1980;255:10464-71.
 39. Kovanen PT, Brown MS, Goldstein JL. Increased binding of low density lipoprotein to liver membranes from rats treated with 17α -ethinyl estradiol. *J Biol Chem* 1979;254:11367-73.

40. Kern F Jr, Everson GT. Contraceptive steroids increase cholesterol in bile: mechanisms of action. *J Lipid Res* 1987;28:828-39.
41. National Cancer Institute of Canada. Canadian cancer statistics 1997. Toronto:National Cancer Institute of Canada, 1997.
42. National Cancer Institute of Canada. Canadian cancer statistics 1993. Toronto:National Cancer Institute of Canada, 1993.
43. Alberta Cancer Board. Alberta cancer registry 1989 annual report. Edmonton:Alberta Cancer Board, 1992.
44. Jass JR. Subsite distribution and incidence of colorectal cancer in New Zealand, 1974-1983. *Dis Colon Rectum* 1991;34:56-9.
45. Cox B, Little J. Reduced risk of colorectal cancer among recent generations in New Zealand. *Br J Cancer* 1992;66:386-90.
46. Kreger BE, Anderson KM, Schatzkin A, Splansky GL. Serum cholesterol level, body mass index, and the risk of colon cancer. *Cancer* 1992;70:1038-43.
47. Nomura AMY, Heilbrun LR, Stemmermann GN. Body mass index as a predictor of cancer in men. *J Natl Cancer Inst* 1985;74:319-23.
48. Lew EA, Garfinkel L. Variations in mortality by weight among 750,000 men and women. *J Chronic Dis* 1979;32:563-76.
49. Fleshner P, Slater G, Aufses AH. Age and sex distribution of patients with colorectal cancer. *Dis Colon Rectum* 1989;32:107-11.
50. Alley PG, McNee RK. Age and sex differences in right colon cancer. *Dis Colon Rectum* 1986;29:227-9.
51. Slater G, Papatestas AE, Tartter PI, Mulvihill M, Aufses AH Jr. Age distribution of right- and left-sided colorectal cancers. *Am J Gastroenterol* 1982;77:63-6.
52. Stewart RJ, Stewart AW, Turnbull PR, Isbister WH. Sex differences in subsite incidence of large-bowel cancer. *Dis Colon Rectum* 1983;26:658-60.
53. Morgan JW, Fraser GE, Phillips RL, Andress MH. Dietary factors and colon cancer incidence among Seventh-day Adventists [abstract]. *Am J Epidemiol* 1988;128:918.

54. Willett WC, Stampfer MJ, Colditz GA, Rosner BA, Spiezer FE. Relation of meat, fat, and fiber intake to the risk of colon cancer in a prospective study among women. *N Engl J Med* 1990;323:1664-72.
55. Shankar S, Lanza E. Dietary fiber and cancer prevention. *Hem-Oncol Clin North Am* 1991;5:25-41.
56. Trock B, Lanza E, Greenwald P. Dietary fiber, vegetables, and colon cancer: critical review and meta-analyses of the epidemiologic evidence. *J Natl Cancer Inst* 1990;82:650-61.
57. Wargovich MJ, Baer AR, Hu PJ, Sumyoshi H. Dietary factors and colorectal cancer. *Gastroenterol Clin North Am* 1988;17:727-45.
58. Reddy BS, Mastromarino A, Wynder EL. Further leads on metabolic epidemiology of large bowel cancer. *Cancer Res* 1975;35:3403-6.
59. Weisburger JH. Causes, relevant mechanisms, and prevention of large bowel cancer. *Sem Oncol* 1991;18:316-36.
60. Burkitt DP. Some diseases characteristic of modern Western civilization. *BMJ* 1973;1:274-8.
61. Burkitt DP. Epidemiology of cancer of the colon and rectum. *Cancer* 1971;28:3-13.
62. Henderson BE, Ross RK, Pike MC. Toward the primary prevention of cancer. *Science* 1991;254:1131-8.
63. Levin KE, Dozois RR. Epidemiology of large bowel cancer. *World J Surg* 1991;15:562-7.
64. Dwyer JT, Ausman LM. Fiber: unanswered questions. *J Natl Cancer Inst* 1992;84:1851-3.
65. Aries VC, Crowther JS, Drasar BS, Hill MJ, Williams REO. Bacteria and the aetiology of cancer of the large bowel. *Gut* 1969;10:334-5.
66. Hill MJ, Drasar BS, Williams REO, Meade TW, Cox AG, Simpson JE, et al. Faecal bile-acids and clostridia in patients with cancer of the large bowel. *Lancet* 1975;1:535-9.
67. Hill MJ. The role of colon anaerobes in the metabolism of bile acids and steroids, and its relation to colon cancer. *Cancer* 1975;36:2387-400.

68. Werner B, de Heer K, Mitschke H. Cholecystectomy and carcinoma of the colon. *Zeitschr Krebsforsch* 1977;88:223-30.
69. Hickman MS, Salinas HC, Schwesinger WH. Does cholecystectomy affect colonic tumorigenesis? *Arch Surg* 1987;122:334-6.
70. Kuniyasu T, Tanaka T, Shima H, Sugie S, Mori H, Takahashi M. Enhancing effect of cholecystectomy on colon carcinogenesis induced by methylazoxymethanol acetate in hamsters. *Dis Colon Rectum* 1986;29(8):492-4.
71. Narisawa T, Sano M, Sato M, Takahashi T, Tanida N, Shimoyama T. The correlation between cholecystectomy and 1, 2-dimethylhydrazine in mice. *Dis Colon Rectum* 1985;28(1):27-30.
72. Bandettini L, Filipponi F, Romagnoli P. Increase of the mitotic index of colonic mucosa after cholecystectomy. *Cancer* 1986;58:685-7.
73. Abrams JS, Anton JR, Dreyfuss DC. The absence of a relationship between cholecystectomy and the subsequent occurrence of cancer of the proximal colon. *Dis Colon Rectum* 1983;26:141-4.
74. Fixa B, Komárková O, Pospíšilová J. Cholecystectomy and right-sided colon cancer. *Neoplasma* 1984;31:223-4.
75. Hyvärinen H, Partinen S. Association of cholecystectomy with abdominal cancers. *Hepatogastroenterology* 1987;34:280-4.
76. Kwai AH. Cholecystectomy and large-bowel cancer. *Mt Sinai J Med* 1983;50:359-63.
77. Vernick LJ, Kuller LH, Lohsoonthorn P, Rycheck RR, Redmond CK. Relationship between cholecystectomy and ascending colon cancer. *Cancer* 1980;45:392-5.
78. Allende HD, Ona FV, Davis HT. Gallbladder disease: risk factor for colorectal carcinoma? *J Clin Gastroenterol* 1984;6:51-5.
79. Breuer NF, Katchinski B, Mörtel E, Leder L-D, Goebell H. Large bowel cancer risk in cholelithiasis and after cholecystectomy. *Digestion* 1988;40:219-26.
80. Eriksson SG, Lindström CG. Lack of relationship between cholecystectomy and colorectal cancer. A case control autopsy study in a defined population.

Scand J Gastroenterol 1984;19:977-82.

81. Hladík V, Nozicka Z, Maslowská H. Colorectal carcinoma and cholecystectomy. *Neoplasma* 1987;34:361-6.
82. Lowenfels AB, Domellöf L, Lindström CG, Bergman F, Monk MA, Sternby NH. Cholelithiasis, cholecystectomy, and cancer: a case-control study in Sweden. *Gastroenterology* 1982;83:672-6.
83. McFarlane MJ, Welch KE. Gallstones, cholecystectomy, and colorectal cancer. *Am J Gastroenterol* 1993;88:1994-9.
84. Turunen MJ, Kivilaakso EO. Increased risk of colorectal cancer after cholecystectomy. *Ann Surg* 1981;194:639-41.
85. Alley PG, Lee SP. The increased risk of proximal colonic cancer after cholecystectomy. *Dis Colon Rectum* 1983;26:522-4.
86. Berkel J, Hombergen DAMA, Hooymayers IE, Faber JAJ. Cholecystectomy and colon cancer. *Am J Gastroenterol* 1990;85:61-4.
87. Lee SS, Cha S, Lee RL. The relationship between cholecystectomy and colon cancer: an Iowa study. *J Surg Oncol* 1989;41:81-5.
88. Lee HP, Gourley L, Duffy SW, Estève J, Lee J, Day NE. Colorectal cancer and diet in an Asian population -- a case-control study among Singapore Chinese. *Int J Cancer* 1989;43:1007-16.
89. Markman M. Cholecystectomy and carcinoma of the colon. *Lancet* 1982;2:47.
90. Moorehead RJ, Kernohan RM, Patterson CC, McKelvey STD, Parks TG. Does cholecystectomy predispose to colorectal cancer? *Dis Colon Rectum* 1986;29:36-8.
91. Paul J, Gessner F, Wechsler JG, Kuhn K, Orth K, Ditschuneit H. Increased incidence of gallstones and prior cholecystectomy in patients with large bowel cancer. *Am J Gastroenterol* 1992;87:1120-4.
92. Sonoda T, Youngman DJ, Reynolds RD. Cholecystectomy and carcinoma of the colon. *Milit Med* 1983;148:721-2.
93. Turnbull PRG, Smith AH, Isbister WH. Cholecystectomy and cancer of the large bowel. *Br J Surg* 1981;68:551-3.

94. Bundred NJ, Whitfield BCS, Stanton E, Prescott RJ, Davies GC, Kingsnorth AN. Cholecystectomy, cholelithiasis and colorectal carcinoma. *J R Coll Surg Edinburgh* 1985;39:115-7.
95. Hoare AM. Carcinoma of the colon and cholecystectomy. *Lancet* 1974;2:1395-6.
96. Kaibara N, Wakatsuki T, Mizusawa K, Sugawara A, Kimura O, Koga S. Negative correlation between cholecystectomy and the subsequent development of large bowel carcinoma in a low-risk Japanese population. *Dis Colon Rectum* 1986;29:644-6.
97. Manousos ON, Gerovassilis F, Papadimitriou C, Tzonou A, Polychronopoulou A, Trichopoulos D. Cholecystectomy and colon cancer. *Lancet* 1981;2:810.
98. Narisawa T, Sano M, Sato M, Takahashi T, Arakawa H. Relationship between cholecystectomy and colonic cancer in a low-risk Japanese population. *Dis Colon Rectum* 1983;26:512-5.
99. Neugut AI, Murray TI, Garbowski GC, Forde KA, Treat MR, Waye JD, et al. Cholecystectomy as a risk factor for colorectal adenomatous polyp and carcinoma. *Cancer* 1991;68:1644-7.
100. Papadimitriou C, Day N, Tzonou A, Gerovassilis F, Manousos O, Trichopoulos D. Biosocial correlates of colorectal cancer in Greece. *Int J Epidemiol* 1984;13:155-9.
101. Spitz MR, Russell NC, Guinee VF, Newell GR. Questionable relationship between cholecystectomy and colon cancer. *J Surg Oncol* 1985;30:6-9.
102. Vlainac H, Jarebinski M, Adanja B. Relationship of some biosocial factors to colon cancer in Belgrade (Yugoslavia). *Neoplasma* 1987;34:503-7.
103. Wynder EL, Shigematsu T. Environmental factors of cancer of the colon and rectum. *Cancer* 1967;9:1520-61.
104. Soltero E, Cruz NI, Nazario CM, López RE, Alonso A, Ríos CF. Cholecystectomy and right colon cancer in Puerto Rico. *Cancer* 1990;66:2249-52.
105. Blanco D, Ross PK, Paganini-Hill A, Henderson BE. Cholecystectomy and colonic cancer. *Dis Colon Rectum* 1984;27:290-2.
106. Fixa B, Komárková O, Zaydlar K, Bures J, Erben J. Is there an increased

- risk of colorectal cancer after cholecystectomy? *Neoplasma* 1985;32:513-7.
107. Friedman GD, Goldhaber MK, Quesenberry CP. Cholecystectomy and large bowel cancer. *Lancet* 1987;1:906-8.
 108. Kune GA, Kune S, Watson LF. Large bowel cancer after cholecystectomy. *Am J Surg* 1988;156:359-62.
 109. Vernick LJ, Kuller LH. Cholecystectomy and right-sided colon cancer: an epidemiological study. *Lancet* 1981;2:381-3.
 110. Vernick LJ, Kuller LH. A case-control study of cholecystectomy and right-side colon cancer. *Am J Epidemiol* 1982;116:86-101.
 111. Weiss NS, Daling JR, Chow WH. Cholecystectomy and the incidence of cancer of the large bowel. *Cancer* 1982;49:1713-5.
 112. Vobecky J, Caro J, DeVroede G. A case-control study of risk factors for large bowel carcinoma. *Cancer* 1983;51:1958-63.
 113. Ekblom A, Yuen J, Adami H-O, McLaughlin JK, Chow W-H, Persson I, et al. Cholecystectomy and colorectal cancer. *Gastroenterology* 1993;105:142-7.
 114. Linos DA, Beard CM, O'Fallon WM, Dockerty MB, Beart RW, Kurland LT. Cholecystectomy and carcinoma of the colon. *Lancet* 1981;2:379-81.
 115. Moorehead RJ, Mills JOM, Wilson HK, McKelvey STD. Cholecystectomy and the development of colorectal neoplasia: a prospective study. *Ann R Coll Surg (Engl)* 1989;71:37-9.
 116. Rundgren A, Mellström D. Cholecystectomy and colon cancer in the elderly. *Age Ageing* 1983;12:44-9.
 117. Adami HO, Krusemo UB, Meirik O. Unaltered risk of colorectal cancer within 14-17 years of cholecystectomy: updating of a population-based cohort study. *Br J Surg* 1987;74:675-8.
 118. Adami HO, Meirik O, Gustavsson S, Nyrén O, Krusemo UB. Colorectal cancer after cholecystectomy: absence of risk increase within 11-14 years. *Gastroenterology* 1983;85:859-65.
 119. Gudmundsson S, Möller TR, Olsson H. Cancer incidence after cholecystectomy -- a cohort study with 30 years follow-up. *Eur J Surg Oncol* 1989;15:113-7.

120. Ichimaya H, Kono S, Ikeda M, Tokudome S, Nadayama F, Kuratsune M. Cancer mortality among patients undergoing cholecystectomy for benign biliary diseases. *Jpn J Cancer Res (Gann)* 1986;77:579-83.
121. Nielsen GP, Theodors A, Tulinius H, Sigvaldason H. Cholecystectomy and colorectal carcinoma: a total-population historical prospective study. *Am J Gastroenterol* 1991;86:1486-90.
122. Wu AH, Paganini-Hill A, Ross RK, Henderson BE. Alcohol, physical activity and other risk factors for colorectal cancer: A prospective study. *Br J Cancer* 1987;55:687-94.
123. Goldbohm RA, van den Brandt PA, van't Veer P, Dorant E, Sturmans F, Hermus RJJ. Cholecystectomy and colorectal cancer: evidence from a cohort study on diet and cancer. *Int J Cancer* 1993;53:735-9.
124. Giovannucci E, Colditz GA, Stampfer MJ. A meta-analysis of cholecystectomy and risk of colorectal cancer. *Gastroenterology* 1993;105:130-41.
125. Bouchier IAD. Bile, bile acids and gallstones. In: Sircus W, Smith AN, editors. *Scientific foundations of gastroenterology*, London:William Heinemann Medical Books, 1980: 565-78.
126. Houghton PWJ, Owen RJ, Henly PJ, Mortensen NJM, Hill MJ, Williamson RCN. Experimental colonic carcinogenesis after gastric surgery. *Br J Surg* 1990;77:774-8.
127. Mullan FJ, Wilson HK, Majury CW, Mills JOM, Cromie AJ, Campbell GR, et al. Bile acids and the increased risk of colorectal tumours after truncal vagotomy. *Br J Surg* 1990;77:1085-90.
128. Pellegrini CA, Patti MG. Motility of the gallbladder and bile ducts and the kinetics of bile flow. In: Way LW, Pellegrini CA, editors. *Surgery of the gallbladder and bile ducts*. Philadelphia:Saunders, 1987: 51-68.
129. Skinner DB. Pathophysiology of gastroesophageal reflux. *Ann Surg* 1985;202:546-56.
130. Geelhoed GW, Burkitt DP. Varicose veins: a reappraisal from a global perspective. *South Med J* 1991;84:1131-4.
131. Hirai M, Naiki K, Nakayoma R. Prevalence and risk factors of varicose veins in Japanese women. *Angiology* 1990;41:228-32.

132. Dunn H. Record linkage. *Am J Public Hlth* 1946;36:1412-6.
133. Cohen MM. Using administrative data for case-control studies: the case of the Papanicolaou smear. *Ann Epidemiol* 1993;3:93-8.
134. Cohen MM, Hammarstrand KM. Papanicolaou test coverage without a cytology registry. *Am J Epidemiol* 1989;129:388-94.
135. Roos LL, Sharp SM. Becoming more efficient at outcomes research. *Int J Technol Assess Health Care* 1988;4:555-71.
136. Fedson DS, Wajda A, Nicol JP, Hammond GW, Kalser DL, Roos LL. Clinical effectiveness of influenza vaccination in Manitoba. *JAMA* 1993;270:1-6.
137. Day NE, Miller AB. Screening for breast cancer. Toronto:Hans Huber Publishers, 1988.
138. Cohen MM. Long-term risk of hysterectomy following tubal sterilization. *Am J Epidemiol* 1987;125:410-9.
139. Roos LL, Roos NP, Hentleff PD. Assessing the impact of tonsillectomies. *Med Care* 1978;16:502-18.
140. Goldberg MS, Carpenter M, Theriault G, Fair M. The accuracy of ascertaining vital status in a historical cohort study of synthetic textiles workers using computerized record linkage to the Canadian Mortality Data Base. *Can J Public Health* 1993;84(3):201-4.
141. Schnatter AR, Acquavella JF, Thomson FS, Donaleski D. An analysis of death ascertainment and follow-up through Statistics Canada's Mortality Data Base system. *Can J Public Health* 1990;81:60-5.
142. Newcombe HB, Smith ME, Howe GR, Mingay J, Strugnell A, Abbatt JD. Reliability of computerized versus manual death searches in a study of the health of Eldorado uranium workers. *Comput Biol Med* 1983;13:157-69.
143. Roos LL, Nicol JP, Cageorge SM. Using administrative data for longitudinal research: comparisons with primary data collection. *J Chronic Dis* 1987;40:41-9.
144. Tennis P, Andrews E, Bombardier C, Wang Y, Strand L, West R, et al. Record linkage to conduct an epidemiologic study on the association of rheumatoid arthritis and lymphoma in the province of Saskatchewan, Canada.

- J Clin Epid 1993;46:685-95.
145. Roos LL, Wajda A. Record linkage strategies: Part I. Estimating information and evaluating approaches. *Meth Inform Med* 1991;30:117-23.
 146. Newcombe HB, Fair ME, Lalonde P. Discriminating powers of partial agreements of names for linking personal records. Part I: The logical basis. *Meth Inform Med* 1989;28:86-91.
 147. Newcombe HB, Fair ME, Lalonde P. Discriminating powers of partial agreements of names for linking personal records. Part II: The empirical test. *Meth Inform Med* 1989;28:92-6.
 148. Newcombe HB. Record linking: the design of efficient systems for linking records into individual and family histories. *Am J Human Genetics* 1967;19:335-59.
 149. Jaro MA. Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida. *J Am Stat Assoc* 1989;84:414-20.
 150. Newcombe HB. Handbook of record linkage. New York:Oxford University Press, 1988.
 151. Young JF. Information theory. New York:Wiley-Interscience, 1971.
 152. Smith ME. Record linkage: present status and methodology. *J Clin Comp* 1984; 13:52-69.
 153. Howe GR, Lindsay J. A generalized iterative record linkage computer system for use in medical follow-up studies. *Computers and Biomedical Research* 1981;14:327-40.
 154. Fellegi IP, Sunter AB. A theory of record linkage. *J Am Stat Assoc* 1969;64:1183-210.
 155. Gill LE, Baldwin JA. Methods and technology of record linkage: some practical considerations. In: Baldwin JA, Acheson ED, Graham WJ, editors. Textbook of medical record linkage. New York:Oxford University Press, 1987: 39-54.
 156. SAS Institute Inc. SAS proprietary software, release 6.11. Cary, North Carolina:SAS Institute Inc., 1995.
 157. Wajda A. LinkPro user's manual, version 1. Winnipeg, Manitoba:InfoSoft,

- 1992.
158. Sackett DL, Haynes RB, Guyatt GH, Tugwell P. *Clinical epidemiology: a basic science for clinical medicine* (second edition). Boston:Little, Brown and Company, 1991.
 159. Bailar JC III, Ederer F. Significance factors for the ratio of a Poisson variable to its expectation. *Biometrics* 1964;20:639-43.
 160. Van den Brandt PA, Schouten LJ, Goldbohm RA, Dorant E, Hunen PMH. Development of a record linkage protocol for use in the Dutch Cancer Registry for epidemiological research. *Int J Epid* 1990;19:553-8.
 161. Bryant H, Brasher P. Breast implants and breast cancer -- reanalysis of a linkage study. *N Engl J Med* 1995;332:1535-9.
 162. Devesa SS, Blot WJ, Stone BJ, Miller BA, Tarone RE, Fraumeni FJ Jr. Recent cancer trends in the United States. *J Natl Cancer Inst* 1995;87:175-82.
 163. Wynder EL, Cohen LA. Correlating nutrition to recent cancer mortality statistics [letter]. *J Natl Cancer Inst* 1997; 89:324.
 164. Chute CG, Willett WC, Colditz GA, Stampfer MJ, Rosner B, Speizer FE. A prospective study of reproductive history and exogenous estrogens on the risk of colorectal cancer in women. *Epidemiology* 1991; 2:201-207.
 165. LaVecchia C, Franceschi S. Reproductive factors and colorectal cancer. *Cancer Causes Control* 1991; 2:193-200.
 166. Gullo L, Pezzilli R, Morselli-Labate AM, and the Italian Pancreatic Cancer Study Group. Diabetes and the risk of pancreatic cancer. *N Engl J Med* 1994; 331:81-84.
 167. Mannes AG, Weinzierl M, Stellaard F, Thieme C, Wiebecke B, Paumgartner G. Adenomas of the large intestine after cholecystectomy. *Gut* 1984;25:863-866.
 168. Moorehead RJ, McKelvey STD. Cholecystectomy and colorectal cancer. *Br J Surg* 1989; 76:250-253.
 169. Parkin DM, Muir CS, Whelan SL, Gao YT, Ferlay J, Powell J. *Cancer incidence in five continents, volume VI*. Lyon:International Agency for Research on Cancer, 1992.

**APPENDIX A. ALBERTA HEALTH CARE INSURANCE PLAN FEE CODES
FOR PROCEDURES**

(Procedure, Effective Years, and Fee code)

1. Biliary Tract Procedures

a. Cholecystectomy

Cholecystectomy	1973-1976	K-114
		K-114A
	1977-1992	K-389
		K-390

b. Other Procedures, "With or Without Cholecystectomy"

Transduodenal	1972-1976	K-117A
Sphincteroplasty	1977-1993	K-393
(with or without cholecystectomy)	1986-1993	K-393A
Choledocho- enterostomy	1972-1976	K-118
	1977-1993	K-394
	1986-1993	K-394A
Choledochostomy	1972-1976	K-117
	1977-1993	K-392
	1986-1993	K-392A

c. Cholecystostomy

Cholecystostomy	1972-1976	K-115
	1977-1993	K-388

2. Stripping and Ligation of Varicose Veins

Ligation of deep vein and stripping saphenous	1973-1976	K-106
Radical multiple ligation of incompetent	1973-1976	K-108

communicating veins of lower leg		
Superficial femoral - ligation	1973-1976	K-109
Post-phlebitic leg, lation of deep vein and stripping of saphenous	1977-1987	K-541
Saphenous ligation, bilateral	1977-1989	K-533
Ligation and stripping of long saphenous, bilateral	1977-1989	K-535
Ligation and stripping of long and short saphenous veins, bilateral	1977-1989	K-537
Saphenous ligation, unilateral	1977-1993	K-532
Ligation and stripping of long saphenous, unilateral	1977-1993	K-534
Ligation and stripping of long and short saphenous veins, unilateral	1977-1993	K-536
Ligation and stripping of short saphenous vein	1977-1993	K-538
Varicose veins (complicated)	1977-1993	K-539
Radical multiple ligation of	1977-1993	K-540

incompetent
communicating veins
of lower leg -
excludes stripping of
long saphenous vein

3. *Gastric Procedures*

Epigastric hernia	1972-1976 1977-1993	K-209 K-368
Anti-reflux procedure	1972-1976 1976-1978 1979-1993	K-69 (oesophagus: cardioplasty) K-143 (esophago-gastric reconstruction) K-426 (cardioplasty) K-427 (esophago-gastric reconstruction) K-426 (anti-reflux procedure) K-427 (esophago-gastric reconstruction)
Anti-reflux procedure for recurrent esophagitis following a previous repair	1972-1979 1980-1993	NA K-426A
Vagotomy: transthoracic or abdominal	1972-1976 1977-1983 1984-1993	K-164 (thoracic) K-165 (abdominal) K-430 (transthoracic) K-431 (transabdominal) K-430
Vagotomy: selective, for denervation of parietal cells	1972-1976 1977-1993	NA K-432
Pyloroplasty	1972-1976 1977-1987 1988-1993	K-147 (adult) K-148 (adult with vagotomy) K-439 (adult without vagotomy) K-440 (adult with vagotomy) K-439
Gastroenterostomy	1972-1976	K-127 (without vagotomy) K-128 (with vagotomy)

	1977-1987	K-441 (without vagotomy) K-442 (with vagotomy)
	1988-1993	K-441
Gastrectomy, sub-total	1972-1976	K-125 (with or without vagotomy) K-125A K-125B K-125D
	1977-1982	K-443 (with or without vagotomy) K-444 K-445 K-446
	1983-1987	K-443A (without vagotomy) K-443B (with vagotomy) K-444 K-445 K-446
	1988	K-443 (without vagotomy ONLY) K-444 K-445 K-446
	1989-1993	K-443 K-444 K-445 K-446
Gastrectomy, total	1972-1976	K-125E K-125F
	1977-1993	K-447 K-448 K-449

Note: Fee codes and full descriptions can be found in the general surgery section of the *Alberta Health Care Insurance Plan Handbook for Claim Submissions*, Edmonton:Alberta Health, 1972-1993.

APPENDIX B. MEASURES OF INFORMATION, AHCIP AND COLON CANCER COHORTS

	Colon Cancer Cohort			AHCIP Cohorts		
	Pocket size	Shannon entropy	Discrim. power [*]	Pocket size	Shannon entropy	Discrim. power [*]
Soundex, both initials, birthyear and birthmonth	1.000	13.650	13.649	1.001	16.973	16.972
Soundex, both initials, and birthyear	1.003	13.645	13.642	1.011	16.953	16.942
Soundex, first initial, and birthyear	1.012	13.627	13.617	1.085	16.808	16.728
Soundex, birthyear and birthmonth	1.018	13.614	13.598	1.109	16.766	16.672
Birthyear, birthmonth, both initials and sex	1.102	13.448	13.347	1.371	16.314	16.009
Soundex and birthyear	1.172	13.323	13.171	1.921	15.602	15.056
Birthyear, birthmonth and both initials	1.170	13.324	13.168	1.539	16.075	15.691
Birthdate	1.381	13.042	12.872	4.701	14.418	14.166
Birthyear, birthmonth, first initial and sex	1.383	13.013	12.794	4.043	14.421	13.983
Birthyear, birthmonth and first initial	1.686	12.658	12.395	6.483	13.746	13.407
Last name	1.668	12.284	10.965	3.264	13.768	11.486
Soundex	5.497	10.170	9.276	31.833	10.449	9.475
AHCIP number	1.093	12.789	7.088	1.000	16.976	16.976
Both initials	25.359	7.207	5.843	204.163	7.918	7.025
First initial	494.500	4.176	3.947	4771.370	4.214	4.024
Middle initial	494.500	3.087	1.953	4954.885	3.744	3.070
Birthyear	127.297	5.879	5.660	1091.754	6.174	6.021
Birthday	401.781	4.976	4.963	4025.844	4.988	4.976
Birthmonth	1071.417	3.583	3.580	9909.769	3.591	3.585
Sex	6428.500	1.000	1.000	64413.500	0.909	0.832

* Discrim. power = discriminating power

APPENDIX C. DETERMINISTIC LINKAGE RESULTS

Variables Used in Merge	Records (Individuals) Available for Merge		Number of Matched Pairs	Number of Individuals Represented in Matched Pairs		Match Ratio (ULI:ACB#)
	AHCIP	Registry		By ULI	By ACB#	
AHCIP#, lastname, first initial, middle initial, birthdate, sex	128,827	12,857	915	915	915	1:1
AHCIP#, alternative lastname, first initial, middle initial, birthdate, sex	127,912	11,942	5	5	5	1:1
AHCIP# only	127,907	11,937	673	673	673	1:1
Soundex ¹ , first initial, middle initial, birthdate, sex	127,234	11,264	43	43	43	1:1
Alternative soundex ² , first initial, middle initial, birthdate, sex	127,191	11,221	0	0	0	1:1
Soundex, first initial, middle initial, birthyear, birthmonth, sex	127,191	11,221	24	24	24	1:1

Alternative soundex, first initial, middle initial, birthyear, birthmonth, sex	127,167	11,197	24	13	14	1:1.08
Soundex, first initial, middle initial, birthyear, birthmonth	127,153	11,184	4	4	4	1:1
Alternative soundex, first initial, middle initial, birthyear, birthmonth	127,149	11,180	10	6	5	1.20:1
Soundex, first initial, middle initial, birthyear, sex	127,144	11,174	139	132	139	1:1.05
Alternative soundex, first initial, middle initial, birthyear, sex	127,005	11,042	175	92	75	1.23:1
Soundex, first initial, middle initial, birthyear	126,930	10,950	71	69	71	1:1.03
Alternative soundex, first initial, middle initial, birthyear	126,859	10,881	50	37	35	1.06:1
Soundex, first initial, birthyear, sex	126,824	10,844	629	553	621	1:1.12

Alternative soundex, first initial, birthyear, sex	126,203	10,291	458	263	253	1.04:1
Soundex, alternative first initial, birthyear, sex	125,950	10,028	7	7	7	1:1
Alternative soundex, alternative first initial, birthyear, sex	125,943	10,021	1	1	1	1:1
Soundex, birthyear, birthmonth, sex	125,942	10,020	615	571	604	1:1.06
Alternative soundex, birthyear, birthmonth, sex	125,338	9,449	243	179	181	1:1.01
Soundex, first initial, birthyear	125,157	9,270	320	297	320	1:1.08
Alternative soundex, first initial, birthyear	124,837	8,973	80	61	68	1:1.11
Soundex, alternative first initial, birthyear	124,769	8,912	7	7	7	1:1
Alternative soundex, alternative first initial, birthyear	124,762	8,905	0	0	0	1:1
Soundex, birthyear, birthmonth	124,762	8,905	481	445	475	1:1.07
Alternative soundex, birthyear, birthmonth	124,287	8,460	119	96	98	1:1.02

First initial, middle initial, birthyear, birthmonth, sex	124,189	8,364	5941	2878	5099	1:1.77
Soundex, birthyear, sex	119,090	5,486	2196	1362	2098	1:1.54
Alternative soundex, birthyear, sex	116,992	4,124	460	223	391	1:1.75
First initial, middle initial, birthyear, birthmonth	116,601	3,901	886	559	836	1:1.50
Soundex, birthyear	115,765	3,342	745	525	727	1:1.38
Alternative soundex, birthyear	115,038	2,817	91	77	90	1:1.17
Birthdate, sex	114,948	2,740	2058	1070	1987	1:1.86
Birthdate	112,961	1,670	527	312	514	1:1.65
First initial, birthyear, birthmonth, sex	112,447	1,358	390	242	372	1:1.54
Alternative first initial, birthyear, birthmonth, sex	112,075	1,116	12	6	12	1:2.00
First initial, birthyear, birthmonth	112,063	1,110	151	108	137	1:1.27
Alternative first initial, birthyear, birthmonth	111,926	1,002	2	1	2	1:2.00

¹ Soundex = soundex code for last name using soundex algorithm in LinkPro

² Alternative soundex = soundex code for alternative last name from the Alberta Cancer Registry using soundex algorithm in LinkPro

APPENDIX D. PROBABILISTIC RESULTS: LINKPRO OUTPUT**Variable abbreviations:**

BTH_YR	Birthyear
SEX	Sex
AHCIPNO	AHCIP registration number
SOUNDEX	Soundex code for last name
LSTNAME	Last name
FSTINIT	First initial
MIDINIT	Middle initial
BTH_MO	Birthmonth
BTH_DY	Birthday

Part 1: Block on BTHVR using all available records.
 Linkpro 1.2: Probabilistic (non-specific weights) Record Linkage Summary
 Linking data set: WORK.ATEMP /n=128997 with data set: WORK.RTEMP /n=17581
 Linked: 6633 (5.15%) Unresolved: 51687 (40.10%) Not Linked: 70577 (54.75%)

BTH_YR	SEX	ANCIPNO	SOUNDEX	LSTNAME	FSTINIT	MIDINIT	BTH_MO	BTH_DY	Matched	Freq.	\$	Min. Weight	Max. Weight
1	0	0	0	0	0	1	1	1	4	10	0.15	-1.01	-1.01
1	0	0	0	0	1	0	1	1	4	160	2.41	-1.01	-1.01
1	0	0	0	0	1	1	0	1	4	6	0.09	-1.01	-1.01
1	0	0	0	0	1	1	1	0	4	24	0.36	-0.80	-0.80
1	0	0	1	0	0	0	1	1	4	4	0.06	5.09	5.09
1	0	0	1	0	0	1	1	0	4	14	0.08	2.57	2.57
1	0	0	1	0	1	0	1	0	4	6	0.21	2.79	2.79
1	0	0	1	0	1	0	0	1	4	16	0.09	5.09	5.09
1	0	0	1	0	1	0	0	0	4	4	0.24	5.30	5.30
1	0	0	1	0	1	1	0	0	4	4	0.06	2.78	2.78
1	0	0	1	0	0	0	1	0	4	33	0.50	10.44	10.44
1	0	0	1	1	0	0	1	0	4	103	1.55	10.44	10.44
1	0	0	1	1	0	0	0	0	4	106	1.60	10.66	10.66
1	0	0	1	1	1	0	0	0	4	59	0.89	10.65	10.65
1	0	0	1	1	1	0	0	0	4	1	0.02	25.58	25.58
1	1	0	0	0	0	0	1	1	4	94	1.42	1.04	1.04
1	1	0	0	0	0	1	0	1	4	10	0.15	-1.48	-1.48
1	1	0	0	0	0	1	1	0	4	36	0.54	-1.27	-1.27
1	1	0	0	0	1	0	0	1	4	42	0.63	-1.27	-1.27
1	1	0	0	0	1	0	1	0	4	355	5.35	1.03	1.03
1	1	0	0	0	1	1	0	0	4	14	0.21	1.25	1.25
1	1	0	0	0	0	0	0	1	4	36	0.54	-1.27	-1.27
1	1	0	0	0	0	0	1	0	4	111	1.67	4.62	4.62
1	1	0	0	0	0	1	0	0	4	89	1.34	4.84	4.84
1	1	0	0	0	1	0	0	0	4	74	1.12	2.32	2.32
1	1	0	0	1	0	0	0	0	4	544	8.20	4.83	4.83
1	0	0	0	0	0	0	0	0	4	1956	29.49	10.19	10.19
1	0	0	0	0	1	1	1	1	5	85	1.28	5.40	5.40
1	0	0	0	0	1	1	1	1	5	2	0.03	8.99	8.99
1	0	0	0	0	1	0	1	1	5	2	0.03	11.50	11.50
1	0	0	0	0	1	1	0	1	5	2	0.03	8.98	8.98
1	0	0	0	0	1	1	1	0	5	6	0.09	9.19	9.19
1	0	0	0	0	1	0	1	1	5	6	0.09	16.86	16.86
1	0	0	0	0	1	0	0	1	5	14	0.21	14.34	14.34
1	0	0	0	0	1	0	0	1	5	34	0.51	14.55	14.55
1	0	0	0	0	1	0	1	0	5	8	0.12	16.85	16.85
1	0	0	0	0	1	1	0	0	5	10	0.15	17.06	17.06
1	0	0	0	0	1	1	0	0	5	10	0.15	14.55	14.55
1	1	0	0	0	0	0	1	1	5	292	4.40	4.93	4.93
1	1	0	0	0	0	0	1	1	5	464	7.00	4.93	4.93
1	1	0	0	0	0	1	0	1	5	144	2.17	7.45	7.45
1	1	0	0	0	0	1	1	0	5	761	11.47	4.93	4.93
1	1	0	0	0	0	0	1	1	5	20	0.30	5.14	5.14
1	1	0	0	0	0	0	1	1	5	43	0.65	11.01	11.01
1	1	0	0	0	0	1	0	1	5	119	1.79	8.52	8.52
1	1	0	0	0	1	0	0	1	5	16	0.24	8.73	8.73
1	1	0	0	0	1	0	0	1	5	43	0.65	11.03	11.03
1	1	0	0	0	1	0	0	0	5	88	1.33	11.24	11.24
1	1	0	0	0	1	1	0	0	5	46	0.69	8.72	8.72
1	1	0	0	0	0	0	0	1	5	192	2.89	16.39	16.39
1	1	0	0	0	0	0	1	0	5	201	3.03	16.60	16.60
1	1	0	0	0	1	0	0	0	5	158	2.38	14.08	14.08
1	1	0	0	0	1	0	0	0	5	2	0.03	16.59	16.59
1	1	0	0	0	0	0	1	0	5	2768	41.73	26.17	26.17

Part 1: Block on BTHVR using all available records (continued)

BTH_YR	SEX	ANICIPNO	SOUNDEX	LSTNAME	FSTINIT	MIDINIT	BTH_MO	BTH_DY	Matched	Freq.	\$	Min. Weight	Max. Weight
1	0	0	1	1	0	1	1	1	6	1	0.02	20.75	20.75
1	0	0	1	1	1	1	0	1	6	1	0.02	20.74	20.74
1	0	0	1	1	1	1	1	0	6	1	0.02	20.96	20.96
1	0	1	1	1	0	0	1	1	6	1	0.02	38.19	38.19
1	0	1	1	1	1	0	1	0	6	1	0.02	38.39	38.39
1	1	0	0	0	1	1	1	1	6	170	2.56	11.34	11.34
1	1	0	1	0	0	1	1	1	6	5	0.08	14.93	14.93
1	1	0	1	0	1	0	0	1	6	2	0.03	17.44	17.44
1	1	0	1	0	1	1	0	1	6	1	0.02	14.92	14.92
1	1	0	1	1	0	1	1	0	6	18	0.27	15.13	15.13
1	1	0	1	1	0	0	1	1	6	7	0.11	22.80	22.80
1	1	0	1	1	0	1	0	1	6	8	0.12	20.28	20.28
1	1	0	1	1	0	1	1	0	6	32	0.48	20.49	20.49
1	1	0	1	1	1	0	0	1	6	12	0.18	22.79	22.79
1	1	0	1	1	1	0	0	1	6	25	0.38	23.00	23.00
1	1	0	0	1	1	1	0	0	6	36	0.54	20.49	20.49
1	1	1	0	0	1	0	1	1	6	1	0.02	28.78	28.78
1	1	1	0	0	1	1	1	0	6	1	0.02	26.47	26.47
1	1	1	0	0	0	0	1	1	6	1	0.02	32.36	32.36
1	1	1	1	1	1	0	0	0	6	4	0.06	37.92	37.92
1	0	1	1	1	1	0	1	1	7	328	4.84	44.59	44.59
1	1	0	1	0	1	1	1	1	7	2	0.03	21.23	21.23
1	1	0	1	1	0	1	1	1	7	1	0.02	26.69	26.69
1	1	0	1	1	1	0	1	1	7	17	0.26	29.20	29.20
1	1	0	1	1	1	1	1	0	7	11	0.17	26.90	26.90
1	1	1	0	0	1	1	1	1	7	6	0.09	32.67	32.67
1	1	1	1	0	1	0	1	0	7	9	0.14	38.77	38.77
1	1	1	1	0	1	1	1	0	7	1	0.02	36.47	36.47
1	1	1	1	1	0	0	1	1	7	28	0.42	44.13	44.13
1	1	1	1	1	1	0	0	1	7	2	0.03	44.12	44.12
1	1	1	1	1	1	0	0	0	7	43	0.65	44.34	44.34
1	1	1	1	1	1	1	0	0	7	18	0.27	41.82	41.82
1	0	1	1	1	1	1	0	0	7	139	2.10	48.49	48.49
1	0	1	1	1	1	1	1	1	8	3	0.05	48.49	48.49
1	1	0	1	1	1	1	1	1	8	44	0.66	33.10	33.10
1	1	1	1	0	1	1	1	1	8	15	0.23	42.66	42.66
1	1	1	1	1	0	1	1	1	8	4	0.06	48.02	48.02
1	1	1	1	1	1	0	1	1	8	322	4.85	50.53	50.53
1	1	1	1	1	1	1	0	1	8	10	0.15	48.02	48.02
1	1	1	1	1	1	1	1	0	8	125	1.88	48.23	48.23
1	1	1	1	1	1	1	1	1	9	523	7.88	54.43	54.43
1	1	1	1	1	1	1	1	1	9	919	13.85	54.43	54.43
1	1	1	1	1	1	1	1	1	9	6633	100.00	54.43	54.43
6633	5893	1519	3958	3203	4378	3551	4782	3143		128897	5.15		
100	88.8	22.9	59.6	48.2	66	53.5	72	47.3					

Part 11: Block on FSTINIT and SEX with remaining records
 LinkPro 1.2
 Probabilistic (non-specific weights) Record Linkage Summary
 Linking data set: WORK.ATEMP /n=122263 with data set: WORK.RTEMP /n=9054
 Linked: 2134 (1.75%) Unresolved: 10482 (8.57%) Not Linked: 109647 (89.68%)

SEX	FSTINIT	BTH_YR	AHCIPNO	SOUNDEX	LSTNAME	MIDINIT	BTH_MO	BTH_DY	Matched	Freq.	#	Min. Weight	Max. Weight
1	1	0	0	0	0	1	1	1	5	608	28.49	8.43	8.43
1	1	0	0	1	0	0	1	1	5	59	2.76	13.15	13.15
1	1	0	0	1	0	1	0	1	5	96	4.50	12.28	12.28
1	1	0	0	1	0	1	0	0	5	254	11.90	11.73	11.73
1	1	0	0	1	1	0	0	1	5	164	7.69	18.72	18.72
1	1	0	0	1	1	0	1	0	5	323	15.14	18.17	18.17
1	1	0	0	1	1	1	0	0	5	343	16.07	17.30	17.30
1	1	0	1	1	1	1	0	0	5	1	0.05	33.11	33.11
1	1	1	0	0	0	0	1	1	5	4	0.19	4.99	4.99
1	1	1	0	0	0	1	1	1	5	32	1.50	4.12	4.12
1	1	1	0	0	0	1	1	0	5	13	0.61	3.57	3.57
1	1	1	0	1	0	1	0	0	5	2	0.09	7.42	7.42
1	1	1	0	1	1	0	0	0	5	1	0.05	13.86	13.86
1	1	0	0	1	0	1	1	1	6	1900	89.03	17.93	17.93
1	1	0	0	1	1	0	1	1	6	12	0.56	24.37	24.37
1	1	0	0	1	1	1	1	1	6	36	1.69	23.50	23.50
1	1	0	0	1	1	1	1	0	6	57	2.67	22.94	22.94
1	1	0	1	1	0	1	1	0	6	87	4.08	32.31	32.31
1	1	0	1	1	1	0	1	0	6	1	0.05	38.75	38.75
1	1	0	1	1	1	1	0	0	6	2	0.09	37.88	37.88
1	1	1	0	0	0	1	1	1	6	4	0.19	9.76	9.76
1	1	0	0	1	1	1	1	1	7	200	9.37	29.14	29.14
1	1	0	0	1	1	1	1	1	7	5	0.23	44.94	44.94
1	1	0	1	1	1	1	1	0	7	5	0.23	43.52	43.52
1	1	0	1	1	1	1	1	0	7	6	0.28	49.72	49.72
1	1	0	1	1	1	1	1	1	8	16	0.75		
1	1	0	1	1	1	1	1	1	8	18	0.84		
1	1	0	1	1	1	1	1	1	8	2134	100.00		
2134	2134	53	37	1476	1052	1539	1434	1097		122263	1.75		
100	100	2.4	1.7	69.1	49.2	72.1	67.1	51.4					

APPENDIX E. PROBABILISTIC LINKAGE THRESHOLDS

Phase I. Block on Birthyear Only

Column abbreviations:

OBS	Observation number (from SAS)
AGREE	Number of variables that agree in candidate pair
CONSTRCT	Construct of variables that agree
MCOUNT	Number of observations in M
MPERCENT	Percent of all observations in M
_WGT	Probabilistic weight
UCOUNT	Number of observations in U
UPERCENT	Percent of all observations in U
MCUMFREQ	Cumulative frequency of M
UCUMFREQ	1 - cumulative frequency of U

Constructs for Phase I of Probabilistic Linkage
Merge of MATCHED and UNMATCHED
(U based on sample of n=136,000)
CONSTRUCT: BYR/SEX/ARCIP#/SNDX/LST/FST/MID/BMO/BDY

OBS	AGREE	CONSTRUCT	MCOUNT	MPERCENT	_WGT	UCOUNT	UPERCENT	MCUMFREQ	UCUMFREQ
LOW (EXCLUDE from review; EXCLUDE from dataset)									
1	0	000000000	.	.	-9999.99	43274	31.6191	0.000	68.1809
2	1	000000001	.	.	-9999.99	1405	1.0331	0.000	67.1478
3	1	000000010	.	.	-9999.99	3932	2.8912	0.000	64.2566
4	2	000000011	.	.	-9999.99	124	0.0912	0.000	64.1654
5	1	000000100	.	.	-9999.99	6509	4.7860	0.000	59.3794
6	2	000000101	.	.	-9999.99	217	0.1596	0.000	59.2199
7	2	000000110	.	.	-9999.99	593	0.4360	0.000	58.7838
8	3	000000111	.	.	-9999.99	15	0.0110	0.000	58.7728
9	1	000001000	.	.	-9999.99	2585	1.9007	0.000	56.8721
10	2	000001001	.	.	-9999.99	72	0.0529	0.000	56.8191
11	2	000001010	.	.	-9999.99	236	0.1735	0.000	56.6456
12	3	000001011	.	.	-9999.99	6	0.0044	0.000	56.6412
13	2	000001100	.	.	-9999.99	437	0.3213	0.000	56.3199
14	3	000001101	.	.	-9999.99	11	0.0031	0.000	56.3118
15	3	000001110	.	.	-9999.99	39	0.0287	0.000	56.2831
16	1	000100000	.	.	-9999.99	41	0.0301	0.000	56.2529
17	2	000100001	.	.	-9999.99	2	0.0015	0.000	56.2515
18	2	000100010	.	.	-9999.99	5	0.0037	0.000	56.2478
19	2	000100100	.	.	-9999.99	8	0.0059	0.000	56.2419
20	2	000101000	.	.	-9999.99	3	0.0022	0.000	56.2397
21	3	000101001	.	.	-9999.99	1	0.0007	0.000	56.2390
22	2	000110000	.	.	-9999.99	21	0.0154	0.000	56.2235
23	3	000110001	.	.	-9999.99	1	0.0007	0.000	56.2228
24	3	000110010	.	.	-9999.99	3	0.0022	0.000	56.2206
25	3	000110100	.	.	-9999.99	1	0.0007	0.000	56.2199
26	3	000111000	.	.	-9999.99	1	0.0007	0.000	56.2191
27	1	010000000	.	.	-9999.99	52673	38.7301	0.000	17.4890
28	2	010000001	.	.	-9999.99	1671	1.2237	0.000	16.2603
29	2	010000010	.	.	-9999.99	4772	3.5088	0.000	12.7515
30	3	010000011	.	.	-9999.99	160	0.1176	0.000	12.6338
31	2	010000100	.	.	-9999.99	9407	6.9169	0.000	5.7169
32	3	010000101	.	.	-9999.99	307	0.2257	0.000	5.4912
33	3	010000110	.	.	-9999.99	863	0.6346	0.000	4.8566
34	4	010000111	.	.	-9999.99	21	0.0154	0.000	4.8412
35	2	010001000	.	.	-9999.99	3706	2.7250	0.000	2.1162
36	3	010001001	.	.	-9999.99	107	0.0787	0.000	2.0375
37	3	010001010	.	.	-9999.99	336	0.2471	0.000	1.7904
38	4	010001011	.	.	-9999.99	15	0.0110	0.000	1.7794
39	3	010001100	.	.	-9999.99	728	0.5353	0.000	1.2441
40	4	010001101	.	.	-9999.99	19	0.0140	0.000	1.2301
41	4	010001110	.	.	-9999.99	69	0.0507	0.000	1.1794
42	5	010001111	.	.	-9999.99	4	0.0029	0.000	1.1765
43	2	010100000	.	.	-9999.99	61	0.0449	0.000	1.1316
44	3	010100001	.	.	-9999.99	2	0.0015	0.000	1.1301
45	3	010100010	.	.	-9999.99	2	0.0015	0.000	1.1287
46	3	010100100	.	.	-9999.99	9	0.0066	0.000	1.1221
47	3	010101000	.	.	-9999.99	7	0.0051	0.000	1.1169
48	4	010101001	.	.	-9999.99	1	0.0007	0.000	1.1162
49	3	010110000	.	.	-9999.99	17	0.0125	0.000	1.1037
50	4	010110001	.	.	-9999.99	1	0.0007	0.000	1.1029
51	4	010110010	.	.	-9999.99	4	0.0029	0.000	1.1000
52	4	010110100	.	.	-9999.99	2	0.0015	0.000	1.0985
53	4	010111000	.	.	-9999.99	2	0.0015	0.000	1.0971
54	5	010111100	.	.	-9999.99	2	0.0015	0.000	1.0956
55	1	100000000	.	.	-9999.99	466	0.3426	0.000	0.7529
56	2	100000001	.	.	-9999.99	14	0.0103	0.000	0.7426
57	2	100000010	.	.	-9999.99	58	0.0426	0.000	0.7000
58	3	100000011	.	.	-9999.99	1	0.0007	0.000	0.6993
59	2	100000100	.	.	-9999.99	85	0.0625	0.000	0.6368
60	3	100000101	.	.	-9999.99	4	0.0029	0.000	0.6338

Constructs for Phase I of Probabilistic Linkage (continued)

61	3	100000110	.	.	-9999.99	3	0.0022	0.000	0.6316
62	2	100001000	.	.	-9999.99	25	0.0184	0.000	0.6132
63	3	100001001	.	.	-9999.99	1	0.0007	0.000	0.6125
64	3	100001010	.	.	-9999.99	5	0.0037	0.000	0.6086
65	3	100001100	.	.	-9999.99	6	0.0044	0.000	0.6044
66	2	100100000	.	.	-9999.99	1	0.0007	0.000	0.6037
67	3	100100001	.	.	-9999.99	1	0.0007	0.000	0.6029
68	2	110000000	.	.	-9999.99	542	0.3985	0.000	0.2044
69	3	110000001	.	.	-9999.99	18	0.0132	0.000	0.1912
70	3	110000010	.	.	-9999.99	53	0.0390	0.000	0.1522
71	3	110000100	.	.	-9999.99	131	0.0963	0.000	0.0559
72	3	110001000	.	.	-9999.99	45	0.0331	0.000	0.0228
73	4	110000101	10	0.1508	-1.48	3	0.0022	0.000	0.0206
74	4	110001100	14	0.2111	-1.27	9	0.0066	0.000	0.0140
75	4	110000110	36	0.5427	-1.27	6	0.0044	0.000	0.0096
76	4	100001101	6	0.0905	-1.01	.	.	0.000	0.0096
77	4	100000111	10	0.1508	-1.01	.	.	1.000	0.0096
78	4	100001110	24	0.3618	-0.80	.	.	1.500	0.0096
79	4	110001001	42	0.6332	1.03	2	0.0015	2.000	0.0081
80	4	110000011	94	1.4172	1.04	1	0.0007	3.558	0.0074
81	4	110001010	355	5.3520	1.25	6	0.0044	8.000	0.0029
82	4	100001011	160	2.4122	1.50	.	.	11.322	0.0029
83	4	110100100	89	1.3418	2.32	.	.	12.664	0.0029
84	4	100100101	5	0.0754	2.57	.	.	12.739	0.0029
85	4	100101100	4	0.0603	2.78	.	.	12.300	0.0029
86	4	100100110	14	0.2111	2.79	.	.	13.011	0.0029
87	4	110100001	36	0.5427	4.62	.	.	13.553	0.0029
88	4	110101000	74	1.1156	4.83	.	.	14.669	0.0029
89	4	110100010	111	1.6735	4.84	.	.	16.343	0.0029
90	5	110001101	144	2.1710	4.93	.	.	18.513	0.0029
91	5	110000111	292	4.4022	4.93	1	0.0007	22.916	0.0022
92	4	100101001	6	0.0905	5.09	.	.	23.006	0.0022
93	4	100100011	4	0.0603	5.09	.	.	23.066	0.0022
94	5	110001110	761	11.4729	5.14	1	0.0007	34.539	0.0015
95	4	100101010	16	0.2412	5.30	.	.	34.781	0.0015
96	5	100001111	85	1.2815	5.40	.	.	36.062	0.0015
97	5	110001011	464	6.9953	7.45	.	.	43.057	0.0015
98	4	100110100	106	1.5981	8.14	.	.	44.656	0.0015
99	5	110100101	43	0.6483	8.52	.	.	45.304	0.0015
100	5	110101100	88	1.3267	8.72	1	0.0007	46.630	0.0007
101	5	110100110	119	1.7941	8.73	.	.	48.425	0.0007
102	5	100101101	2	0.0302	8.98	.	.	48.455	0.0007
103	5	100100111	2	0.0302	8.99	.	.	48.485	0.0007
104	5	100101110	6	0.0905	9.19	.	.	48.575	0.0007
105	4	110110000	544	8.2014	10.19	.	.	56.777	0.0007
106	4	100110001	33	0.4975	10.44	.	.	57.274	0.0007
107	4	100111000	59	0.8895	10.65	.	.	58.164	0.0007
108	4	100110010	103	1.5528	10.66	.	.	59.717	0.0007
109	5	110101001	16	0.2412	11.03	.	.	59.958	0.0007
110	5	110100011	20	0.3015	11.03	.	.	60.259	0.0007
111	5	110101010	43	0.6483	11.24	.	.	60.908	0.0007
112	6	110001111	170	2.5629	11.34	.	.	63.471	0.0007
113	5	100101011	2	0.0302	11.50	.	.	63.501	0.0007
114	5	110110100	201	3.0303	14.08	.	.	66.531	0.0007
115	5	100110101	14	0.2111	14.34	.	.	66.742	0.0007
116	5	100111100	10	0.1508	14.55	.	.	66.833	0.0007
117	5	100110110	34	0.5126	14.55	.	.	67.405	0.0007
118	6	110101101	1	0.0151	14.92	.	.	67.420	0.0007
119	6	110100111	5	0.0754	14.93	.	.	67.496	0.0007
120	6	110101110	18	0.2714	15.13	.	.	67.757	0.0007
121	5	110110001	46	0.6935	16.39	.	.	68.451	0.0007
122	5	110111000	158	2.3820	16.59	.	.	70.843	0.0007
123	5	110110010	192	2.8946	16.60	1	0.0007	73.737	0.0000

Constructs for Phase I of Probabilistic Linkage (continued)

REVIEW:								
124	5	100111001	8	0.1206	16.85	.	.	73.853 0.0000
125	5	100110011	6	0.0905	16.86	.	.	73.349 0.0000
126	5	100111010	10	0.1508	17.06	.	.	74.039 0.0000
127	6	110101011	2	0.0302	17.44	.	.	74.129 0.0000
128	6	110110101	3	0.1206	20.29	.	.	74.250 0.0000
129	6	110111100	36	0.5427	20.49	.	.	74.793 0.0000
130	6	110110110	32	0.4324	20.49	.	.	75.275 0.0000
131	6	100111101	1	0.0151	20.74	.	.	75.230 0.0000
132	6	100110111	1	0.0151	20.75	.	.	75.305 0.0000
133	6	100111110	1	0.0151	20.96	.	.	75.320 0.0000
134	7	110101111	1	0.0151	21.33	.	.	75.335 0.0000
135	6	110111001	12	0.1809	22.79	.	.	75.516 0.0000
136	6	110110011	7	0.1055	22.80	.	.	75.622 0.0000
137	6	110111010	25	0.3769	23.00	.	.	75.939 0.0000
138	4	101110000	1	0.0151	25.58	.	.	76.014 0.0000
139	5	111100010	2	0.0302	26.17	.	.	76.044 0.0000
140	6	111001110	1	0.0151	26.47	.	.	76.059 0.0000
141	7	110110111	1	0.0151	26.69	.	.	76.074 0.0000
142	7	110111110	11	0.1658	26.90	.	.	76.240 0.0000
143	6	111001011	1	0.0151	28.78	.	.	76.255 0.0000
144	7	110111011	17	0.2563	29.20	.	.	76.511 0.0000

HIGH (EXCLUDE from review; INCLUDE in dataset)								
145	6	111100011	1	0.0151	32.36	.	.	76.526 0.0000
146	7	111001111	6	0.0905	32.67	.	.	76.617 0.0000
147	8	110111111	44	0.6633	33.10	.	.	77.280 0.0000
148	7	111101110	1	0.0151	36.47	.	.	77.295 0.0000
149	6	111111000	4	0.0603	37.92	.	.	77.356 0.0000
150	6	101110011	1	0.0151	38.19	.	.	77.371 0.0000
151	6	101111010	1	0.0151	38.39	.	.	77.386 0.0000
152	7	111101011	9	0.1357	38.77	.	.	77.521 0.0000
153	7	111111100	18	0.2714	41.82	.	.	77.793 0.0000
154	8	111101111	15	0.2261	42.66	.	.	78.019 0.0000
155	7	111111001	2	0.0302	44.12	.	.	78.049 0.0000
156	7	111110011	23	0.4221	44.13	.	.	78.471 0.0000
157	7	111111010	43	0.6483	44.34	.	.	79.120 0.0000
158	7	101111011	2	0.0302	44.59	.	.	79.150 0.0000
159	8	111111101	10	0.1508	48.02	.	.	79.300 0.0000
160	8	111110111	4	0.0603	48.02	.	.	79.361 0.0000
161	8	111111110	125	1.8845	48.23	.	.	81.245 0.0000
162	8	101111111	3	0.0452	48.49	.	.	81.291 0.0000
163	8	111111011	322	4.8545	50.53	.	.	86.145 0.0000
164	9	111111111	919	13.8550	54.43	.	.	100.000 0.0000

Phase II. Block on First Initial and Sex**Column abbreviations:**

OBS	Observation number (from SAS)
AGREE	Number of variables that agree in candidate pair
CONSTRUCT	Construct of variables that agree
MCOUNT	Number of observations in M
MPERCENT	Percent of all observations in M
_WGT	Probabilistic weight
UCOUNT	Number of observations in U
UPERCENT	Percent of all observations in U
MCUMFREQ	Cumulative frequency of M
UCUMFREQ	1 - cumulative frequency of U

Constructs for Phase II of Probabilistic Linkage
 Merge of MATCHED and UNMATCHED
 (U based on sample of n=74,000)
 CONSTRUCT: SEX/FST/BYR/AHCIP#/SNDX/LST/MID/BMO/BDY

OBS	AGREE	CONSTRCT	MCOUNT	MPERCENT	_WGT	UCOUNT	UPERCENT	MCUMFREQ	UCUMFREQ
LOW (EXCLUDE from review; EXCLUDE from dataset)									
1	0	000000000	.	.	-9999.99	22408	30.2811	0.000	69.7189
2	1	000000001	.	.	-9999.99	770	1.0405	0.000	69.6784
3	1	000000010	.	.	-9999.99	2013	2.7203	0.000	65.9581
4	2	000000011	.	.	-9999.99	80	0.1081	0.000	65.8500
5	1	000000100	.	.	-9999.99	4122	5.5703	0.000	60.2797
6	2	000000101	.	.	-9999.99	144	0.1945	0.000	60.0851
7	2	000000110	.	.	-9999.99	369	0.4986	0.000	59.5365
8	3	000000111	.	.	-9999.99	11	0.0149	0.000	59.5716
9	1	000010000	.	.	-9999.99	34	0.0459	0.000	59.5257
10	2	000010010	.	.	-9999.99	3	0.0041	0.000	59.5216
11	2	000010100	.	.	-9999.99	5	0.0068	0.000	59.5149
12	2	000011000	.	.	-9999.99	10	0.0135	0.000	59.5014
13	3	000011010	.	.	-9999.99	3	0.0041	0.000	59.4973
14	3	000011100	.	.	-9999.99	1	0.0014	0.000	59.4959
15	1	001000000	.	.	-9999.99	209	0.2824	0.000	59.2135
16	2	001000001	.	.	-9999.99	13	0.0176	0.000	59.1959
17	2	001000010	.	.	-9999.99	20	0.0270	0.000	59.1689
18	3	001000011	.	.	-9999.99	1	0.0014	0.000	59.1676
19	2	001000100	.	.	-9999.99	50	0.0676	0.000	59.1000
20	3	001000101	.	.	-9999.99	5	0.0068	0.000	59.0932
21	3	001000110	.	.	-9999.99	7	0.0095	0.000	59.0838
22	1	010000000	.	.	-9999.99	1321	1.7851	0.000	57.2986
23	2	010000001	.	.	-9999.99	47	0.0635	0.000	57.2351
24	2	010000010	.	.	-9999.99	127	0.1716	0.000	57.0635
25	3	010000011	.	.	-9999.99	6	0.0081	0.000	57.0554
26	2	010000100	.	.	-9999.99	275	0.3716	0.000	56.6838
27	3	010000101	.	.	-9999.99	10	0.0135	0.000	56.6703
28	3	010000110	.	.	-9999.99	33	0.0446	0.000	56.6257
29	2	010010000	.	.	-9999.99	3	0.0041	0.000	56.6216
30	3	010011000	.	.	-9999.99	1	0.0014	0.000	56.6203
31	2	011000000	.	.	-9999.99	16	0.0216	0.000	56.5986
32	3	011000001	.	.	-9999.99	2	0.0027	0.000	56.5959
33	3	011000010	.	.	-9999.99	3	0.0041	0.000	56.5919
34	3	011000100	.	.	-9999.99	5	0.0068	0.000	56.5851
35	4	011000101	.	.	-9999.99	1	0.0014	0.000	56.5838
36	1	100000000	.	.	-9999.99	28589	38.6324	0.000	17.9514
37	2	100000001	.	.	-9999.99	963	1.3014	0.000	16.6500
38	2	100000010	.	.	-9999.99	2538	3.4297	0.000	13.2203
39	3	100000011	.	.	-9999.99	98	0.1324	0.000	13.0878
40	2	100000100	.	.	-9999.99	5625	7.6014	0.000	5.4865
41	3	100000101	.	.	-9999.99	196	0.2649	0.000	5.2216
42	3	100000110	.	.	-9999.99	463	0.6257	0.000	4.5959
43	4	100000111	.	.	-9999.99	17	0.0230	0.000	4.5730
44	2	100010000	.	.	-9999.99	36	0.0486	0.000	4.5243
45	3	100010010	.	.	-9999.99	2	0.0027	0.000	4.5216
46	3	100010100	.	.	-9999.99	8	0.0108	0.000	4.5108
47	4	100010101	.	.	-9999.99	1	0.0014	0.000	4.5095
48	4	100010110	.	.	-9999.99	2	0.0027	0.000	4.5068
49	3	100011000	.	.	-9999.99	9	0.0122	0.000	4.4946
50	4	100011010	.	.	-9999.99	1	0.0014	0.000	4.4932
51	4	100011100	.	.	-9999.99	1	0.0014	0.000	4.4919
52	2	101000000	.	.	-9999.99	280	0.3784	0.000	4.1135
53	3	101000001	.	.	-9999.99	11	0.0149	0.000	4.0986
54	3	101000010	.	.	-9999.99	23	0.0311	0.000	4.0676
55	4	101000011	.	.	-9999.99	1	0.0014	0.000	4.0662
56	3	101000100	.	.	-9999.99	68	0.0919	0.000	3.9743
57	4	101000101	.	.	-9999.99	2	0.0027	0.000	3.9716
58	4	101000110	.	.	-9999.99	5	0.0068	0.000	3.9649
59	2	110000000	.	.	-9999.99	2164	2.9243	0.000	1.0405
60	3	110000001	.	.	-9999.99	63	0.0919	0.000	0.9486

Constructs for Phase II of Probabilistic Linkage (continued)

61	3	110000010	.	.	-9999.99	201	0.2716	0.000	0.6770
62	4	110000011	.	.	-9999.99	7	0.0095	0.000	0.6676
63	3	110000100	.	.	-9999.99	410	0.5541	0.000	0.1135
64	4	110000101	.	.	-9999.99	16	0.0216	0.000	0.0919
65	4	110000110	.	.	-9999.99	31	0.0419	0.000	0.0500
66	3	110010000	.	.	-9999.99	4	0.0054	0.000	0.0446
67	4	110010010	.	.	-9999.99	1	0.0014	0.000	0.0432
68	3	111000000	.	.	-9999.99	26	0.0351	0.000	0.0081
69	4	111000010	.	.	-9999.99	1	0.0014	0.000	0.0068
70	4	111000100	.	.	-9999.99	2	0.0027	0.000	0.0041
71	5	111000110	13	0.6092	3.57	.	.	0.609	0.0041
72	5	111000101	32	1.4995	4.12	.	.	2.109	0.0041
73	5	111000011	4	0.1874	4.99	.	.	2.296	0.0041
74	5	111010100	2	0.0937	7.42	.	.	2.390	0.0041
75	5	110000111	608	28.4911	8.43	2	0.0027	30.381	0.0014
76	6	111000111	1	0.0469	9.76	.	.	30.328	0.0014
77	5	110010110	254	11.9025	11.73	.	.	42.630	0.0014
78	5	110010101	96	4.4986	12.28	.	.	47.329	0.0014
79	5	110010011	59	2.7648	13.15	.	.	50.094	0.0014
80	5	111011000	1	0.0469	13.86	.	.	50.141	0.0014
81	5	110011100	343	16.0731	17.30	1	0.0014	66.214	-0.0000

Constructs for Phase II of Probabilistic Linkage (continued)

REVIEW:

82	6	110010111	12	0.5623	17.93	.	.	66.776	-0.0000
83	5	110011010	323	15.1359	18.17	.	.	81.912	-0.0000
84	5	110011001	164	7.6851	18.72	.	.	89.597	-0.0000
85	6	110011110	87	4.0769	22.94	.	.	93.674	-0.0000
86	6	110011101	57	2.6710	23.50	.	.	96.345	-0.0000
87	6	110011011	36	1.6870	24.37	.	.	98.032	-0.0000
88	7	110011111	5	0.2343	29.14	.	.	98.266	-0.0000

HIGH (EXCLUDE from review; INCLUDE in dataset)

89	6	110110110	1	0.0469	32.31	.	.	98.313	-0.0000
90	5	110111000	1	0.0469	33.11	.	.	98.360	-0.0000
91	6	110111100	4	0.1874	37.88	.	.	98.547	-0.0000
92	6	110111010	2	0.0937	38.75	.	.	98.641	-0.0000
93	7	110111110	6	0.2812	43.52	.	.	98.922	-0.0000
94	7	110111011	5	0.2343	44.94	.	.	99.157	-0.0000
95	8	110111111	18	0.8435	49.72	.	.	100.000	-0.0000

**APPENDIX F. PERSON-YEARS AT RISK FOR THE CHOLECYSTECTOMY
COHORT (5 YEARS INDUCTION): INDIVIDUALS IDENTIFIED BY EITHER
TYPE OF LINKAGE**

Person-Year Matrix for Cohort: Males Only, 5 years induction
(Person-Years at risk and number of contributing persons by agegroup and sex)

AGE YEAR	0 - 9	10 - 14	15 - 19	20 - 24	25 - 29	30 - 34	35 - 39	40 - 44	45 - 49	50 - 54	55 - 59	60 - 64	65 - 69	70 - 74	75 - 79	80+	Year Total
1978	0.8 (3)	0.6 (1)	0.9 (3)	2.0 (7)	9.9 (27)	14.7 (47)	24.2 (72)	34.2 (91)	39.8 (114)	44.7 (126)	48.3 (126)	44.7 (123)	36.5 (99)	25.5 (70)	19.2 (60)	8.7 (26)	354.6 (995)
1979	2.2 (4)	2.3 (3)	3.9 (5)	9.2 (17)	37.2 (56)	77.5 (125)	101.0 (166)	138.5 (238)	183.5 (290)	180.4 (292)	217.8 (334)	190.6 (287)	157.5 (243)	111.8 (176)	80.6 (137)	47.4 (74)	1541.2 (2447)
1980	4.0 (4)	2.1 (6)	5.7 (7)	15.8 (25)	47.8 (68)	126.0 (179)	169.9 (242)	242.7 (335)	302.7 (426)	320.4 (469)	363.7 (504)	297.3 (409)	265.0 (366)	205.3 (297)	151.9 (214)	88.8 (129)	2608.8 (3680)
1981	3.9 (4)	5.1 (6)	4.1 (7)	19.9 (24)	56.2 (75)	165.5 (225)	230.7 (307)	318.6 (425)	383.4 (523)	459.8 (621)	489.7 (646)	392.7 (536)	382.6 (497)	293.8 (396)	212.3 (285)	146.9 (189)	3544.2 (4788)
1982	3.4 (6)	6.7 (9)	4.7 (8)	15.1 (25)	57.1 (81)	174.1 (229)	280.4 (375)	379.5 (501)	456.6 (617)	606.3 (777)	594.5 (775)	536.9 (707)	493.2 (649)	386.9 (515)	266.3 (370)	212.9 (271)	4474.6 (5915)
1983	4.2 (7)	8.7 (12)	6.2 (9)	14.4 (22)	60.8 (81)	184.1 (245)	328.2 (421)	428.6 (552)	532.0 (671)	726.2 (906)	697.5 (914)	679.8 (846)	603.3 (762)	469.4 (622)	344.1 (456)	278.6 (356)	5366.1 (6894)
1984	4.2 (5)	8.4 (13)	7.9 (12)	13.7 (20)	65.7 (88)	188.0 (245)	350.6 (458)	483.5 (621)	588.4 (762)	779.8 (993)	840.3 (1068)	793.6 (1011)	667.5 (864)	553.5 (714)	398.5 (514)	350.3 (422)	6093.9 (7810)
1985	3.0 (3)	8.7 (10)	10.1 (12)	11.4 (16)	69.3 (91)	182.3 (238)	380.2 (480)	513.6 (648)	650.2 (807)	840.9 (1061)	893.5 (1166)	911.3 (1146)	764.0 (978)	644.1 (824)	449.6 (585)	407.7 (487)	6740.0 (8552)
1986	3.5 (6)	7.2 (11)	11.4 (14)	14.3 (17)	63.8 (86)	186.5 (246)	397.1 (520)	559.2 (727)	687.7 (882)	891.1 (1118)	990.4 (1250)	984.8 (1226)	842.2 (1080)	750.1 (901)	518.2 (657)	464.8 (554)	7372.2 (9314)
1987	2.8 (4)	11.4 (13)	12.6 (17)	16.1 (21)	65.5 (89)	193.6 (248)	407.6 (522)	596.1 (766)	749.6 (939)	921.1 (1163)	1087.4 (1366)	1053.6 (1322)	938.8 (1190)	848.0 (1086)	570.9 (735)	526.0 (629)	8001.1 (10090)
1988	1.7 (2)	11.2 (15)	17.7 (23)	12.1 (22)	63.1 (85)	190.0 (262)	425.9 (545)	648.7 (806)	785.5 (980)	957.6 (1185)	1185.3 (1485)	1135.4 (1428)	1054.1 (1310)	923.1 (1158)	654.0 (848)	619.1 (731)	8684.4 (10907)
1989	1.0 (1)	10.6 (12)	13.8 (20)	17.3 (24)	54.1 (70)	192.0 (262)	428.0 (555)	696.9 (902)	845.0 (1090)	1003.7 (1268)	1236.1 (1556)	1272.8 (1584)	1177.9 (1474)	990.9 (1253)	787.7 (983)	732.2 (853)	9459.9 (11907)
1990	1.0 (1)	10.2 (11)	13.8 (18)	20.6 (27)	50.3 (70)	202.5 (264)	422.1 (541)	746.3 (917)	914.0 (1130)	1063.4 (1318)	1333.3 (1640)	1358.8 (1731)	1319.6 (1646)	1105.6 (1390)	906.3 (1135)	881.0 (1027)	10348.6 (12866)
1991	0.3 (0)	10.0 (11)	15.3 (24)	27.1 (34)	49.9 (73)	208.3 (270)	446.1 (579)	747.8 (965)	996.7 (1282)	1115.2 (1420)	1405.3 (1757)	1478.0 (1842)	1422.1 (1753)	1238.3 (1584)	1012.6 (1255)	1033.6 (1168)	11206.6 (14028)
1992	0.9 (1)	3.4 (8)	22.7 (25)	31.6 (40)	56.7 (83)	199.2 (274)	470.9 (611)	780.4 (998)	1055.7 (1331)	1182.6 (1483)	1452.3 (1794)	1606.6 (1993)	1539.2 (1920)	1397.2 (1746)	1148.5 (1412)	1199.7 (1501)	12147.7 (15081)
1993	1.6 (2)	4.7 (6)	20.0 (26)	39.1 (51)	52.4 (80)	206.8 (275)	452.4 (611)	818.1 (1040)	1118.9 (1402)	1250.8 (1567)	1794.3 (1821)	1721.4 (2115)	1633.3 (2051)	1561.3 (1926)	1235.9 (1536)	1393.2 (1603)	12969.3 (16112)
TOTAL	38.4 (54)	110.9 (147)	170.8 (230)	279.6 (392)	859.8 (1209)	2691.2 (3634)	5335.3 (7005)	8132.6 (10532)	10299.7 (13252)	12343.8 (15789)	14294.6 (18202)	14458.4 (18328)	13295.7 (16912)	11505.0 (14637)	8756.5 (11182)	8390.9 (9881)	110933.2 (141396)

Person-Year Matrix for Cohort: Females Only, 5 years induction																		
(Person-years at risk and number of contributing persons by agegroup and sex)																		
AGE	0 - 9	10 - 14	15 - 19	20 - 24	25 - 29	30 - 34	35 - 39	40 - 44	45 - 49	50 - 54	55 - 59	60 - 64	65 - 69	70 - 74	75 - 79	80+	Year Total	
1978	0.5 (2)	0.9 (5)	3.5 (11)	38.0 (106)	127.6 (336)	128.0 (365)	110.9 (298)	95.1 (250)	99.2 (259)	108.0 (299)	111.0 (310)	96.2 (270)	79.9 (216)	55.3 (154)	39.0 (100)	13.9 (40)	1106.9 (3021)	
1979	3.5 (4)	7.1 (9)	11.6 (20)	126.0 (222)	491.1 (759)	571.8 (877)	465.8 (711)	396.5 (613)	424.5 (661)	431.8 (700)	523.5 (795)	412.1 (795)	333.4 (533)	249.2 (381)	160.5 (245)	75.6 (111)	4683.9 (7284)	
1980	5.2 (9)	8.9 (13)	9.0 (20)	162.5 (275)	765.5 (1065)	951.8 (1320)	830.3 (1155)	697.7 (982)	691.4 (978)	788.3 (1101)	848.7 (1201)	722.9 (1000)	611.1 (866)	425.4 (614)	265.6 (382)	149.1 (202)	7033.6 (11183)	
1981	7.2 (10)	11.3 (15)	11.6 (19)	174.0 (271)	974.7 (1300)	1323.9 (1754)	1148.8 (1535)	975.8 (1298)	936.2 (1229)	1033.1 (1365)	1166.0 (1540)	1000.2 (1325)	866.9 (1122)	608.2 (813)	399.2 (536)	244.4 (309)	10881.4 (14441)	
1982	8.6 (9)	11.2 (16)	14.8 (21)	163.7 (255)	1092.5 (1410)	1551.6 (2046)	1499.0 (1961)	1235.6 (1615)	1192.8 (1545)	1257.0 (1618)	1429.9 (1825)	1295.5 (1650)	1058.8 (1379)	798.6 (1045)	543.4 (681)	348.5 (423)	13501.8 (17499)	
1983	7.8 (9)	12.2 (15)	15.4 (20)	154.3 (244)	1140.7 (1492)	1771.3 (2300)	1829.6 (2343)	1520.8 (1976)	1392.2 (1767)	1488.7 (1853)	1610.2 (2045)	1587.8 (2039)	1264.2 (1605)	1000.4 (1273)	683.3 (852)	457.6 (544)	15916.5 (20407)	
1984	3.1 (7)	14.7 (18)	17.9 (25)	142.1 (213)	1102.5 (1478)	1999.2 (2565)	2102.1 (2611)	1778.7 (2232)	1603.9 (2018)	1666.9 (2124)	1743.0 (2258)	1913.2 (2346)	1455.3 (1846)	1177.3 (1486)	786.6 (973)	574.3 (682)	18081.0 (22912)	
1985	3.4 (5)	12.7 (16)	20.3 (28)	131.4 (234)	1058.1 (1411)	2178.3 (2760)	2419.6 (3022)	2007.9 (2506)	1786.4 (2243)	1800.1 (2279)	1960.1 (2461)	2055.1 (2594)	1695.5 (2150)	1377.4 (1715)	898.3 (1137)	697.6 (821)	20102.2 (25392)	
1986	5.0 (5)	10.8 (12)	20.8 (29)	143.3 (224)	1020.5 (1370)	2344.6 (2925)	2699.7 (3327)	2254.9 (2860)	1993.7 (2532)	1939.8 (2419)	2126.2 (2674)	2236.5 (2752)	1914.3 (2369)	1547.9 (1895)	1023.1 (1283)	854.8 (1024)	22135.9 (27700)	
1987	6.1 (8)	10.6 (13)	19.3 (26)	140.4 (227)	1001.7 (1372)	2421.1 (3053)	2830.3 (3636)	2607.8 (3311)	2244.3 (2809)	2094.0 (2626)	2250.5 (2806)	2394.1 (2972)	2133.1 (2654)	1710.6 (2127)	1156.8 (1466)	1040.5 (1201)	24061.4 (30307)	
1988	6.3 (8)	10.5 (15)	26.8 (38)	142.5 (216)	1009.9 (1368)	2489.2 (3139)	3062.6 (3845)	2979.1 (3720)	2523.4 (3187)	2287.9 (2859)	2484.7 (3032)	2509.8 (3134)	2399.4 (3015)	1867.7 (2338)	1356.9 (1711)	1239.5 (1430)	26366.3 (33055)	
1989	3.2 (4)	9.7 (14)	26.8 (35)	150.6 (221)	1004.5 (1360)	2496.7 (3189)	3232.2 (4161)	3298.7 (4079)	2810.7 (3502)	2513.1 (3136)	2610.6 (3253)	2620.2 (3303)	2696.8 (3309)	2019.3 (2542)	1565.2 (1931)	1416.5 (1624)	28565.8 (35463)	
1990	3.1 (6)	10.4 (11)	25.1 (34)	138.3 (209)	995.1 (1344)	2501.3 (3173)	3555.1 (4446)	3705.8 (4559)	3089.4 (3811)	2720.8 (3407)	2755.0 (3448)	2799.7 (3503)	2867.8 (3559)	2281.1 (2846)	1758.5 (2159)	1608.7 (1878)	30815.3 (38393)	
1991	4.6 (5)	9.8 (11)	23.2 (35)	142.2 (211)	973.2 (1316)	2485.5 (3151)	3749.3 (4625)	4036.2 (4938)	3365.1 (4228)	3002.1 (3772)	2919.5 (3613)	2985.8 (3741)	3059.7 (3749)	2502.9 (3115)	1911.1 (2361)	1882.0 (2158)	33062.1 (41035)	
1992	3.0 (4)	11.7 (13)	26.2 (36)	165.2 (262)	955.4 (1329)	2555.1 (3267)	3887.0 (4861)	4213.5 (5275)	3779.9 (4759)	3381.0 (4161)	3094.8 (3863)	3175.2 (3937)	3219.1 (3990)	2801.0 (3446)	2068.4 (2544)	2165.2 (2497)	35471.9 (44241)	
1993	2.9 (5)	8.2 (12)	26.9 (36)	184.1 (275)	970.0 (1262)	2585.5 (3292)	4013.3 (5021)	4423.7 (5541)	4208.3 (5201)	3637.6 (4545)	3325.8 (4111)	3357.5 (4133)	3317.4 (4116)	3103.7 (3834)	2159.2 (2734)	2447.4 (3023)	37731.5 (46820)	
TOTAL	73.6 (100)	160.7 (208)	298.2 (433)	2298.7 (3665)	14632.9 (19992)	30355.0 (39176)	37526.8 (47578)	36227.9 (45755)	32141.2 (40729)	30120.2 (38304)	30929.7 (39235)	31171.8 (39338)	28972.7 (36478)	23526.2 (29624)	16765.2 (21095)	15215.5 (17746)	330417.3 (419456)	

**APPENDIX G. COLON CANCER RATES IN THE ALBERTA POPULATION,
1969-1993**

Note: Population figures from the Electronic Data Dissemination Division, Statistics Canada. Ottawa, Ont:STATSCAN, CANSIM disc, 1993.

Colon Cancer Rate Matrix for Albertans: Males Only		(Number of cases, population, and rates by agegroup and year)																Year Total
AGE		0 - 9	10 - 14	15 - 19	20 - 24	25 - 29	30 - 34	35 - 39	40 - 44	45 - 49	50 - 54	55 - 59	60 - 64	65 - 69	70 - 74	75 - 79	80 +	
YEAR	Popn	Rate	Popn	Rate	Popn	Rate	Popn	Rate	Popn	Rate	Popn	Rate	Popn	Rate	Popn	Rate	Popn	Rate
1969	174100	0.0	89200	0.0	75900	0.0	62600	0.0	52200	0.0	49000	0.0	50500	0.0	47800	0.0	41900	0.0
1970	171800	0.0	91600	0.0	79100	0.0	66800	0.0	56000	0.0	49700	0.0	51000	0.0	48800	0.0	43100	0.0
1971	169700	0.0	93400	0.0	82000	0.0	70600	0.0	59800	0.0	50700	0.0	51300	0.0	49500	0.0	44300	0.0
1972	166300	0.0	94500	0.0	84600	0.0	73300	0.0	62700	0.0	52200	0.0	51300	0.0	50100	0.0	45400	0.0
1973	162700	0.0	96900	0.0	90100	0.0	75700	0.0	67800	0.0	55100	0.0	50300	0.0	50900	0.0	45700	0.0
1974	158900	0.0	97700	0.0	93000	0.0	81000	0.0	71400	0.0	58300	0.0	50200	0.0	51300	0.0	46400	0.0
1975	159500	0.0	97300	0.0	96200	0.0	87800	0.0	77600	0.0	61800	0.0	51700	0.0	51600	0.0	48200	0.0
1976	161900	0.0	95600	0.0	99000	0.0	94600	0.0	84300	0.0	65700	0.0	53600	0.0	52400	0.0	49700	0.0
1977	165700	0.0	93900	0.0	100700	0.0	99200	0.0	88500	0.0	72800	0.0	56500	0.0	53400	0.0	51200	0.0
1978	168000	0.0	91000	0.0	104400	0.0	104800	0.0	93100	0.0	78300	0.0	60100	0.0	52900	0.0	51900	0.0
1979	173800	0.0	88600	0.0	106500	0.0	109700	0.0	100100	0.0	83800	0.0	65200	0.0	53500	0.0	52800	0.0
1980	180200	0.0	91900	0.0	110400	0.0	113800	0.0	103800	0.0	86900	0.0	67600	0.0	55500	0.0	54800	0.0
1981	186000	0.0	92100	0.0	109700	0.0	137000	0.0	126100	0.0	101500	0.0	74600	0.0	59600	0.0	55100	0.0
1982	191400	0.0	92500	0.0	108700	0.0	140000	0.0	135800	0.0	107100	0.0	82300	0.0	62800	0.0	55700	0.0

Colon Cancer Rate Matrix for Allentans: Males Only (continued)
(Number of cases, population, and rates by agegroup and year)

AGE	0 - 9	10 - 14	15 - 19	20 - 24	25 - 29	30 - 34	35 - 39	40 - 44	45 - 49	50 - 54	55 - 59	60 - 64	65 - 69	70 - 74	75 - 79	80+	Year Total
1983	0	1	0	0	1	2	2	6	10	22	24	33	37	45	36	47	266
Popn	193000	91500	105100	135900	140300	111100	87700	65700	55900	53500	45000	36900	28000	21000	15000	14000	1199600
Rate	0.0	1.1	0.0	0.0	0.7	1.8	2.3	9.1	17.9	41.1	53.3	89.4	132.1	214.3	240.0	335.7	22.2
1984	0	0	0	0	2	6	8	6	15	17	24	42	38	59	50	52	319
Popn	197900	89100	99300	125100	137300	112900	90400	67600	55900	53300	45800	38100	28100	21600	15000	14600	1192000
Rate	0.0	0.0	0.0	0.0	1.5	5.3	8.8	8.9	26.8	31.9	52.4	110.2	135.2	273.1	333.3	356.2	26.8
1985	0	0	0	0	1	2	2	5	8	20	27	49	64	45	51	61	335
Popn	200000	86700	96200	118500	132900	115800	93100	69100	56300	57700	47200	38300	28900	22400	15300	15000	1188400
Rate	0.0	0.0	0.0	0.0	0.8	1.7	2.1	7.2	14.2	38.0	57.2	127.9	221.5	200.9	333.3	406.7	28.2
1986	0	0	0	1	2	1	5	3	16	24	27	44	53	54	38	52	320
Popn	199400	88600	92900	112200	130000	118000	96800	71900	57600	53000	48400	39300	30400	23100	15500	15200	1192300
Rate	0.0	0.0	0.0	0.9	1.5	0.8	5.2	4.2	27.8	45.3	55.8	112.0	114.3	233.6	245.2	342.1	26.8
1987	0	0	0	0	2	3	4	6	11	20	39	43	39	46	47	54	314
Popn	202300	88900	93100	105300	126400	121000	97200	76000	58400	52000	48500	39500	31500	23800	16200	16900	1197000
Rate	0.0	0.0	0.0	0.0	1.6	2.5	4.1	7.9	18.8	38.5	80.4	108.9	123.8	193.3	290.1	319.5	26.2
1988	0	0	0	0	2	3	3	10	11	22	40	49	64	65	48	56	373
Popn	203200	89200	92300	102000	123200	123200	99400	79900	60600	52000	49100	40500	32700	24200	16500	17400	1205400
Rate	0.0	0.0	0.0	0.0	1.6	2.4	3.0	12.5	18.2	42.3	81.5	121.0	195.7	268.6	290.9	321.8	30.9
1989	0	0	0	0	0	1	5	6	17	20	38	53	45	72	57	61	375
Popn	204500	89900	92600	97700	119100	122600	103700	85400	64300	53300	50100	42000	34200	24500	17100	16600	1217600
Rate	0.0	0.0	0.0	0.0	0.0	0.8	4.8	7.0	26.4	37.5	75.8	126.2	131.6	293.9	333.3	367.5	30.8
1990	0	0	0	0	0	3	4	8	13	18	39	57	42	47	45	58	334
Popn	208100	91700	92300	97900	116400	123300	108900	90300	67300	54400	50100	43600	34800	25400	18000	17200	1239700
Rate	0.0	0.0	0.0	0.0	0.0	2.4	3.7	8.9	19.3	33.1	77.8	130.7	120.7	185.0	250.0	337.2	26.9
1991	0	0	2	1	2	2	7	8	11	18	31	48	45	63	39	74	351
Popn	211500	95800	90900	97000	117200	129700	116000	95400	70400	56100	50600	45400	36600	26900	18600	18500	1276600
Rate	0.0	0.0	2.2	1.0	1.7	1.5	6.0	8.4	15.6	32.1	61.3	105.7	123.0	234.2	209.7	400.0	27.5
1992	0	0	0	1	4	2	7	7	21	24	40	49	77	59	52	82	475
Popn	216600	99400	94500	101600	118700	133900	123300	99700	77600	58700	51100	46700	37600	28600	19500	19200	1326700
Rate	0.0	0.0	0.0	1.0	3.4	1.5	5.7	7.0	27.1	40.9	78.3	104.9	204.8	206.3	266.7	427.1	32.0
1993	0	0	1	1	1	1	2	7	13	23	31	41	49	66	53	69	358
Popn	216600	102000	96500	101000	114100	133000	127700	103200	82600	61300	51400	47400	38600	30000	20100	19900	1345400
Rate	0.0	0.0	1.0	1.0	0.9	0.8	1.6	6.8	15.7	37.5	60.3	86.5	126.9	220.0	263.7	346.7	26.6
TOTAL	0	3	6	13	24	52	90	142	267	402	621	811	999	1037	881	1143	6491
Popn	4643100	2319000	2386000	2511100	2554800	2317400	1960400	1644300	1393100	1205800	1051300	874800	696500	519500	352200	357200	26786500
Rate	0.0	0.1	0.3	0.5	0.9	2.2	4.6	8.6	19.2	33.3	59.1	92.7	143.4	199.6	240.1	320.0	24.2

Colorectal Cancer Rate Matrix for Alberta: Females Only (Number of cases, population, and rates by agegroup and year)																				
AGE YEAR	0 - 4	5 - 9	10 - 14	15 - 19	20 - 24	25 - 29	30 - 34	35 - 39	40 - 44	45 - 49	50 - 54	55 - 59	60 - 64	65 - 69	70 - 74	75 - 79	80+	Year Total		
1969	0	0	0	2	0	0	2	0	0	2	2	8	4	2	8	10	6	46		
Popn	167100	84800	73800	63400	52400	47000	46100	45300	45300	41700	35500	29900	23700	18700	14200	10700	11400	765700		
Rate	0.0	0.0	0.0	2.7	0.0	0.0	4.3	0.0	0.0	4.8	5.6	26.8	16.9	10.7	56.3	93.5	52.6	6.0		
1970	0	0	0	0	0	0	1	4	4	10	17	19	25	19	25	25	24	173		
Popn	165400	87100	76500	67700	55400	47800	46500	45400	45400	42900	36300	31200	24600	19400	14600	10900	12100	783800		
Rate	0.0	0.0	0.0	0.0	0.0	0.0	2.1	8.6	8.8	23.3	46.8	60.9	101.6	97.9	171.2	229.4	198.3	22.1		
1971	0	0	0	0	0	0	0	3	6	17	10	15	18	21	25	26	26	167		
Popn	162700	88700	78900	71700	58900	49200	46900	45400	45400	43800	37000	32200	25400	20200	15100	11100	12900	800100		
Rate	0.0	0.0	0.0	0.0	0.0	0.0	6.4	6.4	13.2	38.8	27.0	46.6	70.9	104.0	165.6	234.2	201.6	20.9		
1972	0	1	0	0	1	0	1	4	6	1	13	18	19	17	28	16	30	155		
Popn	159100	90000	80900	74300	62000	50700	47200	45600	45600	44500	38200	33100	26500	20800	15600	11300	13600	813400		
Rate	0.0	1.1	0.0	1.3	0.0	2.0	8.5	8.5	13.2	2.2	34.0	54.4	71.7	81.7	179.5	141.6	220.6	19.1		
1973	0	0	0	0	1	0	5	6	3	7	17	17	22	27	32	29	34	191		
Popn	154800	92700	86300	73800	67600	53100	47500	46400	46400	43800	40600	33500	27300	22000	16200	11500	14000	831100		
Rate	0.0	0.0	0.0	0.0	0.0	9.4	12.6	12.6	6.5	16.0	41.9	50.7	80.6	122.7	142.0	252.2	242.9	23.0		
1974	0	0	1	1	3	0	3	5	5	11	19	19	29	26	32	15	41	205		
Popn	151300	93500	89100	77800	70700	55800	48200	46800	46800	44000	42100	33800	28800	22600	17100	11900	14600	848100		
Rate	0.0	0.0	1.1	1.3	4.2	0.0	6.2	6.2	10.7	25.0	45.1	56.2	100.7	115.0	187.1	126.1	280.8	24.2		
1975	0	0	0	1	1	1	1	4	8	11	18	15	18	31	25	26	46	205		
Popn	151700	93300	92100	84300	75400	59600	49700	47300	47300	45100	43000	35200	30200	23300	18000	12700	15000	875900		
Rate	0.0	0.0	0.0	1.2	1.3	1.7	8.0	8.0	16.9	24.4	41.9	42.6	59.6	133.0	138.9	204.7	306.7	23.4		
1976	0	0	0	2	1	0	5	5	5	8	18	21	34	24	30	24	44	216		
Popn	154100	91600	94300	91400	80900	63100	51700	48100	48100	46100	44100	36800	31400	24600	18700	13300	15700	905900		
Rate	0.0	0.0	0.0	2.2	1.2	0.0	9.7	9.7	10.4	17.4	40.8	57.1	108.3	97.6	160.4	180.5	280.3	23.8		
1977	0	0	0	0	0	0	1	3	8	9	17	21	22	31	36	37	54	239		
Popn	157200	90000	95400	96100	85000	70100	54400	49700	49100	47000	44300	39300	32400	25800	19900	14000	16500	936500		
Rate	0.0	0.0	0.0	0.0	0.0	1.4	5.5	5.5	16.3	19.1	38.4	53.4	67.9	120.2	180.9	264.3	327.3	25.5		
1978	0	0	0	0	2	2	0	0	5	9	16	22	29	26	35	30	46	222		
Popn	159700	87100	99800	100800	88600	75500	57200	54000	49700	47600	44200	40800	33100	26600	20600	14600	16900	962800		
Rate	0.0	0.0	0.0	0.0	2.3	2.6	0.0	0.0	10.1	18.9	36.2	53.9	87.6	97.7	189.9	205.5	272.2	23.1		
1979	0	0	0	0	2	3	0	0	4	12	21	25	35	41	49	32	47	271		
Popn	165500	84900	102900	105200	94600	80600	60800	57200	50900	48500	44700	42400	33600	28200	21400	15400	17500	997100		
Rate	0.0	0.0	0.0	0.0	2.1	3.7	0.0	0.0	7.9	24.7	47.0	59.0	104.2	145.4	229.0	207.8	268.6	27.2		
1980	0	0	0	2	0	0	0	5	4	13	14	18	19	50	44	22	60	251		
Popn	170800	87600	106200	108600	97600	83200	62800	52500	52500	50000	46100	43800	34700	29100	22100	15400	18400	1029400		
Rate	0.0	0.0	0.0	1.8	0.0	0.0	8.0	8.0	7.6	26.0	30.4	41.1	54.8	111.8	194.1	138.4	176.1	24.4		
1981	0	0	0	0	1	1	1	2	10	9	18	23	28	41	34	35	49	251		
Popn	176200	87400	104700	127300	113700	94200	69500	55600	55600	50500	47200	43500	35700	30400	23100	16500	19100	1094600		
Rate	0.0	0.0	0.0	0.0	0.9	1.1	7.9	7.9	18.0	17.8	38.1	52.9	78.4	134.9	147.2	212.1	256.5	22.9		
1982	0	0	0	1	0	1	4	5	5	9	17	27	27	40	42	40	60	273		
Popn	181500	87700	102500	131300	122100	98100	77000	58200	58200	51400	48500	43900	38000	31300	23800	17500	20800	1132800		
Rate	0.0	0.0	0.0	0.8	0.0	1.0	5.2	5.2	8.6	17.5	35.1	61.5	71.1	127.8	176.5	228.6	300.0	24.1		

Colon Cancer Rate Matrix for Albertans: Females Only (continued)
(Number of cases, population, and rates by agegroup and year)

AGE	0 - 9	10 - 14	15 - 19	20 - 24	25 - 29	30 - 34	35 - 39	40 - 44	45 - 49	50 - 54	55 - 59	60 - 64	65 - 69	70 - 74	75 - 79	80+	Year Total
1983	0	0	0	0	0	4	5	9	11	13	26	29	41	43	44	74	299
Popn	183100	86900	98500	129100	126600	101000	82300	61100	52200	49200	44700	39800	31700	24700	18500	20900	1150300
Rate	0.0	0.0	0.0	0.0	0.0	4.0	6.1	14.7	21.1	26.4	58.2	72.9	129.3	174.1	237.8	354.1	26.0
1984	0	0	0	0	0	1	3	4	8	21	24	32	51	43	33	64	284
Popn	187400	84800	94000	122500	127500	104100	85800	63600	53000	49300	45000	41100	31900	26100	19000	22000	1157100
Rate	0.0	0.0	0.0	0.0	0.0	1.0	3.5	6.3	15.1	42.6	53.3	77.9	159.9	164.8	173.7	290.9	24.5
1985	0	0	0	0	2	4	2	8	16	19	29	41	41	41	48	82	333
Popn	189200	82500	91100	116100	126300	106800	89100	65900	53700	49300	45800	41500	33000	27300	19600	23300	1160500
Rate	0.0	0.0	0.0	0.0	1.6	3.7	2.2	12.1	29.8	38.5	63.3	98.8	124.2	150.2	244.9	351.9	28.7
1986	0	0	0	0	4	4	7	9	11	16	26	45	36	43	46	68	315
Popn	189300	84000	89300	112700	125800	111100	92700	68800	55000	49700	46100	42000	34800	28300	20300	23700	1173600
Rate	0.0	0.0	0.0	0.0	3.2	3.6	7.6	13.1	20.0	32.2	56.4	107.1	103.4	151.9	226.6	286.9	26.8
1987	0	0	0	0	0	3	3	4	11	16	23	30	40	44	46	78	298
Popn	192100	84500	89200	108400	124900	114100	92400	73100	55600	49200	45900	41500	36000	29100	21100	26200	1183300
Rate	0.0	0.0	0.0	0.0	0.0	2.6	3.2	5.5	19.8	32.5	50.1	72.3	111.1	151.2	218.0	297.7	25.2
1988	0	0	1	0	1	2	5	4	7	16	26	34	35	54	38	78	301
Popn	193000	84700	88300	102800	123000	116400	94400	77100	58200	49400	46400	41800	37300	29700	22000	27500	1192000
Rate	0.0	0.0	1.1	0.0	0.8	1.7	5.3	5.2	12.0	32.4	56.0	81.3	93.8	181.8	172.7	283.6	25.3
1989	0	0	0	0	1	2	2	7	11	11	25	42	44	49	58	90	338
Popn	194500	85200	87900	96100	119300	118200	99700	83000	61800	51400	47800	43000	39100	30100	23400	27600	1208200
Rate	0.0	0.0	0.0	0.0	0.8	1.7	2.0	8.4	11.3	21.4	52.2	97.7	112.5	162.8	247.9	326.1	28.0
1990	0	0	1	0	2	2	2	13	6	10	24	51	54	40	56	72	333
Popn	198200	86700	88200	95000	116900	120100	105000	87800	64700	52400	48100	44000	39700	31500	24200	29000	1232000
Rate	0.0	0.0	1.1	0.0	1.7	1.7	1.9	14.8	9.3	19.1	49.9	115.9	136.0	127.0	226.7	248.3	27.0
1991	0	1	0	1	2	4	8	8	12	9	24	33	45	47	50	95	339
Popn	202100	90700	87000	96400	117600	126100	110800	91800	67900	54000	48500	45000	41100	33000	25700	30700	1268400
Rate	0.0	1.1	0.0	1.0	1.7	3.2	7.2	8.7	17.7	16.7	49.5	73.3	109.5	142.4	194.6	309.4	26.7
1992	0	2	0	3	5	2	2	6	12	19	31	37	43	42	52	104	360
Popn	206100	94100	90300	99300	115000	128700	116600	94500	74200	56100	49200	45800	41400	35100	26600	32600	1305600
Rate	0.0	2.1	0.0	3.0	4.3	1.6	1.7	6.3	16.2	33.9	63.0	80.8	103.9	119.7	195.5	319.0	27.6
1993	0	0	0	0	3	2	4	10	18	15	18	29	50	56	50	84	339
Popn	206200	96800	91900	98300	109500	128800	120900	97900	78900	58900	49700	46500	41900	36700	27200	34500	1324600
Rate	0.0	0.0	0.0	0.0	2.7	1.6	3.3	10.2	22.8	25.5	36.2	62.4	119.3	152.6	183.8	243.5	25.6
TOTAL	0	4	5	13	30	48	86	155	247	382	544	732	876	938	888	1456	6404
Popn	4418300	2207300	2279100	2450400	2457300	2203400	1855200	1550900	1322100	1160700	1036700	897400	750900	592000	435400	515700	26132800
Rate	0.0	0.2	0.2	0.5	1.2	2.2	4.6	10.0	18.7	32.9	52.5	81.6	116.7	158.4	204.0	282.3	24.5