Using Administrative Data to Assemble a Cohort of Women Infected with Hepatitis C

by

Latifa Tse Fung Yeung, MD, FRCPC

A thesis submitted in conformity with the requirements for the Degree of Masters of Science (Clinical Epidemiology), Graduate Department of Health Policy, Management & Evaluation, University of Toronto

© Copyright Latifa T.F. Yeung 2001
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.

0-612-63070-6
ABSTRACT

Title: Using Administrative Data to Assemble a Cohort of Women Infected with Hepatitis C

Degree: Masters of Science (Clinical Epidemiology)

Year of Convocation: 2001

Name: Latifa Tse Fung Yeung

Graduate Department: Health Policy, Management & Evaluation

University: University of Toronto

Objectives:

1. To identify the optimal OHIP code for identifying women infected with HCV using ROC plot, Kappa statistic, and logistic regression.

2. To test combinations requiring more than one submission of a diagnostic code.

3. To test the combination of different codes using serial and parallel methods.

Study Design: Cross-sectional comparative.

Study Subjects: 2372 women

Primary Outcome Measure: ROC plot

Methods: OHIP code combinations (experimental tests) were linked to an abstracted clinic chart diagnosis for presence or absence of hepatitis C (reference standard). The ability of codes to identify infected women was evaluated using ROC plot, Kappa, and logistic regression.

Results: “Viral hepatitis” was the optimal code on ROC plot and predicted HCV infection in logistic regression. “Drug dependence/addiction” had the highest Kappa. Multiple code submissions and parallel codes improved specificity but lost sensitivity. Serial codes increased specificity but lost sensitivity.

Conclusions: “Viral hepatitis” best identified women infected with HCV.
ACKNOWLEDGMENTS

I would like to thank the Canadian Institutes of Health Research/Health Canada Hepatitis C Fellowship Program and the Hospital for Sick Children Trainee Startup Program for helping to fund this research study. I would also like to thank the Institute of Clinical Evaluative Sciences and the Paediatric Gastroentology and Nutrition Research Fellowship Program, University of Toronto for providing the necessary support to conduct this research.

I would like to thank the doctors who generously shared their clinic charts with me for this project. I am also grateful to Dr. Jack Williams who challenged me to embrace the obstacles of database research.

I am greatly indebted to my thesis supervisor, Dr. Teresa To, who provided instrumental advice and support.

I am especially grateful to my research supervisor and mentor, Dr. Eve Roberts. She began training me as a junior resident and offered guidance and feedback through many years. She is both an admirable role model and friend.

Deepest thanks also to Drs. Paul Dick, Jan Hux, and Elizabeth Lin whose course teachings in Health Care Research and Design introduced me to the world of administrative data. Furthermore, Dr. Paul Dick provided regular injections of enthusiasm and perspective to facilitate the pursuit and completion of this project. Many thanks to Dr. Susan King who continued to offer support and feedback while working at the Centre for Diseases Control.

Finally, and most importantly, I thank my parents, my in-laws, and my husband, for their loving support. My husband’s patience and sense of humour continue to be instrumental for me to grow as a person, as well as a researcher and physician.

I dedicate this work to my husband, Dr. Patrick Charn Fai Cheung.
Table of Contents

Chapter 1: INTRODUCTION .......................................................... 1

1.1 CLINICAL BACKGROUND .......................................................... 1
  1.1.1 Terminology of HCV Infection ............................................. 1
  1.1.2 Hepatitis C: Burden of Disease ............................................ 1
    1.1.2.1 Incidence of Acute Hepatitis C in Canada ..................... 2
    1.1.2.2 Prevalence of Chronic Hepatitis C in Canada ................ 2
    1.1.2.3 Other Measures of Prevalence .................................... 3
    1.1.2.4 Future Burden of Disease ......................................... 4
  1.1.3 Mother-to-Infant Transmission of HCV .............................. 4
  1.1.4 Fertility Rate of Women Infected with HCV ....................... 6

1.2 ADMINISTRATIVE DATABASES ................................................... 8
  1.2.1 Canadian Institute for Health Information ......................... 8
  1.2.2 Ontario Health Insurance Program .................................. 9
  1.2.3 Registered Persons Data Base ....................................... 9
  1.2.4 Notifiable Diseases Database ...................................... 10

1.3 USING ADMINISTRATIVE DATA TO STUDY DISEASE INCIDENCE AND
  PREVALENCE ........................................................................... 11

1.4 USING ADMINISTRATIVE DATA TO STUDY HEPATITIS C ............. 12

Chapter 2: STUDY HYPOTHESIS AND OBJECTIVES .................... 14

2.1 STUDY RATIONALE ................................................................. 14
  2.1.1 Sample Size Considerations ............................................ 15

2.2 STUDY HYPOTHESIS ............................................................... 16

2.3 STUDY OBJECTIVES ............................................................... 17

Chapter 3: METHODS ................................................................. 18

3.1 OVERVIEW OF STUDY METHODS ........................................... 18

3.2 STUDY DESIGN ................................................................. 20

3.3 STUDY SUBJECTS ................................................................. 20

3.4 REFERENCE STANDARD ....................................................... 20
  3.4.1 Definition and Rationale .............................................. 20
3.4.2 Choice of Physicians’ Clinics........................................................................... 21
3.4.3 Inclusion of Criteria for Reference Standard.................................................... 22
3.4.4 Chart Abstraction............................................................................................. 22
3.4.5 Description of Study Subjects.......................................................................... 23

3.5 EXPERIMENTAL TEST......................................................................................... 23
3.5.1 OHIP Code Selection....................................................................................... 23
  3.5.1.1 Diagnostic Codes......................................................................................... 24
  3.5.1.2 Procedure Codes......................................................................................... 25
3.5.2 Panel of Experimental Tests............................................................................. 26
3.5.3 Identifying All Patients with OHIP Test Codes............................................... 27
3.5.4 Incorporating Experimental Test Codes into the Sample Population................ 28

3.6 ANALYSIS: DETERMINING THE OPTIMAL EXPERIMENTAL TEST................. 29
3.6.1 Overview of Analysis....................................................................................... 29
3.6.2 Basic Test Characteristics.............................................................................. 30
3.6.3 Receiver Operating Characteristic (ROC) Plot.............................................. 32
3.6.4 Test Reliability: Cohen’s Kappa Statistic....................................................... 32
3.6.5 Logistic Regression......................................................................................... 32
3.6.6 Multiple Submissions of a Diagnostic Code.................................................... 33
3.6.7 Serial and Parallel Test Strategies................................................................. 34

3.7 FEASIBILITY STUDY.......................................................................................... 34

3.8 SAMPLE SIZE CALCULATION.......................................................................... 37

3.9 RESEARCH ETHICS........................................................................................... 39

Chapter 4: RESULTS................................................................................................. 40

4.1 REFERENCE STANDARD: DESCRIPTION OF STUDY SUBJECTS.................... 40
  4.1.1 Proportion of Clinic Charts Assessed for Inclusion......................................... 40
  4.1.2 Prevalence of Hepatitis C.............................................................................. 41
  4.1.3 Age Distribution............................................................................................ 41
  4.1.4 Death............................................................................................................. 41
  4.1.5 Multiple Comparisons.................................................................................. 42


4.3 ANALYSIS: DETERMINING THE OPTIMAL EXPERIMENTAL TEST............... 44
  4.3.1 Overview of Results....................................................................................... 44
  4.3.2 Single Codes (Binary Predictor Variable)....................................................... 46
    4.3.2.1 Basic Test Characteristics and ROC Plot................................................. 46
    4.3.2.1.1 Impact of Adding Procedure Code Liver Biopsy............................... 47
    4.3.2.2 Test Reliability: Cohen’s Kappa Statistic.............................................. 48
Chapter 5: DISCUSSION

5.1 SUMMARY OF RESULTS

5.2 RELEVANCE OF FINDINGS

5.3 STRENGTHS OF STUDY DESIGN

5.4 LIMITATIONS OF STUDY DESIGN

5.5 FUTURE DIRECTIONS
# List of Tables

1-1. Canadian Fertility Rate by Age Group and Year............................................ 7

3-1. Clinics chosen for Chart Abstraction................................................................. 22
3-2. Candidate Codes for Identifying Patients with Hepatitis C.............................. 24
3-3. Definitions of Experimental Test Codes............................................................ 26
3-4. Measures of Test Performance and Test Reliability........................................ 31
3-5. Number of Test Code Combinations – Fiscal 1996........................................ 36
3-6. Sample Size Determination – Variables............................................................ 37
3-7. Sample Size Determination – 25% Prevalence............................................... 38
3-8. Sample Size Determination – 60% Prevalence............................................... 39

4-1. Results of Chart Abstraction............................................................................... 41
4-2. Number of Test Code Combinations – 1995-2000.......................................... 43
4-3. Optimal OHIP Code Combinations Determined by ROC plot, Kappa statistic, and Logistic Regression.......................................................... 44
4-4. Test Characteristics (%) for Sites Toronto - 3°, London - 3°, Toronto - 2°........ 46
4-6. Goodness of Fit for Models with Single Codes as Binary and Discrete Variables... 49
4-7. Comparing Models E2 and D2 with and without Liver Biopsy....................... 50
4-8. Goodness of Fit for Models with Multiple Codes............................................... 51
4-9. Test Characteristics (%) – Code Submitted >1................................................ 53
4-10. Kappa statistics – Code Submitted >1.............................................................. 55
4-11. Test Characteristics (%) for Codes Combined by Serial and Parallel Methods..... 56

# List of Figures

3-1. Overall Strategy................................................................................................. 65
3-2. Number of Female Case of HCV Reported in Ontario by Year of Reporting........ 66

4-1.(a) Female Cases of HCV in Ontario by Age Group Reported to Notifiable Diseases... 67
4-1.(b) Number of Women with Hepatitis C in Sample Population by Decade of Birth..... 68
4-2. Sensitivity and Specificity of Codes..................................................................... 69
4-3. Sensitivity and Specificity of Codes (%) – Code Submitted >1.......................... 70
4-4. Sensitivity and Specificity of Codes Combined by Serial and Parallel Methods..... 71

# List of Appendices

Appendix 1 Glossary and Abbreviations................................................................. 72
Appendix 2 Addiction Research Foundation Methadone Clinic (Site Toronto – 1°)..... 74
Appendix 3 Ethics Approval Letters............................................................................ 75
Appendix 4 Feasibility Study: SAS Programming.................................................. 80
Appendix 5 Testing of Code Accuracy: SAS Programming...................................... 89
Chapter 1: Introduction

The objective of this chapter is to provide the background and rationale for estimating the fertility rate of women infected with hepatitis C virus (HCV) and its impact on mother-to-infant transmission of HCV. The chapter also introduces uses of administrative databases for this project and discusses the challenges of using administrative data to study the disease incidence and prevalence for hepatitis C.

1.1 Clinical Background

1.1.1 Terminology of HCV Infection

"Hepatitis C" indicates inflammation of the liver as a result of infection with hepatitis C virus (HCV). Hepatitis C may be classified as acute or chronic. By convention, “acute hepatitis” refers to recent infection of less than six months duration whereas “chronic hepatitis” refers to prolonged infection of greater than six months duration (1). As most cases of hepatitis C are initially asymptomatic, cases of acute hepatitis C often escape detection. In practice, the majority of identified cases of hepatitis C are considered chronic (2). For purposes of this study, no attempt will be made to distinguish between acute and chronic hepatitis C. Rather, “hepatitis C” will be used to represent past or current infection with HCV.

1.1.2 Hepatitis C: Burden of Disease

Hepatitis C virus was discovered in 1989 (3). Subsequently it was established that approximately 90% of “non-A non-B (NANB) post-transfusional hepatitis” were caused by HCV. The disease spectrum of hepatitis C has been derived from the experience of patients identified with HCV infection over the last decade and past understanding of NANB hepatitis (4-6).
1.1.2.1 Incidence of Acute Hepatitis C in Canada

The incidence of acute and chronic hepatitis C infection is difficult to quantify because many cases are initially asymptomatic and the date of acquiring infection is often not known. No laboratory test can differentiate the time HCV was first acquired (acute infection from remotely acquired infection). In October 1998 an enhanced surveillance project was initiated at four health regions (in Calgary, Winnipeg, Ottawa, and Edmonton) to make special effort to detect cases of acute HCV infection. From 1998 to 1999 the peak incidence rates of acute hepatitis C in these four regions were 8 per 100,000 men and 5.7 per 100,000 women (7).

1.1.2.2 Prevalence of Chronic Hepatitis C in Canada

Hepatitis C virus is largely acquired through blood-borne routes including blood transfusion and intravenous drug injection. Other routes of transmission include mother-to-infant transmission, transmission through contaminated needles (health care exposure), or through sexual or household contact (8). Available estimates of the prevalence of chronic HCV infection in Canada were derived from the incidence of new cases of hepatitis C acquired following blood transfusion, a situation where the date of acquiring infection is better known. Based on data from the number of transfusion-acquired hepatitis C, an estimated 240,000 Canadians are positive for anti-HCV antibodies and of these, 192,000 have chronic hepatitis C (9). Approximately one-third of all HCV-infected individuals are female (7, 10).

Although available Canadian prevalence estimates of chronic hepatitis C were based on data from transfusion-acquired hepatitis C, the demographics of this disease have since evolved. Since 1990, the incidence of transfusion-acquired hepatitis C has decreased to 1/103,000 due to the implementation of blood product screening (11). Consequently, from 1993 to 1995, the Laboratory Centre for Disease Control (LCDC) conducted a Sentinel Hepatitis Surveillance
Study. This study took place at eight health units across Canada and found that injection drug use is now the major risk factor for acquiring HCV infection (12-14). However, given the slowly progressive nature of HCV infection, new infections are unlikely to have an immediate impact on the disease burden. As a result, the fact that injection drug use has replaced blood transfusion as the main route of acquiring acute hepatitis C is unlikely to impact on the estimated prevalence of chronic hepatitis C in the near future (7).

1.1.2.3 Other Measures of Prevalence

Other attempts to measure the prevalence of hepatitis C (usually defined as positive for anti-HCV antibody) tend to involve selected patient groups or volunteer blood donors. The World Health Organization estimates that the prevalence of hepatitis C ranges from 0.003% (in Sweden) to 18.1% (in Egypt) (15). Among US hospitalized veterans, 11.8% - 35% proved positive for anti-HCV antibody and a history of intravenous drug use was a major risk factor (16, 17). By contrast, a 1990 survey of a single Canadian hospital found that 0.5% (95% confidence interval 0.1% - 0.9%) of all weekday admissions were positive for anti-HCV antibody (18). A French population-based survey of volunteer blood donors found that 1.15% of 61,283 donors were anti-HCV antibody positive (19). Comprehensive population-based determination of the prevalence of hepatitis C has been limited to a single population-based U.S. study using serum samples from the third National Health and Nutrition Examination Survey which studied 21,241 individuals six years old and older. This large-scale study found that the overall prevalence of anti-HCV antibody was 1.8%, corresponding to an estimated 3.9 million Americans with HCV infection. Seventy-four percent were positive for HCV RNA, indicating that an estimated 2.7 million Americans were chronically infected. Among subjects 17 to 59 years of age, the strongest factors independently associated with HCV infection were illegal drug use and high-
1.1.2.4 Future Burden of Disease

The significant burden of disease attributable to chronic infection with HCV has become increasingly apparent. Within the next decade, the prevalence of cases of cirrhosis due to hepatitis C is expected to increase by 92% and the prevalence of cases of end-stage liver disease and hepatocellular carcinoma related to hepatitis C is predicted to increase by 126% and 102% respectively. The prevalence of liver-related deaths associated with hepatitis C is predicted to increase by 126% (7). In North America, hepatitis C is already the leading indication for liver transplantation. The World Health Organization estimates that 170 million people have HCV infection worldwide (20). These results indicate the importance of controlling disease progression in those infected with hepatitis C as well as the need for primary prevention of HCV infection. The five-year Health Canada / CIHR Research Initiative on Hepatitis C initiative reflects recent government efforts to respond to this important burden of disease (21).

The burden of disease of hepatitis C in women and children is just beginning to be elucidated. Children tend to acquire hepatitis C from contaminated blood products or mother-to-infant transmission. Since the implementation of blood product screening for HCV in 1990, the proportion of prevalent children infected with HCV through blood transfusion has diminished (11, 22). Mother-to-infant transmission of HCV is now the increasingly important route of transmission for children.

1.1.3 Mother-to-Infant Transmission of HCV

Determining the rate of mother-to-infant transmission is important for counseling pregnant women infected with HCV and for predicting the burden of disease in pediatric populations. Small sample sizes, variations in the definition of mother-to-infant transmission, and heterogeneity in the reporting of risk factors hinder precise determination of this rate. Taken
altogether, available studies of mother-to-infant transmission provide a weighted rate of mother-to-infant transmission in the range of 1.0% to 5.0%. Among viremic (HCV RNA-positive) women, this rate increases to 3.1% to 6.9%. The rate appears higher in Japanese studies of viremic women. Maternal risk factors for increased mother-to-infant transmission include maternal viremia, co-infection with human immunodeficiency virus (HIV), and history of intravenous drug use (IVDU). By contrast, mode of delivery and breastfeeding do not appear to result in significantly different rates of mother-to-infant transmission. For details regarding the weighted rates of mother-to-infant transmission of HCV, please see reference #22. More precise measurements of the rate of mother-to-infant HCV transmission will require larger studies and better tests for HCV. No Canadian data exist regarding mother-to-infant transmission of HCV.

Although the rate of maternal-infant transmission is numerically small, its impact is substantial. Of the estimated 170 million people worldwide who are chronically infected with HCV, if 35% of these are females of child-bearing age, and assuming an annual fertility rate of 2%, conservative estimates suggest that 10,000-60,000 newborn babies will be infected with HCV each year (23). Indeed, mother-to-infant transmission of hepatitis C virus will be an important determinant of the burden of HCV disease in future generations.

The utility of antenatal screening of pregnant women for HCV infection is currently a subject of debate. Advocates for screening suggest that asymptomatic women who are identified with a chronic disease are likely to benefit from therapy with interferon and ribavirin (when they are not pregnant) and that such diagnosis occurs during a time when they are most receptive to medical intervention. Those identified with HCV infection may be counseled to avoid alcohol intake, offered hepatitis A and B vaccination, and reminded of body substance precaution practices. On the other hand, others are uncertain whether the benefits of screening to reduce
mortality outweigh the psychological harm caused by the screening procedure. They further argue that the relatively low rate of transmission and lack of treatment options for hepatitis C during pregnancy limit the benefit achieved by screening pregnant women (24-26).

In contrast to maternal screening, the screening of all infants born to women infected with HCV is recommended (27). Infants with HCV infection deserve ongoing monitoring and management. Although most cases of pediatric hepatitis C are slowly progressive (22), only careful monitoring can identify children with more severe disease who would benefit from early treatment. If children with fibrosis do not receive treatment, progressive cirrhosis would leave eventual liver transplantation as the only remaining option (28). Furthermore, treatment options for pediatric hepatitis C are expected to improve in the near future. Replacement of standard interferon-alfa with pegylated-interferon-alfa has increased the sustained treatment response rate from 20% to 45% (29, 30). The addition of ribavirin to standard interferon-alfa offers a treatment response rate of almost 50% (31). Thus it is anticipated that trial of pegylated-interferon-alfa combined with ribavirin will yield even more promising results in adults with hepatitis C. Pediatric trials will follow. Compared with adults, children appear to respond better to treatment due to their shorter duration of infection (32-34).

1.1.4 Fertility Rate of Women Infected with HCV

The arguments concerning mother-to-infant transmission assume that women infected with HCV have a fertility rate comparable to that of the general population. In Canada, the age-specific-fertility rate (defined as the annual number of live births per 1000 women aged 15-49 years) ranged from 0.2 to 103.88 per 1000 women in 1997 (Table 1-1) (35). The fertility rate of women infected with HCV has never been established.

Determination of the fertility rate of women infected with HCV has two important consequences. Firstly, the assumption that the fertility rate of women infected with HCV
matches that of the general population may be tested. If the fertility rate of women with hepatitis C is greater than that of the general population, then the estimated burden of disease resulting from maternal-infant transmission of HCV will be even larger than previously thought. On the other hand, if the fertility rate of women with hepatitis C is less than that of the general population, then the estimated burden of disease resulting from mother-to-infant transmission of HCV may actually be reduced. Secondly, measurement of fertility rate is instrumental during the provision of obstetrical counseling to women with hepatitis C.

<table>
<thead>
<tr>
<th>Age Group (years)</th>
<th>Year (July 1 - June 30)</th>
<th>1995</th>
<th>1996</th>
<th>1997</th>
</tr>
</thead>
<tbody>
<tr>
<td>15-19</td>
<td>24.5</td>
<td>22.3</td>
<td>20.2</td>
<td></td>
</tr>
<tr>
<td>20-24</td>
<td>70.5</td>
<td>67.3</td>
<td>64.1</td>
<td></td>
</tr>
<tr>
<td>25-29</td>
<td>109.7</td>
<td>105.8</td>
<td>103.9</td>
<td></td>
</tr>
<tr>
<td>30-34</td>
<td>86.8</td>
<td>85.5</td>
<td>84.4</td>
<td></td>
</tr>
<tr>
<td>35-39</td>
<td>31.3</td>
<td>32.2</td>
<td>32.5</td>
<td></td>
</tr>
<tr>
<td>40-44</td>
<td>4.8</td>
<td>5.1</td>
<td>5.2</td>
<td></td>
</tr>
<tr>
<td>45-49</td>
<td>0.2</td>
<td>0.2</td>
<td>0.2</td>
<td></td>
</tr>
</tbody>
</table>

The fertility rate of women infected with HCV has never been measured. While some speculate that the fertility rate of women infected with HCV is lower than the population fertility rate, others argue the converse. The use of intravenous drugs or cocaine (intravenous or use of "crack" cocaine) is a significant risk factor for acquiring HCV infection (2, 36). Women who use cocaine have an increased rate of spontaneous abortion (37-39). Extrapolation of these relationships suggests that women with hepatitis C are more likely to have spontaneous abortions. As a result, the fertility rate of women with infected with HCV may be lower than the population fertility rate. On the other hand, some women with infected with HCV may take part in high-risk sexual practices (higher sexual promiscuity without contraception) resulting in an increased pregnancy – and hence, fertility – rate.
Once pregnant, the obstetrical course of women with hepatitis C resembles that of the general population. Women with other forms of liver disease, such as non-cirrhotic portal hypertension, have fertility rates and pregnancy outcomes comparable to controls (40, 41). Furthermore, the child-bearing years of most women usually precede the time frame for progression to advanced liver disease from hepatitis C. It is only during the later, advanced stages of liver disease that amenorrhea and anovulation tend to occur and compromise fertility (42).

The direct impact of hepatitis C on pregnancy has also been examined. In an Irish cohort of rhesus-negative women who became infected with HCV from contaminated anti-D immunoglobulin, 36 women were studied for pregnancy outcomes. When compared with controls, women infected with HCV showed neither increase in spontaneous miscarriage rate nor significant differences in obstetric complications or birth weights (43). Other studies have also failed to identify increased complications of pregnancy as a result of maternal HCV infection (44, 45), although a higher incidence of cholestasis of pregnancy has been reported among women positive for anti-HCV antibodies compared with those who were negative for anti-HCV antibody (46). From the perspective of liver disease, small case-control studies suggest that serum ALT decreases, HCV RNA titres increase, and liver histology worsens as a result of pregnancy in women infected with hepatitis C (47, 48). Taken together, no study of pregnant women with hepatitis C reports significant complications in pregnancy.

1.2 Administrative Databases

1.2.1 Canadian Institute for Health Information

In Ontario, Canada, all discharges from acute care facilities are recorded in the Canadian Institute for Health Information (CIHI) database. This population-based, non-profit, non-government, federally-chartered company was established to collect information on health for
purposes of health-care planning, resource allocation, utilization management, and health services research. The accuracy of CIHI data compared to the hospital record has been assessed. While demographic and procedure information were highly accurate, coding of in hospital comorbidities and complications had high false negative rates (63% - 70%) (49).

1.2.2 Ontario Health Insurance Program

The Ontario Health Insurance Program (OHIP) database contains administrative information generated from physician billings. It is estimated that 95% of all physicians in Canada are reimbursed in a fee-for-service manner. This database does not capture health care provided by physicians on salary, physicians on Alternate Funding Plans, and services paid for by patients, Workers’ compensation, or other third party payers. Records with unique health care numbers are available from 1991 onwards. The accuracy of OHIP data is maintained with regular audits. The demographic data (patient age, sex, and place of residence) on physician’s claims are 99% complete (50-53).

The accuracy of OHIP data is considered poorer than hospital discharge data with respect to procedure codes and even worse with regards to diagnosis codes because completion of this field is optional. In 1989, the accuracy of diagnostic codes compared to hospital chart data ranged from 37%-81% in a study by the Ontario Hospital Association of 3000 records from 43 hospitals. Procedure codes accuracy was 88%-95% in the same study. By comparison, a 1992 study abstracted 300 records at Doctors Hospital and found 95%-96% accuracy for diagnosis codes and 96% for procedure codes. However, studies comparing OHIP diagnosis data with patient charts have all been hospital-based (54-56). No studies have specifically assessed the accuracy of OHIP diagnostic codes when compared to charts in physicians’ office.

1.2.3 Registered Persons Data Base

The Registered Persons Data Base (RPDB) contains data on all individuals registered
within the OHIP plan. OHIP coverage is available to individuals who are Canadian citizens, landed immigrants, or refugees who make Ontario their permanent and principal home and are present in Ontario for at least 153 days in any 12-month period. Tourists, transients, or visitors are not eligible for OHIP coverage. No premium is required for Ontario health coverage; federal and provincial taxes fund health care. Each individual has a unique health care number (HCN).

Using patient HCN as the unique identifier, patient gender, date of birth, and death date (if applicable) are recorded. This database dates back to April 1, 1990. Prior to that date, Ontario assigned an OHIP number that was shared by all members of a family.

1.2.4 Notifiable Diseases Database

The Notifiable Diseases Database is maintained by the Population and Public Health Branch of Health Canada. It contains surveillance data on notifiable disease in Canada, including hepatitis C since 1991. Each disease under national surveillance has a case definition. A confirmed case of hepatitis C requires laboratory confirmation of infection with (1) HCV RNA PCR positive regardless of results of anti-HCV antibody testing if >1 year of age, (2) anti-HCV antibody positive, or (3) HCV RNA positive if anti-HCV antibody negative. The provinces and territories have mandated reporting of notifiable diseases. When a provincial or commercial laboratory finds a positive anti-HCV antibody result, it notifies the local health authority. Subsequently the local health authority notifies the provincial health authority. If the provincial health authority determines that a positive anti-HCV antibody result meets the case definition of a confirmed case, aggregated data is then provided to the national database. This database does not capture individuals infected with HCV who have never been tested (i.e. unaware of infection) and Canadians who tested positive for anti-HCV antibody outside of Canada. The main limitation of this database is the failure to determine the date of acquiring infection, since the first positive test result may occur years following the time of initial infection.
1.3 **Using Administrative Data To Study Disease Incidence and Prevalence**

Large computerized databases have become an integral feature of health care delivery and administration. Although primarily intended for administrative purposes, these databases are proving increasingly useful to researchers for several reasons. Firstly, administrative data offer a breadth of information regarding events that involve the utilization of health care, including medical office visits, surgical procedures, and hospitalizations. Using predefined codes, hospital-based procedures such as liver biopsy and childbirth are readily identified. Secondly, unlike research using primary data collection, administrative data does not rely upon individual subjects to consciously participate in the research process. As a result, research using administrative data permits the inclusion of a large and diverse population and these data tend to be highly generalizable. Thirdly, linkage of administrative data systems allows ongoing accrual over long periods without the need to contact individuals, thereby yielding continuous longitudinal follow-up with a relatively low drop-out rate. Such population-based research often enables the capture of rare events. Fourthly, research using administrative data does not involve patients directly and is therefore less prone to the Hawthorne effect (subjects may systematically alter their behavior when they know that they are being observed). Fifth, administrative data is readily available, comprehensive, and rapidly accessible. As data have already been collected, analysis may occur with little delay. Finally, compared with the cost of conducting clinical trials, research using administrative data is inexpensive to analyze. (57-60).

For some purposes of research, the accuracy of coding for inpatient administrative data is deemed reasonable (61). Although comparison of hospital records and Canadian Institute for Health Information (CIHI) data highlighted a tendency for under-coding of co-morbidities and in-hospital complications (49), such inaccuracy still compared favorably to chart review (62). In
the United States, diagnostic codes in hospital charts have been studied for their accuracy in predicting complications of diabetes and found to have modest sensitivity (73% - 95%) and positive predictive value (80% - 88%) compared with chart data (57).

1.4 **Using Administrative Data To Study Hepatitis C**

Although administrative data offers multiple advantages, research in HCV disease using administrative data has several limitations. In Canada, *ambulatory* diagnostic codes are considered to be inaccurate and unreliable when compared to inpatient hospital charts (54). Such standards are extrapolated from abstraction of inpatient hospital charts; ambulatory clinic charts have never been evaluated for accuracy in diagnostic coding. Secondly, no specific diagnostic code exists for hepatitis C. In the Ontario Health Insurance Program (OHIP) administrative database, hepatitis A, B, and C share the ICD-9 diagnostic code “070 viral hepatitis”. Thirdly, although the quality of inpatient (CIHI discharge data) is considered more reliable than ambulatory data, it will only capture a limited spectrum of patients with hepatitis C; HCV infection seldom requires hospitalization until end-stage liver disease occurs. Fourthly, laboratory testing for HCV infection (for example, serum antibody to hepatitis C virus, anti-HCV antibody) cannot be linked to OHIP administrative data because testing is not uniformly billed to OHIP. Tests performed within hospitals do not generate OHIP billings. While commercial laboratories submit OHIP billings, their billing codes are not specific for hepatitis C. Finally, results of laboratory testing are not accessible due to rules of patient confidentiality. The Health Canada Notifiable Diseases database receives reports of new cases of HCV through mandatory reporting at the level of the provincial laboratory. Data collection is only in the form of aggregated data, and confidentiality laws prevent follow-up of individual patients.

Fischer et al. used administrative data to diagnose hepatitis C. This U.S. study developed
a risk algorithm for diagnosing HCV infection based on lifestyle factors and specific medical diagnoses. Serological evidence of HCV infection obtained during an exercise of anonymous HCV screening was compared against patient lifestyle factors and medical diagnoses data from the internal database of a Health Maintenance Organization. The strong correlation between risk factors and the diagnosis of HCV infection suggested that a database risk algorithm for screening for hepatitis C was a useful approach. However, this study was possible because the internal database possessed a unique diagnostic code for hepatitis A, B, and C, whereas Canadian administrative databases cannot differentiate between diagnoses of hepatitis A, B, and C (63).

In order to use administrative data to study hepatitis C, particularly in mother-to-infant transmission, a valid cohort of HCV-infected individuals must be created. If such a cohort can be assembled successfully from OHIP billing data, then it may be linked to the CIHI database. By calculating the number of live births from this cohort, the fertility rate of women with hepatitis C can be measured.
Chapter 2: Study Hypothesis and Objectives

The objective of this chapter is to explain the rationale for the study, to state the study hypothesis, and to list the objectives of the study.

2.1 **Study Rationale**

Maternal-infant transmission of HCV will soon become the major route of transmission of HCV for children with chronic hepatitis C. The incidence of perinatal HCV transmission must be estimated to enable appropriate counseling of child-bearing women who are infected with HCV and to predict the burden of illness of HCV on future generations. Current estimates of maternal-infant transmission of HCV assume that women infected with HCV have a fertility rate comparable to that of the general population. This assumption has never been proven. Rather, women infected with HCV may have a fertility rate different from the general population. If women infected with HCV have a fertility rate that is higher than that of the general population, then the estimated number of children infected by maternal-infant HCV transmission may be greater than previously thought. On the other hand, if women infected with HCV have a fertility rate that is lower than that of the general population, then the estimates of maternal-infant HCV transmission may be less than previously thought and the anticipated burden of illness of HCV on future generations may be reduced. This burden of illness from HCV must be properly characterized in order for strategic allocation of health care resources for people infected with HCV.

Measurement of the fertility rate of women infected with HCV would require that a large number of infected women be identified, enrolled, and followed prospectively through their child-bearing years. Such a study design is subject to sampling bias because women would likely be sampled from tertiary care facilities, where recruitment would be relatively efficient.
given the higher prevalence of disease. The long follow-up duration required would be prohibitively expensive and a significant number of subjects may be lost to follow-up (64).

In view of the sampling bias, high cost, and prolonged duration required by a prospective study, a population-based method for determining the fertility rate of women infected with HCV deserves consideration. If a cohort of women infected with HCV can be assembled, then the number of live births can be determined using administrative data from hospital discharges. Using this strategy, sampling bias from tertiary care clinics can be minimized, and the fertility rate can be determined at relatively low cost.

This thesis will address the assembly of a cohort of women with HCV infection through the use of administrative ambulatory OHIP diagnostic codes. A selection of OHIP diagnostic codes are compared for their accuracy in identifying women infected with HCV. As patients with hepatitis C may be candidates for undergoing liver biopsy, diagnostic codes are combined with the procedural codes for liver biopsy. After the accuracy of such diagnostic codes are characterized, then the code with highest sensitivity and specificity, as reflected by plotting as close as possible to the left upper corner of a ROC plot, will be used to assemble the desired cohort. Supplementary measures to characterize the reliability and predictive value of different diagnostic codes will be measured using Kappa statistic and logistic regression. Furthermore, additional experimental tests will be created by (a) requiring more than one submission of a code and (b) combining test codes using serial ("and") and parallel ("or"), and evaluated using ROC plot and Kappa statistic.

2.1.1 Sample Size Considerations

In order to ensure that the resultant cohort is representative of women with hepatitis C, higher specificity (lower false positive rate) was desired at the sacrifice of lower sensitivity (higher false negative rate). Although some women infected with HCV may be missed, the
women captured by the experimental code would likely be infected with HCV. If cases of hepatitis C were not captured as a result of low sensitivity, the final fertility rate would underestimate the true fertility rate of women with hepatitis C. Based on these considerations, an upper limit of 15% for the false positive rate (calculated using a 95% binomial confidence interval) and a specificity of 80% to 95% were chosen. Furthermore, the prevalence of hepatitis C among women of child-bearing age in the chosen clinics was estimated to range from 25% to 60%. This range was in agreement with a published prevalence of 47% of acute and chronic viral hepatitis in an urban, Canadian hospital-based practice (65).

2.2 Study Hypothesis

The combination of OHIP diagnostic code “viral hepatitis” together with procedure code “liver biopsy” is optimal for identifying women of child-bearing age infected with HCV as characterized by plotting closest to the left upper corner on ROC plot.
2.3 **Study Objectives**

1. To identify the optimal OHIP code combination (experimental test) for identifying women infected with HCV as characterized by:

   A. **ROC plot** (a measure of sensitivity and specificity) where the OHIP code plotting closest to the left upper corner is the optimal test (*primary objective*).

   B. **Kappa statistic** where the OHIP code with the highest Kappa statistic value is the most reliable.

   C. **logistic regression** using the binary outcome of HCV-infected or not infected:

      i) **univariate regression** using OHIP codes as

         a) binary predictor variables, or

         b) discrete predictor variables

      ii) **multiple regression** using OHIP codes as binary predictor variables.

      The predictor variables of the most parsimonious model will be deemed the most predictive.

2. To create additional OHIP experimental code combinations by requiring **more than one submission of a diagnostic code** and evaluate using

   A. ROC plot and

   B. Kappa statistic.

3. To create additional OHIP experimental code combinations by combining different codes using parallel ("or") and serial ("and") methods and evaluate using

   A. ROC plot and

   B. Kappa statistic.
Chapter 3: Methods

The objective of this chapter is to describe the overall strategy, the study design, the study subjects, the development of the reference standard (chart diagnosis of HCV infection), the development of the experimental test (OHIP code combinations), and the analysis of OHIP codes using various strategies to determine the optimal experimental test code. The feasibility study, the sample size calculation, and approval from research ethics boards are also included.

3.1 Overview of Study Methods

The strategy of comparing an experimental test against a reference standard was used to determine the accuracy of OHIP codes in identifying women infected with HCV. OHIP codes served as various “experimental tests” and were tested for their ability to identify women infected with HCV. The patient’s ambulatory chart record of HCV infection served as the reference standard for determining whether or not a woman is infected with HCV.

The process entailed five main steps (Figure 3-1). First, OHIP diagnostic and procedural codes were selected. Second, these codes were used in isolation or in combination to form a panel of experimental test codes. Third, all OHIP billings containing any of the experimental test codes were identified over a five-year period. All patients in this group were positive for at least one of the experimental test codes. The accuracy of the experimental test codes was determined in a subset of these code-positive patients. This subset, or sample population, must have a known reference standard. That is, the chart record of HCV infection (positive or negative) was required for this subset of patients. To obtain this gold standard, the fourth step required abstraction of systematically sampled patient charts. During chart abstraction at four medical clinics, ambulatory clinic charts were assessed for the presence or absence of HCV infection. The patients whose charts were abstracted formed the sample population. Finally, this
sample population was identified within the larger group of code-positive patients through linkage using the patient's health care number. These patients with known HCV status (presence or absence of the diagnosis of hepatitis C) and known experimental test status (OHIP code combination positive or negative) thus formed the sample population for evaluating the accuracy of OHIP code combinations in identifying women infected with HCV.

The optimal experimental test (OHIP code combinations) was characterized by plotting closest to the left upper corner on ROC plot (primary analysis), by a high Kappa statistic (a measure of test reliability), and by contributing significantly as a predictor variable in a logistic regression model for predicting diagnosis of HCV infection. As well, additional experimental tests were created by (a) by requiring more than one submission of a diagnostic code and (b) combining different codes using serial ("and") and parallel ("or") methods. These additional experimental tests were each evaluated using ROC plot and Kappa statistic.
3.2 **Study Design**

The study design is a cross-sectional comparative study.

3.3 **Study Subjects**

The study subjects are women of child-bearing age, defined by Statistics Canada as 15-49 years of age inclusive (born between January 1, 1941 and December 31, 1984), with valid OHIP health care number who were seen at one of four physicians’ clinics between January 1, 1995 and March 31, 2000.

3.4 **Reference Standard**

3.4.1 **Definition and Rationale**

The reference standard for diagnosis of HCV infection ("hepatitis C present") was defined as any positive serum anti-HCV antibody or detectable serum HCV RNA recorded in the patient chart. Given the limited level of clinical detail offered by administrative databases, no attempts were made characterize the precise timing and duration of infection. Similarly, resolution of HCV infection through spontaneous clearance or following medical therapy was not pursued. Acute infection was not distinguished from chronic infection. For purposes of this study, the ability of administrative codes to reflect chart status of ever having been infected with HCV was assessed. This definition, which does not differentiate viremic (detectable serum HCV RNA) from non-viremic women, will yield a more conservative estimate of the fertility rate for women infected with HCV.

Serum HCV RNA (measured by RT-PCR) is the laboratory gold standard for diagnosing infection with HCV. However, anti-HCV antibody (measured by ELISA) is a comparable test with sensitivity of 97.2% and specificity of 100% when used for screening of the general population. If applied to individuals with chronic liver disease, the sensitivity increases to 98.9% (66). Limitations of anti-HCV antibody testing may be due to a 5-6 week window between date
of acquiring infection and detection of seroconversion, delayed seroconversion in immunocompromised hosts, and the inability of serological tests to confirm active HCV infection. In practice, infrequent cases of false-positive and false-negative anti-HCV antibody results are addressed by incorporating HCV RNA testing with clinical and histological findings. Most patients are tested using anti-HCV antibody. When HCV RNA is also ordered, its presence confirms ongoing infection with HCV (67).

The definition for diagnosis of HCV infection used in this study agrees with the case definition of a confirmed case of hepatitis C that is used by the Health Canada Notifiable Diseases national surveillance program (35). The presence of positive anti-HCV antibody has also been used in other chart reviews as diagnosis of infection with HCV (65).

3.4.2 Choice of Physicians' Clinics

The selection of ambulatory clinics for chart abstraction was determined by two main criteria. Firstly, to ensure an efficient process of chart abstraction, the physician should have a reputation for caring for patients with hepatitis C. Such physicians were expected to have a reasonable prevalence of hepatitis C in their practice. Furthermore, the chosen clinics should differ in their type of practice to ensure some heterogeneity in the final chart sample. Tertiary (3°), secondary (2°), and primary care (1°) practices were chosen.

Based on these criteria, the following four physicians were chosen: Dr. A is an international authority on hepatitis C who works at the University Health Network (site Toronto – 3°). Dr. C is a secondary care gastroenterologist at Etobicoke General Hospital who is recognized in the Greater Toronto community for his interest in hepatitis C (site Toronto – 2°). Dr. D cares for many individuals with drug addiction, particularly pregnant women in the city’s methadone program at the Addiction Research Foundation (site Toronto – 1°). As the patient population in Toronto may be unique from other areas in Ontario, a site with high prevalence of
hepatitis C in London, Ontario was included. Dr. B is a London hepatologist with a strong reputation in hepatitis C (site London – 3°) (Table 3-1).

Table 3-1. Clinics chosen for Chart Abstraction

<table>
<thead>
<tr>
<th>Clinic</th>
<th>Physician</th>
<th>Specialty</th>
<th>City</th>
<th>Practice Type</th>
<th>Site</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>A1</td>
<td>Hepatologist</td>
<td>Toronto</td>
<td>Tertiary – University Health Network</td>
<td>Toronto – 3°</td>
</tr>
<tr>
<td>B</td>
<td>B1</td>
<td>Hepatologist</td>
<td>London</td>
<td>Tertiary</td>
<td>London – 3°</td>
</tr>
<tr>
<td>C</td>
<td>C1</td>
<td>Gastroenterologist</td>
<td>Toronto</td>
<td>Secondary – Etobicoke General Hospital</td>
<td>Toronto – 2°</td>
</tr>
<tr>
<td>D</td>
<td>D1</td>
<td>Family Physician</td>
<td>Toronto</td>
<td>Primary – Addiction Research Foundation</td>
<td>Toronto – 1°</td>
</tr>
</tbody>
</table>

3.4.3 Inclusion Criteria for Reference Standard

Only charts meeting the following criteria were eligible for inclusion: (1) female gender (2) date of birth between January 1, 1941 and December 31, 1984 inclusive (3) valid OHIP number and (4) seen by a physician between January 1, 1995 and March 31, 2000 inclusive.

3.4.4 Chart Abstraction

At each site, a sample of patient charts of females of childbearing age was manually abstracted for the diagnosis of hepatitis C, as defined as chart evidence of any positive serum anti-HCV antibody or HCV RNA result. If the clinical notes or laboratory values did not contain any evidence of a diagnosis of hepatitis C, then the patient was not considered to have a diagnosis of hepatitis C. Systematic sampling of patient charts was obtained in the following manner: At each office, shelves of charts were numbered by row and column. A random numbers table was used to generate a row number and column number to determine the shelf from which charts would be sampled. Numbers from the random numbers table were taken consecutively until a shelf was identified. Charts of the chosen shelf were then consecutively reviewed. When a shelf was completed, a new shelf was chosen from the random numbers table.

With consent from the responsible physician, a record of all charts composed of the health care number, birth date, and presence or absence of hepatitis C were compiled. The 10-digit OHIP health care number, date of birth, hepatitis C status (0 for absence of diagnosis, 1 for
presence of diagnosis), and clinic site were entered into an Excel datasheet using a laptop computer. The datasheet required password access, used coded column headings (with no reference to hepatitis C), and was stored on floppy diskette and kept in a locked facility.

Organizational limitations prevented pseudo-randomized chart abstraction from Dr. Selby’s clinic. As a result, the charts of all active patients from his methadone clinic at the Addiction Research Foundation were abstracted. However, the number of active patient charts was limited. Therefore, non-active patients from the Addiction Research Foundation database were also included to maximize the sample size from this site (site Toronto – 1°).

3.4.5 Description of Study Subjects

Among the study subjects, the prevalence of hepatitis C was established. To verify that the sample population reflected the true spectrum of women infected with HCV, the age distribution of the study subjects infected with HCV was determined and compared with the known age distribution of women infected with HCV from Notifiable Diseases. As well, to confirm that the severity of HCV disease in the sample population was similar to that of most women infected with HCV, the proportion of sample subjects who had died was compared among those infected with HCV and those not infected with HCV. As death from hepatitis C is uncommon for most women of child-bearing age, the death rates of women infected with HCV was expected to resemble that of women not infected with HCV.

3.5 Experimental Test

By combining OHIP diagnostic and procedure codes, a panel of experimental test codes was created.

3.5.1 OHIP Code Selection

The OHIP Schedule of Benefits was reviewed to select candidate diagnostic and procedure codes that might be used by a physician assessing a patient with hepatitis C. The
codes were presented to physicians who care for patients with hepatitis C to obtain their expert opinion regarding the appropriateness of the chosen codes. Following discussion with physicians, a broad range of diagnostic codes was included for analysis (Table 3-2). Two procedural codes (Z551, Z554) for liver biopsy were also used.

Table 3-2. Candidate Codes for Identifying Patients with Hepatitis C

<table>
<thead>
<tr>
<th>Type of OHIP Code (68)</th>
<th>Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>OHIP diagnostic code</td>
<td>070 viral hepatitis</td>
</tr>
<tr>
<td></td>
<td>571 cirrhosis of the liver (e.g. alcoholic or biliary cirrhosis)</td>
</tr>
<tr>
<td></td>
<td>573 other diseases of the liver</td>
</tr>
<tr>
<td></td>
<td>304 drug dependence, drug addiction</td>
</tr>
<tr>
<td></td>
<td>079 other viral diseases</td>
</tr>
<tr>
<td></td>
<td>K026, K027 Ontario Hepatitis C Assistance Program (OHCAP)</td>
</tr>
<tr>
<td>OHIP procedure code</td>
<td>Z551 needle liver biopsy</td>
</tr>
<tr>
<td></td>
<td>Z554 incisional liver biopsy</td>
</tr>
</tbody>
</table>

3.5.1.1 Diagnostic Codes

The diagnostic code **070 viral hepatitis** is likely to encompass a significant proportion of hepatitis B, in addition to hepatitis C. This code may also be applicable to less frequent causes of viral hepatitis (for example, hepatitis A, hepatitis D, and hepatitis E). The diagnostic code **571 cirrhosis of the liver** is expected to include a significant proportion of other causes of cirrhosis such as hepatitis B, alcoholic cirrhosis, and biliary cirrhosis. It may apply to advanced cases of hepatitis C as an indicator of more severe disease. It may also be the “convenient” diagnostic code used by tertiary hepatologists (who are referred the most difficult, often cirrhotic, cases of liver disease). However, a low incidence of cirrhosis is anticipated in females of childbearing age. The diagnostic code **573 other diseases of the liver** similarly may include all cases of liver disease. The diagnostic code **304 drug dependence, drug addiction** will capture the high-risk drug addict subset of patients with hepatitis C. For billing purposes, this code is often used to represent the multiple health issues that are addressed during one medical
assessment. In particular, a patient with this code who also undergoes liver biopsy will have a significant probability of having hepatitis C. The diagnostic code 079 other viral diseases may take account of any case of viral illness, not necessarily just hepatitis.

Effective December 1, 1998, the Ontario Hepatitis C Assistance Program (OHCAP) provides financial compensation to patients infected with hepatitis C resulting from a blood transfusion in an Ontario hospital prior to 1986 and after 1990. Those who became infected with HCV following blood or blood product transfusions, as well as those infected from a partner or parent who received blood or blood products are eligible. This program is not applicable to individuals without a history of blood transfusion who likely acquired their HCV infection from intravenous drug use. There is strong financial incentive for the patient ($10,000 – $25,000) to submit an application. Every OHCAP application must be completed in part by the patient’s attending physician who certifies the patient’s history of hepatitis C infection. These OHCAP fee codes are used by the physician for completing the OHCAP application form. Both family physicians and specialists may claim these fees. K026 covers any insured service and completion of the OHCAP application whereas K027 covers form completion without an associated consultation or visit on the same day.

3.5.1.2 Procedure Codes

The procedure code Z551 is used for percutaneous needle liver biopsy whereas code Z554 is used for incisional liver biopsy. Internists commonly use the former code whereas interventional radiologists favour the latter code. These codes will apply to a proportion of patients with hepatitis C. Tertiary care hepatologists estimate that 67-80% of all liver biopsies involves cases of hepatitis C (69). Current Canadian Consensus Guidelines recommend that all patients undergoing treatment have a liver biopsy (70). Treatment rates for hepatitis C patients
have increased from 30% to 50% while treatment rates for hepatitis B patients have increased from 5% to 10%. A 1997 U.S. survey found that 46% of physicians would obtain a liver biopsy if HCV infection were confirmed with normal ALT, but only 15% would treat. In a similar scenario but abnormal ALT, >90% physicians would obtain liver biopsy, and 60% would treat with interferon (71). Thus a patient with a diagnosis code of viral hepatitis and a procedure code of liver biopsy will be more likely to have hepatitis C rather than hepatitis B. In Ontario liver biopsies are performed by hepatologists, gastroenterologists, and a less commonly, by interventional radiologists. Most procedures occur in day surgery and do not necessarily involve an overnight stay in hospital. CIHI may be used to verify this procedure code, as it captures data on day surgery procedures.

3.5.2 Panel of Experimental Tests

A panel of experimental test codes was formed. Individual diagnostic codes as well as combinations of a diagnostic code with the procedure codes for “liver biopsy” were included (Table 3-3).

<table>
<thead>
<tr>
<th>Experimental Test</th>
<th>Presence of any one of the following:</th>
</tr>
</thead>
<tbody>
<tr>
<td>E1</td>
<td>070 viral hepatitis and Z551, Z554 liver biopsy</td>
</tr>
<tr>
<td>E2</td>
<td>070 viral hepatitis</td>
</tr>
<tr>
<td>E3</td>
<td>571 cirrhosis of liver</td>
</tr>
<tr>
<td>E4</td>
<td>573 other diseases of liver</td>
</tr>
<tr>
<td>E5</td>
<td>304 drug dependence, drug addiction</td>
</tr>
<tr>
<td>E6</td>
<td>079 other viral diseases</td>
</tr>
<tr>
<td>E7</td>
<td>571 cirrhosis of liver and Z551, Z554 liver biopsy</td>
</tr>
<tr>
<td>E8</td>
<td>573 other diseases of liver and Z551, Z554 liver biopsy</td>
</tr>
<tr>
<td>E9</td>
<td>304 drug dependence, drug addiction and Z551, Z554 liver biopsy</td>
</tr>
<tr>
<td>E10</td>
<td>079 other viral diseases and Z551, Z554 liver biopsy</td>
</tr>
<tr>
<td>E11</td>
<td>K026/7 Ontario Hepatitis C Assistance Program (OHCAP)</td>
</tr>
</tbody>
</table>
3.5.3 Identifying All Patients with OHIP Test Codes

All patients with OHIP billing data from January 1, 1995 to March 31, 2000 containing any of the experimental test codes E1-E8 inclusive and E11 OHCAP were identified. Test codes E9 and E10 were not used because they did not identify any patients during a feasibility study (please see section 3.7). The five-year time frame was chosen to reflect the onset of a markedly increased number of hepatitis C cases reported to the Notifiable Diseases Database (Figure 3-2) and cover as much recent OHIP data as possible. OHIP billings were limited to those with a fee for service code with suffix ‘A’ (feesuff ‘A’) to avoid multiple counts of a single procedure. Only those billings submitted by family physicians, internists, and gastroenterologists were included (specialty code = 00, 13, or 41, respectively). Patients with hepatitis C who may see any other types of specialists are likely to have concomitant assessment by family physicians, internists, or gastroenterologists. Duplicate visits (a patient with the same service date and same fee code) were eliminated. The dataset was further reduced to make individual patients the unit of analysis. This was accomplished by using an encrypted version of each patient’s health care number, called the ICES key number (IKN). Each entry corresponded to a unique IKN. Variables were introduced to (a) count the total number of times a unique IKN satisfied a code combination and (b) track in binary fashion whether or not a unique IKN had satisfied a code combination. The resulting dataset was merged with the Registered Persons Data Base by unique encrypted IKN to permit further limitation to female sex. Limitation of the dataset to women aged 15-49 years included women who were born after January 1, 1941 and before December 31, 1984.
3.5.4 Incorporating Experimental Test Codes into the Sample Population

The reference standard, the presence or absence of a chart diagnosis of HCV infection, was known for all patients in the sample population. In order to obtain the experimental test (the presence or absence of OHIP test codes) for all patients of the sample population, linkage of the chart records to the OHIP test code records was achieved using the patients' health care number. The unique health care numbers obtained from chart abstraction were encrypted to an equivalent IKN. The unique IKN was used to link OHIP billing data test codes to the abstracted chart diagnosis of HCV infection. Mr. Don Deboer at the Institute of Clinical Evaluative Sciences (ICES) performed the encryption. The encryption process used a confidential, predefined ICES algorithm that was not divulged to the investigators of this study.
3.6 Analysis: Determining the Optimal Experimental Test

3.6.1 Overview of Analysis

The definition of the “optimal test” must acknowledge the interdependent relationship between sensitivity and specificity, positive and negative predictive value. For example, excellent specificity may often be achieved at significant loss of sensitivity (72). In order to recognize the “relative cost” that each characteristic of a test may bear, three separate methods were applied separately to determine the optimal test. In the principal analysis (primary objective 1A), the optimal experimental test was determined by plotting closest to the left upper corner of a ROC plot of sensitivity by 1-specificity. As a secondary analysis, test performance was expressed in terms of test reliability, as characterized by measuring Cohen’s Kappa statistic. The test with the highest Kappa statistic value was deemed optimal (objective 1B). As well, the most parsimonious model from logistic regression was used to identify the experimental tests that were most accurate in predicting the diagnosis of hepatitis C. Experimental codes were treated as binary predictor variables (objective 1Cia) or discrete predictor variables (objective 1Cib). Multiple logistic regression using binary predictor variables was also performed (objective 1Cii).

An additional secondary objective was to evaluate the accuracy of experimental tests created by requiring more than one submission of a diagnostic code (objective 2) or by combining different codes using serial (“and”) and parallel (“or”) methods (objective 3). These additional experimental test strategies were also evaluated using ROC plot and Kappa statistics.

The 95% confidence interval was calculated for all estimates of sensitivity, specificity, positive predictive value, negative predictive value, and Kappa statistic. However, for ease of reading, these confidence intervals were presented only when necessary for purposes of
comparison. In instances where the confidence intervals were omitted, the ranges of the estimates were narrow and did not affect the stated conclusions.

3.6.2 Basic Test Characteristics

The basic test characteristics include sensitivity, specificity, positive predictive value, and negative predictive value. They may be derived from the traditional 2 x 2 table of experimental test (OHIP code combinations) and truth (diagnosis of hepatitis C), which has separate boxes for true positives, false negatives, false positives, and true negatives. By dividing these boxes by the total number of women (i.e. set N=1), the prevalence of hepatitis C is represented by P (where P=TP+FN). Table 3-4 forms the basis for calculating the basic test characteristics of a diagnostic or experimental test relative to a reference standard (diagnosis of hepatitis C) (73, 74).
Table 3-4. Measures of Test Performance and Test Reliability

<table>
<thead>
<tr>
<th>Test Code: +</th>
<th>Hepatitis C+</th>
<th>Test Code: -</th>
<th>Hepatitis C-</th>
</tr>
</thead>
<tbody>
<tr>
<td>TP (true positive)</td>
<td>FP (false positive)</td>
<td>FN (false negative)</td>
<td>TN (true negative)</td>
</tr>
<tr>
<td>P</td>
<td></td>
<td>Q</td>
<td>Q' = 1-P</td>
</tr>
</tbody>
</table>

**Test Performance**

- **Sensitivity** (SE) = \( SE = \frac{TP}{P} \)
- **Specificity** (SP) = \( SP = \frac{TN}{P'} \)
- **Positive predictive value** (PPV) = \( PPV = \frac{TP}{Q} \)
- **Negative predictive value** (NPV) = \( NPV = \frac{TN}{Q'} \)

**Test Reliability**

- **Cohen's Kappa statistic (K)**

\[
K = \frac{\text{Observed agreement} - \text{Agreement by chance}}{1 - \text{Agreement by chance}}
\]

\[
= \frac{(\text{EFF} - PQ - P'Q')}{(1 - PQ - P'Q')}
\]

Where

- **Efficiency** = \( \text{EFF} = TP + TN \)
- \( P = TP + FN \)
- \( P' = 1 - P = FP + TN \)
- \( Q = TP + FP \)
- \( Q' = 1 - Q = FN + TN \)

**Sensitivity** (SE) is defined as the probability of having a positive test result among those patients who have a positive diagnosis. This reflects the probability that a diagnosed case of hepatitis C will be detected by the test. **Specificity** (SP) is the probability of having a negative test result among those patients who have a negative diagnosis. This reflects the likelihood that a case that is not diagnosed as hepatitis C will not be detected by the test. The **positive predictive value** (PPV) of a test is the probability of having a positive diagnosis among patients having a positive test. The **negative predictive value** (NPV) of a test is the probability of having a negative diagnosis among patients having a negative test. These terms may be expressed in equation form (Table 3-4) (73). All statistical analyses were performed using the Statistical Analysis Software (SAS Institute, Cary, NC, USA) version 8.0.
3.6.3 **Receiver Operating Characteristic (ROC) Plot**

One common method for evaluating the performance of diagnostic systems based on the basic test characteristics is the receiver operating characteristic (ROC) plot, a discontinuous graph of sensitivity (the true positive rate) versus 1-specificity (the false positive rate) for a set of experimental tests compared against the same reference diagnosis. These tests are plotted to illustrate different sensitivities and specificities. The left upper corner of the graph is considered the ideal test point, where both sensitivity and specificity are 100%. The experimental test that plots closest to the left upper corner is therefore the optimal test (75).

The discontinuous ROC plot, a two-dimensional graph of sensitivity and 1-specificity, is distinct from the continuous ROC curve, which demonstrates the relationship between sensitivity and 1-specificity across a range of some continuous variable (a “diminishing signal”) (72). In this study, no continuous variable existed to permit the use of a ROC curve.

3.6.4 **Test Reliability: Cohen’s Kappa Statistic**

The degree of agreement between the experimental test and the reference standard, or the overall reliability of a test may be expressed using Cohen’s Kappa statistic (K). As an alternative measure of test performance, Kappa statistic quantifies the *degree of reliability* of a test while correcting for the degree of agreement expected by chance alone. A formula for calculating Kappa statistic is shown in Table 3-4. By convention, qualitative interpretation of Kappa statistic values are as follows: $<0.20$, poor agreement; $0.21-0.40$, fair agreement; $0.41-0.60$, moderate agreement; $0.61-0.80$, good agreement; $0.81-1.00$, very good agreement. The most reliable experimental test has the highest Kappa statistic value (75).

3.6.5 **Logistic Regression**

Logistic regression offers another strategy for evaluating the accuracy of experimental tests. By identifying the most significant predictor variables in a model, logistic regression
focuses on the *predictive value* of a test. The diagnosis of hepatitis C (present or absent) is used as the dichotomous outcome variable. Logistic regression offers a mathematical model for evaluating test codes separately (univariate analysis) as well as in combination (multiple regression). Also, both binary (present or absent) and discrete (1, 2, 3, etc.) forms of predictor variables may be evaluated. Logistic regression facilitates the evaluation of these variations of the predictor variables.

The different experimental test codes, codes E2 – E6 inclusive, with or without the procedure of liver biopsy, were each evaluated as predictor variables. For each predictor variable, both binary (Ex) and discrete (Dx) values were tested in separate models. For example, if a patient had 5 billing records with the diagnosis of viral hepatitis (E2), then E2=1 as a binary variable, and D2=5 as a discrete variable. Models containing multiple test codes as predictor variables were also constructed using multiple logistic regression through stepwise elimination of non-significant predictor variables.

The most parsimonious model of predictor variables was determined by comparing the log likelihood and deviance values (measures of goodness of fit) of each model. By calculating the difference of deviance and degrees of freedom between two models that differ by one predictor variable, a chi-square value may be generated. If the chi-square value corresponded to a p-value <0.05 then the predictor variable was considered to belong in the most parsimonious model. The variables in the most parsimonious model represented the experimental codes that were most useful in predicting the diagnosis of hepatitis C (76, 77).

3.6.6 Multiple Submissions of a Diagnostic Code

Hepatitis C is generally a chronic condition requiring multiple visits to a physician for adequate assessment. A patient with hepatitis C is likely to be labeled with a diagnostic code more than once, in contrast to a patient with a mild viral illness who may only require a single
assessment. The lag time required between initial blood work and subsequent discussion of blood test results suggests that a minimum of two medical visits will be required. Thus patients with repetition of code entries may be more likely to have hepatitis C. In attempt to improve the specificity of the experimental test, test characteristics were evaluated with the requirement that a “positive” diagnostic code must be submitted more than once for a patient. These experimental tests were evaluated using ROC and Kappa statistic (73).

3.6.7 Serial and Parallel Test Strategies

Once the optimal codes were identified, serial assembly of codes (code X “and” code Y) and parallel assembly of codes (code X “or” code Y) were tested as additional test strategies in attempt to improve the specificity of the experimental test. These experimental tests were also evaluated using ROC and Kappa statistic (73).

3.7 Feasibility Study

The possible unreliability of diagnostic codes in OHIP billing data raised the concern that the diagnostic codes may be omitted in a significant number of billing claims. In order to ensure that the candidate codes would identify a reasonable number of patients, a feasibility study was conducted using data from Notifiable Diseases as an estimate of disease prevalence. If the diagnostic code was never completed in a claim, then a candidate code would identify close to zero patients and the study would not be feasible.

Newly detected cases of hepatitis C are reported to Notifiable Diseases. As many cases of hepatitis C are asymptomatic, the majority of newly diagnosed cases represent the prevalence of hepatitis C rather than the true incidence. For most cases, the date of acquiring infection likely preceded the first positive test for HCV. Nevertheless, if a candidate code is to be feasible, its ability to detect cases of hepatitis C that are being monitored by physicians would identify at least a proportion of the number of newly reported cases of hepatitis C. To avoid using codes
that fail to detect a reasonable number of patients, at least 50% of the number of cases reported to Notifiable Diseases should be identified by a candidate code. This cutoff was increased to 60% to guard against an insufficient number of patients.

From July 1, 1996 to June 30, 1997, the Notifiable Diseases Database reported 1507 newly reported cases of hepatitis C for women 15-39 years of age and 916 newly reported cases of hepatitis C for women 40-59 years of age (10) (Figure 3-2). It was assumed that half of the 916 cases represented women aged 40-49 years of age, yielding a total of 1965 new cases of women aged 15-49 years with hepatitis C. In order to justify the feasibility of using diagnostic codes to assemble a cohort of women with hepatitis C, a candidate code would be expected to generate at least 60% of these cases, or 1200 women aged 15-49 years.

The proposed experimental test codes E1-E10 (from Table 3-3) were pulled from fiscal 1996 OHIP data. The Ontario Hepatitis C Assistance Program [OHCAP (E1)] code was not used because it was not in effect until December 1, 1998. These OHIP billing data were reduced in the same manner as the five-year OHIP billing data described in section 3.5.3.

The feasibility study of fiscal 1996 OHIP data identified a total of 125,345 hits based on the proposed experimental test codes E1-E8 (Table 3-5). Test codes E9 (drug dependence/addiction & liver biopsy) and E10 (other viral diseases and liver biopsy) did not identify any patients. Thus the code combinations of “drug dependence” and “liver biopsy” as well as “other viral diseases” and “liver biopsy” were eliminated from further study. The experimental test code viral hepatitis alone found 17,500 patients, almost nine-fold the 1965 cases of women aged 15-49 years with hepatitis C reported by Notifiable Diseases. The large number of patients identified with the proposed codes supported the feasibility of this study.
### Table 3-5. Number of Test Code Combinations – Fiscal 1996

<table>
<thead>
<tr>
<th>Test Code Combination</th>
<th>Number of Submitted Claims*</th>
</tr>
</thead>
<tbody>
<tr>
<td>E1 viral hepatitis &amp; liver biopsy</td>
<td>188</td>
</tr>
<tr>
<td>E2 viral hepatitis</td>
<td>17500</td>
</tr>
<tr>
<td>E3 cirrhosis of liver</td>
<td>1709</td>
</tr>
<tr>
<td>E4 other diseases of liver</td>
<td>4074</td>
</tr>
<tr>
<td>E5 drug dependence/addiction</td>
<td>21255</td>
</tr>
<tr>
<td>E6 other viral diseases</td>
<td>80415</td>
</tr>
<tr>
<td>E7 cirrhosis of liver &amp; liver biopsy</td>
<td>38</td>
</tr>
<tr>
<td>E8 other diseases of liver &amp; liver biopsy</td>
<td>166</td>
</tr>
<tr>
<td>E9 drug dependence/addiction &amp; liver biopsy</td>
<td>0</td>
</tr>
<tr>
<td>E10 other viral diseases &amp; liver biopsy</td>
<td>0</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>125,345</strong></td>
</tr>
</tbody>
</table>

* an individual may have more than one experimental test code combination

As the code of "viral hepatitis" (E2) alone was able to identify nine-times more cases of hepatitis C than the Notifiable Diseases Database, this reinforced the different nature of the Notifiable Diseases and OHIP data. The Notifiable Diseases Database records newly identified cases of hepatitis C (an incidence of disease detection, rather than disease incidence), whereas OHIP data reflects the prevalence of individuals with hepatitis C seen by physicians.
3.8 Sample Size Calculation

For this study, rather than calculate the number of cases required to have a 95% confidence interval of a fixed distance for a given point estimate (such as the proportion of women with hepatitis C who test positive), the sample size calculation was based on the desire to build a clinically valid 2 x 2 table of experimental test (OHIP code combinations) and truth (diagnosis of hepatitis C). For OHIP test codes to be useful, a high false positive rate was prohibitive. As a result, as discussed in the study rationale (section 2.1.1), the sample size was required to exclude a false positive rate of no higher than 15% and provide a range of specificity from 80% - 95%, across a range of prevalence from 25% - 60%. Based on these assumptions, the 2 x 2 table was generated to yield the total number of charts required for abstraction.

To clarify the calculations for estimating an appropriate sample size, please refer to the following 2 x 2 table:

<table>
<thead>
<tr>
<th>Test Code +</th>
<th>Hepatitis C+</th>
<th>Hepatitis C-</th>
<th>(a + b)</th>
</tr>
</thead>
<tbody>
<tr>
<td>a (true positive)</td>
<td>b (false positive)</td>
<td>(a + b)</td>
<td></td>
</tr>
<tr>
<td>Test Code -</td>
<td>c (false negative)</td>
<td>d (true negative)</td>
<td>(c + d)</td>
</tr>
<tr>
<td>(a + c)</td>
<td>(b + d)</td>
<td>N = (a + b + c + d)</td>
<td></td>
</tr>
</tbody>
</table>

A sample of 180 patients defined by the algorithm as having hepatitis C would provide an upper limit of 15% for the false positive rate (for a 95% binomial confidence interval). That is, if the maximum false positive rate = b/(a + b) = 0.15 and (a + b) = 180, then based on the formula for the 95% upper limit of any proportion, (maximum p) = p + 1.96*{square root [p*(1-p)/N]}, in this case:

0.15 = b/180 + 1.96*{square root [(b/180)*(1- (b/180))/N]} which rearranges to:
0.000030864b^2 - 0.50446b + 13.8 = 0. By using the quadratic formula, \( b = 27 \) or \( 16315 \). Since \( b \) must be \( < (a + b) \), \( b = 27 \) is used. Since \( (a + b) = 180 \) and \( b = 27 \), then \( a = 180 - 27 = 153 \). Assuming a specificity of 95\%, and given that specificity = \( d/(b + d) \), \( 0.95 = d/(27 + d) \), and \( d = 513 \). Thus \( (b + d) = 27 + 513 = 540 \). Assuming a prevalence of 25\% (in the medical clinic), and given that prevalence = \( (a + c)/(a + b + c + d) \), then \( 0.25 = (153 + c)/(153 + 27 + c + 540) \) and thus \( c = 27 \). Therefore, \( N = (a + b + c + d) = (153 + 27 + 27 + 540) = 720 \) total charts. This would provide a sensitivity \( a/(a + c) \) of 85\%.

By similar calculations, the algorithm that defines 540 patients as not having hepatitis C would permit a maximum false negative rate of 5\%, based on a 95\% binomial confidence interval (Table 3-7).

<table>
<thead>
<tr>
<th>Table 3-7. Sample Size Determination – 25% Prevalence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chart Diagnosis</td>
</tr>
<tr>
<td>Test Code</td>
</tr>
<tr>
<td>Positive</td>
</tr>
<tr>
<td>Negative</td>
</tr>
<tr>
<td>Total</td>
</tr>
</tbody>
</table>

Differences in the prevalence of hepatitis C among clinics affect sample size estimates. If prevalence increased from 25\% to 60\%, a sample of 320 patients defined by the algorithm as having hepatitis C would give an upper limit of 15\% for the false positive rate (for a 99.9\% binomial confidence interval). Assuming a specificity of 80\%, 600 total charts would provide a sensitivity of 76\%. 280 patients defined by the algorithm as not having hepatitis C would permit a maximum false negative rate of 30\%, based on a 95\% binomial confidence interval (Table 3-8).
Table 3-8. Sample Size Determination – 60% Prevalence

<table>
<thead>
<tr>
<th>Test Code</th>
<th>Chart Diagnosis</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>HCV+</td>
<td>HCV-</td>
<td>Total</td>
</tr>
<tr>
<td>Positive</td>
<td>272</td>
<td>48</td>
<td>320</td>
</tr>
<tr>
<td>Negative</td>
<td>88</td>
<td>192</td>
<td>280</td>
</tr>
<tr>
<td>Total</td>
<td>360</td>
<td>240</td>
<td>600</td>
</tr>
</tbody>
</table>

The number of total charts required for abstraction was inflated to 800 per site to accommodate variability in prevalence and false positive rate across the sites.

3.9 Research Ethics

Ethical approval for this study was obtained from the Research Ethics Boards of the University of Toronto (protocol #6381), the University Health Network (protocol #00-0295-U), the Centre for Addiction and Mental Health (protocol #128/2000), and the Hospital for Sick Children (protocol #2000/324) (please see Appendix 3).
Chapter 4: Results

The objective of this chapter is to characterize the sample population and the reference standard, to describe the prevalence of the OHIP code combinations (experimental tests), and to present the analysis of the experimental tests as outlined in the study objectives.

4.1 Reference Standard: Description of Study Subjects

The medical charts of 2548 women were abstracted from the four physicians’ clinics. Among the 2548 records, no duplication of health care numbers occurred, indicating 2548 unique study subjects. The health care numbers were encrypted into a corresponding ICES key number (IKN). For successful record linkage, the IKN’s needed to be recognized by the ICES database of validated IKN’s. This validation process eliminated sixteen invalid records (Table 4-1). These sixteen mismatched records likely represented errors in data entry during the chart abstraction. By linkage of valid IKN’s, the remaining sample population of 2532 unique records with chart diagnosis of hepatitis C status was combined with OHIP billing data to form the sample population for evaluation of OHIP diagnostic code accuracy.

4.1.1 Proportion of Clinic Charts Assessed for Inclusion

The number of charts in a clinic and the number of charts assessed for this study were estimated by counting shelves. In Dr. A’s clinic (site Toronto – 3°) 25 out of 31 shelves (80.6%) were assessed for inclusion into the study. In Dr. B’s clinic (site London – 3°) all 22 shelves (100%) were assessed for inclusion into the study. In Dr. C’s clinic (site Toronto – 2°) 12 out of 21 shelves (57.1%) were assessed for inclusion into the study. Each shelf contained approximately 200-400 charts. In Dr. D’s clinic (site Toronto – 1°) all active charts were assessed.
4.1.2 **Prevalence of Hepatitis C**

The prevalence of hepatitis C ranged from 2% - 60% among the 5 various sources of patient charts (Table 4-3). The low prevalence of hepatitis C in the secondary care clinic (site Toronto - 2°) resulted from the heavy presence of other diagnoses in that practice among women of child-bearing age. The prevalence of HCV infection in the total sample was 27.2%.

<table>
<thead>
<tr>
<th>Site: Description</th>
<th>Practice Type</th>
<th>Physician</th>
<th># charts</th>
<th>Valid IKN</th>
<th>Prevalence of HCV+</th>
</tr>
</thead>
<tbody>
<tr>
<td>Toronto - 3°</td>
<td>Toronto</td>
<td>Hepatology - tertiary care</td>
<td>A</td>
<td>798</td>
<td>796</td>
</tr>
<tr>
<td>London - 3°</td>
<td>London</td>
<td>Hepatology - tertiary care</td>
<td>B</td>
<td>757</td>
<td>751</td>
</tr>
<tr>
<td>Toronto - 2°</td>
<td>Etobicoke</td>
<td>Gastroenterology - secondary care</td>
<td>C</td>
<td>828</td>
<td>825</td>
</tr>
<tr>
<td>Toronto - 1°</td>
<td>ARF database</td>
<td>Family Practice - primary care</td>
<td>D</td>
<td>87</td>
<td>83</td>
</tr>
<tr>
<td>Toronto - 1°</td>
<td>ARF active</td>
<td>Family Practice - primary care</td>
<td>D</td>
<td>78</td>
<td>77</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td></td>
<td>2548</td>
<td>2532</td>
</tr>
</tbody>
</table>

4.1.3 **Age Distribution**

The age distribution of women with hepatitis C reported to Notifiable Diseases (Figure 4-1a) and abstracted charts of women with hepatitis C (Figure 4-1b) shared similar peaks for women born in the 1940s to 1960s (aged 30-60 years in 1995-2000). This suggested that the sampled charts reflect the national database’s distribution of women with hepatitis C. This peak of hepatitis C cases also agreed with U.S. estimates that persons born between 1940 and 1965 have the highest lifetime risk of acquiring the infection (2, 78).

4.1.4 **Death**

The date of death, if applicable, was available from the RPDB dataset. Of 2532 women in the sample population, 29 (1.1%) had died. Ten of these women had been infected with HCV. At time of death, the mean age of these women was 45.1 years for those not infected with HCV.
43.6 years for those infected with HCV (p>0.1).

4.1.5 Multiple Comparisons

Particularly when large administrative databases are utilized, the large sample size tends to facilitate the ability to demonstrate statistical significance. As well, the comparison of multiple test codes for different test characteristics are likely to identify a statistically significant outcome due to random chance. To address these issues, the traditional alpha value (type I error) of 0.05 may be decreased to lower the threshold for detection of significance. The most conservative approach would be to utilize the Bonferroni correction, whereby the alpha value of 0.05 is divided by the number of possible combinations (79). As the results from logistic regression reveal that the p-values of the test codes E2 and E5 are <0.000001 (table 4-8), any adjustment of the type I error to correct for multiple comparisons is unlikely to impact on the study results.
A total of 561,801 submitted claims containing at least one of the codes E1-E8 and E11 between January 1, 1995 and March 31, 2000 (Table 4-2) were identified from 521,750 women of child-bearing age. A woman is counted more than once if she has multiple submissions of test codes. For example, a woman may see a physician on ten separate visits for “viral hepatitis” and later develop liver fibrosis, resulting in three physician visits for “cirrhosis of liver”. Such a woman would be represented as ten submitted claims of “viral hepatitis” and three submitted claims of “cirrhosis of liver”. As a woman may have multiple claims, the following table cannot be used to determine the prevalence of each test code among women in Ontario.

Table 4-2. Number of Test Code Combinations – 1995-2000

<table>
<thead>
<tr>
<th>Experimental Test Code</th>
<th>Number of Claims Submitted*</th>
</tr>
</thead>
<tbody>
<tr>
<td>E1 viral hepatitis &amp; liver biopsy</td>
<td>992</td>
</tr>
<tr>
<td>E2 viral hepatitis</td>
<td>69,372</td>
</tr>
<tr>
<td>E3 cirrhosis of liver</td>
<td>6,421</td>
</tr>
<tr>
<td>E4 other diseases of liver</td>
<td>18,119</td>
</tr>
<tr>
<td>E5 drug dependence/addiction</td>
<td>102,091</td>
</tr>
<tr>
<td>E6 other viral diseases</td>
<td>363,296</td>
</tr>
<tr>
<td>E7 cirrhosis of liver &amp; liver biopsy</td>
<td>273</td>
</tr>
<tr>
<td>E8 other diseases of liver &amp; liver biopsy</td>
<td>908</td>
</tr>
<tr>
<td>E11 OHCAP</td>
<td>329</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>561,801</strong></td>
</tr>
</tbody>
</table>

* an individual may have more than one experimental test code combination
4.3 **Analysis: Determining the Optimal Experimental Test**

4.3.1 **Overview of Results**

Table 4-3 summarizes the results for the optimal codes determined using ROC plot, Kappa statistic, and logistic regression for the various experimental test strategies: individual codes, multiple submissions of a diagnostic code, and serial and parallel methods.

Table 4-3. Optimal OHIP Code Combinations Determined by ROC plot, Kappa statistic, and Logistic Regression

<table>
<thead>
<tr>
<th>Experimental Test Strategy</th>
<th>ROC plot</th>
<th>Kappa statistic</th>
<th>Logistic Regression</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Univariate Regression</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Binary predictor variable (Ex)</td>
</tr>
<tr>
<td>2. Multiple submissions of a diagnostic code</td>
<td>E2</td>
<td>E5</td>
<td></td>
</tr>
<tr>
<td>3. Serial and Parallel Methods</td>
<td>E2§ (E2 or E5)§</td>
<td>(E2 and E5)</td>
<td></td>
</tr>
</tbody>
</table>

* The model using counts of a diagnostic code (Dx) was *inferior* to the corresponding model using the binary form (Ex).

§ Models E2 and (E2 or E5) are comparable on ROC plot.

The ROC plot is the traditional, widely utilized measure of test performance characterized by measures of sensitivity and specificity. It identifies E2 “viral hepatitis” as the optimal code. As a supplementary evaluation of test performance, Kappa statistic is utilized. Although Kappa statistic is also derived from the same traditional 2 x 2 table of experimental test (OHIP code combinations) and truth (diagnosis of hepatitis C), it focuses on the degree of agreement between the true positives and true negatives. Using this measure, Kappa statistic finds E5 “drug dependence/addiction” to be the optimal code. However, by convention, the range of Kappa values from 70% - 80% indicate good agreement, and the Kappa values of the
majority of the diagnostic codes (including E2 “viral hepatitis”) fall within this range. Finally, logistic regression uses modeling to evaluate the ability of different test codes to predict an outcome of hepatitis C infection. The model containing E2 “viral hepatitis” and E5 “drug dependence/addiction” is deemed the most parsimonious. Taken together, these three different approaches to evaluating test performance yield consistent findings – codes E2 “viral hepatitis” and E5 “drug dependence/addiction” are most useful for identifying a case of HCV infection.
4.3.2 Single Codes (Binary Predictor Variable)

4.3.2.1 Basic Test Characteristics and ROC Plot

The accuracy of OHIP diagnostic codes was evaluated using the sample population of 2372 women (from sites Toronto – 3°, London – 3°, and Toronto – 2°). As pseudo-randomized sampling of charts was not feasible at site Toronto – 1°, only a small number of charts (160) were available for abstraction. Such a small sample size would generate estimates of test characteristics such as sensitivity and specificity with large confidence intervals. Thus, these women from ARF were excluded in the following results and described separately in Appendix 2.

The chart diagnosis of hepatitis C served as the reference standard while various OHIP code combinations served as the experimental tests. The prevalence of HCV infection in this sample of three clinics was 25.3%. Table 4-4 reports the basic test characteristics of each experimental test code. Each test code was a binary variable. For example, a patient who has ever been diagnosed as “viral hepatitis” would have E2=1. If a patient has never been diagnosed as “viral hepatitis”, then E2=0.

Table 4-4. Test Characteristics (%) for Sites Toronto – 3°, London – 3°, and Toronto – 2°*

<table>
<thead>
<tr>
<th>Test Code Combination</th>
<th>SP</th>
<th>SE</th>
<th>PPV</th>
<th>NPV</th>
</tr>
</thead>
<tbody>
<tr>
<td>E11 OHCAP</td>
<td>99.9</td>
<td>6.0</td>
<td>97.3</td>
<td>75.8</td>
</tr>
<tr>
<td>E1 viral hepatitis &amp; liver biopsy</td>
<td>98.5</td>
<td>8.7</td>
<td>65.8</td>
<td>76.1</td>
</tr>
<tr>
<td>E7 cirrhosis of liver &amp; liver biopsy</td>
<td>98.4</td>
<td>3.5</td>
<td>42.0</td>
<td>75.0</td>
</tr>
<tr>
<td>E5 drug dependence/addiction</td>
<td>95.8</td>
<td>23.5</td>
<td>65.3</td>
<td>78.7</td>
</tr>
<tr>
<td>E8 other diseases of liver &amp; liver biopsy</td>
<td>92.8</td>
<td>11.8</td>
<td>35.9</td>
<td>75.6</td>
</tr>
<tr>
<td>E6 other viral diseases</td>
<td>85.9</td>
<td>14.5</td>
<td>25.8</td>
<td>74.7</td>
</tr>
<tr>
<td>E3 cirrhosis of liver</td>
<td>76.5</td>
<td>49.3</td>
<td>41.6</td>
<td>81.6</td>
</tr>
<tr>
<td>E2 viral hepatitis</td>
<td>74.5</td>
<td>81.0</td>
<td>51.9</td>
<td>92.1</td>
</tr>
<tr>
<td>E4 other diseases of liver</td>
<td>54.7</td>
<td>65.2</td>
<td>32.8</td>
<td>82.3</td>
</tr>
</tbody>
</table>

*Bolded results indicate values > 50%.

To address objective 1A, the ROC plot of sensitivity and 1-specificity is shown in Figure
4-2. The combination of diagnostic OHIP code “viral hepatitis” together with procedure code “liver biopsy” does not plot closest to the left upper corner on ROC plot, disproving the study hypothesis. Instead, E2 (viral hepatitis) is the optimal code because it lies closest to the left upper corner on the ROC plot (Figure 4-2). It has the highest sensitivity (81.0%) and offers a specificity of 74.5%.

4.3.2.1. **Impact of Adding Procedure Code Liver Biopsy**

In Figure 4-2, the triangular points represent combinations of a diagnostic code with the addition of liver biopsy. While E2 (viral hepatitis), E3 (cirrhosis of liver), and E4 (other diseases of liver) plot in the left upper quadrant, all corresponding codes which include liver biopsy (triangles) plot toward the left lower corner, reflecting a small increase in specificity with great loss of sensitivity. Indeed, the combination of diagnostic and procedural administrative codes (e.g. viral hepatitis & liver biopsy) has greater specificity than a single diagnostic code (e.g. viral hepatitis). However, the gain in specificity is accompanied by a substantial loss in sensitivity.
4.3.2.2 Test Reliability: Cohen’s Kappa Statistic

To address objective 1B, table 4-5 reports the Kappa statistic of each experimental test code.

Table 4-5. Kappa statistics for Sites Toronto – 3°, London – 3°, and Toronto – 2°*

<table>
<thead>
<tr>
<th>Test Code Combination</th>
<th>Kappa statistic (%)</th>
<th>95% Confidence Interval (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>E5 drug dependence/addiction</td>
<td>76.5</td>
<td>74.8 - 78.2</td>
</tr>
<tr>
<td>E11 OHCAP</td>
<td>76.0</td>
<td>74.3 - 77.7</td>
</tr>
<tr>
<td>E1 viral hepatitis &amp; liver biopsy</td>
<td>75.4</td>
<td>73.6 - 77.1</td>
</tr>
<tr>
<td>E7 cirrhosis of liver &amp; liver biopsy</td>
<td>74.1</td>
<td>72.3 - 75.8</td>
</tr>
<tr>
<td>E8 other diseases of liver &amp; liver biopsy</td>
<td>71.2</td>
<td>69.4 - 73.1</td>
</tr>
<tr>
<td>E2 viral hepatitis</td>
<td>71.1</td>
<td>69.3 - 73.0</td>
</tr>
<tr>
<td>E6 other viral diseases</td>
<td>65.6</td>
<td>63.7 - 67.5</td>
</tr>
<tr>
<td>E3 cirrhosis of liver</td>
<td>64.9</td>
<td>63.0 - 66.9</td>
</tr>
<tr>
<td>E4 other diseases of liver</td>
<td>45.2</td>
<td>43.2 - 47.2</td>
</tr>
</tbody>
</table>

*Bolded results indicate values sharing 95% confidence interval of E5 drug dependence/addiction.

Evaluation using Kappa statistic identifies **E5 (drug dependence/addiction)** as the optimal code with a Kappa statistic of 76.5%. If the 95% confidence interval of the Kappa statistic values is taken into account, the accuracy of E5 (drug dependence/addiction) is comparable to E11 (OHCAP), E1 (viral hepatitis & liver biopsy), and E7 (cirrhosis of liver & liver biopsy).

With the exception of E4 (other diseases of liver), the predominance of weighted Kappa statistic values within the 60% - 80% range indicates good agreement. Compared to the other codes, E4 (other diseases of liver) has a low true negative rate (41%). In other words, a patient without a diagnosis of “other diseases of liver” may still be infected with HCV. The low true negative rate results in a lower efficiency (57.4%) and consequently, a low weighted Kappa statistic (45.2%).
4.3.3 Logistic Regression

4.3.3.1 Univariate Regression of Single Codes

To address objective 1Ci, table 4-6 displays the variables of each univariate model and the corresponding negative log likelihood value.

Table 4-6. Goodness of Fit for Models with Single Codes as Binary and Discrete Variables

<table>
<thead>
<tr>
<th>Predictor Variables</th>
<th>-Log Likelihood for binary Ex</th>
<th>Predictor Variables</th>
<th>-Log Likelihood for discrete Dx</th>
</tr>
</thead>
<tbody>
<tr>
<td>E2</td>
<td>958</td>
<td>D2</td>
<td>1060</td>
</tr>
<tr>
<td>E3</td>
<td>1086</td>
<td>D3*</td>
<td>1096</td>
</tr>
<tr>
<td>E4</td>
<td>1093</td>
<td>D4</td>
<td>1091</td>
</tr>
<tr>
<td>E5</td>
<td>1050</td>
<td>D5</td>
<td>1070</td>
</tr>
<tr>
<td>E6</td>
<td>1089</td>
<td>D6</td>
<td>1086</td>
</tr>
</tbody>
</table>

* not significant in model (p<0.05)

4.3.3.1.1 Binary Predictor Variable

Of all the models without liver biopsy, the model using E2 (viral hepatitis) had the smallest (least negative) log likelihood value (-958), suggesting that this code provided the best model for predicting the diagnosis of hepatitis C. The model using E5 (drug dependence/addiction) had the next smallest log likelihood value (-1050).

4.3.3.1.2 Discrete Predictor Variable

The models using the corresponding discrete values, D2 and D5, had larger (more negative) log likelihood values (-1060 and -1070, respectively). Therefore the binary form of a diagnostic code (Ex) provided a better model than multiple counts of a diagnostic code (Dx). In all of the models, each variable remained significant (p<0.05) except D3 (cirrhosis of liver) in the model of D3.

4.3.3.1.3 Impact of Adding Procedure Code Liver Biopsy

To formally assess the contribution of liver biopsy to the logistic model, Table 4-7 compares these two models of D2 and E2 with and without liver biopsy. The calculated chi-
square values are not significant (p=0.22208 and 0.14234), supporting the lack of predictive utility offered by the inclusion of liver biopsy.

Table 4-7. Comparing Models E2 and D2 with and without Liver Biopsy

<table>
<thead>
<tr>
<th>Predictor Variables</th>
<th>-Log Likelihood</th>
<th>Deviance (measures goodness-of-fit)</th>
<th>Degrees of freedom</th>
<th>Δ Deviances$</th>
<th>Δ Degrees of freedom$</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>E2 LBx</td>
<td>957</td>
<td>1913.9727</td>
<td>1691</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E2</td>
<td>958</td>
<td>1915.4636</td>
<td>1692</td>
<td>1.4909</td>
<td>1</td>
<td>0.22208</td>
</tr>
<tr>
<td>D2 LBx</td>
<td>1059</td>
<td>2118.2766</td>
<td>1691</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D2</td>
<td>1060</td>
<td>2120.4290</td>
<td>1692</td>
<td>2.1524</td>
<td>1</td>
<td>0.14234</td>
</tr>
</tbody>
</table>

$ (deviance of given model) – (deviance of preceding model)
$ (degrees of freedom of given model) – (degrees of freedom of preceding model)
LBx = liver biopsy

**Bolded variables** were eliminated in the subsequent model.
4.3.3.2 Multiple Regression of Binary Codes

To address objective 1Cii, each subsequent model in table 4-8 has one less variable compared to the preceding model. The difference of deviances and the difference of degrees of freedom between two models yield a chi-square value (deviance/degrees of freedom). The p-value of the chi-square reflects the contribution of the eliminated variable to the model.

Table 4-8. Goodness of Fit for Models with Multiple Codes

<table>
<thead>
<tr>
<th>Predictor Variables</th>
<th>-Log Likelihood</th>
<th>Deviance (measures goodness of fit)</th>
<th>Degrees of freedom</th>
<th>( \Delta ) Deviances</th>
<th>( \Delta ) Degrees of freedom</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>E2 E3 E4 E5 E6 LBx</td>
<td>909</td>
<td>1817.2030</td>
<td>1687</td>
<td>0.4813</td>
<td>1</td>
<td>0.48783</td>
</tr>
<tr>
<td>E2 E4 E5 E6 LBx</td>
<td>909</td>
<td>1817.6843</td>
<td>1688</td>
<td>0.4813</td>
<td>1</td>
<td>0.48783</td>
</tr>
<tr>
<td>E2 E5 E6 LBx</td>
<td>910</td>
<td>1819.5775</td>
<td>1689</td>
<td>1.8932</td>
<td>1</td>
<td>0.16884</td>
</tr>
<tr>
<td>E2 E5 E6</td>
<td>911</td>
<td>1822.3911</td>
<td>1690</td>
<td>2.8136</td>
<td>1</td>
<td>0.09347</td>
</tr>
<tr>
<td>E2 E5</td>
<td>915</td>
<td>1829.0549</td>
<td>1691</td>
<td>6.6638</td>
<td>1</td>
<td>0.00984</td>
</tr>
<tr>
<td>E2</td>
<td>958</td>
<td>1915.4636</td>
<td>1692</td>
<td>86.4087</td>
<td>1</td>
<td>&lt;0.00001</td>
</tr>
<tr>
<td>E2 E5</td>
<td>915</td>
<td>1829.0549</td>
<td>1691</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E5</td>
<td>1050</td>
<td>2100.6340</td>
<td>1692</td>
<td>271.5791</td>
<td>1</td>
<td>&lt;0.00001</td>
</tr>
<tr>
<td>E2 E5 LBx*</td>
<td>913</td>
<td>1825.7798</td>
<td>1690</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E2 E5</td>
<td>915</td>
<td>1829.0549</td>
<td>1691</td>
<td>3.2751</td>
<td>1</td>
<td>0.070339</td>
</tr>
</tbody>
</table>

\( \$ \) (deviance of given model) – (deviance of preceding model)
\( \| \) (degrees of freedom of given model) – (degrees of freedom of preceding model)

LBx = liver biopsy

**Bolded variables** were eliminated in the subsequent model.

* This model is equivalent to predictor variables E1 and E5; liver biopsy borderline significant p=0.0696

By multiple regression, the most parsimonious model included **E2 (viral hepatitis)** or **E5 (drug dependence/addiction)**. The addition of **E6 (other viral diseases)** to the model of [E2 (viral hepatitis) or E5 (drug dependence/addiction)] significantly improved the model (p-value of 0.00984). However, its p-value was remarkably greater than that value offered by E2 (viral hepatitis) and E5 (drug dependence/addiction), each with p-value of <0.00001. As a result, E6 was removed from the final model.
4.3.3.2.1 Impact of Adding Procedure Code Liver Biopsy

The addition of liver biopsy to the model containing E2 (viral hepatitis) or E5 (drug dependence/addiction) minimally improved the model's predictive value (-log likelihood changes from 915 to 913) and did not significantly improve the model (p=0.070339). Since E1 represented viral hepatitis (E2) and liver biopsy, these results echoed earlier findings that E2 (viral hepatitis) and E5 (drug dependence/addiction) were most predictive for the diagnosis of hepatitis C.
4.3.4 Multiple Submissions of a Diagnostic Code (Binary Predictor Variable)

4.3.4.1 Basic Test Characteristics and ROC Plot

For a code to be counted, the diagnostic code must have been submitted more than once. Table 4-9 reports the basic test characteristics of each experimental test code. To address objective 2A, the ROC plot of sensitivity and 1-specificity is shown in Figure 4-3.

Table 4-9. Test Characteristics (%) – Code Submitted >1*

<table>
<thead>
<tr>
<th>Test Code Combination</th>
<th>SP</th>
<th>SE</th>
<th>PPV</th>
<th>NPV</th>
</tr>
</thead>
<tbody>
<tr>
<td>E11 OHCAP</td>
<td>100.0</td>
<td>0.3</td>
<td>100.0</td>
<td>74.7</td>
</tr>
<tr>
<td>E1 viral hepatitis &amp; liver biopsy</td>
<td>99.9</td>
<td>0.8</td>
<td>83.3</td>
<td>74.8</td>
</tr>
<tr>
<td>E7 cirrhosis of liver &amp; liver biopsy</td>
<td>99.8</td>
<td>0.0</td>
<td>0.0</td>
<td>74.6</td>
</tr>
<tr>
<td>E8 other diseases of liver &amp; liver biopsy</td>
<td>99.2</td>
<td>0.8</td>
<td>26.3</td>
<td>74.7</td>
</tr>
<tr>
<td>E5 drug dependence/addiction</td>
<td>98.1</td>
<td>18.3</td>
<td>76.4</td>
<td>78.0</td>
</tr>
<tr>
<td>E6 other viral diseases</td>
<td>95.1</td>
<td>4.7</td>
<td>24.3</td>
<td>74.6</td>
</tr>
<tr>
<td>E3 cirrhosis of liver</td>
<td>83.6</td>
<td>30.4</td>
<td>38.7</td>
<td>78.0</td>
</tr>
<tr>
<td>E2 viral hepatitis</td>
<td>82.4</td>
<td>62.9</td>
<td>54.9</td>
<td>86.8</td>
</tr>
<tr>
<td>E4 other diseases of liver</td>
<td>68.3</td>
<td>40.9</td>
<td>30.4</td>
<td>77.3</td>
</tr>
</tbody>
</table>

*Bolded results indicate values > 50%.

Comparison of the plot of codes submitted more than once (Figure 4-3) to the previous plot of single codes (Figure 4-2) shows that when a code must be submitted more than once in order to count as a “positive test”, the sensitivity of all codes decreases while specificity increases. The overall ranking of codes remains similar to initial analyses that required that a code be submitted only once. **E2 (viral hepatitis)** continues to have highest sensitivity (62.9%) and a specificity of 82.4%; it plots closest to the left upper corner in Figure 4-3.

4.3.4.1.1 Impact of Adding Procedure Code Liver Biopsy

While **E2 (viral hepatitis)**, **E3 (cirrhosis of liver)**, and **E4 (other diseases of liver)** plot in the upper half, all corresponding codes that include liver biopsy (triangles) crowd the left lower corner, reflecting an improved specificity with loss of sensitivity. Similar to the experimental codes requiring only one submission of a diagnostic code, the combination of
diagnostic and procedural administrative codes (e.g. viral hepatitis & liver biopsy) has greater specificity than a single diagnostic code (e.g. viral hepatitis).
4.3.4.2 Test Reliability: Cohen's Kappa Statistic

To address objective 2B, table 4-10 reports the Kappa statistic of each experimental test code (submitted more than once).

Table 4-10. Kappa statistics - Code Submitted >1*

<table>
<thead>
<tr>
<th>Test Code Combination</th>
<th>Kappa statistic (%)</th>
<th>95% Confidence Interval (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>E5 drug dependence/addiction</td>
<td>77.3</td>
<td>75.6 – 78.9</td>
</tr>
<tr>
<td>E1 viral hepatitis &amp; liver biopsy</td>
<td>74.8</td>
<td>73.1 – 76.5</td>
</tr>
<tr>
<td>E11 OHCAP</td>
<td>74.7</td>
<td>73.0 – 76.4</td>
</tr>
<tr>
<td>E7 cirrhosis of liver &amp; liver biopsy</td>
<td>74.5</td>
<td>72.8 – 76.3</td>
</tr>
<tr>
<td>E8 other diseases of liver &amp; liver biopsy</td>
<td>74.2</td>
<td>72.4 – 76.0</td>
</tr>
<tr>
<td>E2 viral hepatitis</td>
<td>74.2</td>
<td>72.4 – 75.9</td>
</tr>
<tr>
<td>E6 other viral diseases</td>
<td>71.6</td>
<td>69.8 – 73.4</td>
</tr>
<tr>
<td>E3 cirrhosis of liver</td>
<td>67.3</td>
<td>65.4 – 69.2</td>
</tr>
<tr>
<td>E4 other diseases of liver</td>
<td>54.5</td>
<td>52.5 – 56.5</td>
</tr>
</tbody>
</table>

*Bolded results indicate values sharing 95% confidence interval of E5 drug dependence/addiction.

The Kappa statistic coefficients of codes requiring multiple submissions indicate that E5 (drug dependence/addiction) is optimal. Similar to the findings of Kappa statistic coefficients on once-submitted codes in section 4.3.2.2, if the 95% confidence interval of the Kappa statistic values is taken into account, the accuracy of E5 (drug dependence/addiction) is comparable to most of the other codes, once multiple submissions of a diagnostic code are required.

With the exception of E4 (other diseases of liver), the predominance of weighted Kappa statistic values within the 60% - 80% range indicates good agreement. Compared to the other codes, multiple submissions of E4 (other diseases of liver) bear a low true negative rate (51.0%). In other words, a patient without a diagnosis of “other diseases of liver” may still be infected with HCV. The low true negative rate results in a lower efficiency (61.3%) and consequently, a low weighted Kappa statistic (54.5%).
4.3.5 Serial and Parallel Methods (Binary Predictor Variable)

Earlier results using ROC plot, Kappa statistic, and logistic regression identified codes E2 (viral hepatitis) and E5 (drug dependence/addiction) as the most predictive variables for identifying hepatitis C. Based on these results, additional algorithms were created using serial (“and”) and parallel (“or”) methods and compared with each individual code using ROC plot and Kappa statistic. To address objective 3, table 4-11 reports the basic test characteristics and Kappa statistic of each code combination and the ROC plot of sensitivity and 1-specificity is shown in Figure 4-4.

Table 4-11. Test Characteristics (%) for Codes Combined by Serial and Parallel Methods*

<table>
<thead>
<tr>
<th>Test Code Combination</th>
<th>SP (95% CI)</th>
<th>SE (95% CI)</th>
<th>PPV (95% CI)</th>
<th>NPV</th>
<th>Kappa statistic (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>E2 and E5</td>
<td>98.9 (98.4-99.4)</td>
<td>20.0 (16.8-23.2)</td>
<td>85.7 (79.9-91.5)</td>
<td>78.5</td>
<td>78.3 (76.7-80.0)</td>
</tr>
<tr>
<td>E5 drug dependence/addiction</td>
<td>95.8 (94.8-96.7)</td>
<td>23.5 (20.1-26.8)</td>
<td>65.3 (58.9-71.6)</td>
<td>78.7</td>
<td>76.5 (74.8-78.2)</td>
</tr>
<tr>
<td>E2 viral hepatitis</td>
<td>74.5 (72.5-76.6)</td>
<td>81.0 (77.9-84.2)</td>
<td>51.9 (48.7-55.1)</td>
<td>92.1</td>
<td>71.1 (69.3-73.0)</td>
</tr>
<tr>
<td>E2 or E5</td>
<td>71.4 (69.3-73.5)</td>
<td>84.5 (81.6-87.4)</td>
<td>50.1 (47.0-53.2)</td>
<td>93.2</td>
<td>68.9 (67.0-70.7)</td>
</tr>
</tbody>
</table>

*Bolded results indicate values > 50%.

4.3.5.1 Basic Test Characteristics and ROC Plot

Serial combination of codes [E2 and E5 (viral hepatitis or drug dependence/addiction)] resulted in improved specificity but significant loss of sensitivity. Parallel combination of codes [E2 or E5 (viral hepatitis or drug dependence/addiction)] had maximal sensitivity of 84.5% with specificity of 71.4% and PPV of 50.1%. In fact, E2 (viral hepatitis) and [E2 or E5 (viral hepatitis or drug dependence/addiction)] were remarkably comparable: E2 (viral hepatitis) offered slightly better specificity (74.5% compared to 71.4%) and PPV (51.9% compared to 50.1%) than [E2 or E5 (viral hepatitis or drug dependence/addiction)]. On the other hand, [E2 or E5 (viral hepatitis or drug dependence/addiction)] offered mildly higher sensitivity (84.5% compared to 81%), and NPV (93.2% compared to 92.1%). Therefore parallel combination of
these codes increased sensitivity without significant loss of specificity. These findings agree with the results of multiple regression from Table 4-9 that also found that the model containing [E2 or E5 (viral hepatitis or drug dependence/addiction)] to be the most parsimonious.

Figure 4-4 shows that both E2 (viral hepatitis) and [E2 or E5 (viral hepatitis or drug dependence/addiction)] most closely approach the left upper corner. All other algorithms cluster around the left lower corner, indicating high specificity but low sensitivity.

4.3.5.2 Test Reliability: Cohen’s Kappa Statistic

The code combination of E2 and E5 (viral hepatitis or drug dependence/addiction) has the highest Kappa statistic of 78.3%, although the 95% confidence interval of this estimate overlapped with that of E5 (drug dependence/addiction) alone. By convention, the fact that all weighted Kappa statistic values were within the 60% - 80% range indicates good agreement between the code combinations and the reference standard.
Chapter 5: Discussion

The objective of this chapter is to summarize the results of the study, discuss the relevance of the findings, discuss the strengths and limitations of the study design, and describe future directions.

5.1 Summary of Results

The OHIP diagnostic code “viral hepatitis” (E2) is the optimal test for identifying women of child-bearing age infected with HCV as characterized by plotting closest to the left upper corner of the ROC plot. “Viral hepatitis” is also the most significant predictor variable in a logistic regression model for predicting the diagnosis of hepatitis C. The OHIP diagnostic code for “viral hepatitis” has a sensitivity of 81%, specificity of 75%, and positive predictive value of 52% when used to assemble a cohort of women infected with HCV. Alternate codes with improved positive predictive value and specificity incur a loss of sensitivity.

The OHIP diagnostic code “drug dependence/addiction” (E5) is worthy of mention because parallel combination of “E2 or E5” (viral hepatitis and drug dependence/addiction) is comparable to E2 alone by ROC plot. The serial combination of “E2 and E5” (viral hepatitis and drug dependence/addiction) is most reliable with the highest Kappa statistic estimate. As well, “viral hepatitis,” together with “drug dependence/addiction,” are significant predictors of a diagnosis of hepatitis C by multiple logistic regression.

Multiple submissions of a diagnostic code improve the specificity of a code combination, but often result in a loss of sensitivity. Similarly, the addition of the procedure code liver biopsy improves specificity at a loss of sensitivity. Kappa statistics, a measure of test reliability, tend to be high when a test has high specificity.
5.2 Relevance of Findings

Characterization of the accuracy of the assembled cohort is a necessary step before outcome measures derived from the cohort can be interpreted. The positive predictive value of 52% for OHIP diagnostic code “viral hepatitis” (E2) suggests that more than half of all women identified in the OHIP database using this diagnostic code will have a diagnosis of hepatitis C. Other causes of viral hepatitis, aside from hepatitis C, will contribute to the proportion of false positives generated by administrative data using OHIP diagnostic codes. As hepatitis C and hepatitis B are the most common forms of viral hepatitis (80), it is conceivable that hepatitis B accounts for the majority of the remaining cohort. In fact, the cohort assembled using E2 might be better described as “women of child-bearing age with viral hepatitis”.

Fortunately, combination codes can be selected, based on their test characteristics, to form a cohort to meet the requirements of any particular research question. Indeed, for the purpose of assembling a cohort of women of child-bearing age to measure their fertility rate, higher positive predictive value and specificity at a cost of less sensitivity is necessary. The most accurate code combination of “E2 and E5” (viral hepatitis and drug dependence/addiction) offers a sensitivity of 20%, specificity of 99%, and positive predictive value of 86%.

In fact, the higher accuracy offered by including the diagnostic code “drug dependence/addiction” (E5) was not foreseen but parallels the current epidemiology of hepatitis C. Since the testing of donor blood for HCV was implemented in 1990, the incidence of post-transfusional hepatitis C has decreased (81). Injection drug use is now the major risk factor for HCV infection in Canada (14). It is likely that many individuals with hepatitis C are primarily being treated for their drug dependence. Only some of these individuals will see a specialist to be assessed for treatment.
5.3 **Strengths of Study Design**

In Canada, all previous work assessing the accuracy of diagnostic codes has focused on inpatient hospital charts and CIHI codes. This is the first study to evaluate the ability of OHIP diagnostic codes to identify a clinic chart diagnosis of hepatitis C. For this study, a panel of experimental tests (OHIP code combinations) was analyzed against a reference standard (ambulatory chart diagnosis of hepatitis C).

Experimental tests were formed using multiple strategies – single codes of at least one submission, codes submitted more than once, and combinations of codes using serial and parallel methods. These multiple strategies provided insight into the behavior of diagnostic codes under different circumstances. While “viral hepatitis” is deemed the optimal code, the inclusion of “drug dependence/addiction” or “liver biopsy” can be used to adjust the measures of sensitivity and specificity.

The reference standard consisted of over 2300 abstracted clinic charts for the diagnosis of hepatitis C. The large number of abstracted charts improves the precision of the test characteristics and Kappa statistics. By comparison, other studies of administrative data abstracted less than 500 charts (49, 57).

Although the primary analysis evaluated OHIP codes using ROC plot, alternate measures of test performance including Kappa statistic and logistic regression also found comparable results. Whereas the ROC plot measured sensitivity and specificity, Kappa statistic measured the degree of agreement, and logistic regression measured predictive value. These different measures consistently ranked experimental tests containing the diagnostic codes E2 (viral hepatitis) and/or E5 (drug dependence/addiction) highest.
5.4 **Limitations of Study Design**

5.4.1 **Sampling Bias**

Despite the attempt to assemble a population-based cohort of women with hepatitis C, this project is subject to the same limitations of all studies of HCV prevalence: namely, the background prevalence of hepatitis C is unknown, particularly because many infected individuals are asymptomatic and have yet to be identified and tested. However, the age distribution of the chart cohort and Notifiable Diseases data are comparable, suggesting that the chart cohort reflects the population of women already identified to have hepatitis C.

5.4.2 **Spectrum Bias**

For purposes of analysis, only three clinic sites were used. Yet diagnostic indices can vary substantially, depending on the composition of tested patients (82). The patients of the sample population used in this study yielded a range of diagnostic indices that may be subject to change when OHIP test codes are applied to the provincial population. Namely, the decreased prevalence of hepatitis C within the provincial population will result in a lower positive predictive value and lower specificity. The degree to which the findings from the three sites reflect the population of clinics remains unknown. Although the potential spectrum bias can distort the study results, the ideal sampling approach of systematically gathering all eligible cases from all clinical facilities through continued surveillance (64) was not feasible.

In attempt to capture a broad spectrum of women identified to have hepatitis C, chart abstraction was performed at a diversity of sites, including primary, secondary, and tertiary care clinics. A clinic in London was chosen to sample beyond the unique multi-cultural composition of Toronto. The Addiction Research Foundation Methadone Clinic with its particularly high prevalence of hepatitis C was utilized to include patients with drug dependence or addiction. The choice of the secondary care clinic was made based on its reputation as a popular referral
destination for cases of hepatitis C; even then the actual prevalence of hepatitis C among women of child-bearing age was 2%. Other clinics are expected to have an even lower prevalence of hepatitis C, limiting the feasibility of a more exhaustive chart abstraction at additional sites. Given the lower prevalence of hepatitis C, the inclusion of additional sites is unlikely to contribute further to the sample population.

Further attempts to diminish sampling bias of charts within the clinic were made by sampling charts using a pseudo-randomization procedure. However, pseudo-randomized sampling was not possible at the Addiction Research Foundation Methadone Clinic. The difficulties of pseudo-randomized chart sampling, coupled with a low prevalence of hepatitis C in other clinics, suggest that current degree of sampling bias is unavoidable.

5.4.3 Selection of OHIP Codes

OHIP codes were selected without available data regarding the nature of OHIP billing practices for diagnostic codes. Several factors may contribute to wide variation in the use of diagnostic codes during claims submission. Diagnostic codes are not mandatory and do not directly affect the amount the health care provider is reimbursed. Furthermore, the familiarity with the variety of eligible diagnostic codes listed in the OHIP Schedule of Benefits may also vary between clinics. In fact, OHIP billing is often delegated to a medical secretary whose appreciation of an accurate diagnosis may be limited. Despite these concerns, it may be argued that the current electronic OHIP billing system encourages the use computerized billing programs. These billing programs offer key-word search capabilities that facilitate the completion of diagnostic codes.

5.5 Future Directions

The assembled cohort of women infected with HCV can be linked to other databases to provide population-based data on women infected with HCV. In particular, linkage to hospital
discharge data, using the unique health care number, will permit measure of the number of live births for these women. The resultant age-specific fertility rates can help refine estimates of maternal-infant transmission of HCV. If the fertility rate of the cohort is significantly higher than that of the general population, than the estimated number of infants who will acquire hepatitis C from their infected mother will be greater than previously thought. These estimates can influence public health planning decisions as the government prepares to meet the increasing burden of disease from hepatitis C.

As the burden of disease from hepatitis C gains attention, the unmet needs of individuals infected with hepatitis C have become more apparent. Patients with hepatitis C have followed the lead of AIDS activists in raising awareness of their disease in public and professional domains. These patient-driven groups (e.g. HepCure, Canadian Hemophila Society) are highly motivated to cultivate an international network for promoting hepatitis C education, support, and research. Their efforts strongly suggest that patients with hepatitis C do not have adequate access to medical care and support (83).

By raising such concerns, these patients succeeded in organizing the 1st Canadian Conference on Hepatitis C in May 2001 with full support from the Canadian government. This multi-disciplinary meeting was not only attended by experts in the field of medicine and public health, but also by HCV-infected individuals. The priority of the conference was to emphasize the accessibility and practicality of information regarding hepatitis C and to ensure the widest dissemination of such information to the community of persons infected with and affected by hepatitis C. This meeting highlighted the need for Canadian population-based research to assess the needs of patients with hepatitis C (84).

An administrative cohort of patients infected with hepatitis C can help characterize the
burden of disease of hepatitis C in Canada. Such a cohort can be used to evaluate regional variation of health care services to patients with hepatitis C. Longitudinal follow-up of an assembled cohort can provide measures of health care utilization for infected women. Linkage through health care number to the National Population Health Survey can yield disease-specific utility measures for Canadians with viral hepatitis to facilitate measures of health-related quality of life. Future linkage to drug dispensing and laboratory databases can provide further detail to help characterize the disease spectrum of patients with of hepatitis C. Such population-based data are relatively inexpensive and are a necessary adjunct to Canadian surveillance studies of hepatitis C.

Finally, the resultant basic test characteristics of OHIP diagnostic codes measured in this study suggest that certain diagnostic codes may be more useful than previously thought. It is invalid to assume that all diagnostic codes have poor accuracy. Rather, the combination of different codes can offer a spectrum of accuracy. Diagnostic code combinations can be strategically manipulated to assemble different groups of patients with different characteristics. By understanding the behavior of test characteristics such as sensitivity and specificity when codes are combined, one can harness the information in diagnostic codes to enrich the opportunities for research using administrative data.

Diagnostic codes represent an uncharted frontier in the realm of population-based research. Population-based cohorts have been recognized for their many advantages. They offer a relatively inexpensive, efficient mechanism for tracking province-wide disease burden, selecting high risk patient groups for intensive intervention, and evaluating the effect of changes in disease management on outcomes over time. The time is ripe for expanding the utility of diagnostic codes in medical conditions above and beyond hepatitis C.
1. Evaluate Experimental Tests (code combinations) using:
   A. ROC plot
   B. Kappa statistic
   C. logistic regression (binary and discrete predictor variables)

2. Create additional Experimental Tests by requiring more than one submission of a code and evaluate using:
   A. ROC plot
   B. Kappa statistic

3. Create additional Experimental Tests by combining test codes using serial or parallel method and evaluate using:
   A. ROC plot
   B. Kappa statistic

Analysis
Figure 3-2. Number of Female Cases of HCV Reported in Ontario by Year of Reporting

Number of Female Cases of HCV

![Bar chart showing number of female cases of HCV by year of reporting.](chart)

Figure 4-1. (a) Female Cases of HCV in Ontario by Age Group Reported to Notifiable Diseases

Number of Female Cases of HCV


This age distribution of women with hepatitis C reported to Notifiable Diseases has peaks for women aged 30-60 years in 1995-1998. These peaks match the peaks of Figure 4-1b of women aged 30–59 years from abstracted charts of women with hepatitis C (Figure 4-1b), suggesting that the sampled charts reflect the national database’s distribution of women with hepatitis C.
Figure 4-1. (b) Number of Women with Hepatitis C in Sample Population by Age Group

Number of Female Cases of HCV

Women's Age Groups (years)

Toronto – 3° = tertiary care hepatologist, Dr. A
London – 3° = tertiary care hepatologist, Dr. B
Toronto – 2° = secondary care gastroenterologist, Dr. C
Toronto – 1° = primary care family physician in methadone clinic, Dr. D
Figure 4-2. Sensitivity and Specificity of Codes for Sites Toronto – 3°, London – 3°, and Toronto – 2°

Test Codes

E1 Viral hepatitis & liver biopsy
E2 Viral hepatitis
E3 Cirrhosis of liver
E4 Other diseases of liver
E5 Drug dependence, addiction
E6 Other viral diseases
E7 Cirrhosis of liver & liver biopsy
E8 Other diseases of liver & liver biopsy
E11 OHCAP

★ = Ideal Test Location

Sensitivity = probability of having a given test code among those patients with a diagnosis of hepatitis C. Specificity = probability of not having a given test code among those patients without a diagnosis of hepatitis C.

Toronto – 3° = University Health Network tertiary care hepatologist: Dr. A
London – 3° = London tertiary care hepatologist: Dr. B
Toronto – 2° = Etobicoke General Hospital secondary care gastroenterologist: Dr. C

E2 Viral hepatitis plots closest to the left upper corner.
Figure 4-3. Sensitivity and Specificity of Codes (%) – Code Submitted >1 for Sites Toronto – 3°, London – 3°, and Toronto – 2°

Test Codes
E1 Viral hepatitis & liver biopsy
E2 Viral hepatitis
E3 Cirrhosis of liver
E4 Other diseases of liver
E5 Drug dependence, addiction
E6 Other viral diseases
E7 Cirrhosis of liver & liver biopsy
E8 Other diseases of liver & liver biopsy
E11 OHCAP

Ideal Test Location

Sensitivity = probability of having a given test code among those patients with a diagnosis of hepatitis C.
Specificity = probability of not having a given test code among those patients without a diagnosis of hepatitis C.

Toronto – 3° = University Health Network tertiary care hepatologist: Dr. A
London – 3° = London tertiary care hepatologist: Dr. B
Toronto – 2° = Etobicoke General Hospital secondary care gastroenterologist: Dr. C

E2 Viral hepatitis plots closest to the left upper corner.
Figure 4-4. Sensitivity and Specificity of Codes Combined by Serial and Parallel Methods for Sites Toronto – 3º, London – 3º, and Toronto – 2º

Ideal Test Location

Sensitivity = probability of having a given test code among those patients with a diagnosis of hepatitis C.
Specificity = probability of not having a given test code among those patients without a diagnosis of hepatitis C.

Toronto – 3º = University Health Network tertiary care hepatologist: Dr. A
London – 3º = London tertiary care hepatologist: Dr. B
Toronto – 2º = Etobicoke General Hospital secondary care gastroenterologist: Dr. C

E2 (viral hepatitis) and E2 or E5 (viral hepatitis or drug dependence/addiction) both plot closest to the left upper corner.
### Appendix 1: Glossary and Abbreviations

<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>anti-HCV</td>
<td>antibody to hepatitis C virus</td>
</tr>
<tr>
<td>ARF</td>
<td>Addiction Research Foundation</td>
</tr>
<tr>
<td>child-bearing age</td>
<td>15-49 years of age inclusive (This age range agrees with the definition of child-bearing age used by Statistics Canada) (35).</td>
</tr>
<tr>
<td>CIHI</td>
<td>Canadian Institute for Health Information</td>
</tr>
<tr>
<td>CMG</td>
<td>Case Mix Group</td>
</tr>
<tr>
<td>EFF</td>
<td>efficiency</td>
</tr>
<tr>
<td>ELISA</td>
<td>enzyme linked immunosorbent assay</td>
</tr>
<tr>
<td>HCN</td>
<td>health care number</td>
</tr>
<tr>
<td>HCV</td>
<td>hepatitis C virus</td>
</tr>
<tr>
<td>hepatitis C</td>
<td>serum anti-HCV antibody or HCV RNA positive on at least one occasion</td>
</tr>
<tr>
<td>ICES</td>
<td>Institute of Clinical Evaluative Sciences</td>
</tr>
<tr>
<td>IKN</td>
<td>ICES key number</td>
</tr>
<tr>
<td>LCDC</td>
<td>Laboratory Centre for Disease Control</td>
</tr>
<tr>
<td>OHCAP</td>
<td>Ontario Hepatitis C Assistance Program</td>
</tr>
<tr>
<td>OHIP</td>
<td>Ontario Hospital Insurance Program</td>
</tr>
<tr>
<td>NPV</td>
<td>negative predictive value</td>
</tr>
<tr>
<td>P</td>
<td>prevalence</td>
</tr>
<tr>
<td>PPV</td>
<td>positive predictive value</td>
</tr>
<tr>
<td>Q</td>
<td>level of a test</td>
</tr>
<tr>
<td>RIBA</td>
<td>recombinant immunoblot assay</td>
</tr>
<tr>
<td>RNA</td>
<td>ribonucleic acid</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Definition</td>
</tr>
<tr>
<td>--------------</td>
<td>-------------------------------------</td>
</tr>
<tr>
<td>ROC</td>
<td>receiver operator characteristic</td>
</tr>
<tr>
<td>RPDB</td>
<td>Registered Persons Data Base</td>
</tr>
<tr>
<td>RT-PCR</td>
<td>reverse transcriptase polymerase chain reaction</td>
</tr>
<tr>
<td>SE</td>
<td>sensitivity</td>
</tr>
<tr>
<td>SP</td>
<td>specificity</td>
</tr>
</tbody>
</table>
Appendix 2: Addiction Research Foundation Methadone Clinic (Site Toronto – 1°)

The small number of charts available for abstraction precluded the incorporation of this site into the chart cohort for evaluation. Similarly, the small sample size prohibits valid determination of basic statistics of test performance such as sensitivity and specificity. Nonetheless, these 160 charts provide a descriptive overview of hepatitis C among women from a methadone clinic. Ninety-nine percent of all women from ARF received a diagnosis of “drug dependence or drug abuse” (E5).

According to the clinic chart, 54% of women in this group had hepatitis C. Studies from other addiction clinics and young injection drug users also report high prevalence of HCV infection (85, 86), reinforcing the significance of drug use as a risk factor for acquiring hepatitis C (2, 17). Although these cases from ARF were excluded in the chart cohort, the diagnostic code of “drug dependence or drug abuse” (E5) still emerged as a significant predictor of the chart diagnosis of hepatitis C. If these 160 cases are added to the chart cohort, the sensitivity of E5 increases from 23.5% to 33.1% with mild loss in specificity (95.8% to 92.0%) and PPV (65.3% to 60.8%). The trend toward improved sensitivity and high prevalence of hepatitis C among drug users suggest that code E5 will detect additional cases of hepatitis C in the population, arguing for the use of [E2 or E5] as the optimal algorithm for assembling a cohort of women with hepatitis C rather than E2 alone. The age distribution of the cases from ARF (site Toronto – 1°) peaked in the 1960s, reflecting a younger group.
Appendix 3: Ethics Approval Letters

University of Toronto

University Health Network

Centre for Addiction and Mental Health

Hospital for Sick Children
University of Toronto

OFFICE OF RESEARCH SERVICES

PROTOCOL REFERENCE #6381

August 11, 2000

Dr. Teresa To
Hospital for Sick Children
555 University Ave.
Toronto, ON
fax: 813-5979

Dear Dr. To:

Re: Your research protocol entitled, “Assembling a Cohort of Women with Hepatitis C in Ontario”

We are writing to advise you that a Review Committee composed of Dr. J. Hux and Prof. R. Hutchinson has granted approval to the above-named research study.

During the course of the research, any significant deviations from the approved protocol (that is, any deviation which would lead to an increase in risk or a decrease in benefit to human subjects) and/or any unanticipated developments within the research should be brought to the attention of the Office of Research Services.

Best wishes for the successful completion of your project.

Yours sincerely,

Tom Fleming
Ethics Review Officer

TF/dn
Enclosure

cc: Dr. L. Yeung fax: 813-6531
    Hospital for Sick Children
    Dr. H. O’Brodovich

Simcoe Hall, 27 King’s College Circle, Room 10A, Toronto Ontario M5S 1A1
Telephone 416/ 946-3389  Fax 416/ 946-5763  email tom.fleming@utoronto.ca
August 29, 2000

UHN # 00-0295-U

Dr. Latifa Yeung
Hospital for Sick Children
555 University Avenue
Paediatrics Department
Black Wing, 8th Floor
Toronto, Ontario
M5G 1X8

Dear Dr. Yeung:

The chart review protocol entitled "Assembling a Cohort of Women with Hepatitis C in Ontario" has been reviewed by the University Health Network Research Ethics Board. The proposal is approved from an ethical standpoint for the next 12 months.

If, during the course of the research, there are any serious adverse events, substantial changes in the approved protocol or any new information or developments which must be considered with respect to the study, these should be brought to the attention of the Board.

Yours sincerely,

Ronald Heslegrave, Ph.D.
Chair, University Health Network Research Ethics Board

RH/bh

28 August, 2000
Date of Approval

28 August, 2001
Expiry Date of Protocol
PROTOCOL REFERENCE #128/2000

September 8, 2000

Latifa Yeung, MD, FRCPC
Division of Gastroenterology and Nutrition
Department of Paediatrics
The Hospital for Sick Children
555 University Avenue
Toronto, ON M5G 1X8
FAX: 416-813-6531

Dear Dr. Yeung:

Re: Research protocol #128/2000 entitled "Assembling a cohort of women with Hepatitis C in Ontario" by Dr. Latifa Yeung

We are writing to advise you that the Centre for Addiction and Mental Health Research Ethics Board has granted approval to the above-named research protocol (received August 18, 2000).

We understand that consent documents are not applicable, as no subjects will be contacted.

During the course of the research, any significant deviations from the approved protocol (that is, any deviation which would lead to an increase in risk or a decrease in benefit to human subjects) and/or any unanticipated developments within the research should be brought to the attention of the Research Ethics Office.

Best wishes for the successful completion of your project.

Yours sincerely,

Susan Pilon
Executive Officer, Research Ethics Board
Centre for Addiction and Mental Health, ARF site

cc: A. Irani  P. Darby  J. Simpson
    F. Vaccarino  P. Garfinkel

Better understanding, prevention and care
Mieux comprendre - prévenir - soigner

** TOTAL PAGE .02 **
RESEARCH ETHICS BOARD

May 9, 2001

Dr. Latifa Yeung
Division of Gastroenterology & Nutrition
The Hospital for Sick Children

Dear Dr. Yeung,

Your study Assembling a Cohort of Women with Hepatitis C in Ontario

REB File No. 2000/324

On behalf of the Research Ethics Board, I am writing to provide approval for the above noted study until September 2001.

Yours sincerely

Max Perlman
Chair, Research Ethics Board
Appendix 4: Feasibility Study: SAS Programming
## Feasibility Study

<table>
<thead>
<tr>
<th>Dataset</th>
<th># observations</th>
<th># variables</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data.96b</td>
<td>977041</td>
<td>7</td>
<td>Original data</td>
</tr>
<tr>
<td>Data.ofh96b</td>
<td>977041</td>
<td>17</td>
<td>Added combos 1-10</td>
</tr>
<tr>
<td>Data.some</td>
<td>975860</td>
<td>17</td>
<td>Limit to feesuf=A</td>
</tr>
<tr>
<td>Data.someemd</td>
<td>783734</td>
<td>17</td>
<td>Limit to spec 00 13 41</td>
</tr>
<tr>
<td>Data.nodbs</td>
<td>685929</td>
<td>17</td>
<td>Each line = unique DATE, specialist, combo</td>
</tr>
<tr>
<td></td>
<td>685929</td>
<td>18</td>
<td>Added cons (0 if seen by fmd, 1 if ever seen by specialist)</td>
</tr>
<tr>
<td></td>
<td>685929</td>
<td>23</td>
<td>Cons2, cons3, cons4, cons5, cons6 (=1 if seen by specialist for combo; 0 if never)</td>
</tr>
<tr>
<td>Data.oneikn</td>
<td>385838</td>
<td>16</td>
<td>Added tx= total #times combo per pt</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Tx&gt;0 implies combo assigned at least once</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>T1 t2 t3 t4 t5 t6 t7 t8</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Tcons=0 means pt only s/b fmd</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Tcons&gt;0 shows #times pt saw specialist</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Tconsx&gt;0 shows #times pt saw specialist</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>For combo</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>tcons2 tcons3 tcons4 tcons5 tcons6</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>freq shows # entries per ikn</td>
</tr>
<tr>
<td>Work.justfmd</td>
<td>358255</td>
<td>16</td>
<td>385838-358255=27583 (7%) s/b spec</td>
</tr>
<tr>
<td>Data.refcombo</td>
<td>27583</td>
<td>16</td>
<td>Limit to tcons &gt;= 1 (has been s/b spec)</td>
</tr>
<tr>
<td>Data.count</td>
<td>27583</td>
<td>24</td>
<td>Create BINARY cx</td>
</tr>
<tr>
<td>data.cnoneikn</td>
<td>385838</td>
<td>24</td>
<td>Add BINARY cx (no subsetting to spec/fmd)</td>
</tr>
<tr>
<td>work.druggrp</td>
<td>385838</td>
<td>29</td>
<td>Dhb dh dc do dho</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Drug also given liver dx (5%)</td>
</tr>
<tr>
<td>work.virgrp</td>
<td>385838</td>
<td>29</td>
<td>Vhb vh vc vo vheo</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Viral dx also given liver dx (1%)</td>
</tr>
<tr>
<td>Rpdg.rpdbdemo</td>
<td>13553800</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data.link</td>
<td>385838</td>
<td>28</td>
<td></td>
</tr>
<tr>
<td>Work.women</td>
<td>201560</td>
<td>28</td>
<td>Limit to female</td>
</tr>
<tr>
<td>Work.final</td>
<td>201560</td>
<td>31</td>
<td>Calculate age in years</td>
</tr>
<tr>
<td>Data.final</td>
<td>121187</td>
<td>31</td>
<td>Limit to yrage between 15-49</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Link.sas**

**Finalsum.sas**
This file will read in OHIP fiscal 96 data for viral hep and liver biopsy */

%include '/moh/macos/getohip.sas';

%name data '~/data';

tohip{
    start=199604,
    end=199703,
    subset=%str(
        if dxcode in: ('070','571','573','304','079') or
        feecode in ('Z551','Z554');
    ),
    dir=data,
    out=ohip96,
    keep=jname servdate feecode dxcode feesuff spec totpaid
}
libname data '-/data';
/* classify billings into different combo */
1. take original data ohip96 97041 obs, 7 vars
2. form data ohip96b - includes combos 977041 obs 17 vars /*
data set data shipt96b;
if dxcode=:'070' and fee code in ('2551', '2554') then combo1=1; else combo1=0;
if dxcode=:'071' then combo2=1; else combo2=0;
if dxcode=:'573' then combo4=1; else combo4=0;
if dxcode=:'079' then combo6=1; else combo6=0;
if dxcode=:'571' and fee code in ('2551', '2554') then combo7=1; else combo7=0;
if dxcode=:'573' and fee code in ('2551', '2554') then combo9=1; else combo9=0;
if dxcode=:'079' and fee code in ('2551', '2554') then combo10=1; else combo10=0;
run;
/* majority of billings for pertinent codes are limited to 3 spec groups */
data set data somemd;
if spec='00' or spec='13' or spec='41';
run;
/* keep spec: cons=0 = only seen by fmd; cons>1: if ever seen by int med/GI */
data set data nodbls;
if cons=0 or cons>1 then cons=1; else cons=0;
run;
/* pick out the combos 2-6 which are associated with a specialist */
data set data nodbls;
if (combo2>0 and cons>0) then cons2=1; else cons2=0;
if (combo3>0 and cons>0) then cons3=1; else cons3=0;
if (combo4>0 and cons>0) then cons4=1; else cons4=0;
if (combo5>0 and cons>0) then cons5=1; else cons5=0;
if (combo6>0 and cons>0) then cons6=1; else cons6=0;
run;
/* next step will yield the SMALLEST DATASET so far on oneikn shrinks dataset so each line = one ikn */
libname data '-/data';
run;
/* thus far, haven't removed erroneous duplicate entries */
libname data '-/data';
proc sort data=data.somemd;
by ikn servdate combo1 combo2 combo3 combo4 combo5 combo6 combo7 combo8 spec cons;
run;
/* delete multiple entries for the same encounter */
run;
set data.shipt96b;
in the following only deletes subsequent entries if the listed variables all match the resulting data.nodb1s has 685929 obs and 17 variables, less entries than if we merely deleted entries where totpaid<5 */
data set data.somemd;
by ikn servdate combo1 combo2 combo3 combo4 combo5 combo6 combo7 combo8 spec cons;
run;
/* let's repeat above counts of combos by spec: similar conclusions */
/* keep spec: cons=0 = only seen by fmd; cons>1: if ever seen by int med/GI */
data set data nodbls;
if cons=0 or cons>1 then cons=1; else cons=0;
run;
/* pick out the combos 2-6 which are associated with a specialist */
data set data nodbls;
if (combo2>0 and cons>0) then cons2=1; else cons2=0;
if (combo3>0 and cons>0) then cons3=1; else cons3=0;
if (combo4>0 and cons>0) then cons4=1; else cons4=0;
if (combo5>0 and cons>0) then cons5=1; else cons5=0;
if (combo6>0 and cons>0) then cons6=1; else cons6=0;
run;
/* next step will yield the SMALLEST DATASET so far oneikn shrinks dataset so each line = one ikn */
TX = total #times a combo was assigned to a patient
TX > 0 represents that a given combo was assigned at least once to that patient

no frequency of visits noted, since this info difficult to interpret
tcons=0 means patient only seen by MD
tcons=0 shows #times patient saw a specialist (Int Med/Liver)
tconsx=0 shows #times patient saw a specialist for boxcom

data oneikn has 385388 obs and 16 vars
_freq will represent #separate entries per ikn */

proc means sum noprint data=data nodbiss:
by ikn;
var combo combo2 combo3 combo4 combo5 combo6 combo7 combo8 cons cons2 cons3 cons s4 cons5 cons6;
output out=data oneikn
sum=c1 t2 t3 t4 t5 t6 t7 t8 tcons tcons2 tcons3 tcons4 tcons5 tcons6;
run;
data;
set data oneikn;
drop _type_;
run;

/* re-summarize data. now that each line = one ikn
calculate total #times combos are hit
will see that combos 6 & 7 (other viral & drug) still most popular...
maybe should limit these to codes only to specialists? KEEP IN MIND... */
libname data '.' data;
proc means sum data=data oneikn;
var t1 t2 t3 t4 t5 t6 t7 t8 tcons tcons2 tcons3 tcons4 tcons5 tcons6;
title 'A. #times a combo was invoked amongst our patients (each line=1 ikn)';
runchsum_cl t2 t3 t4 t5 t6 t7 t8 tcons tcons2 tcons3 tcons4 tcons5 tcons6;
run;

/* how much redundancy in the system? That is, if seen by MD then also by
specialist then can restrict to specialist entries only
i.e. ASSUME that any cases of hep C get referred for consultation
find that just from had 358255 obs and 16 vars
i.e. 385388-358255 = 27133 (7%) pts seen by specialist
which still gives a lot of hits, much more than 1200 women */
data;
set data oneikn;
if tcons=0;
run;
proc means sum data=justmd:
var t1 t2 t3 t4 t5 t6 t7 t8 tcons tcons2 tcons3 tcons4 tcons5 tcons6;
title 'B. total #times a combo was invoked amongst patients seen only by MD';
runchsum_cl t2 t3 t4 t5 t6 t7 t8 tcons tcons2 tcons3 tcons4 tcons5 tcons6;
run;

/* so reasonable (optional) to only care if combos2-6 are from specialists
create data refcombo which as data oneikn w/ refined combos (2-6 ONLY from specialists)
note: tconsx no longer =0 */
data;
set data oneikn;
if tcons %%= 1;
run;
proc means sum data=data refcombo;
var t1 t2 t3 t4 t5 t6 t7 t8 tcons tcons2 tcons3 tcons4 tcons5 tcons6;
title 'C. total #times a combo was invoked amongst patients NOT seen only by MD';
runchsum_cl t2 t3 t4 t5 t6 t7 t8 tcons tcons2 tcons3 tcons4 tcons5 tcons6;
run;

/* above 'sum' will total ALL entries. Not count them for combo for that patient
cx refers to at least one dx of combo for that patient
data count = summary data for refcombo, where cx is BINARY */
data;
set data refcombo;
if t1>0 then cl=1; else cl=0;
it2>0 then c2=1; else c2=0;
it3>0 then c3=1; else c3=0;
it4>0 then c4=1; else c4=0;
it5>0 then c5=1; else c5=0;
it6>0 then c6=1; else c6=0;
it7>0 then c7=1; else c7=0;
it8>0 then c8=1; else c8=0;
run;
proc means sum data=data count;
var c1 c2 c3 c4 c5 c6 c7 c8 ;
title 'D. count #times a combo occurred for an ikn';
runchsum_cl c1 c2 c3 c4 c5 c6 c7 c8;
run;

/* what if we DIDN'T refine combos 2-6 and included MD-related combos 2-6?
repeat above WITHOUT refined combos
i.e. stick with oneikn
given the 2190 viral heps pulled (WITHOUT refined codes) this looks better */
data;
set data oneikn;
if t1>0 then cl=1; else cl=0;
it2>0 then c2=1; else c2=0;
it3>0 then c3=1; else c3=0;
it4>0 then c4=1; else c4=0;
it5>0 then c5=1; else c5=0;
it6>0 then c6=1; else c6=0;
it7>0 then c7=1; else c7=0;
it8>0 then c8=1; else c8=0;
run;
proc means sum data=data c4 oneikn;
var c1 c2 c3 c4 c5 c6 c7 c8 ;
title 'E. count #times a combo occurred for an ikn - NO REFINED COMBOS';
runchsum_cl c1 c2 c3 c4 c5 c6 c7 c8;
run;

/* How many drug addicts were also described as 070, 571, 573, or had Lbx?
check ikins' with 304 who also had 070, 571, 573, or Lbx */
data;
set data c4 oneikn;
if c5>0 and c7>0 then dbh=1; else dbh=0;
it5>0 and c7>0 then dh=1; else dh=0;
it5>0 and t3>0 then dc=1; else dc=0;
if t5>0 and t4>0 then do = 1; else do=0;
if t5>0 and (t2>0 or t3>0 or t4>0) then dhco=1; else dhco=0;
run;

proc means sum data=druggrp;
var dhb dh dc do dhco;
title1 'F1. total # times drug use also billed as other liver disease ':;
run;

*******************************************************************************
/\ 304 with 070 and Lbx...found 31 hits, out of 57395 entries of 304             
/\ 304 with 070 ...got 1943 hits, out of 57395 entries of 304                   
/\ 304 with 571 ...got 397 hits, out of 57395 entries of 304                   
/\ 104 with 571 ...got 637 hits, out of 57395 entries of 304                   
/\ 304 with any ...got2667 hits, out of 57395 entries of 304 (5%)              
/\ 5% of 304 drug users get subsequent "liver" diagnosis/followup /*
/\ code 304 (combo5) targets largely FMD drug using hepC group other codes may m
iss */

/\ How many other virals were also described as 070, 571, 573, or had Lbx?
count #x's w/079 who also had 070, 571, 573, or Lbx /*

data countx;
set data.ctomeikn;
if t6>0 and t1>0 then vhb = 1; else vhb=0;
if t6>0 and t2>0 then vh = 1; else vh=0;
if t6>0 and t3>0 then vc = 1; else vc=0;
if t6>0 and t4>0 then vco = 1; else vco=0;
if t6>0 and (t2>0 or t3>0 or t4>0) then vhco=1; else vhco=0;
run;

proc means sum data=virgrp;
var vhb vh vc vo vhco;
title1 'G. total # times viral illness also billed as other liver disease ':;
run;

/\ 079 with 070 and Lbx...found 19 hits, out of 264420 entries of 079           
/\ 079 with 070 ...got 2190 hits, out of 264420 entries of 079                 
/\ 079 with 571 ...got 241 hits, out of 264420 entries of 079                 
/\ 079 with 573 ...got 552 hits, out of 264420 entries of 079                 
/\ 079 with any ...got2719 hits, out of 264420 entries of 079 (1%)             
/\ Only 1% of 079 other virals got "liver" diagnosis/followup /*
/\ code 079 (combo6) largely FMD who may code hepC as other viral disease other
codes may miss */
libname data '*/data';
libname rdp /mh/rdpdb;

data icn sex bdate_freq c1 t2 t3 t4 t5 t6 t7 t8 tcons tcons2 tcons3 tcons4 tcons5 tcons6 c1 c2 c3 c4 c5 c6 c7 c8 dthdate dth;
  by icn;
  if ini;
run;

/* limit to women
   has 201560 obs and 28 vars */
data women;
  set data.icn:
  if sex='F';
run;

/* need to limit dataset to those aged<50y on Jan 1, 1991 (i.e. DOB AFTER 01JAN1991)
   and those aged >15y on Dec 31, 1999 (i.e. DOB BEFORE 31DEC1980)
   set today = latest date allowed, 31DEC1984 */
data icn;
  set women;
  today = mdy(12,31,1984);
  age=floor((linck('month'.bdate.tod) -
            (day(tod)+day(bdate))/12)); /* gives age in months */
  yrage=age/12; /* gives age in years */
run;

data icn;
  set work.final:
  if yrage< 44 and yrage>0;
run;

/* data.final represents females aged 15-49y from icn
   121187 obs 31 vars */
proc sort data=data.final:
  by bdate;
run;

/* limit to aged 20-39y (rather than 15-49y)
   need to limit dataset to those aged<40y on Jan 1, 1991 (i.e. DOB AFTER 01JAN1991)
   and those aged >20y on Dec 31, 1999 (i.e. DOB BEFORE 31DEC1979)
   set today = latest date allowed, 31DEC1979 */
data icn;
  set work.final:
  if yrage< 30 and yrage>0:
run;
proc means sum
  data=data rfinal;
  var c1 c2 c3 c4 c5 c6 c7 c8;
  title1 'count #times a combo occurred for an ikm - these are refined combos ';
run:

/* comparing p7 with pl [refined combos vs. not]
using refined combos (=only specialists coding codes 2-6 matter; other codes are only
coded by specialists anyway)
still c1..c6 yields more codes!! (Much more than expected 1200 hep C cases)

BUT to restrict only to specialists would assume that all codes are pretty accur
ate, and would MISS cases of hep C seen only by FMD and not specialist
So no restriction for now, but keep in mind for later the possibility of limitin
g c2-c6 only to specialists
*/
Appendix 5: Testing of Code Accuracy: SAS Programming
Flowchart of Programming

<table>
<thead>
<tr>
<th>Dataset</th>
<th>#obs</th>
<th>#var</th>
<th>description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data.pull</td>
<td>5,586,450</td>
<td>9</td>
<td>from OHIP 1995-2000</td>
</tr>
<tr>
<td>Data.combop</td>
<td>5,586,450</td>
<td>20</td>
<td>apply combos 1-11 to represent algorithms</td>
</tr>
<tr>
<td>Data.noAp</td>
<td>5,582,395</td>
<td>20</td>
<td>Keep if fecode=A / exclude B, C</td>
</tr>
<tr>
<td>Data.somempd</td>
<td>4,555,971</td>
<td>20</td>
<td>Keep if spec = 00, 13, 41</td>
</tr>
<tr>
<td>Nodblesp</td>
<td>4,019,526</td>
<td>20</td>
<td>Delete multiple entries using some encounter</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(same IKN servdate combo1…combo11) introduce cons, cons2…cons6, cons11</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(0=never seen by specialist/non-FMD)</td>
</tr>
<tr>
<td>Data.oneiknp</td>
<td>1,637,384</td>
<td>19</td>
<td>Each line = ikn</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>( \sum ) combo = tx</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>( \sum ) consx = tconsx</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Dataset</th>
<th>#obs</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Data.refcombop</td>
<td>104,065</td>
<td>18</td>
<td>Refined combos</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Limits to cases s/b specialists (at least once)</td>
</tr>
<tr>
<td>Data.countp</td>
<td>104,065</td>
<td>27</td>
<td>BINARY cx</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(If tx&gt;0 then cx=1)</td>
</tr>
<tr>
<td>Justfmdp</td>
<td>1,533,319</td>
<td>18</td>
<td>Limits to cases where tcons=0 (never s/b specialist)</td>
</tr>
<tr>
<td>Data.ctoneiknp</td>
<td>1,637,384</td>
<td>27</td>
<td>All codes kept (no limit to specialists), counted</td>
</tr>
<tr>
<td>Druggrpp</td>
<td>1,637,384</td>
<td>33</td>
<td>Considers drug use → liver dx overlap (8%)</td>
</tr>
<tr>
<td>Virgrpp</td>
<td>1,637,384</td>
<td>33</td>
<td>Considers viral dx → liver dx overlap (3%)</td>
</tr>
<tr>
<td>Rdpd.rpdbdemo</td>
<td>1,355,800</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Dataset</th>
<th># obs</th>
<th># variables</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data.cleanedy</td>
<td>2548</td>
<td>8</td>
<td>Original dataset from D. Deboer</td>
</tr>
<tr>
<td>Data.validikn</td>
<td>2532</td>
<td>8</td>
<td>Eliminated invalid ikn</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Work.agegrps</td>
<td>2532</td>
<td>9</td>
<td>Added agegrps variable</td>
</tr>
<tr>
<td>Data.finalp</td>
<td>521,750</td>
<td>34</td>
<td>Aged 15-49 OHIP set</td>
</tr>
<tr>
<td>Druggrpp</td>
<td>521,750</td>
<td>40</td>
<td>drug use → liver dx overlap (6%)</td>
</tr>
<tr>
<td>Virgrpp</td>
<td>521,750</td>
<td>40</td>
<td>viral dx → liver dx overlap (3%)</td>
</tr>
<tr>
<td>justfmd</td>
<td>493,570</td>
<td>34</td>
<td>tcons=0</td>
</tr>
<tr>
<td>rfinalp</td>
<td>28,180</td>
<td>34</td>
<td>Tcons&gt;0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Dataset</th>
<th># observations</th>
<th># variables</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data.core*§</td>
<td>2532</td>
<td>32</td>
<td>Gold std + test together!!</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Work.calculate</td>
<td>2532</td>
<td>59</td>
<td>Calc boxes &quot;A/B/C&quot; for each test code (e.g. totc2A = box A for combo 2)</td>
</tr>
<tr>
<td>Work.caletble</td>
<td>Varies by site</td>
<td>Varies</td>
<td>Calculate test characteristics</td>
</tr>
</tbody>
</table>

validikn.sas
9500link.sas
9500final1549.sas
validikn.sas
JCM.sas
Codes125.sas
/* This file will read OHIP data
Jan 1, 1995 - March 31, 2000

dxcodes
070 viral hep
571 cirrhosis of liver
573 other diseases of liver
304 drug dependence, drug addiction
079 other viral diseases

feecodes
2551 liver biopsy
2554 liver biopsy
K026 OHCAP
K027 OHCAP */

%include '~/moh/macros/getohip.sas';

libname data '~/data';

%getohip( start=199501,
       end=200003,
       subset=tstrf
       if dxcode in: ('070','571','573','304','079') or
       feecode in: ('2551','2554','K026','K027');
       dir=data,
       out=+pull,
       keep=+kn servdate feecode dxcode feesuff spec totpaid numser
   vphysnum
   );
/* classify billings into different combo */

data data combos;
  if dxcode='270' and feecode in ('2551', '2554') then comb1=1; else comb1=0;
  if dxcode='270' then comb2=1; else comb2=0;
  if dxcode='271' then comb3=1; else comb3=0;
  if dxcode='271' then comb4=1; else comb4=0;
  if dxcode='272' then comb5=1; else comb5=0;
  if dxcode='272' then comb6=1; else comb6=0;
  if dxcode='273' and comb1=1 then comb7=1; else comb7=0;
  if dxcode='273' and comb1=1 then comb7=1; else comb7=0;
  if dxcode='273' then comb8=1; else comb8=0;
  if dxcode='273' then comb8=1; else comb8=0;
  if dxcode='273' then comb9=1; else comb9=0;
  if dxcode='273' then comb9=1; else comb9=0;
  if dxcode='273' then comb10=1; else comb10=0;
  if comb1=1 then comb1=1; else comb1=0;
  if comb2=1 then comb2=1; else comb2=0;
  if comb3=1 then comb3=1; else comb3=0;
  if comb4=1 then comb4=1; else comb4=0;
  if comb5=1 then comb5=1; else comb5=0;
  if comb6=1 then comb6=1; else comb6=0;
  if comb7=1 then comb7=1; else comb7=0;
  if comb8=1 then comb8=1; else comb8=0;
  if comb9=1 then comb9=1; else comb9=0;
  if comb10=1 then comb10=1; else comb10=0;
run;

proc sort by kln feesuf; run;

data data nodisp;
set data combos;
if feesuf = 'A';
run;

/* majority of billings for pertinent codes are limited to 3 spec groups */
data data somedp;
set data nodisp;
if spec='08' or spec='13' or spec='41';
run;

proc freq data=data somedp;
tables spec*comb1;
tables spec*comb2;
tables spec*comb3;
tables spec*comb4;
tables spec*comb5;
tables spec*comb6;
tables spec*comb7;
tables spec*comb8;
tables spec*comb9;
tables spec*comb10;
tables spec*comb11;
run;

/* dropped combos 9 and 10 since no hits */

proc means sum data=data somedp;
var comb1 comb2 comb3 comb4 comb5 comb6 comb7 comb8 comb11;
run;

/* this lets you pick out one kln and look at its data */
data perusep;

/* set data somedp;
if kln = 0000016856 ;
run;
*/

/* thus far, haven't removed erroneous duplicate entries
some lines are redundant (totpaid=0)
try to delete duplicate day entries by diff totpaid but same all else */

libname data '~/data';
proc sort data=data somedp;
by kln servdate comb1 comb2 comb3 comb4 comb5 comb6 comb7 comb8 comb11;
spec descending totpaid;
run;

/* want to delete multiple entries for the same encounter
i.e. each line unique date/specialist/combo
the following only deletes subsequent entries if the listed variables all match
the resulting data.nodisp = 889529 obs and 17 variables.
less entries than if we merely deleted entries where totpaid=5 */
data data nodisp;
set data somedp;
by kln servdate comb1 comb2 comb3 comb4 comb5 comb6 comb7 comb8 comb11 spec;
if first.spec;
run;

/* let's repeat above counts of combos by spec
find not a lot of overall difference in general conclusions */
proc means sum data=data nodisp;
var comb1 comb2 comb3 comb4 comb5 comb6 comb7 comb8 comb11;
run;

/* need to keep spec = cons=0 means only seen by fmd, cons=1 if ever seen by int
ed/OP */
data data nodisp;
set data nodisp;
if spec='11' or spec='41' then cons=1; else cons=0;
run;

/* want to be able to pick out the combos 2-6, 11 that are a/w a specialist
where cons=1 means combbox was associated with a specialist
data.nodisp now = 4019526 obs with 27 vars */
data data nodisp;
set data nodisp;
if (comb2>0 and cons>0) then cons2=1; else cons2=0;
if (comb3>0 and cons>0) then cons3=1; else cons3=0;
if (comb4>0 and cons>0) then cons4=1; else cons4=0;
if (comb5>0 and cons>0) then cons5=1; else cons5=0;
if (comb6>0 and cons>0) then cons6=1; else cons6=0;
if (comb7>0 and cons>0) then cons7=1; else cons7=0;
if (comb8>0 and cons>0) then cons8=1; else cons8=0;
if (comb11>0 and cons>0) then cons11=1; else cons11=0;
run;
/*
the next step will
yield the SMALLEST DATASET so far
oneknp shrinks dataset down such that each line is one ikn
tx = total #times a combo was assigned to a patient
x > 0 represents that a given combo was assigned at least once to that patient
no frequency of visits noted, since this info difficult to interpret
tcons=0 means patient only seen by FMD
tcons=0 shows #times patient saw a specialist (int med/iver)
tcons>0 shows #times patient saw a specialist for combo
data oneknp = 383 obs and 16 vars
_tick will represent separate entries per ikn
*/

/*
proc means sum nopsprint data=data.nodbsp;
by ikn;
var combo combo2 combo3 combo4 combo5 combo6 combo7 combo8 combo1;
cons cons2 cons3 cons4 cons5 cons6 cons11;
output out=data.oneknp;
ttile t2 t3 t4 t5 t6 t7 t8 t11 tcons tcons2 tcons3 tcons4 tcons5 tcons6 tcons11;
run;
data data oneknp;
set data oneknp;
drop _type_
run;
*/

/*
if need to re-summarize data, now that each line represents one ikn
first, calculate total times combos are hit
will see that combo 5 & 6 (other viral & drug) still most popular....
maybe should limit these to codes only to specialists? KEEP IN MIND....
*/

/*
libname data 'C:/data';
proc means sum data=data oneknp;
var t1 t2 t3 t4 t5 t6 t7 t8 t11 tcons tcons2 tcons3 tcons4 tcons5 tcons6 tcons11;
ttitle 'total #times a combo was invoked amongst our patients';
run;
*/

/*
how much redundancy in the system? That is, if seen by fmd and then also by
specialist can only look at specialist entries.
i.e. ASSUME that any cases of hep C get referred for consultation
find that justfndp = 153319 obs 18 vars ; i.e. 1637354/153319 = 0.10465 seen b
by specialist which STILL gives a lot of hits, much more than 1200 women
However, this is a huge assumption to make...
*/

/*
data justfndp;
set data oneknp;
if tcons=0;
run;
*/

/*
proc means sum data justfndp;
var t1 t2 t3 t4 t5 t6 t7 t8 t11 tcons2 tcons3 tcons4 tcons5 tcons6 tcons11;
ttitle 'total #times a combo was invoked amongst patients seen only by FMD';
run;
*/

/*
so if looks pretty reasonable (optional) to only care if combos 2-6, 11 from specialists
create data refcombo which = data oneknp with refined combos = 2-6, 11 only!
specialists
notice that tcons no longer >0;
data data refcombo;
set data oneknp;
if tcons = 2;
run;
*/

/*
above "sum" will total ALL entries. NOT count them
ex refers to (at least one dx) of combo for that patient
data count = summary data for refcombo, where cx is BINARY
*/

/*
data data count;
set data refcombo;
run;
*/

/*
proc means sum data data count;
var t1 t2 t3 t4 t5 t6 t7 t8 t11 tcons2 tcons3 tcons4 tcons5 tcons6 tcons11;
ttitle 'total #times a combo was invoked amongst patients NOT seen only by FMD';
run;
*/

/*
proc means sum data data count;
var t1 t2 t3 c4 c5 c6 c7 c8 c11;
ttitle 'count #times a combo occurred for an ikn';
run;
*/

/*
what if we DIDN'T refine combos 2-6, 11 & kept fmd-related combos 2-6, 11 ??
repeat above WITHOUT refined combos
i.e. stick with oneknp
given the 24564-6537 viral heps (c2)
pulled (WITHOUT refined codes) this looks better
*/
if t7>0 then c7=1; else c7=0;
if t8>0 then c8=1; else c8=0;
if t1>0 then c1=1; else c1=0;
run;

proc means sum data=data.ctoneiknp;
var cl c2 c3 c4 c5 c6 c7 c8 c11:
title 'count #times a combo occurred for an ikn - NO REFINED COMBOS ';
run;

/*How many drug addicts were also described as 070, 571, 573, or had LBx or OHCAP?
count #ikns with 304 who also had 070, 571, 573, or LBx or OHCAP */
data druggrgp;
set data.ctoneiknp;
if t5>0 and t1>0 then dbh = 1; else dbh=0;
if t5>0 and t2>0 then dh = 1; else dh=0;
if t5>0 and t3>0 then dc = 1; else dc=0;
if t5>0 and t4>0 then do = 1; else do=0;
if t5>0 and t11>0 then dp=1; else dp=0;
if t5>0 and (t2>0 or t3>0 or t4>0 or t11>0) then dall=1; else dall=0;
run;

proc means sum data=druggrgp;
var dbh dh dc do dp dall;
run;

/* 304 with 070 and LBx... = 500 hits. out of 266032 entries of 304
304 with 070...15908
304 with 571... 3525
304 with 573... 5162
304 with OHCAP... 174
304 with any... 20419 (6%)
8% of 304 drug users get subsequent "liver" diagnosis/followup code 304 (combo5) targets largely PAD drug using hepC group other codes may miss */

/*How many other virals were also described as 070, 571, 573, or had LBx or OHCAP?
count #ikns with 079 who also had 070, 571, 573, or LBx or OHCAP */
data virgrgp;
set data.ctoneiknp.
if t6>0 and t1>0 then vhb = 1; else vhb=0;
if t6>0 and t2>0 then vh = 1; else vh=0;
if t6>0 and t3>0 then vc = 1; else vc=0;
if t6>0 and t4>0 then vo = 1; else vo=0;
if t6>0 and t11>0 then vp = 1; else vp=0;
if t6>0 and (t2>0 or t3>0 or t4>0 or t11>0) then vall=1; else vall=0;
run;

proc means sum data=virgrgp;
var vhb vh vc vo vp vall;
run;

/*079 with 070 and LBx...got 438 hits. out of 1178488 entries of 079
079 with 070...24564 hits
079 with 571... 2947 hits
079 with 073... 7363 hits
079 with OHCAP... 93 hits
079 with any...30922 hits (3%)
Only 3% of 079 other virals got "liver" diagnosis/followup code 079 (combo5) largely PPM who may code hepC as other viral disease other codes may miss */
program your code here
/* Get summary info on data.finalp: females of ctnonekn aged 15-49y */ 
libname data '*/data';
proc means sum

data=data.finalp;
var cl c2 c3 c4 c5 c6 c7 c8 c11;
title 'count times a combo occurred for an ikn - NO REFINED COMBOS';
run;

/* compared with 9500combos.lst. p.8 section G, c1-c6 have 992+9501=10493 hits already (feasibility estimates total algorithm will pull ~1200 subjects x 5, 5y 1 = 
* females = cx/rcx (from 9500combos.lst p.8 section G)
1/29.7, Generally 1/3 of all hits are females.
2 39% 3 17% 4 26% 5 38% 6 31% 7 25% 8 24.7% */

/* How many drug addicts were also described as 070, 571, 573. OHCAP or had Lbx? count ikn's with 304 who also had 070, 571, 573, OHCAP or Lbx */
data drugrpp;
set data.finalp;
if t5>0 and t1>0 then dhb = 1; else dhb=0;
if t5>0 and t2>0 then dh = 1; else dh=0;
if t5>0 and t6>0 then dc = 1; else dc=0;
if t5>0 and t4>0 then do = 1; else do=0;
if t5>0 and t8>0 then dp = 1; else dp=0;
if t5>0 and t2>0 or t3>0 or t4>0 or t1>0 then dali=1; else dali=0;
run;
proc means sum data=drugrpp;
var dhb dh dc do dp dali;
run;

/* 304 with 070 and Lbx...found 132hits. out of 10209entries of 304
304 with 070 ... 4794 hits
304 with 571 ... 799 hits
304 with 573 ... 1408 hits
304 with OHCAP ... got 71 hits
304 with any ... got 5905hits (6%) 
6% of 304 drug users get subsequent 'liver' diagnosis/followup 
code 304 (combo5) targets largely hepC drug using hepC group other codes may mis 
s */

/* How many other viruses were also described as 070, 571, 573. OHCAP or had Lbx? count ikn's with 079 who also had 070, 571, 573, OHCAP or Lbx */
data virgpp;
set data.finalp;
if t5>0 and t1>0 then vhb = 1; else vhb=0;
if t6>0 and t2>0 then vh = 1; else vh=0;
if t5>0 and t3>0 then vc = 1; else vc=0;
if t6>0 and t4>0 then vo = 1; else vo=0;
if t6>0 and t11>0 then vp = 1; else vp=0;
if t6>0 and t2>0 or t3>0 or t4>0 or t11>0 then vall=1; else vall=0;
run;
proc means sum data=virgpp;
var vhb vh vc vo vp vall;
run;

/* 079 with 070 and Lbx... 138 hits. out of 163296entries of 079
079 with 070 ... 9501 hits
079 with 571 ... 810 hits
079 with 573 ... 2359 hits
079 with OHCAP ... 39 hits
079 with any ... 11491 hits (30%)
Only 30% of 079 other viruses get a 'liver' diagnosis/followup 
code 079 (combo6) largely FMD who may code hepC as other virus disease other cor 
es may miss */

proc means sum data=data.finalp;
var t1 t2 t3 t4 t5 t6 t7 t8 t11
titons1 tconsin4 tconsin5 tconsin6 tconsin11;
title 'total times a combo was invoked amongst our patients';
run;

/* how much redundancy in the system? That is, if seen by fnd and then also by specialist 
then can only look at specialist entries 
I.e. assume that any cases of hep C get referred for consultation. 
find that justfmd had 345055 obs and 33 vars: i.e. hepC hepD hepE t4 t5 t6 t7 t8 t9 t10 
which STILL gives a lot of hits. much more than 1200 women */

data justfmd;
set data.finalp;
if tconsin=0;
run;
proc means sum data=justfmd;
var t1 t2 t3 t4 t5 t6 t7 t8 t11
titons1 tconsin4 tconsin5 tconsin6 tconsin11;
title 'total times a combo was invoked amongst patients seen only by FMD';
run;

/* it looks pretty reasonable (optional) to only care if combo2-6,11 are for 
m specialists 
create data.rfinalp which = data.finalp with refined combos - 2-6,11 ONLY for 
m specialists 
notice that tconsin no longer >0 */

data rfinalp;
set data.finalp;
if tconsin >1;
run;
proc means sum data=rfinalp;
var t1 t2 t3 t4 t5 t6 t7 t8 t11
titons2 tconsin3 tconsin4 tconsin5 tconsin6 tconsin11;
title 'total times a combo was invoked amongst patients NOT seen only by FMD';
/* count total #patients hit with given combos - only refined combos */
proc means sum
   data=refinalp;
var cl c2 c3 c4 c5 c6 c7 c8 cll;
title 'count #times a combo occurred for an ikn - these are refined combos';
run;

/* comparing pl with pl (refined combos vs. not)
using refined combos (only specialists coding codes 2-6,11 matter; other codes
only coded by specialists anyway)
still c2..c6,cll yields cases!! (Much more than expected 1200x6y hep C cases)
BUT to restrict only to specialists would assume that all codes are pretty accurate,
and would MISS cases of hep C seen only by FMD and not specialist
So no restriction for now, but keep in mind for later the possibility of limiting
r2-6,cll only to specialists */
Start with data set from D. Deboer
1. eliminate invalid IKN's
2. summarize resulting dataset
3. link with larger OHIP dataset of combos data.finalp

/*****************************/
libname data 'C:/data';
proc contents data=data.cleanedy;
run;
data data.validikn;
set data.cleanedy;
if valikn
= 'V';
run;
proc means range min max n data=data.validikn;
var by bd cond bdate;
run;
proc sort;
by site;
run;
proc freq data=data.validikn;
tables cond'site;
run;
data aegegrps;
set data.validikn;
if by <50 then aeggrp = 4;
if by <60 and by >49 then aeggrp = 5;
if by <70 and by >59 then aeggrp = 6;
if by <80 and by >69 then aeggrp = 7;
if by <90 and by >79 then aeggrp = 8;
run;
proc freq data=aegegrps;
tables cond'aeggrp;
run;
proc freq data=aegegrps;
tables site=cond'aeggrp;
run;
proc sort data=data.validikn;
by site bdate;
run;
proc sort data=data.finalp;
by site bdate;
run;
proc contents data=data.finalp;
run;
/*****************************/
Here we link ...
How many seen only by FMD?
*************************************************/
data justfmd;
  set data.core;
  if tcons=0;
run;
proc means sum data=justfmd;
  var t1 t2 t3 t4 t5 t6 t7 t8 t11 tcons2 tcons3 tcons4 tcons5 tcons6 tcons11;
title1 'total #times a combo was invoked amongst patients seen only by fmd';
run;  *****************************************************
/*FMD's don't do liver biopsies!!
*****************************************************
How many NOT seen by FMD?
*****************************************************
data speconly;
  set data.core;
  if tcons=1;
run;
proc means sum data=speconly;
  var t1 t2 t3 t4 t5 t6 t7 t8 t11 tcons2 tcons3 tcons4 tcons5 tcons6 tcons11;
title1 'total #times a combo was invoked amongst patients NOT seen by fmd';
run;  *****************************************************
/*Gosh, only 84 t1 hits!! But 5005 t2 hits!!
*****************************************************
* now repeat site analysis on J+M+C (specialists only);

```
/* Site J+N+C - results comparable to all sites data (lose 160 patients)
* ---------------------------------*/
libname data ‘~/data’;
data calculate;
  set data.core;
  if site in ('A', 'M', 'C');
  if cond=1 and c1=1 then c1A=1; else c1A=0;
  if cond=1 and c2=1 then c2A=1; else c2A=0;
  if cond=1 and c3=1 then c3A=1; else c3A=0;
  if cond=1 and c4=1 then c4A=1; else c4A=0;
  if cond=1 and c5=1 then c5A=1; else c5A=0;
  if cond=1 and c6=1 then c6A=1; else c6A=0;
  if cond=1 and c7=1 then c7A=1; else c7A=0;
  if cond=1 and c8=1 then c8A=1; else c8A=0;
  if cond=1 and c11=1 then c11A=1; else c11A=0;
  if cond=1 and c1=0 and c1A=0; else c1A=0;
  if cond=1 and c2=0 and c2A=0; else c2A=0;
  if cond=1 and c3=0 and c3A=0; else c3A=0;
  if cond=1 and c4=0 and c4A=0; else c4A=0;
  if cond=1 and c5=0 and c5A=0; else c5A=0;
  if cond=1 and c6=0 and c6A=0; else c6A=0;
  if cond=1 and c7=0 and c7A=0; else c7A=0;
  if cond=1 and c8=0 and c8A=0; else c8A=0;
  if cond=1 and c11=0 and c11A=0; else c11A=0;
  proc means sum data=calculate noprint;
  var c1A c1B c1C
c2A c2B c2C
c3A c3B c3C
c4A c4B c4C
c5A c5B c5C
c6A c6B c6C
c7A c7B c7C
c8A c8B c8C
c11A c11B c11C
output out=calctble
sum=A1 B1 C1
A2 B2 C2
A3 B3 C3
A4 B4 C4
A5 B5 C5
A6 B6 C6
A7 B7 C7
A8 B8 C8
A11 B11 C11;
run;
```

```
data calctble;
  set calctble;
  combo=1; A=A1; B=B1; C=C1; output;
  combo=2; A=A2; B=B2; C=C2; output;
  combo=3; A=A3; B=B3; C=C3; output;
  combo=4; A=A4; B=B4; C=C4; output;
  combo=5; A=A5; B=B5; C=C5; output;
  combo=6; A=A6; B=B6; C=C6; output;
  combo=7; A=A7; B=B7; C=C7; output;
  combo=8; A=A8; B=B8; C=C8; output;
  combo=9; A=A9; B=B9; C=C9; output;
  combo=10; A=A10; B=B10; C=C10; output;
  drop _type_ _freq_ A1 A2 A3 A4 A5 A6 A7 A8 A9 A10 B1 B2 B3 B4 B5 B6 B7 B8 B9 B10 C1 C2 C3 C4 C5 C6 C7 C8 C9 C10;
run;
```

```
data calctble;
  set calctble;
  NO= 2372;
  D = NO-A-B-C;
  TP = A/NO;
  FN = B/NO;
  FP = C/NO;
  TN = (NO-P)/D;
  P = TP/FN;
  Q = TP/FP;
  PZ = FP/FN;
  QZ = FN/FP;
  SP = TP/(TP+FP);
  LLSP = SP - 1.96*sqrt((VARSP)/D);
  USLP = SP + 1.96*sqrt((VARSP)/D);
  LLPP = PP - 1.96*sqrt((VARPP)/(1-PP));
  USPP = PP + 1.96*sqrt((VARPP)/(1-PP));
  VARPV = PP*FF/(1-PP);...
```
proc print; run;
/******************************
Limit test=1 to pts w/more than once for any code - 2372 obs kept
*******************************/
data multicode;
  set data.core;
  if site in ('J', 'M', 'C');
  if cond1 and t1>1 then c1A=1; else c1A=0;
  if cond1 and t2>1 then c2A=1; else c2A=0;
  if cond1 and t3>1 then c3A=1; else c3A=0;
  if cond1 and t4>1 then c4A=1; else c4A=0;
  if cond1 and t5>1 then c5A=1; else c5A=0;
  if cond1 and t6>1 then c6A=1; else c6A=0;
  if cond1 and t7>1 then c7A=1; else c7A=0;
  if cond1 and t8>1 then c8A=1; else c8A=0;
  if cond1 and t11>1 then c11A=1; else c11A=0;
  if cond1 and t1 not t1 then c1B=1; else c1B=0;
  if cond1 and t2 not t1 then c2B=1; else c2B=0;
  if cond1 and t3 not t1 then c3B=1; else c3B=0;
  if cond1 and t4 not t1 then c4B=1; else c4B=0;
  if cond1 and t5 not t1 then c5B=1; else c5B=0;
  if cond1 and t6 not t1 then c6B=1; else c6B=0;
  if cond1 and t7 not t1 then c7B=1; else c7B=0;
  if cond1 and t8 not t1 then c8B=1; else c8B=0;
  if cond1 and t11 not t1 then c11B=1; else c11B=0;
  if cond0 and t1>1 then c1C=1; else c1C=0;
  if cond0 and t2>1 then c2C=1; else c2C=0;
  if cond0 and t3>1 then c3C=1; else c3C=0;
  if cond0 and t4>1 then c4C=1; else c4C=0;
  if cond0 and t5>1 then c5C=1; else c5C=0;
  if cond0 and t6>1 then c6C=1; else c6C=0;
  if cond0 and t7>1 then c7C=1; else c7C=0;
  if cond0 and t8>1 then c8C=1; else c8C=0;
  if cond0 and t11>1 then c11C=1; else c11C=0;
run;
proc means sum data=multcode noprint:
  var c1A c1B c1C
c2A c2B c2C
c3A c3B c3C
c4A c4B c4C
c5A c5B c5C
c6A c6B c6C
c7A c7B c7C
c8A c8B c8C
c11A c11B c11C
;output out=calctble
sum=A1 B1 C1
A2 B2 C2
A3 B3 C3
A4 B4 C4
A5 B5 C5
A6 B6 C6
A7 B7 C7
A8 B8 C8
A11 B11 C11
run;
now repeat site analysis on J+M+C (specialists only);
last revised: May 15/01 - added details to genmod;

Site JMC - results comparable to all sites data (lose 160 patients)

need to flag whether patient ever had Lbx
left with combo 2, 3, 4, 5, 6 /< Lbx

******************************************************************************
libname data1:/'data';
data calculate;
set data.core ;
if site in ('J','M','C');
if c1>0 or c7>0 or c8>0 then Lbx=1; else Lbx=0;
run;

proc genmod; model cond = c2 c3 c4 c5 c6 Lbx / dist=binomial link=logit type3; run;
proc genmod; model cond = c2 / dist=binomial link=logit type3; run;
proc genmod; model cond = c3 / dist=binomial link=logit type3; run;
proc genmod; model cond = c4 / dist=binomial link=logit type3; run;
proc genmod; model cond = c5 / dist=binomial link=logit type3; run;
proc genmod; model cond = c6 / dist=binomial link=logit type3; run;
proc genmod; model cond = c2 Lbx / dist=binomial link=logit type3; run;
proc genmod; model cond = c3 Lbx / dist=binomial link=logit type3; run;
proc genmod; model cond = c4 Lbx / dist=binomial link=logit type3; run;
proc genmod; model cond = c5 Lbx / dist=binomial link=logit type3; run;
proc genmod; model cond = c6 Lbx / dist=binomial link=logit type3; run;
proc genmod; model cond = t2 t3 t4 t5 t6 Lbx / dist=binomial link=logit type3; run;
proc genmod; model cond = t2 / dist=binomial link=logit type3; run;
proc genmod; model cond = t3 / dist=binomial link=logit type3; run;
proc genmod; model cond = t4 / dist=binomial link=logit type3; run;
proc genmod; model cond = t5 / dist=binomial link=logit type3; run;
proc genmod; model cond = t6 / dist=binomial link=logit type3; run;
proc genmod; model cond = t2 Lbx / dist=binomial link=logit type3; run;
proc genmod; model cond = t3 Lbx / dist=binomial link=logit type3; run;
proc genmod; model cond = t4 Lbx / dist=binomial link=logit type3; run;
proc genmod; model cond = t5 Lbx / dist=binomial link=logit type3; run;
proc genmod; model cond = t6 Lbx / dist=binomial link=logit type3; run;

******************************************************************************
Comparison of c vs. t variables show better LL with c's
C3 and C4 NS in complete model
try simplifying model
******************************************************************************
proc genmod; model cond = c2 c4 c5 c6 Lbx / dist=binomial link=logit type3; run;
proc genmod; model cond = c2 c5 c6 Lbx / dist=binomial link=logit type3; run;
proc genmod; model cond = c2 c6 / dist=binomial link=logit type3; run;
proc genmod; model cond = c2 / dist=binomial link=logit type3; run;

******************************************************************************
Focus on combos 1, 2, 5
******************************************************************************

******************************************************************************
Create new combo combinations:
combo 15n= combo0 AND combo5 aka combo 151
combo 15r= combo0 OR combo5 152
combo 25n= combo0 AND combo5 251
combo 25r= combo2 OR combo5 252
******************************************************************************

******************************************************************************
if c1>0 and c1<1 then c1A=1, else c1A=0;
if c1>0 and c2<1 then c2A=1, else c2A=0;
if c1>0 and c3<1 then c3A=1, else c3A=0;
if c1>0 and c4<1 then c4A=1, else c4A=0;
if c1>0 and c5<1 then c5A=1, else c5A=0;
if c1>0 and c6<1 then c6A=1, else c6A=0;
if c2>0 and c1<1 then c1B=1, else c1B=0;
if c2>0 and c2<1 then c2B=1, else c2B=0;
if c2>0 and c3<1 then c3B=1, else c3B=0;
if c2>0 and c4<1 then c4B=1, else c4B=0;
if c2>0 and c5<1 then c5B=1, else c5B=0;
if c2>0 and c6<1 then c6B=1, else c6B=0;
******************************************************************************

******************************************************************************
proc means sum data=calculate nopy nprint;
VAR c1A c1B c1C
c2A c2B c2C
c5A c5B c5C
c15A c15B c15C
c15RA c15RB c15RC
c25A c25B c25C
c25RA c25RB c25RC

output out=calctable;
sum=A1 B1 C1
    A2 B2 C2
    A5 B5 C5
A15n B15n C15n
A15r B15r C15r
A25n B25n C25n
A25r B25r C25r

run;
proc print data=calcble;
title "site=JCM with combos of 125";
run;

data calcble;
set calcble;
combo1; A=A1; B=B1; C=C1; output;
combo2; A=A2; B=B2; C=C2; output;
combo5; A=A5; B=B5; C=C5; output;
combo151; A=A15n; B=B15n; C=C15n; output;
combo152; A=A15r; B=B15r; C=C15r; output;
combo251; A=A25n; B=B25n; C=C25n; output;
combo252; A=A25r; B=B25r; C=C25r; output;
drop _type_ _freq_ A1 A2     A5 A15n A15r A25n A25r
     B1 B2     B5 B15n B15r B25n B25r
     C1 C2     C5 C15n C15r C25n C25r;
run;

data calcble;
set calcble;
NO= 3372;
D = NO-A-B-C;
TP = A/NO;
FN = B/NO;
FP = C/NO;
TN = D/NO;
P = TP-FN;
P2 = FP*TN;
Q = TP*FP;
Q2 = TN*FP;
SP = TN*P2;
VARSP = (SP* (1-SP))/ (NO*P2);
ULSP = SP + 1.96*sqrt(VARSP);
LLSP = SP - 1.96*sqrt(VARSP);
SE = TP/P;
VARSE = (SE* (1-SE))/ (NO*P);
ULSE = SE + 1.96*sqrt(VARSE);
LLSE = SE - 1.96*sqrt(VARSE);
PPV = TP/Q;
References


27. Par A. Diagnosis and management of chronic hepatitis C. Can J Gastroenterol 2000;14:83B-88B.


