DOES CULTURE AFFECT THE LOCATION OF DEATH AMONG CANCER PATIENTS IN BRITISH COLUMBIA?: A PILOT STUDY UTILIZING STANDARD AND NOVEL STATISTICAL METHODS IN END-OF-LIFE RESEARCH.

by

MICHAEL DAVID REGIER

B.A.R. Hons., Canadian Bible College, 1997
B.Sc., The University College of the Fraser Valley, 2003

A THESIS SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

in

THE FACULTY OF GRADUATE STUDIES

(Statistics)

THE UNIVERSITY OF BRITISH COLUMBIA

August 2005

© Michael David Regier, 2005
Abstract

Objective
Understanding the clinical, cultural, social, demographic and economic landscape that affects end-of-life health service utilization for British Columbian cancer patients is novel work. The location of death was taken as the initial use of health services to investigate. The goals for this thesis are to understand the trends in the place of death from 1997 to 2003 for adult British Columbians that died of cancer, understand which predictive factors of dying out of hospital, develop new indicators of culture, and evaluate the utility of data mining for end-of-life research.

Methods
The subjects were all adults, age 18 years and older who died of cancer, in British Columbia as identified from the death certificates, from 1997 to 2003. Data from the British Columbia Cancer Registry and the British Columbian Vital Statistics were linked. To this the Postal Code Conversion File program was used to associate each subject with a dissemination area. This was used to impute ecological measures of income, ethnicity, language and religion to each patient. The culture indicators were based upon an indicator developed using principal component analysis and k-means.

When considering the utility of linear discriminant analysis, logistic regression, neural networks, classification trees, and nearest neighbours for end-of-life research, a sub set comprised of all adults aged 20 years and older from 1999-2003 was obtained from the previously described data set. Receiver Operating Characteristic Curve, Area Under the Curve (AUC), Misclassification Curve and Hit curve and cross-validation were used.

Conclusions
After all other significant factors have been taken into consideration; people who were more likely to die out of hospital were females, lived in neighbourhoods in a higher income quintile, lived in either the Interior or Northern Health Authorities, lived in communities with less than 500,000 people, died from breast, colorectal, pancreatic or prostate cancer, had longer survival times, and were older. It was found that people associated with the derived the Aboriginal Presence, British Isles Presence and the Americas and Europe ethnic origins cluster and those people associated with the Chinese Presence and the Punjabi Presence mother tongue cluster were less likely to die out of hospital.
# Table of Contents

Abstract ........................................................................................................... ii  
Table of Contents ........................................................................................... iii  
List of Tables .................................................................................................... vii  
List of Figures ................................................................................................... ix  
List of Abbreviations ....................................................................................... xi  
Acknowledgements .......................................................................................... xii  

Chapter 1: Introduction .................................................................................... 1  
Introduction and Background .......................................................................... 1  
Health Outcomes in End-of-Life Research ...................................................... 1  
Use of Administrative Databases in End-of-Life Research .............................. 3  
Data Mining in End-of-Life Research .............................................................. 7  
Cancer, End-Of-Life Research, and Data Mining .......................................... 8  
Appropriateness of Data Mining for End-of-Life Research ......................... 9  
Ensuring Model Generalization ..................................................................... 10  
Purpose ........................................................................................................... 11  
Context ........................................................................................................... 11  
Thesis Structure .............................................................................................. 12  
References ....................................................................................................... 14  

Chapter 2: Manuscript 1 - Assessing the Necessity and Appropriateness of Mixed-Level Ecological Studies in End-of-Life Research ................................................. 23  
Introduction ................................................................................................... 24  
Census Data as an Ecological Measure ........................................................... 24  
Balancing the Two Fallacies ......................................................................... 24  
Understanding the Variables, Study, and Design .......................................... 26  
Types of variables ......................................................................................... 26  
Prescriptive versus Descriptive Construction of Culture indicators .............. 28  
Types of Studies ............................................................................................. 30  
Reasonability of an Ecological Study .............................................................. 32  
Current Situation ............................................................................................ 33  
New Horizons ................................................................................................. 34  
Prescriptive Approaches ............................................................................... 34  
Specification ................................................................................................ 34  
Specification and Quantification ................................................................... 34  
Descriptive Approaches ............................................................................... 34  
Classification Trees to Define Cut-off Values .............................................. 34  
Cluster Algorithms for Indicator Development .......................................... 35  
Conclusion ..................................................................................................... 35  
References ..................................................................................................... 36  

Chapter 3: Manuscript 2 - Trends in the Place of Death of British Columbia Cancer Patients, 1997-2003 ................................................................. 39  
Background .................................................................................................... 40
Chapter 4: Manuscript 3 - Developing a Multivariate Defined Indicator of Culture: A Novel Synthesis of Unsupervised Learning Techniques

Background ......................................................... 57
Methodology .................................................... 59
  Data Source and Population .................................. 59
  Study Variables ................................................. 59
  Data Transformation ........................................... 59
  Statistical Tools – An Overview
     Principal Component Analysis (PCA) ...................... 64
     K-Means ......................................................... 65
Results ................................................................. 66
  Ethnic Origins ................................................... 66
  Mother Tongue .................................................. 73
  Religion ............................................................ 76
Indicator Validation ................................................. 78
Discussion ........................................................... 81
Limitations .......................................................... 82
Conclusions ......................................................... 82
Appendix ............................................................. 83
References ........................................................ 86
List of Tables

Table 1-1: The variables used in this thesis and the source for each variable ...................................................... 5
Table 2-1: Comparison between Susser’s variables and MacIntyre-Duncan spatially based variables .............................. 26
Table 2-2: Table for developing a new nomenclature by which to specify predictor and response variables ....................... 28
Table 3-1: Characteristics of adults in British Columbia who died of cancer from 1997 to 2003, by place of death ............................................. 43
Table 3-2: Crude and adjusted odds of dying out of hospital, by predictors, for British Columbia, 1997-2003 ............................................. 45
Table 3-3: Nova Scotia Groupings for ICD codes. The British Columbia groups were identical to these groups .................. 52
Table 4-1: Conceptual grouping of ethnic origins .............................................................. 60
Table 4-2: Conceptual groupings of mother tongue .......................................................... 61
Table 4-3: Conceptual groupings of religion ........................................................................ 62
Table 4-4: PCA proportion of variance and cumulative proportion of variance for the conceptually grouped ethnic origins census variables ................................. 67
Table 4-5: Statistical descriptions of the population proportions for the six ethnic origin groups expressed by only the minimum value, the median and the maximum value 69
Table 4-6: Percentage of the six ethnic origins clusters in 22 British Columbian municipalities ........................................... 72
Table 4-7: Percentage of the five mother tongue clusters in 22 British Columbian municipalities ............................................ 75
Table 4-8: Percentage of the six religion clusters in 22 British Columbian municipalities ...................................................... 78
Table 4-9: Percentage of clusters for mother tongue groups given the ethnic origin group ............................................. 79
Table 4-10: Percentage of clusters for religion groups given the ethnic origin group ........................................... 80
Table 4-11: Percentage of clusters for religion groups given the mother tongue group ............................................... 80
Table 4-12: Statistical descriptions of the population proportions for the five mother tongue groups expressed by only the minimum value, the median and the maximum value ...................................................... 83
Table 4-13: Statistical descriptions of the population proportions for the six five mother tongue groups expressed by only the minimum value, the median and the maximum value ...................................................... 85
Table 5-1: Advantages and disadvantages specific to each classifier .......................................................... 107
Table 5-2: Misclassification table .................................................................................. 110
Table 5-3: 0-1-c loss function; the truth is the real state (what was observed) and the decision is the decision made by the classifier. “H” indicates death in a hospital. “O” indicates death out of a hospital .......................................................... 110
Table 5-4: Tabular representation of the hit curve where “O” is out of hospital death and “H” is hospital death ...................................................... 114
Table 5-5: $2^3$ design for determining the best set of parameters for a neural network ...................................................... 124
List of Figures

Figure 1-1: Relationship between the databases and census information used in this thesis. The British Columbia Vital Statistics and the British Columbia Cancer Registry are linked and the study subjects are obtained. To this research database, dissemination area information is attributed to each patient. .......................................................... 4

Figure 3-1: Proportion of death out of hospital and in hospital among British Columbian adults who died of cancer, 1997-2003 ........................................................................ 45

Figure 3-2: Percentage of patients who died out of hospital over the study period 1997-2003 by British Columbia Health Region ......................................................... 47

Figure 4-1: Structure of the 2001 Canada Census Ethnic Origins characteristic ............. 60

Figure 4-2: Scree plot for the ethnic origins census PCA .................................................. 65

Figure 4-3: Total within-cluster sum of squares (SS) versus the number of clusters ...... 68

Figure 4-4: Scatter plots of the first five principal components with the centres for the six ethnic origins clusters: Heterogeneous (1), Aboriginal Presence (2), Chinese Presence (3) British Isles Presence (4), East Indian Presence (5) and the Americas and Europe (6) ...................................................................................... 71

Figure 4-5: Scatter plots of the first three principal components with the centres for the five mother tongue clusters: Punjabi Presence (1), Dominant English (2), Heterogeneous with Dominant English (3), Chinese Presence (4), and Strong English Presence (5) ...................................................................................................... 73

Figure 4-6: Scatter plots of the first four principal components with the centres for the six religion clusters ............................................................................................. 76

Figure 5-1: Graph for study set variable completeness ...................................................... 96

Figure 5-2: Example of a one layer architecture for a neural network with three input variables. The input variables are $x_1$, $x_2$, and $x_3$ with a constant input $x_0$. The transformation functions (perceptrons) for the hidden layer are $h_1$ and $h_2$. The output layer has a transformation function $f$. A skip allows $x_0$ to skip over the hidden layer and be one of the weighted inputs to the output layer. .................................................. 101

Figure 5-3: Example of how the classification tree algorithm partitions the explanatory variable space. There are two explanatory variables, $x_1$ and $x_2$. The binary splits are represented by a, b, c, and d .................................................................................. 103

Figure 5-4: The classification tree for Figure 5-3. If the condition is true, go down the left branch. If the condition is false, go down the left branch. “H” indicates death in a hospital. “O” indicates death out of a hospital. The terminal nodes are considered pure. There are only members of one class in each terminal node......................... 104

Figure 5-5: The ROC curve graphically represents the relationship between the sensitivity and the specificity. Line (A) represents a classifier that performs no better than flipping a coin. The associated AUC would be 0.5. Line (B) represents a classifier that performs well and has the ability to separate the two classes. The AUC would be near 1.0 ........................................................................................................ 104

Figure 5-6: The ideal hit curve (A). The curve rises at a 45 degree angle to the number of selected patients axis until it reaches the total number of patients who died out of hospital (n1). After this the curve is horizontal. The curve (B) is what is commonly seen with hit curves. The curve rises more slowly and needs more than n1 patients selected to find the n1 patients who died out of the hospital ........................................ 113
Figure 5-7: The partitioning of data into three sets: Training set for model construction, validation set for model selection, and testing set for error estimation (generalization) ........................................... 117

Figure 5-8: An example of cross-validation illustrated by 10-fold cross-validation .... 120

Figure 5-9: An example of two-layer cross-validation illustrated by 10-fold two layer cross-validation ........................................................................................................ 122

Figure 5-10: Proportion of death out of hospital and in hospital among adults in British Columbia who died of cancer, 1999-2003 ................................................................. 128

Figure 5-11: Receiver Operating Characteristic Curves for linear discriminant analysis, logistic regression, neural networks, classification trees, and nearest neighbours. Neural network has the best ROC curve and nearest neighbours has the worst. Linear discriminant analysis, logistic regression, and classification trees are similar. ........................................................................................................... 129

Figure 5-12: Linear discriminant analysis hit curve ................................................. 131

Figure 5-13: Logistic regression hit curve for the full model .................................. 132

Figure 5-14: Logistic regression hit curve for the parsimonious model ................ 133

Figure 5-15: Neural network hit curve ................................................................... 134

Figure 5-16: Classification tree hit curve ................................................................. 135

Figure 5-17: Nearest neighbour hit curve ............................................................. 136

Figure 5-18: British Columbia Health Authorities .............................................. 156

Figure 5-19: Agreement between the diagnosis of cancer and the cancer cause of death ......................................................................................................................... 157

Figure 5-20: Misclassification curve for linear discriminant analysis model ........ 158

Figure 5-21: Misclassification curve for the full logistic regression model ............ 159

Figure 5-22: Misclassification curve for the parsimonious logistic regression model 160

Figure 5-23: Misclassification curve for neural networks model .......................... 161

Figure 5-24: Misclassification curve for classification trees model ..................... 162

Figure 5-25: Misclassification curve for the nearest neighbours model ............... 163
# List of Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arg</td>
<td>Argument</td>
</tr>
<tr>
<td>AUC</td>
<td>Area Under the Curve</td>
</tr>
<tr>
<td>BC</td>
<td>British Columbia</td>
</tr>
<tr>
<td>BCCA</td>
<td>British Columbia Cancer Agency</td>
</tr>
<tr>
<td>BCCR</td>
<td>British Columbia Cancer Registry</td>
</tr>
<tr>
<td>CI</td>
<td>Confidence Interval</td>
</tr>
<tr>
<td>CIHR</td>
<td>Canadian Institutes of Health Research</td>
</tr>
<tr>
<td>CT</td>
<td>Classification Tree</td>
</tr>
<tr>
<td>CV</td>
<td>Cross-Validation</td>
</tr>
<tr>
<td>DA</td>
<td>Dissemination Area</td>
</tr>
<tr>
<td>DCI</td>
<td>Death Certificate Identified</td>
</tr>
<tr>
<td>DCO</td>
<td>Death Certificate Only</td>
</tr>
<tr>
<td>Err</td>
<td>Error</td>
</tr>
<tr>
<td>ICD-10</td>
<td>International Classification of Diseases: Tenth Revision</td>
</tr>
<tr>
<td>ICD-9</td>
<td>International Classification of Diseases: Ninth Revision</td>
</tr>
<tr>
<td>ICD-O</td>
<td>International Classification of Disease for Oncology</td>
</tr>
<tr>
<td>IPPE</td>
<td>Income Per Person Equivalent</td>
</tr>
<tr>
<td>KNN</td>
<td>K Nearest Neighbours</td>
</tr>
<tr>
<td>LDA</td>
<td>Linear Discriminant Analysis</td>
</tr>
<tr>
<td>MSE</td>
<td>Mean Squared Error</td>
</tr>
<tr>
<td>NAACCR</td>
<td>North American Association of Central Cancer Registries</td>
</tr>
<tr>
<td>NET</td>
<td>New Emerging Team</td>
</tr>
<tr>
<td>NN</td>
<td>Neural Networks</td>
</tr>
<tr>
<td>NS</td>
<td>Nova Scotia</td>
</tr>
<tr>
<td>OR</td>
<td>Odds Ratio</td>
</tr>
<tr>
<td>PCA</td>
<td>Principal Component Analysis</td>
</tr>
<tr>
<td>PCCF</td>
<td>Postal Code Conversion File</td>
</tr>
<tr>
<td>PHSA</td>
<td>Provincial Health Services Authority</td>
</tr>
<tr>
<td>ROC</td>
<td>Receiver Operating Characteristic</td>
</tr>
<tr>
<td>SDEC</td>
<td>Socio-demographic and economic characteristics</td>
</tr>
<tr>
<td>SE</td>
<td>Standard Error</td>
</tr>
<tr>
<td>SES</td>
<td>Socioeconomic Status</td>
</tr>
<tr>
<td>VS</td>
<td>British Columbia Vital Statistics</td>
</tr>
</tbody>
</table>
Acknowledgements

I would like to thank the Sociobehavioural Research Center at the British Columbia Cancer Agency for supporting me in the completion of this thesis. The professional, moral and financial support has been an asset to do the work that was set before me.

I would like to acknowledge both my supervisors, Dr. Ruben Zamar and Dr. Maria Cristina Barroetavena for their support. I would like to thank Dr. R. Zamar whose help, stimulating suggestions and encouragement have been invaluable as I have been working on this thesis. I would like to thank Dr. M.C. Barroetavena whose mentorship in the area of epidemiology, stimulating suggestions and encouragement have been invaluable to this investigation. I am indebted to both Dr. R. Zamar and Dr. M.C. Barroetavena.

Finally, I would like to thank my wife Christine, who read through many drafts, spent many weekends at home as I worked, and who encouraged me throughout the past two years.
Chapter 1: Introduction

Introduction and Background

I was fourteen when I attended my first funeral. It was for my grandfather. My family did not talk about his illness. They did not talk about death. One day the phone rang. I answered. He had died in the hospital. I told my parents when they arrived at home. They left immediately. I was left alone.

Death is inevitable. Some people see it as a point of transition from one form of existence to another. For others it is the terminus. Some are surrounded by family and friends. Others die alone. Some people desire to involve their family and friends as they move towards it. Others seek a solitary path. Regardless of the choices that are made, death is the one thing all people must face.

Research may never be able to pull back the veils of mystery that surround death, but it will help researchers understand the choices people make at the end of their life. Questions that are of interest include:

- whether culture play a role in the way people understand death;
- whether income affect the quality of life as one prepares for death;
- whether the place of residence determine the resources that will be utilized during the end-of-life period.

Answers to these and other questions will assist health care providers in enhancing the quality of end-of-life care for people.

This thesis focuses upon the theme of end-of-life health service utilization. As this is novel work for British Columbia, the point of origin was to investigate the location of death of cancer patients in British Columbia. By establishing firm research foundations, further investigations can occur that will help epidemiologists and social scientists to understand the clinical, cultural and economic landscape that affects end-of-life health service utilization. By identifying the patterns of health service utilization among vulnerable populations, those marginalized from the health care system, interventions can be developed and implemented, through programs and legislation that will ensure equal access to and use of health services at the end-of-life for all British Columbians.

Health Outcomes in End-of-Life Research

Death can occur in a variety of places, yet for many it occurs in a hospital, care facility or the home. Current research indicates that if a patient is given a choice, up to 88% (Connecticut, United States) of patients dying of cancer would choose to die at home, yet only a smaller proportion realizes this preference. If congruence between the desired place of death and the actual place of death is to occur, end-of-life services must be established that would enable people to die where they choose. For the purpose of this research, the location of death, defined as “In Hospital” and “Out of Hospital”, is investigated in order to provide a foundation of research upon which further investigations in to the location of death of cancer patients.
In Canada, it is estimated that 69,500 deaths due to cancer will occur in 2005. For women and men, lung cancer will be the leading cause of death. Approximately 8,700 deaths due to cancer will occur in BC in 2005. Of the top three cancer causes of death for men and women, the majority of deaths correspond to people aged 50 years and older.

In Nova Scotia (NS) Canada, out of hospital deaths were 25.4% the period from 1992 to 1997. This difference may have as much to do with how health services are delivered as with individual choice. Out of hospital deaths ranged from 19.8% in 1992 to 30.2% in 1997. In Ontario, out-of-hospital deaths have ranged from 31% in 1980 to 21.5% in 1994 to 34% in 1997. In Manitoba, 57.4% of cancer patients die in the hospital during the fiscal year of 2000/01. In Vancouver, approximately 15% died in non-institutional settings from 1990 to 1992 (home or other locations), yet in 1993 this number increased to just under 30%.

It is not only important to understand where people die, but the predictors for the location of death are of great interest in understanding the underlying factors that may influence where a person dies. In the United States, much emphasis is placed upon conventional measures of socioeconomic status (SES) while the sociodemographic aspect is generally limited to race, age, gender and marital status. The consideration of cultural context extends mostly to region of residence which is summarised in a rural versus urban indicator. Similar themes emerge when considering international research into the cultural determinants of the utilization of health services and the location of death for persons dying of cancer.

Research upon populations from Toronto and Nova Scotia presents further challenges as it suggests there is no SES connection to these health outcomes. Roos and Veugelers imply that the lack of association between health outcomes and SES is due to Canada's universal health care system. If this is true, then the conventional measures of SES (e.g. income, education, and occupation) may not be discriminants of health outcomes. Health Canada claims that “universal coverage for medically necessary health care services [is] provided on the basis of need, rather than the ability to pay” thus economic related measures of SES may not capture any systematic marginalization of peoples from health services.

By investigating the indicators used to understand the non-clinical factors that are associated to the use of health services and the place of death within cancer research, it became apparent that international comparisons must be made within the context of the health system in place in each of the countries being considered. This suggests that indicators that are relevant in the United States and other countries with different health care systems may not have the same influence in Canada. Since there is not a consensus on the role that SES measures play in understanding health outcomes, perhaps there is another mechanism which creates vulnerable populations within Canada – culture.

British Columbia is a microcosm of the Canadian mosaic. Many different ethnic groups, such as the Chinese, South-Asian, Dutch, English, and German, are found within the provincial borders. The Department of Canadian Heritage states that “Multiculturalism
ensures that all citizens can keep their identities, can take pride in their ancestry and have a sense of belonging” 36. Furthermore, people are “free to choose for themselves, without penalty, whether they want to identify with their specific group or not” 36. By incorporating the spirit of multiculturalism, this research has posited a novel approach for defining dissemination area derived indicators of culture. By utilizing the multivariate aspect of the census categories, a descriptive approach to the construction of cultural indicators has been developed and constitutes one of the four manuscripts of this thesis. The motivation for this development is given in the first manuscript where the difference between the current prescriptive approach 10 and the proposed descriptive approach is discussed. The current methodology suggests that culture fits into a predetermined structure or identity whereas this new approach suggests that culture is defined by a complex web of relationships. Grouping dissemination areas based upon these webs of interconnectivity allows for a shift from modern to post-modern understandings of culture. In doing so, this research will be poised to capture the complex interplay of cultures that constitutes the mosaic of Canadian society.

**Use of Administrative Databases in End-of-Life Research**

Administrative databases represent an efficient means by which a sketch of a population’s health can be obtained. They can be accessed with relative ease and the data is typically available within reasonable time period and the associated costs are manageable. A disadvantage to using an administrative database is that the information desired from the databases does not naturally extend from the underlying motivation for collecting the information. As well, the data sets obtained are often very large in terms of observations, variables or both. For the purposes of this research, an end-of-life database was constructed. The British Columbia Cancer Registry (BCCR) and British Columbia Vital Statistics (VS) databases were linked by to add information on place of death for each study subject (Figure 1-1).
The BC Cancer Registry, having been in existence from 1969, and maintained by the BC Cancer Agency (BCCA) since 1980, collects and generates cancer statistics on the BC population \(^3\). This database is used for cancer surveillance and research, which provides information on the magnitude of the cancer problem within BC, assists in the development and monitoring of programs for reduced mortality and morbidity, and provides information for future planning \(^3\). By providing a population based registry, the BCCA has access to a source of information which avoids a source of bias due to non-representative participation. This translates into better quality data than that used in non-population-based sources \(^3\). The BC Cancer Registry has attained an overall silver certification status from the North American Association of Central Cancer Registries (NAACCR) \(^5\). The registry has reached gold standard achievement for specific items such as timeliness and death certificate only cases \(^5\).

The original database contained just over 60,000 patients who resided in and died in BC and over 100 fields. From this database, the study sets were obtained. These were smaller in terms of the number of patients and number of fields as the study cohort was further defined by age and cause of death, yet they were still sizable. After ascertaining the completeness and utility of the variables for this investigation, the variables to be used in this research were selected (Table 1-1). A small number of variables were used, but the number of patients was still very large, over 30,000 for the smallest data set.
Table 1-1: The variables used in this thesis and the source for each variable.

<table>
<thead>
<tr>
<th>Study Variables</th>
<th>Units of Measure</th>
<th>Description</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sex</td>
<td>M=Male, F=Female, X=Missing</td>
<td>Indicates the patient’s gender</td>
<td>BCCR</td>
</tr>
<tr>
<td>Age at death</td>
<td>Years</td>
<td>Death date – Birth date</td>
<td>BCCR</td>
</tr>
<tr>
<td>Year of death</td>
<td>Calendar Year</td>
<td>Obtained from Death date</td>
<td>BCCR</td>
</tr>
<tr>
<td>Region of Residence</td>
<td>1=Interior, 2=Fraser, 3=Vancouver Coastal, 4=Vancouver Island, 5=Northern, 6=Provincial Health Authority</td>
<td>Provincial Health Authority excluded as it does not represent a geographic region of BC</td>
<td>VS</td>
</tr>
<tr>
<td>Primary Cause of Death</td>
<td>ICD-9 and ICD 10 groupings</td>
<td>The patient’s primary or underlying cause of death as determined by the BC Vital Statistics Agency or another province/country’s vital statistics department.</td>
<td>BCCR</td>
</tr>
<tr>
<td>Survival time from diagnosis</td>
<td>Days</td>
<td>Death date – Diagnosis date</td>
<td>BCCR</td>
</tr>
<tr>
<td>Place of death</td>
<td>H=Hospital, O=Out of Hospital</td>
<td>The outcome variable for the place of death was obtained from VS. The location of death was derived from a field which indicated hospital death. If a person died in hospital, this was identified with the hospital code. If a person died out of the hospital, there was no code. Place of death was dichotomized as in hospital or out of hospital.</td>
<td>VS</td>
</tr>
<tr>
<td>Rural vs. Urban</td>
<td>R/U</td>
<td>Postal code of the usual place of residence at death will be classified according to the Canada Post designation where the second digit indicates rural or urban. If the number is 1-9, then the mail is destined to an urban</td>
<td>VS</td>
</tr>
</tbody>
</table>
Dissemination level, census based information was associated with each of the study sets using the Statistics Canada Postal Code Conversion File (PCCF) program. The census information was linked with SAS based upon a unique identifier for each dissemination area.

Although there is much information contained in these data sources, the data has been collected for purposes other than the understanding of the location of death for cancer patients. Research based upon a secondary use of administrative or corporate databases bears the characteristics of observational studies. From a classical statistical perspective, observational studies are the most problematic type of study as there is, in general, no reference to any experimental design. It is this lack of reference which puts observational studies in such low regard as “design [is] at the base of statistics” which has strong mathematical support. This lack of design is most commonly seen in non-randomized data, thus observational, non-randomized studies tend to carry lower evidential weight than randomized, experimental ones where the researcher has control over the interventions under study. It is because of this diminished evidential weight that a well-planned and credible method for analyzing data sets of this nature is needed. Data mining has emerged in response to the desire to make sense of the vast amounts of data housed in administrative databases and will be discussed in the following section.
Data Mining in End-of-Life Research

The two branches of data mining are unsupervised and supervised learning. With both, the main objective is to learn about the data and discover the patterns within it. Unsupervised learning is used when only the features of the data are known and there is no expressed outcome variable. Having only the predictor variables, the central objective is to understand how the data is organized based upon the known features. Principal Component Analysis is an example.

Supervised learning has an outcome variable which guides the process. Here, the goal is to understand how the data is organized with respect to both the predictor and the outcome variables. A familiar example of this type of learning is logistic regression.

Over the past decade, data mining techniques have become more common within medical research. As awareness of these techniques has grown the applications have been used to study health related conditions such as:

- medication error and adverse drug events,
- Binswanger's disease,
- anophthalmia and microphthalmia,
- cystic fibrosis,
- myocardial infarction,
- pregnancy outcomes,
- stroke,
- cardiac ischemia,
- meningococcal disease,
- carpal tunnel syndrome,
- survival and mortality,
- cancer,
- schistosomiasis,
- pneumococcal disease,
- critical care,
- and asthma.

A survey of these applications reveals two primary motivations for bringing data mining techniques into medical research: diagnosis and the comparison of how different techniques perform to find a better tool for diagnosis. The first concerns the diagnosis of a disease given a set of patient descriptors. Here, the primary focus is upon the ability to diagnose (or predict) an outcome. The second motivation is to determine the best classifier for a data set when the objective is to predict an outcome. Here, the motivation is to find the best classifier in order to have more reliable diagnosis. Although diagnosis is a goal, it is subordinate finding the best predictor.

Taking the lead from the wider medical literature, a reasonable group of classification techniques have been selected. Linear discriminant analysis, logistic regression, neural networks, classification trees, and nearest neighbours have been chosen as they are common choices within the literature. The final manuscript seeks to compare the
performance of these five classifiers and determine if they are appropriate for end-of-life cancer research.

Cancer, End-Of-Life Research, and Data Mining

The set of articles reviewed represent a search for data mining techniques within cancer research and epidemiology focusing upon research of the last 15 years. Research which used data mining techniques other than logistic regression was sought. Even with these constraints, logistic regression was prevalent.

Outside of the use of logistic regression, a majority of the articles related to cancer diagnosis have utilized neural networks with cluster analysis. When comparing techniques, neural networks are present in almost all the papers reviewed. In these papers, neural networks are compared against linear discriminant analysis, logistic regression, and classification trees, and nearest neighbours.

A diversity of views in relation to the use of data mining techniques was present. When the research focus was diagnosis, the majority of the reviewed articles were supportive of the use of data mining techniques. Some simply used the techniques with no secondary comparison, while others sought to extend the techniques. Baxt was the only article to exhibit any caution about the techniques and viewed the body of current research using neural networks within a medical context as being inconclusive to its usefulness.

When a comparison of techniques was the focus, a majority of the reviewed papers supported the data mining technique citing that the chosen technique outperformed a more standard one, such as logistic regression. The remaining articles viewed these as potentially beneficial techniques, but were cautious about how they should be applied and about the methodology surrounding their assessment.

Although the majority of reviewed articles were enthusiastic about the techniques, the sober judgement of the few was compelling. Without a clear methodology which incorporates appropriate validation and assessment of comparable techniques, it is questionable if a preferred modelling technique can be objectively chosen. Schwarzer et al. conclude their survey of the uses of neural networks within cancer research by stating that, "In our opinion, there is no evidence that artificial neural networks have provided real progress in the field of diagnosis and prognosis in oncology." This sentiment is echoed by Snoer et al. and Dybowski et al.

When narrowing the focus to end-of-life research, the range of techniques used also narrows. Logistic regression features in a majority of the reviewed articles. Of these there is one instance of neural networks, and one instance of classification trees. Ottenbacher compares logistic regression and neural networks and views neural networks as an aid to clinical prognosis, but still prefers logistic regression. Abu-Hanna does not compare classification trees and logistic regression, rather a two stage modelling procedure is used where classification trees are used to identify...
disjoint groups of patient and then logistic regression is used as a local model for these
groups. The implicit view is that the more novel data mining techniques should not
supplant logistic regression, but rather function as an added tool to better understand the
underlying structure of the data.

By reviewing the medical literature for uses of data mining in cancer and end-of-life
research, some trends are apparent.

- Logistic regression is being used as a base model against which the newer or less
  familiar techniques are being compared.
- With logistic regression as the standard of performance, it is often compared with
  one other classifier. Some studies employ a larger battery of classifiers for the
  comparison.
- Neural networks and classification trees are commonly used.
- Many practitioners present a favourable view towards data mining techniques
  while few offer insightful comments about the misuse and poor comparisons of
  these techniques.

**Appropriateness of Data Mining for End-of-Life Research**

In order to predict the location of death for a cancer patient, we need a tool that will help
classify the patient into one of the predetermined categories. Beyond this classification,
we want to determine the set of predictor variables that are most useful in correctly
classifying the patient. It is readily apparent why logistic regression has been widely
used within this research. The outcome of the model is a probability that the person will
die in a certain location and a set of useful explanatory variables. The results from
logistic regression are easily put into an epidemiological framework through the use of
the odds ratio.

The logistic regression models employed in location-of-death end-of-life research
assumes a linear model. No interaction terms are included in the models, nor are any
higher order terms. This linearity may be an overly simplistic representation of the
underlying phenomenon. Although the linear assumption provides simplicity of
explanation, it may not be an accurate representation of the underlying sociological
phenomena. If a researcher chooses to relinquish the preference for a linear model,
then models which provide a high degree of flexibility may be required. This does not
eclipse the use of logistic regression. The need for more complex models suggests that
either more complex logistic regression models should be used or other classification
techniques should be used. As the current research is observational, large administrative
databases to predict a health outcome, data mining techniques would be appropriate.

If a researcher prefers the logistic model, other data mining techniques should not be
excluded from the model development process. They can be used to gain insights which
could then be applied to the logistic model in terms of interactions or higher powers of
predictors. A classifier could be used to find an interesting subset of the data and then
logistic regression used to understand that subset. Furthermore, setting logistic
regression in a classification context will allow researchers to gain an understanding of
how it performs when compared with other techniques. This may be valuable in assessing the epidemiological utility of the model itself.

At this stage of the research, data mining is appropriate for understanding the location of death of cancer patients. Understanding the cultural and clinical predictors for where cancer patients die is predicated upon correctly classifying the place-of-death for British Columbian cancer patients. Once a classifier has been fit and assessed, the explanatory variables used for the classification can be identified and explored.

By expanding the range of perspectives by which classification and exposition occurs, a more complete understanding may emerge. By participating in the comparison of classification techniques, location-of-death for cancer patients will emulate the current research occurring in many areas of cancer research. Within this exploration, caution must be exerted, as novelty must not eclipse sound methodology 80.

**Ensuring Model Generalization**

One aspect of fitting classifiers that bears discussion is the assessment of how a classifier performs. When fitting a classifier the method of assessment must be determined. In this thesis four methods will be used: misclassification error, receiver operating characteristic curve, the area under the receiver operating characteristic curve and the hit curve. These will give an indication of how the fit classifier will perform when an independent data set is used.

The measure of performance associated with an independent set of data is an indication of how well the model will generalize. A naïve approach is to use all the data to fit and assess the model. When the classifier is fit to the data, it will adapt to some of the random noise in the data. When the same data is used to assess the performance, the performance measure will give an overly optimistic impression of the fit classifiers ability to generalize.

A classifier can be tailored to fit the data if overly complex models, with interactions and higher order terms, are used without a good means of assessing their performance. This will result in over-fitting. Over-fit models often give an overly optimistic of performance, but this generally happens when the data is used for both the fitting of the classifier and its assessment.

To overcome this problem the data could be split into a training set used to fit the classifier and a test set used to assess the performance of the fit classifier. In reality, this represents an inefficient use of data. Data is limited and typically costly, in terms of time and money, to obtain.

One way to overcome this is to have a sophisticate reuse of the data. Cross-validation is one such method. It is a non-analytic method used to approximate the test step and give an estimate of the performance of the classifier and will be explained in chapter 5.
Purpose

The investigation into end-of-life health service utilization for cancer patients is novel work in British Columbia. Beyond serving as a platform from which future investigations can proceed, this thesis serves several epidemiological and statistical purposes.

- To understand the implications upon research when utilizing ecological variables.
- To fulfill the one of the objectives of a Canadian Institutes of Health Research (CIHR) grant entitled “Palliative Care in cross-cultural context: A NET for equitable and quality cancer care for culturally diverse populations” through:
  - collaborative work with a Nova Scotia research team, and
  - seminal work to understand the place-of-death for British Columbian cancer patients.
- To extend the Nova Scotia work through the development of a novel approach for the construction of cultural indicators.
- To bring the topic of data mining into a new area of investigation by exploring the utility of data mining for end-of-life research.
- To identify areas where future epidemiological and statistical work can be done.
- To identify areas where departure from the Nova Scotia methodology must occur for British Columbian based research (e.g. characterizing predictors of place of death that apply to BC in particular).

Through these varied purposes an initial understanding of how health services are utilized at the end of life will be formed. As much of this research is part of a collaborative investigation, many of the variables will be based upon methodology established by the Nova Scotia research team. This thesis serves to build upon the work which has been started in Nova Scotia by exploring several novel applications of statistical methodology. The identification of predictors for the location of death will use both the Nova Scotia methods and data mining techniques which represent novelty in this field of investigation.

Although the place-of-death is a gross measure of how end-of-life health services are used, it serves as first step towards characterizing where British Columbia die and what are the predictors for those different locations. This information will allow to uncover patterns of service use among culturally diverse groups. As these populations are identified, remedial actions can be taken.

Context

In July 2004, the Canadian Institutes of Health Research (CIHR) approved the grant application entitled “Palliative Care in cross-cultural context: A NET for equitable and quality cancer care for culturally diverse populations”. This five year grant allows researchers in British Columbia, Nova Scotia, and Saskatchewan to focus upon three themes: Access to health services, caregivers and complementary and alternative therapies. The access theme seeks to understand the cultural constructs that influence the use of health services by cancer patients at the end-of-life. Not only does this
involve the construction and interpretation of predictive models, but it is also concerned with the construction of cultural indicators to be used in the models and in the quality of the data used for the modelling process. This thesis is part of the access theme as it relates to the development of cultural indicators and the determination of the cultural predictors for the location-of-death for British Columbians. The use ecological variables for end-of-life research are considered. The cultural predictors of ethnicity, language and religion are developed utilizing novel methodology and supplement the SES dissemination area level measurement of income. The utility of these predictors for end-of-life research is determined through an investigation of the usefulness of data mining techniques for end-of-life research. As this is novel work for British Columbia, it establishes a base upon which future access work can be built.

**Thesis Structure**

This thesis has six chapters, four of which represent manuscripts to be submitted to journals. The first chapter is an overall introduction to manuscript. Brief descriptions of the remaining manuscript section are subsequently presented.

**Chapter 2 – Manuscript 1: Assessing the Necessity and Appropriateness of Mixed-Level Ecological Studies in End-of-Life Research.**

The aim of this paper is to assess the necessity and appropriateness of ecological research in end-of-life studies. The two fallacies, ecological and individualistic are presented. From this integral, contextual and conceptual ecological variables are considered and a new nomenclature for variable specification is proposed. The proposed nomenclature indicates both the type of ecological variable and the type of spatial relationship that occurs between groups.

As the ecological variables to be used in the research are derived from the Canadian Census, the creation of ecological variables is considered. A prescriptive approach, which searches for a predetermined cultural aspect, such as French as the mother tongue, are the primary method used for developing dissemination-level census-based ecological variables. A descriptive approach is proposed as an alternative. With this approach, unsupervised learning techniques are used to group the dissemination areas in a meaningful, data driven way. The groups are described based upon the features of the group. Group identity is discovered rather than imposed.

The reasonability of an ecological study is considered. It is determined that if a researcher is interested in exploring the underlying cultural constructs, then an ecological study is both appropriate and necessary. It is concluded that although Canadian investigation into the cultural indicators for the place-of-death for cancer patients is appropriate and necessary, overcoming this ecological critique may not be simple.

As no prior work has been done in British Columbia for understanding where cancer patients die, this study serves as point of origin for such an investigation. As part of a collaborative effort with Nova Scotia, this study is reproduction of seminal study by Johnston et al. The goal was to understand the trends in the place of death from 1997 to 2003 for adult British Columbians that died of cancer. A corollary to this was to understand which factors were predictive of dying out of hospital. The variables investigated were the year of death, age, sex, region of residence, residence, income, tumour group and survival time.


Despite the variety of socioeconomic measures, few measures of culture exist. In Canada, researchers have employed a method by which binary cultural variables are constructed from univariate explorations of census based data. Although this represents a first step in understanding the role of culture, an indicator of culture that is based upon the multivariate nature of the data will allow the subtle nuances of a cultural construct to be retained.

Chapter 5 – Manuscript 4: Identifying Socioeconomic, Demographic and Clinical Predictors for the Location of Death of Cancer Patients in British Columbia: A Data Mining Investigation.

In Canada, the role of socio-economic measures in predicting health outcomes has mixed results. Although a health gradient exists with respect to income, the general absence of income as a predictor for health service utilization is confusing. As Canada has a universal health system, it is thought that culture may play an important role in how health services are used. Within this context, it has been noticed that logistic regression has been used as the preferred modelling technique. Given that the research uses administrative health service databases which are large in terms of the number of patients and variables, a data mining perspective is proposed as an alternative analytic tool. This paper utilizes a data mining approach to compare five data mining techniques. It compares the performance of the techniques and draws health related insights from them.

Chapter 6 – Concluding Chapter

This chapter is a summary across all the manuscripts. A brief discussion of the limitations, strengths, contributions and future recommendations follows.
References


Introduction

In a society that creates and stores vast amounts of data every day, it seems peculiar that many researchers are faced with the problem of insufficient data for their research. For many, the omission of socioeconomic and demographic information from a data set can present a serious challenge, especially when it is believed that this information may have an impact upon the outcome being studied. Within the context of cancer research, much information is collected about the disease, yet little is collected about the patient and the patient’s social context. In many clinical situations, these factors are of little interest, but they are of interest when studying how cancer patients use health resources at the end of their life. Collecting the desired information about a patient’s socioeconomic position can be a lengthy and expensive process, if it can even be done. Given these challenges many investigators have turned their attention to census based information. Although it is a pragmatic solution, it must be asked whether the choice of census information is necessary and appropriate for the desired research. The aim of this paper is to assess the necessity and appropriateness of ecological research in end-of-life studies.

Census Data as an Ecological Measure

Current cancer patient place of death population based research, within Canada, strives to describe the socioeconomic factors that determine the location of death for a cancer patient. By understanding these factors, researchers will be able to address the desires of those who are in the final stages of life and identify potential areas where end-of-life services are not connected with the surrounding people and culture. In order to achieve these goals, investigators are associating aggregate demographic and socioeconomic information obtained from census data with cancer registry data.

In Canada, end-of-life research pertaining to the place of death uses the dissemination area level data (DA) from Statistics Canada as a reasonable substitute for individual socioeconomic information. The dissemination area represents a small aggregation of the population for which information is made public and represents a compact set of blocks with a target population of 500 people. As this is the smallest area for which information is available, it has served as the geographic region for which a characteristic is imputed for each person in that local. The use of such data implies that the study will be an ecological one, thus the researcher must clearly understand not only the reasons for using such a design and the limitations of the design, but also the role the data takes within that design.

Balancing the Two Fallacies

A researcher involved in an ecological study can encounter two types of fallacy: the ecological fallacy and the individualistic fallacy. In many situations the risk of participating in either of the fallacies is minimal, but in place of death research, the investigator has the risk of participating in both fallacies. A tension results from balancing the use of aggregate data, such as census level data, and the desire for individual level data.
The ecological fallacy is the inference of phenomena from the group to the individual level\(^3\). Simply, this is individual-level conclusions based upon aggregate data. Three subsequent fallacies are related to the ecological one: individual level models are better specified, ecological correlations are always substitutes for individual ones and aggregate level variables do not cause disease\(^3\). In contrast to this concern is the individualistic fallacy which is the notion that individual level data is sufficient for epidemiological studies\(^3,6,7,12\). This suggests that the individual-level analysis is able to capture the group dynamics of a health related phenomena\(^8\). Simply, this is aggregate-level conclusions based upon a generalization from individual-level data.

A tension emerges for any researcher that seeks to understand population health. In end-of-life research pertaining to the place-of-death of cancer patients, this tension is exacerbated by having predictors which are both individual-level and group-level. Inference in either direction may lead to the participation in one of the two fallacies.

For cancer research into the place-of-death, this problem is further compounded by the inability to obtain individual-level demographic and socioeconomic data. The data that is used is derived from the census and is only available at an aggregate level. If the research conceives of the investigation in terms of individual-level data and uses census data because it is easily obtained, then participation in the these fallacies may be inevitable as the individual-level specified model has conceptual precedence over any group-level analysis. In this context, the census data serves as a proxy for the unattainable individual level socioeconomic level data. As suggested, the use of the census data as a proxy causes the researcher to participate in this tension. The desire is for individual level data, but group level data is used as a sufficient substitute. The group level data is the second choice, resulting in the subtle implication that it is an inferior type of data. If this conceptual bias is not made explicit, then a lucid interpretation may be eclipsed.

Not only will the investigator risk fallacy, but the reader will also encounter a similar risk. For example, if the reader believes that the particular study is only credible with individual level data, the reader will suspect any results of ecological studies, even if it has been shown by the investigator that the constructs of interest are best served by the group-level variables. By making explicit the conceptual structure of the variables, study and the tenability of the investigation, the researcher gives the reader a tool by which unrealized notions can be made explicit and aptly addressed. This denotative process will help sieve out unwarranted criticism, thus enabling a constructive dialogue to emerge. By clearly specifying the type of variable, the construct for which an ecological measure is being used, the type of study and the reason for the design, the investigator and reader will have the tools to avoid fallacy and clearly present the strengths and weakness of the investigation.
Understanding the Variables, Study, and Design

Types of variables

Many researchers will be familiar with the four levels of data which can be used to describe individual level data: nominal, ordinal, interval, and ratio. For group level data, several descriptions have been suggested. Susser proposes two added dimensions: integral and contextual (Table 2-1)\(^8\),\(^4\),\(^10\). An integral variable “affects all or virtually all members of a group”\(^8\). This is a non-derived variable and is a measure of some pre-existing condition pertaining to the group. An example of this may be the number of hospitals in a well defined geographic region or the presence of a breast screening campaign. This measure applies uniformly to the entire group, thus implying no within group variation. If within group variation does occur it is expected to be small\(^8\). This implies that this type of variable will have larger variation between groups than within groups.

Table 2-1: Comparison between Susser's variables and Macintyre-Duncan spatially based variables

<table>
<thead>
<tr>
<th></th>
<th>Susser(^8)</th>
<th>Macintyre-Duncan(^11)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Integral</td>
<td>A variable which affects all or virtually all members of a group</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Expect to have high variability between groups and low variability within groups</td>
<td></td>
</tr>
<tr>
<td></td>
<td>e.g. Number of hospitals in a region, number of doctors in a region</td>
<td></td>
</tr>
<tr>
<td>Contextual</td>
<td>A variable derived from a measured attribute of individuals within each group</td>
<td>Aggregate effect of social, cultural, and environmental characteristics.</td>
</tr>
<tr>
<td></td>
<td>e.g. Mean, median, proportion</td>
<td>Similar types of people will have different experiences in different neighbourhoods</td>
</tr>
<tr>
<td>Compositional</td>
<td></td>
<td>Aggregate of all individual characteristics in a neighbourhood.</td>
</tr>
</tbody>
</table>
Similar types of people will have similar experiences no matter where they live

A contextual variable “characterizes the group and not the individual” (Table 2-1)⁸. This measure is derived from individual level measurement. This characterization of the group cannot be applied uniformly to all members of the group as within group variation will exist. Larger internal variation than the integral variable is implied. This measure can be associated with a distribution which will allow for a dynamic characterization of the group. For a contextual variable, there may be no a priori assumption about how distributions will vary across groups.

Two additional types of variables have been suggested: compositional and contextual ¹¹. These variables convey a spatial aspect of the data which is relevant to much of the research that utilizes ecological variables. The compositional variable is defined as the “aggregate of all individual characteristics in a neighbourhood, that is, similar types of persons will have similar illness experiences no matter where they live” ¹¹. The compositional variable can be integral in that the variable does not vary with geography or it can be contextual in that it is a derived measure which does not vary with geography.

In contrast to the compositional variable is the unfortunately named contextual variable. This is “the aggregate effect of social, cultural and environmental characteristics of the neighbourhood” or otherwise stated as similar types of people having illness experiences dependent upon where they live ¹⁰,¹¹. As with the compositional variable, the contextual can be either integral or contextual in Susser’s terms.

For research into the cultural constructs that predicate health outcomes, a nomenclature can be established that describes not only the type of variables but the geographic aspects of them. Four possible labels emerge: integral-compositional, integral-contextual, contextual-compositional and contextual-contextual (Table 2-2). The structure is illustrated as follows. The integral-compositional variable is one where the groups have measure which applies uniformly to the group and this measure varies with respect to location. An example of this would be the number of doctors in a dissemination area. A dissemination area would have a single count and it could be conjectured that all people in that dissemination area would have access to these doctors. As medical professionals tend to be concentrated in certain locations, it is reasonable to posit that this measure would not be uniform over all dissemination areas. These measures would have little to no variation within the geographically defined group (dissemination area), but would have higher variation between dissemination areas. This suggests that the group is homogeneous with respect to the variable of interest and similar people in different groups have similar health outcomes (e.g. access to health services).
Table 2-2: Table for developing a new nomenclature by which to specify predictor and response variables

<table>
<thead>
<tr>
<th>Susser</th>
<th>Integral</th>
<th>Macintyre-Duncan</th>
<th>Compositional</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contextual</td>
<td>Groups have uniform measure that does not vary with location</td>
<td>Groups have uniform measure which varies with location</td>
<td></td>
</tr>
<tr>
<td>Contextual</td>
<td>Groups have derived measure (or associated distribution), which does not vary with location</td>
<td>Groups have derived measure (or associated distribution), which does vary with location</td>
<td></td>
</tr>
</tbody>
</table>

The contextual-compositional variable has a derived measure or associated distribution, which does vary with location. This suggests that there would be variation within groups and between groups. It suggests that the group is heterogeneous (has an associated probability distribution) with respect to the variable of interest and similar people in different groups will have different health outcomes. The income for a dissemination area would be an example of this. A dissemination area in the Point Grey community of Vancouver will have an associated income distribution, as would a dissemination area in the Downtown East Side area. The income distributions of a wealthy area and an economically frustrated one will be different. As well, variation between the two areas is expected.

By using the nomenclature derived, specification of group-level variables can occur. This specification will assist the researcher in understanding the construction of the variable and any underlying constructs that may be associated with it. Also, it will assist the reader in coming to understand how the researcher has understood and has used the variables. Through this both the researcher and the reader will move towards a common interpretation.

**Prescriptive versus Descriptive Construction of Culture indicators**

Understanding the group level data will assist the researcher towards a clear understanding of the underlying social construct that is being measured. There will be cases where the independent variable is clearly contextual, but due to methodological limitations, the distributional property of the variable cannot be utilized.

Consider the situation for the culture indicator of religion. Within any region, it is reasonable to posit that not all people, in a well defined geographic region, may be characterised by the same religious belief. The religious practice of a small region can be best represented by a probabilistic distribution, a multinomial distribution. If this was incorporated into the model, then the model would allow all religions in the area to influence the outcome. This would be a contextual-compositional variable as the proportions for each area would have locative variations.
Suppose that there are three major religions, A, B, and C, being practiced, each of which has a strong presence. There are other religions, but their presence is not predominant. Due to methodological constraints five categories are defined. If any religion has at least 50% of the population as adherents, then the region is described as either A, B, or C. If a region exists where at least two religions each have at least 30% of the people describing themselves as adherents, then the region is described as being “Religiously Mixed” (M). If a region does not satisfy any of these, then it is called “Indeterminate” (I). The new religion variable has five categories.

Each region is no longer associated with the appropriate probability distribution for religions in that region, but instead with a descriptor of some predetermined structure imposed by the researcher. For categories A, B, and C, the region is described by the dominant religious influence. All subtlety of measurement is lost. For example, a region with 55% of its adherents to religion A and a region with 95% of its adherents to religion A are now equivalent with respect to their designation. Arguments can be made for the equivalence of the two regions, but this would be based upon the notion that all regions with majority of adherents to a specific religion are equivalent which negates other influences. In terms of the specification, this would be an integral-contextual variable. All people in a dissemination area would be characterised as coming from a type of community to which there is no individual antecedent.

Two concerns emerge with prescribing the structure of a culture indicator. The first, as described above, is that all subtlety is lost by the imposition of such cut-points. By investigating a small set of viable cut-points premised upon a univariate exploration of the data, the researcher will negate the subtle influences that are produced by the interplay of cultural forces within a small well defined geographic region. The second is the role of the researcher in defining the culture indicators. The data is explored based upon a predetermined understanding of how culture is to be defined. By prescribing an operational definition of culture upon the data, the researcher assumes a position of privileged cognitive access with respect to defining a cultural construct.

With the current approach, the type of measurement is prescribed. By imposing structure upon the data, it is more difficult to ascertain the intent of the researcher. Both proxies and cultural constructs could be obtained through such a prescriptive means. This means that the variables could be integral or contextual in Susser’s terms. Without specification, the reader has no indications of the intent of the researcher and is left in an interpretive quandary.

By shifting the development of culture indicators from a prescriptive stance to that of a descriptive one, clarity can be obtained. Developing the culture indicators as a result of a multivariate exploration of the data will result in groups that are naturally defined through the data itself. By allowing the data to drive the construction of the groups, the researcher will be creating an indicator of an underlying cultural construct rather than a proxy.
Univariate constructions of culture indicators suggest that a single aspect is important independent from all other aspects associated with that cultural construct, for example the use of French as mother tongue suggests that this aspect of the mother tongue census characteristic is the defining aspect apart from all other languages that could occur. It suggests the search for French speaking people rather than the search for a full understanding of the mother tongue languages that are influential in a region.

The multivariate descriptive approach, allows for all the languages to influence the eventual description of the underlying cultural construct. Researchers do not seek out speakers of a particular language with this approach. They allow all the languages to work together to obtain groups. In the case of end-of-life research, the result will be groups of dissemination areas. These groups will be based upon similarities for all languages, not just one particular language or language group (e.g. official languages versus non-official languages). The best descriptor of the group is the associated distribution of languages, but evocative names can be assigned to assist communication. By constructing indicators of culture by this means, contextual-compositional and contextual-contextual variables would be constructed using census based data. On a conceptual level, the researcher relinquishes any position of privilege over to the data and allows the data to determine the underlying structure of the cultural construct.

The point at which the culture indicator is identified may appear trivial, but it has an effect upon the nature of the construct being considered. As suggested earlier, prescriptive approaches will leave the nature of the indicator in an ambiguous state if not explicitly stated by the researcher. When utilising descriptive approaches, a clear indicator for the underlying cultural construct will be developed. The indicator is premised upon a multivariate distribution, thus is will be more natural to develop a culture indicator for the underlying cultural construct than a simple proxy. The indicator would be a succinct representation of a complex underlying multivariate description of a cultural construct.

By using prescriptive rather than descriptive approaches to developing the culture indicators a trade-off occurs. If the categories are pre-specified, majority-minority relationships could evolve resulting in unexpected social perceptions. The descriptive approach can be used in negative manner names must be attached to each of the derived distributions to ease discussion, but this is best left to a discussion concerning ethical research. Although, from the investigative perspective, this is a semantic issue, it must not be neglected. Through the specification of the type of variables that are being used, the use of prescriptive or descriptive approaches, and the identification of the true construct being measured, researchers will be assisted in the development, presentation, interpretation and application of the research while recognizing its limitations.

**Types of Studies**

As with the variables, when ecological variables are used, an extra dimension must be considered which goes beyond understanding the modelling procedures used for more routine epidemiological studies. This aspect incorporates the ways in which the levels of variables can be combined. Following Susser’s notation, let x and y represent
individual level independent and dependent variables respectively and let X and Y represent the group level independent and dependent variable respectively. An unmixed study is defined to be one where the relationship is defined on the same level that is from x to y or from X to Y. A mixed study is defined as one where the inference crosses from one level to another, from x to Y or from X to y. In the Canadian context of studying the place of death of cancer patients, the relationship is from \{x, X\} to y, where \{x, X\} is the set of individual and ecological independent variables. For completeness, a study of the form \{x, X\} to Y is possible. With such mappings specified, it is apparent that any guideline for interpretation based solely on Susser's supposition is not immediately transferable to this situation unless the individual level predictor variables or the ecological variables do not contribute to the parsimonious model. These special cases reduce to either an unmixed relationship if the group level variables are not significant or a mixed relationship if the individual variables are not present in the final model.

When the model does not reduce to one of Susser's relationships, the interpretation must contain aspect of both the individual and the group. If the set \{x, X\} is found to be predictive of y, then the investigator would have information that individuals with characteristic x from areas defined by X have result y. When this structural relationship occurs, lucid interpretation may be precluded. Individual level interpretations will force the researcher to participate in the ecological fallacy, yet group level interpretations will result in the individualistic fallacy. When this is encountered, the common interpretation is of the form, individuals with characteristic x from an area defined as X have result y.

Morgenstem opposes the common interpretation with his concept of cross-level bias. Cross-level bias is the sum of the aggregation bias due to groupings and the specification bias due to the confounding effect of the group. It is suggested that any study where the parsimonious model maps \{x, X\} to y contains cross-level bias. Mathematically, Geronimus and Susser verify this notion and identify the structure of the bias.

Morgenstem and Susser then claim that any study which has cross-level bias participates in the ecological fallacy. From the perspective of the individual, the cross-level bias occurs from the aggregation effects of the variables, the misspecification of the underlying group construct, or from both. From the perspective of the group, the cross-level bias results from the disregard for group effects or from confounding. Even though the components of cross-level bias can cancel each other, it is rare. Morgenstem illustrates that there will be no cross-level bias when the coefficient for the group-level variables are equal to zero when the response is at the individual level. The parsimonious model is a mapping from \{x\} to \{y\}. Susser describes cross-level bias as a result from the "atomistic fallacy" where individual observations ignore group effects or from specification bias where confounding occurs due differential distribution of individual characteristics among groups. The atomistic fallacy resulting from aggregation bias implies that the between group relationship differs from the between individual relationship within group and ignoring groups. The specification bias results from within group relationship differing from the between group relationship and the individual relationship ignoring groups.
It may seem that the researcher is trapped, but Greenland gives balanced insight into the problem, "it is important to remember that the possibility of bias does not demonstrate the presence of bias, and that a conflict between ecological and individual-level estimates does not by itself demonstrate that the ecological estimates are more biased". Unfortunately this does not suggest a more acceptable interpretation for the researcher who is constrained to a \{x, X\} to y mixed-level ecological study, yet given the complexity that multiple levels introduces, the current interpretation attempts to be sensitive to individual level variables and group level variables.

It is at this stage that much discussion over the methodology has occurred. It is within the context of this discussion that alternate measures are suggested in order to compensate for the bias which occurs with ecological studies. As well, the concept of confounding between the individual level variable and the ecological variable is addressed. In order to handle these two issues, several alternate models have been suggested ranging from modifications of standard regression techniques in order to identify how the bias affects the estimates to the proposal of variations of random effects models.

Reasonability of an Ecological Study

By identifying the types of variables, the construction of indicators and the type of study, it will be apparent if an investigator is utilizing an ecological design. If an ecological study results, then the relevance of the study needs to be assessed. A researcher may be obligated to use an ecological variable because of the research question, concerns about group level actions, or because of the inability to obtain individual level data. If the study is apt, then the use of ecological variables is logically appropriate.

Within the context of understanding the place of death for cancer patients, if the ecological variables are included in the study because they represent a relevant construct then the study is obligate and apt. If they are included in the study because the individual level data is unattainable, then the study is obligate, not apt and convenient. It is not apt because the proxy may not be measuring the same construct. If through the specification of the variable, the proxy is found to be measuring the same thing as the original variable, then the study is apt and convenient.

The situation where the study is obligate, not apt and convenient could fundamentally alter the investigation. If through the specification of the variable, the proxy is found to be measuring a different construct, then an exchange of convenience for logical appropriateness has occurred. Although this study may have the same model as in the other two situations and employ the same analysis, they are not equivalent studies because the underlying constructs are not equivalent. Two studies may be addressing similar issues, say the cultural predictors for the location of death of Canadian cancer patients, but the apt study is predicated upon potentially different constructs than the not apt one. It is through the specification of the variables, that the equivalence of the studies will be best understood.
Current Situation

Current cancer patient place of death population based research, within Canada suffers as there has been no explicit specification of the study variables, type and reasonableness. Stating the variables, the method by which they were constructed and modelling procedure does not address what constructs are actually being measured by ecological variables. An excellent example of this is the current use of income quintiles as a proxy for socioeconomic status. Income will fluctuate over time and over short and long periods and as a result may not be a sufficient indicator of household purchasing power, personal wealth, the ability to respond to a health crisis, or the resources available in the surrounding area. If two individuals are compared with the same income, nothing is revealed about how much of that income is designated to savings, mortgage, food, or lifestyle activities which translates into the ability for the individual to weather a health crisis. Inference is made based upon this measure, but the knowledge gained is limited.

Current findings serve to further compound the issue. There exists evidence that small differences in income are related to larger changes in health status. This evokes a second problem with the current use of ecological proxies. Given that small changes in income have been found to be associated with larger changes in health status, the income quintile may be too gross a measure to elicit any significance regardless of how the model is specified. As this has not been addressed, the lack of significance of income in current research may be due to the construction of the variable as opposed to its true significance. If this is the case, then the model is highly sensitive to how the underlying constructs are defined which would cast a shadow over legitimacy of the research itself.

Without specifying the underlying social construct that is being measured by such a proxy, it is apparent that the use of such a variable will entertain discussions about bias, confounding and misspecification. Such discussions must occur as there has been no clear presentation of what is being measured and the utility of that measurement. Even when specification has occurred, bias and confounding may occur, but in this context the discussion of such technical aspects will be premised upon a clear understanding of the measures employed in the analysis.

A secondary issue that emerges in the current Canadian research is the construction of the ecological variables. As the studies have not specified the intent of the culture indicators, there is not reason to assume they are operating as anything other than a proxy to some individual cultural attribute. As the studies are mixed-level ecological ones, interpretive fallacies are avoided through careful semantics and mathematical limitations are noted. Unfortunately this does not negate the inherent problems with using indicators predicated upon a univariate exploration of the data. Specification would illuminate research intent and illuminate areas of concern within an ecological study, but not address the subtle problems surrounding the development of the indicator.

Even though the current research attempts to advance knowledge on the cultural forces influencing patient’s decisions at the end-of-life, it is limited by the current method for specifying indicators of culture. Without specification, research intent can only be
inferred. This inference is obtained by the body of related publications, model specification, the culture indicators used and the methodology associated with the development of the indicators. With a shift to descriptive indicators, the intent of the researcher would be made more explicit. In attempting to indicate the contextual aspects of culture, the researcher will be using apt ecological variables, thus validating the ecological approach to the research.

**New Horizons**

Faced with the spectre of frustration, there are avenues researchers can take in order to begin addressing the issues at hand. The following collection of potential strategies serves to illustrate a variety of methodologies and directions that could be taken.

**Prescriptive Approaches**

**Specification**

The simplest and immediate solution is to have the variables completely specified. This will enable other researchers to produce comparable research in different geographic areas and in different time periods. Unfortunately this leaves the research open to discussion about the construct being measured and its validity as well as the usual issues of bias and confounding. Unfortunately, it requires that the researcher accept all the limitations, both conceptually and mathematically.

**Specification and Quantification**

This is the same as the previous situation except that the bias is quantified. Geronimus, Bound and Neidert use the example of linear regression to illustrate a method for quantifying the bias due to using an ecological variable. Although the bias can be assessed, it is contingent upon knowledge of how the individual level measure is related to its ecological counterpart. This approach parallels Susser’s quantification of the bias. Morgenstem provides a methodology for the quantification of the bias inherent in ecological studies by comparing the risk ratio and correlation coefficient for the grouped data and the individual level data using exposure status, disease status and group identity. The immediate challenge with this approach is the need to develop an estimate for the bias when individual level data is not available or when the data can not be manipulated into the structure used by Morgenstem.

**Descriptive Approaches**

**Classification Trees to Define Cut-off Values**

This approach assumes a clearly specified set of variables and addresses the issue of constructing the cut points when transforming independent contextual variables into independent integral variables. The central issue was the effect of the choice of cut-point. This data driven method uses classification trees to define potential cut points based upon the organization of the information with respect to some specified outcome. Given a set of possible cut-points for the data, the researcher then applies area specific knowledge to select the collection of data selected and clinically reasonable cut-
off values. The advantage of this method is that any unspecified perception based researcher bias may be avoided due to the presence of a data driven set of cut-off values.

**Cluster Algorithms for Indicator Development**

The multivariate distribution of the cultural characteristic can be used to organize regions based upon a measure of similarity\(^1\). Unsupervised learning hierarchical or non-hierarchical clustering algorithms could be used. This is a data driven approach would geographic regions without a pre-specified outcome. The resultant groups could then be statistically described and given a label in order to facilitate communication. Each patient could be assigned to a group based upon their postal code, thus an indicator of culture, premised upon a multivariate, would result.

**Conclusion**

How does a researcher even begin to address the issues of end-of-life studies in Canada using a mixture of ecological and individual level predictor variables and which approaches should be used? A reasonable approach can be taken from literary theory. The hermeneutical spiral is the method of refining a textual interpretation through a continual reflexive critique of the current interpretation. Transferring this to the end-of-life research context, the hermeneutical process would be the continual critique of previous research and methodologies. Within this paradigm, a point of origin for this self-referential critic has been established with the current, yet limited, body of work. This article would represent a reflexive critique of the research, as it sits within the frame work of end-of-life research, yet is attempting to extract learning from the current body of work to improve this next round of research. By understanding the limitations of the current methodology for the development of culture indicators new approaches can be derived which may allow for the researcher to avoid the two fallacies. By moving end-of-life cancer research in Canada beyond the traditional forms of indicating culture, new methodologies and understandings may emerge.
References


Background

Current research indicates that if a patient is given a choice, up to 88% of patients dying of cancer would choose to die at home, yet only a smaller proportion realizes this preference. International estimates of the proportion of cancer patients who die at home range from 23% in Antwerp, Belgium to 35% in Houston, United States. In Nova Scotia (NS) Canada, out of hospital deaths were 25.4% the period from 1992 to 1997 when measured in terms of calendar years and 31.6% for the fiscal period from 1992 to 1997. Out of hospital deaths in Ontario have ranged from 19.8% in 1992 to 30.2% in 1997. Out-of-hospital deaths in Canada have ranged from 31% in 1980 to 21.5% in 1994 to 34% in 1997. In Manitoba, 57.4% of cancer patients die in the hospital during the fiscal year of 2000/01. Vancouver has experienced a range of out-of-institutional cancer deaths. From 1990 to 1992, approximately 15% died in non-institutional settings (home or other locations), but in 1993 this number increased to just under 30%.

When the patient characteristics of those who are more likely to die out of hospital are considered, congruencies across the studies can be seen. These commonalities include old age, long term relationship, ethnicity, health care provider, education and the geographic region of residence. As well, sex was found to relate to the location of death. Burge (2003) and Constantini (2000) found that women were more likely to die out of hospital.

In July 2004, the Canadian Institutes of Health Research (CIHR) approved the grant application entitled “Palliative Care in cross-cultural context: A NET for equitable and quality cancer care for ethnically diverse populations”. This five year grant allows researchers in British Columbia, Nova Scotia, and Saskatchewan to focus upon three themes: Access to health services, caregivers and complementary and alternative therapies. The Access theme seeks to understand the cultural constructs that influence the use of health services by cancer patients at the end-of-life. Not only does this involve the construction and interpretation of predictive models, but it is also concerned with the construction of culture indicators to be used in the models and in the quality of the data used for the modelling process.

This study is the first to analyze the trend in the place of death for BC cancer patients (1997 to 2003) and identify predictors of end-of-life care. The objective was to understand which factors were predictive of dying out of hospital. As this study is the first step towards a comparative study between BC and NS, the factors taken were those used in the original NS study.

Methods

Data Sources

For this retrospective population-based study, a secondary data set was created from the linkage of two administrative data bases: The British Columbia Vital Statistics and the British Columbia Cancer Registry (BCCR). The Statistics Canada Postal Code Conversion File (PCCF) programs 3J and 4D were used along with the 1996 and 2001...
Canadian censuses to obtain the quintiles for the neighbourhood income per person equivalent (IPPE). The response, location of death, was obtained from the VS database. The predictors were obtained from the British Columbia Cancer Agency (BCCA) database. The cause of death was taken from the BCCA. It should be noted that the BCCA database obtained this information from BC Vital Statistics.

**Study Population**

The subjects were all adults, age 18 years and older who died of cancer, in British Columbia as identified from the death certificates, from 1997 to 2003. The death certificate information was obtained from the BC Vital Statistics. Death Certificate Only (DCO) patients were excluded from the formation of the data set. These patients were those who had a cancer diagnosis at the time of death, although no prior information relating to a diagnosis of cancer was found after following up the diagnosis at the time of death. The age was derived from the BCCR database as the difference between the date of death and the date of birth. The postal code for the usual place of residence was obtained from VS. If this information was unavailable, then the postal code for the place of residence was obtained from the BCCR. Vital Statistics provided the location of death. Deaths due to cancer, primary or secondary, were identified using the BCCA database.

**Study Variables**

From this data set, the year of death, sex, age, neighbourhood IPPE, diagnosis (coded according to the International Classification of Diseases, ninth and tenth revisions \(^{16,17}\)), cause of death (coded according to the International Classification of Diseases O, third revision \(^{18}\)), postal code of usual residence, health region associated with the postal code of usual residence, survival time, and location of death (dichotomized as in hospital or out of hospital) as recorded on the death certificate were obtained.

The categories for age and survival were obtained from the Nova Scotia research. The age groups were constructed to reflect the skewed age distribution of people who died from cancer in Nova Scotia \(^{10}\). The patient’s age was taken as the difference between the date of death as indicated on the death certificate and the date of birth as reported in the BCCR. For survival, the categories were constructed to capture the varying lengths of survival \(^{10}\). The survival time was taken as the difference between the date of death on the death certificate and the date of diagnosis as reported in the BCCR.

The neighbourhood income per person equivalent is a household size-adjusted measure of income based upon the dissemination area level information \(^{19}\). In the PCCF program, then quintiles were defined in such a manner as to reflect the relative nature of the measure of income across regions, to minimize the effect of large differences in housing costs upon the overall household welfare, and have equal proportions of the population in each income quintile \(^{19}\). This measure was selected for two reasons. The first is the relative ease with which it could be obtained, thus it was an opportunistic choice. The second was due to the adjustments used in the measure. As the final intent is to compare across regions it is desirable to have a measure of income that does adjust for regional variation.
The International Classification of Diseases (ICD) codes were grouped as in the Nova Scotia research (Appendix). As statistical power was a concern of the Nova Scotia team, cancer types were aggregated to produce more information in each group. As inter-provincial comparisons are intended, this structure was adopted.

The region of death taken to be the health authority associated with the patient’s usual residence. In BC, the five health authorities are:

- Fraser,
- Vancouver Coastal,
- Vancouver Island,
- Interior, and
- Northern.

A provincial health authority exists, but was not used as it is not geographically defined. The Provincial Health Services Authority’s (PHSA) primary role is to ensure access to network of specialized health care services. It operates agencies such as BC Children’s Hospital, the BC Transplant Society, and Riverview Hospital. It is responsible for specialized services such as chest surgery and cardiac services.

An indicator of rural or urban residence was constructed using the postal code of usual residence, obtained from the death certificate, employed in the PCCF process. This follows the methodology of the Nova Scotia research. If the second number was a 0, then a rural delivery point was indicated. If no address was given on the death certificate, the address from the BCCR was used.

BC has two postal codes which were reused. The distance between the old and new area is considerable enough to pose geocoding problems. As the data set spanned the vintage date, 1 April 1999, for these postal codes a supplemental program was used which handles this peculiarity. The identified postal codes were verified for accuracy and the appropriate re-coding was performed when the reported postal code and the current postal code did not agree.

**Analysis**

As an inter-provincial comparison is an eventual goal, the groupings for age, survival, and cause of death were the same as those used by the Nova Scotia research. Cross tabulations were used to understand the relationship between the predictors and the location of death. Only those with a significant association were selected for entry into the multivariable logistic regression. Logistic regression was used to understand how the selected set of predictors was related to the location of death. The predictors were year of death, sex, age, neighbourhood IPPE, aggregate tumour group (lung as reference group), region of residence (BC Health Authority), and survival group. Residence (rural versus urban) was excluded from the regression as no association was found. Multivariable logistic regression was conducted utilizing backwards stepwise elimination. Stepwise elimination that allowed backwards and forward steps was used as a comparative technique. An exclusion p-value of 0.05 was used to find a
parsimonious model for the prediction of out-of-hospital death. All the variables that were used in the full model were retained.

**Results**

A total of 52,295 adults died of cancer in British Columbia in the period from 1997 to 2003. 2,366 patients were excluded due to missing information on: Neighbourhood IPPE and region of death. Of the remaining 49,929 patients 33,409 (66.91%) died in hospital (Table 3-1). Not all the characteristics were found to be significantly associated with the place of death. The residence indicator (urban versus rural) had insufficient evidence for an association (p=.24), thus it never entered the multivariable logistic regression model.

<p>| Table 3-1: Characteristics of adults in British Columbia who died of cancer from 1997 to 2003, by place of death. |
|-----------------------------------------|-----------------|-----------------|-----------------|-----------------|</p>
<table>
<thead>
<tr>
<th>Characteristic</th>
<th>In Hospital</th>
<th>Percent in Hospital</th>
<th>Out of Hospital</th>
<th>Percent out of Hospital</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Year</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1997</td>
<td>4964</td>
<td>74.09%</td>
<td>1736</td>
<td>25.91%</td>
</tr>
<tr>
<td>1998</td>
<td>5052</td>
<td>74.54%</td>
<td>1726</td>
<td>25.46%</td>
</tr>
<tr>
<td>1999</td>
<td>4232</td>
<td>61.83%</td>
<td>2613</td>
<td>38.17%</td>
</tr>
<tr>
<td>2000</td>
<td>4518</td>
<td>66.40%</td>
<td>2286</td>
<td>33.60%</td>
</tr>
<tr>
<td>2001</td>
<td>4568</td>
<td>64.72%</td>
<td>2490</td>
<td>35.28%</td>
</tr>
<tr>
<td>2002</td>
<td>4898</td>
<td>64.49%</td>
<td>2697</td>
<td>35.51%</td>
</tr>
<tr>
<td>2003</td>
<td>5177</td>
<td>63.53%</td>
<td>2972</td>
<td>36.47%</td>
</tr>
<tr>
<td><strong>Sex</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>18079</td>
<td>68.19%</td>
<td>8434</td>
<td>31.81%</td>
</tr>
<tr>
<td>Female</td>
<td>15330</td>
<td>65.61%</td>
<td>8034</td>
<td>34.39%</td>
</tr>
<tr>
<td><strong>Age, yr</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18-44</td>
<td>1275</td>
<td>70.36%</td>
<td>537</td>
<td>29.64%</td>
</tr>
<tr>
<td>45-64</td>
<td>8283</td>
<td>68.47%</td>
<td>3815</td>
<td>31.53%</td>
</tr>
<tr>
<td>65-74</td>
<td>9641</td>
<td>69.27%</td>
<td>4278</td>
<td>30.73%</td>
</tr>
<tr>
<td>75-84</td>
<td>10297</td>
<td>66.54%</td>
<td>5179</td>
<td>33.46%</td>
</tr>
<tr>
<td>≥85</td>
<td>3913</td>
<td>59.07%</td>
<td>2711</td>
<td>40.93%</td>
</tr>
<tr>
<td><strong>Neighbourhood IPPE</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td>8353</td>
<td>68.34%</td>
<td>3870</td>
<td>31.66%</td>
</tr>
<tr>
<td>Lower Middle</td>
<td>6990</td>
<td>67.89%</td>
<td>3306</td>
<td>32.11%</td>
</tr>
<tr>
<td>Middle</td>
<td>6520</td>
<td>66.95%</td>
<td>3218</td>
<td>33.05%</td>
</tr>
<tr>
<td>Upper Middle</td>
<td>5992</td>
<td>66.24%</td>
<td>3054</td>
<td>33.76%</td>
</tr>
<tr>
<td>Upper</td>
<td>5554</td>
<td>64.37%</td>
<td>3074</td>
<td>35.63%</td>
</tr>
</tbody>
</table>
Over the study period the general trend was towards an increase in the proportion of deaths occurring out of hospital from 25.9% in 1997 to 36.5% in 2003 (Figure 3-1). Logistic regression confirmed this trend. Compared with 1997, the adjusted odds ratio (OR) for 2003 was 1.6 (95% confidence interval [CI] 1.5-1.7) (Table 3-2). Figure 3-2 shows this same trend occurring over the regional health authorities. From 1997 to 1999 there is a clear increase in the proportion of out of hospital deaths for all the health authorities. The same anomalous spike is found in 1999 for all health authorities, and then from 1999 to 2003 the trends are less obvious. Despite the unclear trends, there is an increasing trend for the Northern Health Authority.
Figure 3-1: Proportion of death out of hospital and in hospital among British Columbian adults who died of cancer, 1997-2003.

Table 3-2: Crude and adjusted odds of dying out of hospital, by predictors, for British Columbia, 1997-2003

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Crude OR</th>
<th>Adjusted OR</th>
<th>95% CI</th>
<th>Lower</th>
<th>Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Year</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1997*</td>
<td>1.0</td>
<td>1.0</td>
<td></td>
<td>0.9</td>
<td>1.1</td>
</tr>
<tr>
<td>1998</td>
<td>1.0</td>
<td>1.0</td>
<td></td>
<td>0.9</td>
<td>1.1</td>
</tr>
<tr>
<td>1999</td>
<td>1.8</td>
<td>1.8</td>
<td>1.7</td>
<td>1.0</td>
<td>2.0</td>
</tr>
<tr>
<td>2000</td>
<td>1.4</td>
<td>1.5</td>
<td>1.3</td>
<td>1.0</td>
<td>1.6</td>
</tr>
<tr>
<td>2001</td>
<td>1.6</td>
<td>1.6</td>
<td>1.5</td>
<td>1.0</td>
<td>1.7</td>
</tr>
<tr>
<td>2002</td>
<td>1.6</td>
<td>1.6</td>
<td>1.5</td>
<td>1.0</td>
<td>1.7</td>
</tr>
<tr>
<td>2003</td>
<td>1.6</td>
<td>1.6</td>
<td>1.5</td>
<td>1.0</td>
<td>1.7</td>
</tr>
<tr>
<td><strong>Sex</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male*</td>
<td>1.0</td>
<td>1.0</td>
<td></td>
<td>0.9</td>
<td>1.1</td>
</tr>
<tr>
<td>Female</td>
<td>1.1</td>
<td>1.1</td>
<td>1.1</td>
<td>1.0</td>
<td>1.2</td>
</tr>
<tr>
<td><strong>Age, yr</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18-44*</td>
<td>1.0</td>
<td>1.0</td>
<td></td>
<td>0.9</td>
<td>1.1</td>
</tr>
<tr>
<td>45-64</td>
<td>1.1</td>
<td>1.1</td>
<td>1.0</td>
<td>1.0</td>
<td>1.2</td>
</tr>
<tr>
<td>65-74</td>
<td>1.1</td>
<td>1.1</td>
<td>1.0</td>
<td>1.0</td>
<td>1.2</td>
</tr>
<tr>
<td>75-84</td>
<td>1.2</td>
<td>1.2</td>
<td>1.1</td>
<td>1.0</td>
<td>1.4</td>
</tr>
<tr>
<td>≥85</td>
<td>1.6</td>
<td>1.7</td>
<td>1.6</td>
<td>1.0</td>
<td>2.0</td>
</tr>
<tr>
<td>Neighbourhood IPPE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------------------</td>
<td>----------</td>
<td>----------</td>
<td>----------</td>
<td>----------</td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td>1.0</td>
<td>1.0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lower Middle</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.1</td>
<td></td>
</tr>
<tr>
<td>Middle</td>
<td>1.1</td>
<td>1.1</td>
<td>1.0</td>
<td>1.1</td>
<td></td>
</tr>
<tr>
<td>Upper Middle</td>
<td>1.1</td>
<td>1.1</td>
<td>1.1</td>
<td>1.2</td>
<td></td>
</tr>
<tr>
<td>Upper</td>
<td>1.2</td>
<td>1.2</td>
<td>1.1</td>
<td>1.3</td>
<td></td>
</tr>
<tr>
<td>Aggregate Tumour Group</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Breast</td>
<td>1.2</td>
<td>1.0</td>
<td>0.9</td>
<td>1.1</td>
<td></td>
</tr>
<tr>
<td>Colorectal</td>
<td>1.2</td>
<td>1.1</td>
<td>1.0</td>
<td>1.1</td>
<td></td>
</tr>
<tr>
<td>Lung</td>
<td>1.0</td>
<td>1.0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prostate</td>
<td>1.3</td>
<td>1.0</td>
<td>0.9</td>
<td>1.1</td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td>1.0</td>
<td>0.9</td>
<td>0.9</td>
<td>1.0</td>
<td></td>
</tr>
<tr>
<td>Region of Death</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interior</td>
<td>1.2</td>
<td>1.2</td>
<td>1.2</td>
<td>1.3</td>
<td></td>
</tr>
<tr>
<td>Fraser</td>
<td>0.8</td>
<td>0.8</td>
<td>0.7</td>
<td>0.8</td>
<td></td>
</tr>
<tr>
<td>Vancouver Coastal</td>
<td>1.0</td>
<td>1.0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vancouver Island</td>
<td>1.4</td>
<td>1.4</td>
<td>1.3</td>
<td>1.5</td>
<td></td>
</tr>
<tr>
<td>Northern</td>
<td>1.3</td>
<td>1.4</td>
<td>1.2</td>
<td>1.5</td>
<td></td>
</tr>
<tr>
<td>Length of Survival</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>≤60</td>
<td>1.0</td>
<td>1.0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>61-120</td>
<td>1.7</td>
<td>1.8</td>
<td>1.7</td>
<td>1.9</td>
<td></td>
</tr>
<tr>
<td>≥121</td>
<td>1.9</td>
<td>2.0</td>
<td>1.9</td>
<td>2.1</td>
<td></td>
</tr>
</tbody>
</table>

Note: OR=Odds Ratio, CI=Confidence Interval
† Adjusted for year of death, sex, age, neighbourhood IPPE, aggregate tumour group, region of death and length of survival
* Reference group
The adjusted results for the multivariable logistic regression are given in Table 3-2. After accounting for all the other significant factors in the model, patients who died out of hospital were more likely to be female, over 45 years of age, from a neighbourhood in the upper three neighbourhood IPPE quintiles, live outside of the Fraser and Vancouver Coastal Health Authorities, and survive longer than 60 days after diagnosis.

The adjusted OR for females was 1.1 (95% CI 1.1-1.2), yet 53.2% of all the deaths were men.

In relation to age, the adjusted OR was 1.1 (95% CI 1.0-1.2) for both the 45-64 age group and the 65-74 age group compared with those who were under 45 years of age. For the 75-84 age group the OR was 1.2 (95% CI 1.1-1.4) and the 85+ age group had an OR of 1.7 (95% CI 1.6-2.0) compared with those who were under 45 years of age.

Compared to the lowest income quintile, the highest had an OR of 1.2 (95% CI 1.1-1.3). The middle income quintile had an OR of 1.1 (95% CI 1.0-1.1) and the upper-middle income quintile had an OR of 1.1 (95% CI 1.1-1.2).
When the region of death, as defined by the BC Health Authorities, was considered, it was found that the non-Fraser Valley health authorities were more likely to have out of hospital deaths. When compared to the Vancouver Coastal Health Authority, the Interior Health Authority had an OR of 1.2 (95% CI 1.2-1.3), the Vancouver Island Health Authority had an OR of 1.4 (95% CI 1.3-1.5) and the Northern Health Authority had an OR of 1.4 (95% CI 1.2-1.5). When compared with those who survived less than 60 days, those who survived longer than 120 days had an OR of 2 (95% CI 1.9-2.1).

The method of obtaining a parsimonious model did produce any differences in the final model. As such, the results for the stepwise method were not reported.

**Discussion**

We found that the proportion of out of hospital deaths among adults with cancer increased over the study period. The proportion rose from 25.9% in 1997 to 36.5% in 2003. It was found that women, patients over 45 years of age, patients in neighbourhoods in the Middle to Upper IPPE quintiles, patients with colorectal cancer, patients who lived outside the Fraser Valley, and those who have survival times greater than 60 days were more likely to die outside of a hospital.

The difference in the location of death as seen in Figure 3-1 was unexpected and warrants further investigation. This may reflect an implementation of a commitment for home related health care delivery systems. This puzzling event warrants further investigation into the health policies being implemented in the 1998/1999 time period.

It was found that women are more likely than men to die out of hospital. This is congruent with Nova Scotia and Italian results. Even though this is true, Tang (2001) reveals that there is conflicting evidence for the relevance of sex as a predictor for the location of death. As well, it was found studies where women are less likely to die out of hospital.

The British Columbia and Nova Scotia have complimentary results with respect to the role age plays in determining the location of death. In BC, it was found that patients 45 years of age and older were more likely to die out of the hospital. In Nova Scotia, it was found that patients 75 years and older were more likely to die out of the hospital. These findings concur with those of Constantini et al. (2000) and Bruera et al. (2002). On the other hand, Tang (2001) finds that the relationship between age and the location of death is yet unsettled.

BC has a median age of 38.4 years. Nova Scotia, which also reported that age was a significant predictor for the location of death has a median age of 38.8 years. Both of these are higher than the Canadian level of 37.6 years of age. On the other hand, Manitoba had a median age of 36.8 years, yet it has the second highest proportion of seniors aged 65 and over (14%) exceeded only be Saskatchewan (15%). Manitoba saw the largest increase in median age over the last decade. This was due to the increase in the numbers of elderly people.
BC has seen an increase in the numbers of older people in the province between the 1996 and 2001 censuses. There was a 49% increase in the 45 to 64 age group, a 27% increase in the 70 to 79 age group, and a 54% increase in the 80 and over group. The Fraser Valley and Vancouver Island regions saw their median age increase from 36.0 years in 1996 to 38.1 years in 2001.

BC has 13 of the top 25 municipalities with the oldest populations. The oldest BC municipality was Qualicum Beach (Median age: 58.1 years) followed by White Rock (Median age: 50.9 years). Victoria is the second oldest Canadian metropolis areas. It has a median age of 41.0 years. Vancouver, on the other hand, with a median age of 37.4 years, is close to the Canadian figure. With such demographic shifts, age may begin to play an important role in the observed health outcome and in the decisions made concerning end-of-life issues.

Income was found to be predictive of the location of death. In Nova Scotia and Toronto the opposite was found. The difference between the studies is the measure of income. The measure used in this investigation incorporated regional variations; it is unclear if this was done in the Nova Scotia and Toronto studies. This suggests that the means by which income is assessed may contribute to the role it plays in predicting health outcomes. This may be especially true for countries that have universal health care systems. Following this avenue of thought, Roos (2004) and Veugelers (2001) imply that the lack of association may be due to Canada’s universal health care system.

Unlike the Nova Scotia results, the aggregate tumour group was found to be predictive of the location of death. Those patients with colorectal cancer were more likely to die out of the hospital compared with patients who had lung cancer. There was no significant difference in the likelihood of dying out of hospital between breast cancer, prostate cancer, and all other cancers with reference to lung cancer.

Patients who reside in the Fraser Health Authority are less likely to die out of hospital when compared with the region comprising the Vancouver Coastal Health Authority. The Fraser Health Authority has the largest proportion of the BC population comprising 36.4% of it and is the most urbanised. It has the second lowest number of acute care days (503 days/1000 pop) with the Vancouver Coastal having the lowest (496 days/1000 pop).

Patients who reside in the Northern, Interior, and Vancouver Island Health Authorities are more likely to die out of hospital when compared to Vancouver Coastal. The Northern Health Authority has the highest number of acute care days (734 days/1000 pop). The Interior has 602 days/1000 acute care days and the Vancouver Island Health Authority has 596 days/1000.

In BC, the regions with higher odds of dying out of hospital are those regions where the length of stay in acute care facilities is longest. The regions were the odds of dying out of hospital are lower than the Vancouver Coastal having short lengths of stay in acute care facilities. For British Columbians, these results may be related to the access of health services if people are required to travel substantial distance for health services.
For patients outside the Fraser and Vancouver Health Authorities, hospitalization may occur when other local options have been exhausted. This unexpected relationship posits a challenge for BC researchers as this relationship may be intimately connected to utilization, geographic distribution and the administration of health services as well as the influence of other non-institutional support systems.

The use of administrative databases allows for a rough sketch of population health to be obtained. Although they represent an efficient means of obtaining health related data, they are limited as their primary intent is for administration not research. This aspect does limit the number of usable fields and constrains the researcher into using predetermined constructs and coding. Fortunately the knowledge gained outweighs the limitations. Our data was taken from the administrative databases used at the British Columbia Cancer Agency and the British Columbia Vital Statistics. In the future, it is hoped that more administrative databases pertaining to palliative care, hospital separations and hospice care will be added to the existing database. It is believed that with this information a better understanding of how palliative care and primary care services are used by cancer patients at the end of their life.

**Limitations**

Beyond the limitation of using administrative databases as previously mentioned, other limitations do exist with using administrative databases. The VS and BCCA databases do not collect socioeconomic and demographic information. Due to this, the income measure used in this study was obtained from dissemination area census data. The use of such data does have ramifications upon the utility of the model in the form of the ecological fallacy. Also, researchers are limited to the categories that have been specified by the data providers. Even with these and the aforementioned limitations, administrative databases do represent an efficient means, in terms of cost and time by which population based health data can be obtained. This allows for a rough sketch of population health to be made with relative ease.

The response, death out of the hospital may be too vague. These deaths are not a hospital death, but they may occur in an institution such as a nursing home or a palliative care facility. Although technically not a hospital, locations such as these represent a mixed environment as medical services and place of usual residence are intermingled. What is needed is clearer information on the location of death so that pseudo-residential locations can be identified.

The use of the postal code for an indicator of rural versus urban residence is questionable. The zero in the second position of the postal code indicates delivery from a rural post office. This indicates the point of service, but not the rural or urban feature of the area. Some rural areas are urbanised and suburbanised, yet they have a zero in the second position, thus it may not be the best way to define the indicator. As well, rural route services and sub-urban route services are provided from urban post offices. For example, the entire province of New Brunswick is considered urban as no postal code has a zero in the second position. A definition must be established in order to use this type of indicator with precision and clarity. A measure based upon population
size or on population density and distance to regional center may serve as starting points for such a definition.

The problem of re-used postal codes presents a limitation in that only those patients who had a recorded address could undergo postal code verification. When a postal code was identified as correctly identifying an institution, then all instances of that postal code were taken as correct. Unfortunately patients with an old postal code that was not associated with an institution could not be followed up if they did not have a recorded address. Also, the address was taken from the BCCR and may not represent the place of usual residence. Although a limitation, this was restricted to 207 patients in the data set.

This research only describes where people have died. As administrative databases were used, no comment can be made on the congruence between the location of death and the patient's preferred place of death. Understanding the reasons for dying in a certain location are also eclipsed.

As this paper is a first step in an inter-provincial comparison, the categories for age and survival from the Nova Scotia research were used. As these categorizations may not be appropriate for the BC population, this research is limited by the conceptual framework implemented for the population in Nova Scotia. A common framework must be established in order to facilitate lucid comparisons.

**Conclusion**

A shift has occurred in the place-of-death for BC cancer patients between 1997 and 2003. As the population of BC continues to age, understanding the nature of these choices will be an essential component in meeting the needs of those who are completing their life. It is recommended that future research includes data from hospital separations, continuing care and the BC Medical Services Plan to improve our understanding of how health services are utilized at the end-of-life. Based upon the insights gained through the integration of these databases into the current research, qualitative approaches should be developed to capture how and why people make certain decisions about service utilization at the end of their life. These insights will be necessary to develop effective health policy and service delivery programs for all.
Appendix

ICD codes used to generate disease groups for CR and VS data for up to 2002 for End-of-Life with cancer research

Table 3-3: Nova Scotia Groupings for ICD codes. The British Columbia groups were identical to these groups.

<table>
<thead>
<tr>
<th>Disease Group</th>
<th>Vital Statistics (VS) Deaths</th>
<th>Cancer Registry (CR) cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Head and Neck</td>
<td>140-149,160,161</td>
<td>C00-C14,C30-C32</td>
</tr>
<tr>
<td>Gastrointestinal</td>
<td>150-159</td>
<td>C15-C26,C48</td>
</tr>
<tr>
<td>Thoracic</td>
<td>162-165</td>
<td>C33,C34,C37-C39,C45</td>
</tr>
<tr>
<td>Breast</td>
<td>174,175</td>
<td>C50</td>
</tr>
<tr>
<td>Gynecologic</td>
<td>179-184</td>
<td>C51-C58</td>
</tr>
<tr>
<td>Genitourinary</td>
<td>185-189</td>
<td>C60-C68</td>
</tr>
<tr>
<td>CNS</td>
<td>190-192</td>
<td>C69-C72</td>
</tr>
<tr>
<td>Hematologic</td>
<td>200-208</td>
<td>C81-C85,C88,C90-C95</td>
</tr>
<tr>
<td>Unknown Primary</td>
<td>196-199</td>
<td>C77-C80</td>
</tr>
</tbody>
</table>

† This table was constructed by the Nova Scotia Research team."
References


30. Martin, S. Canada Post [Personal communication 2004 April 20]


Background

Understanding the underlying cultural constructs which are associated with a cancer patient's utilization of health services and the location of death can be critical to identifying vulnerable populations. Research in Manitoba, Nova Scotia and Vancouver suggests that the place of residence is associated with health-related outcomes such as service utilization or location of death. This definition of the place of residence is too gross a measure (e.g. rural versus urban) to identify a coherent underlying social construct that gives insight other than proximal discrimination. Research upon populations from Toronto and Nova Scotia presents further frustrations as it suggests there is no socioeconomic status (SES) connection to these health outcomes.

Roos and Veugelers imply that the lack of association between health outcomes and SES is due to Canada's universal health care system. If this is true, then the conventional measures of SES (e.g. income, education, and occupation) may not be discriminants of health outcomes. Health Canada claims that "universal coverage for medically necessary health care services [is] provided on the basis of need, rather than the ability to pay" thus economic related measures of SES may not capture any systematic marginalization of peoples from health services.

This understanding will have repercussions upon the body of work from which Canadian researchers can draw inference. Consider the current work in the United States which investigates cancer-related health outcomes. Within the known literature, much emphasis is placed upon conventional measures of SES while the sociodemographic aspect is limited to race, age, gender and marital status. The consideration of cultural context extends only to region of residence which is summarised in a rural versus urban indicator.

Similar themes emerge when considering international research into the cultural determinants of the utilization of health services and the location of death for persons dying of cancer. By investigating the indicators used to understand the non-clinical factors that are associated to the use of health services and the place of death within cancer research, it became apparent that international comparisons must be made within the context of the health system in place in each of the countries being considered. This suggests that indicators that are relevant in the United States may not have the same influence in Canada, thus the use of the SES measures may be the source of the mixed results for Canadian research. In order to overcome this limitation and seek understanding from the international community, Canadian research should draw from sources that share similar health service ideologies, such as Sweden, Italy and Belgium. The challenge is that research upon these populations has focused upon the conventional measures of SES.

Although leadership in developing a more complete understanding of how people dying of cancer use end-of-life services based upon sociodemographic and economic characteristics (SDEC) is absent within the known international body of research for universal health care systems, Canadian research has not remained inactive. Research in Nova Scotia has utilised Canada Census data to create culture indicators other than SES.
measures for small geographic regions called enumeration areas (EA). Within the context of understanding the utility of letters concerning Pap smear information and the use of cervical cancer screening facilities, census based ecological indicators were created for ethnicity and language. Within the investigation, it was found that areas where French was the dominant mother tongue had a decreased risk of an abnormal smear. Unfortunately, the use of census derived SDEC for understanding the social context and effects upon health service utilization in Canada does not extend much beyond this research.

Although representing a departure from the exclusive use of SES measures in cancer research to indicate the SDEC context, the methodology employed in Nova Scotia serves only to highlight a predetermined aspect of a cultural characteristic. The researcher chooses French as mother tongue over others due to the large French community. An EA was considered to be Francophone if 50% or more of the people in the EA had French as their mother tongue. Although this appears to be a reasonable means by which to describe a small geographic region in order to infer a cultural context, they are inadequate in two respects: reliance upon a univariate description of the regions and the choice of cut-points.

The first is in the choice of a univariate decision versus a multivariate one. The current approach deconstructs the concept of culture into a series of discrete, well-defined partitions such as French, Black and Aboriginal identities. Each is a binary indicator that is constructed without reference to any other cultural feature. In the construction of these indicators, only a small amount of the information is used. Although useful for an initial investigation in order to build understanding of the underlying cultural constructs of a region, they do not represent the underlying multivariate nature of a cultural identity. The complex interplay of the cultural characteristics is eclipsed by this approach.

The second is the choice of cut points. The cut point of 50% for French as the mother tongue was chosen to define a Francophone community. This may appear sensible as it would identify EAs with a majority of people with French as their mother tongue. Although this presents a clear cut-point, it is not easily defendable against ambiguous cases. Consider two EAs. One has 52% French as mother tongue and the other has 48%. Is it reasonable to suggest that they should be classified differently without any other evidence of their dissimilarity? The researcher-determined cut-point of 50% suggests that these are very different EAs. Yet they may share similar underlying social constructs due to the similarity in mother tongues and in the proximal relationship to one another. An analogous situation would arise when the second EA has a proportion that is distinctly different than the first. If the second had 90% French as mother tongue then it may not be appropriate to group the two because of the differences, yet a 50% cut suggests that they are identical in terms of the underlying construct being measured. Choosing cut-points on a univariate exploration of the data casts a shadow over the underlying cultural constructs that obscures all cultural nuances for a geographically constrained construct resulting in a myopic view of culture.

Although providing "universal, comprehensive coverage for medically necessary hospital and physician services" is the federal mandate, SES measures may not successfully
assess whether or not this egalitarian representation of health services in Canada is equitably applied amongst all peoples. In order to understand the non-economic mechanisms of health service marginalization, SES measures must be supplemented with non-economic measures of sociodemographic and economic characteristics. Although cancer research upon Canadian populations has begun this endeavour, much development is needed in order to capture the multivariate nature of the underlying cultural constructs which may be serving as non-economic based forces of health service utilization marginalization. This research presents a preliminary step in the development of such an indicator. This multi-stage approach uses the empirical multivariate distribution for a chosen census characteristic in order to construct an indicator for the underlying cultural context. The objective of this study is to explore a novel approach for the development of census-based culture indicators.

Methodology

Data Source and Population

The data for this investigation came from the 2001 Canadian Census as reported by Statistics Canada for dissemination areas (DA). The dissemination area is the lowest level of data that is released by Statistics Canada. It is defined as a standard geographic area designed for data output which is predicated upon temporal stability, reduced area suppression, uniformity, intuitive boundaries and a compact shape. The dissemination areas were constructed to respect the Census Sub-Division and the Census-Tract boundaries and to be comparable to the 1996 census enumeration areas. Within the construction of the DAs, uniformity of numerical size and geographic shape was sought. 75% of the DAs had populations within the 400 to 700 respondents range. 10% had populations less than 400 with approximately 2% experiencing data suppression. Data suppression occurs when a DA has less than 40 respondents. 15% had populations over 700. Approximately 90% of the DAs retained a compact shape (square) with 10% exhibiting horseshoe, strip, cookie bit and donut shapes. These pertained to cul-de-sacs, DAs that boarder upon geographic boundaries, such as rivers, or municipal infrastructure boundaries, such as roads, and high density regions such as apartment blocks. The observational unit was the DA. In BC there are 6669 DAs. 97 of the DAs experienced data suppression and were removed from the exploration.

Study Variables

The census categories chosen for the investigation were ethnic origins (total responses), mother tongue (single response only) and religion. Ethnic origins has 186 sub-categories. Mother tongue has 80 sub-categories. Religion has 35 categories. The structure of the census data is such that the sub-categories may be further broken down, for example, the total population by mother tongue (category) can be broken down into single responses, multiple responses, and total responses (sub-categories) (Figure 4-1). Total responses can be further broken down into English, French and non-official languages. The non-official languages can be further broken down into specific languages, for example Chinese, Arabic, and Polish. For the purposes of this investigation, the variables were taken to be the most specific level reported for the category (e.g. a specific language, ethnic group, or religion).
As there were a large number of variables, they were first grouped according to some internal logic such as geography and ethnic groups for ethnic origins (Table 4-1), geography and language families for mother tongue (Table 4-2), and theological positions, historical evolution, and religious practices for religion (Table 4-3).

Table 4-1: Conceptual grouping of ethnic origins

<table>
<thead>
<tr>
<th>Conceptually Grouped Variables</th>
<th>Original Census Variables – Ethnic Origins (Total Response)</th>
</tr>
</thead>
<tbody>
<tr>
<td>French</td>
<td>Quebecois, Acadian, French</td>
</tr>
<tr>
<td>Canadian</td>
<td>Canadian</td>
</tr>
<tr>
<td>British Isles</td>
<td>English, Scottish, Irish, Welsh, British not indicated elsewhere</td>
</tr>
<tr>
<td>North-West European</td>
<td>German, Dutch, Norwegian, Swedish, Danish, Austrian, Belgian, Finnish, Swiss, Icelandic</td>
</tr>
<tr>
<td>Southern Europe and Mediterranean</td>
<td>Italian, Portuguese, Greek, Spanish</td>
</tr>
<tr>
<td>Eastern European</td>
<td>Ukrainian, Polish, Russian, Hungary-Magyar, Romanian, Croatian, Czech, Yugoslav not indicated elsewhere, Slovak, Serbian, Armenian</td>
</tr>
<tr>
<td>East Asia</td>
<td>Filipino, Vietnamese, Korean, Japanese, South Asian not indicated elsewhere</td>
</tr>
<tr>
<td>USA</td>
<td>American – USA</td>
</tr>
<tr>
<td>Caribbean</td>
<td>Jamaican, Haitian, West Indian, Trinidadian-Tobagonian</td>
</tr>
<tr>
<td>Latin America</td>
<td>Latin America not indicated elsewhere, Guyanese</td>
</tr>
<tr>
<td>Conceptually Grouped Variables</td>
<td>Original Census Variables – Mother Tongue (Single Response)</td>
</tr>
<tr>
<td>-------------------------------</td>
<td>-----------------------------------------------------------</td>
</tr>
<tr>
<td>Chinese</td>
<td>Cantonese, Mandarin, Hakka, Chinese not otherwise specified</td>
</tr>
<tr>
<td>Aboriginal</td>
<td>Cree, Inuktitut–Eskimo, Ojibway, Montagnais-Naskapi, Micmac, South Slave, Dogrib, Chipewyan, Tlingit, Kutchin-Gwich in Loucheux, Dakota Sioux, Blackfoot</td>
</tr>
<tr>
<td>Southern European</td>
<td>Italian, Greek, Maltese</td>
</tr>
<tr>
<td>Eastern European</td>
<td>Polish, Ukrainian, Hungarian, Croatian, Russian, Czech, Romanian, Slovak, Macedonian, Serbian, Slovenian, Bulgarian, Serbo-Croatian</td>
</tr>
<tr>
<td>North-West European</td>
<td>Dutch, Finnish, Danish, Norwegian, Swedish, Flemish, Gaelic Languages</td>
</tr>
<tr>
<td>Baltic</td>
<td>Estonia, Lithuanian, Latvian-Lettish</td>
</tr>
<tr>
<td>Middle East</td>
<td>Arabic, Persian-Farsi, Armenian, Hebrew, Turkish, Kurdish, Pashto</td>
</tr>
<tr>
<td>Indian Sub-Continent</td>
<td>Gujarati, Hindi, Tamil, Urdu, Bengali</td>
</tr>
<tr>
<td>East Asia</td>
<td>Tagalog-Filipino, Vietnamese, Korean, Japanese, Khmer-Cambodian, Lao, Malay-Bahasa, Malayalam, Thai</td>
</tr>
<tr>
<td>Spanish-Portuguese</td>
<td>Spanish, Portuguese</td>
</tr>
<tr>
<td>English</td>
<td>English</td>
</tr>
<tr>
<td>French</td>
<td>French</td>
</tr>
<tr>
<td>German</td>
<td>German</td>
</tr>
<tr>
<td>Punjabi</td>
<td>Punjabi</td>
</tr>
<tr>
<td>Other</td>
<td>Creoles, Yiddish, Other languages</td>
</tr>
</tbody>
</table>

Table 4-2: Conceptual groupings of mother tongue
Table 4-3: Conceptual groupings of religion

<table>
<thead>
<tr>
<th>Conceptually Grouped Variables</th>
<th>Original Census Variables for Religion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Catholic</td>
<td>Roman Catholic, Ukrainian Catholic</td>
</tr>
<tr>
<td>Orthodox</td>
<td>Greek Orthodox, Orthodox not included elsewhere, Ukrainian Orthodox, Serbian Orthodox</td>
</tr>
<tr>
<td>Protestant</td>
<td>United Church, Anglican, Christian not included elsewhere, Baptist, Lutheran, Protestant not included elsewhere, Presbyterian, Pentecostal, Salvation Army, Christian Reformed Church, Evangelical Missionary Church, Christian and Missionary Alliance, Adventist, Non-denominational, Hutterite, Methodist, Brethren in Christ, Mennonite</td>
</tr>
<tr>
<td>None</td>
<td>No religion</td>
</tr>
<tr>
<td>Muslim</td>
<td>Muslim</td>
</tr>
<tr>
<td>Sikh</td>
<td>Sikh</td>
</tr>
<tr>
<td>Jewish</td>
<td>Jewish</td>
</tr>
<tr>
<td>Other</td>
<td>Buddhist, Hindu, Jehovah's Witnesses, Church of Jesus Christ of Latter-day Saints, Aboriginal spirituality, Pagan</td>
</tr>
</tbody>
</table>

Each conceptual group measure was the sum of the individual ones. In several situations, the underlying constructs prevented a clear allocation to a group, for example the Church of Jesus Christ of Latter-day Saints shares themes with Protestant churches, yet there are theological distinctions between the two which would negate the identification of Mormon Church to be with those churches that ascribe to the basic tenets of faith as upheld by the other Protestant churches identified in the Protestant group. In these situations, either a separate conceptual variable was created or they were included in the “Other” category. In this case, the Mormon variable was allocated to the “Other” group (Table 4-3).

**Data Transformation**

The prevalence of a cultural attribute within a DA was expressed as the percentage of individuals in the DA that had that particular attribute. This transformation, applied to the conceptually grouped variables, provides a measure of cultural density. The new culture indicators will be premised upon the prevalence of the attribute rather than the numbers of people who share the attribute.

**Statistical Tools – An Overview**

For each census category, a set of variables is obtained. The data did not contain a variable which classified the dissemination areas or any information to organize the raw data. Given this situation, statistical methods are needed to discover a useful structure. In order to discover the underlying structure of the data, a series of descriptive and data
mining techniques were employed. Univariate descriptions and bivariate scatter plots were used to establish an initial conceptual framework for the data. This schema served as a guide for the subsequent data mining exploration.

Data mining can be divided into two general areas: supervised and unsupervised learning. Supervised learning trains a data set upon an expressed outcome (e.g. logistic regression). Unsupervised learning is an exploration to understand the structure of the data for the intent of generating data-related hypotheses (statistical or otherwise). The development of the multivariate-defined culture indicator was predicated upon unsupervised learning techniques. These techniques were chosen as there was no explicit response or outcome variable.

As previously indicated, three cultural constructs were explored: ethnicity, mother tongue and religion. Each cultural characteristic had many sub-categories, thus each census category represented a high dimensional space, for example the ethnic origins census characteristic had 61 ethnicities; each represented one dimension (variable or axis) in the ethnic origins space. A succinct expression of these cultural characteristics was desired, thus the number of variables (dimensions) was reduced. The conceptual grouping served as the first step of reduction. For example, the conceptual grouping of ethnic origins reduced the space from 61 dimensions to 18; one dimension for each conceptual group.

Principal component analysis was applied to the conceptually reduced data for several reasons. The conceptual groups represented one way of organizing data, based upon known similarities of between various groups. With principal component analysis, the data could be further organized based upon the multivariate structure of these groups. By using principal components for this second stage of organization, further data reduction could be obtained. This allows for the data to be compressed into a smaller set of variables while still retaining much of the original variability. As well, this organization of the data helps to reduce the noise in the data. As the data is in a lower dimensional variable space it is easier to visualize the data. As there is less noise in the data, better results can be obtained by using the principal components as input variables for subsequent analysis.

With the data adequately compressed, the DAs were grouped based upon similarities as expressed by the principal component variables. K-means was used to group the DAs and the number of groups was chosen through an exploration based upon the knowledge gained from the univariate and bivariate explorations. Once the number of groups was determined, using the criteria of within-group sum of squares and the ease of interpreting the groups, the groups were statistically described and given an evocative label.

Finally, the reasonability of the groups was explored. British Columbian municipalities were selected to assess how well the indicator described the cultural composition of each municipality. Agreement between the indication of cultural groups within a city and the current understanding of the cultural composition of the city was sought. As well, the groups were tested for internal consistency through a series of chi-squared tests. SAS was used to generate the data sets and to explore group validity. R was used for the univariate, bivariate, and unsupervised learning procedures.
Principal Component Analysis (PCA)

A brief overview of PCA will be given. For a more complete discussion see Venables and Ripley\textsuperscript{34} or Johnson and Wichern\textsuperscript{35}. PCA is a data-driven method by which data in a high dimensional variable space is transformed into a smaller dimensional variable space. This is done by creating linear combinations of the original variables that have maximal variance. The specific linear combinations are determined by the structure of the covariance-variance matrix.

For a given data set, let $\Sigma$ represent the covariance-variance matrix. As the covariance-variance matrix is positive definite, spectral decomposition splits $\Sigma$ into two portions: a diagonal matrix of eigenvalues, $\Lambda$, and an orthogonal one of eigenvectors, $E$, thus $\Sigma = E^T \Lambda E$. The columns of $E$ are the loadings. These are the coefficients for the linear combinations of original variables. The size of the loading indicates the weight given to each of the original variables. The entries of $\Lambda$ are a measure of variation associated with each of the columns of $E$.

As the primary objective for the use of PCA in this study was for variable reduction, the set of principal components which explained much of the variation in the data was determined. Three measures, based upon the eigenvalue for the $i^{th}$ component, $\lambda_i$, were used to identify the desired number of components: the proportion of variance explained by the component, the overall cumulative proportion of variance explained and the scree plot. The $i^{th}$ component explains,

$$\frac{\lambda_i}{\sum \lambda_i},$$

of the variation. The variation explained by the first $i$ components is,

$$\frac{\sum_{1}^{i} \lambda_i}{\sum \lambda_i}.$$  

The scree plot is a useful aid to determine the number of principal components that are desired. With the components ordered from the one explaining the most variation in the data to the one explaining the least, the scree plot is a plot of the variation in the data, $\lambda$, versus $I$, the index. An elbow in the curve is desirable (Figure 4-2). The principal components chosen are those that precede the elbow, which are all the components that capture much of the variation in the data.
**K-Means**

K-means is a clustering algorithm which chooses a predetermined number of group or cluster centres that minimises the within-class sum of squares for each cluster. This iterative process begins with the random selection of k initial centres. The distance from each observational unit, DA, to each centre was found. Each point was associated with the centre to which it is closest resulting in k clusters. New centres (means) were calculated for each cluster. This would comprise the first iteration. The second iteration began with the removal of the associations for each cluster. The distance between each DA and each of the k new centres was determined. The DA were now associated with one of the k centres resulting in k new clusters. The clusters were different from the previous k clusters. New centres were calculated for each of the clusters. This would comprise the second iteration. This process continued until the k centres became stable.

By selecting the minimum distance as the measure by which each observational unit is associated with a centre, the algorithm sought to minimize the within-cluster sum of squares,
where \( z_i \) is the \( i^{th} \) DA, \( \mu_i \) is the centre of the \( j^{th} \) cluster, and \( C_j \) is the \( j^{th} \) cluster. By minimizing the within-cluster sum of squares, the clusters exhibit a desired level of homogeneity within each cluster. Cluster homogeneity can be further increased or decreased by increasing or decreasing the number of clusters that are initially chosen. For the purposes of this exploration, the Euclidean distance, as implied in (0.1), was used.

In order to assess the quality of the clustering two criteria were considered. The first was the total within-cluster sum of squares. This is the sum of (0.1) over all clusters and is defined as,

\[
\sum_{i \in C_j} \left\| z_i - \mu_j \right\|^2.
\] (0.2)

The second was the ability to interpret reasonably the results. A balance was desired between a low total within-cluster sum of squares and the ability to give an evocative label to the clusters.

**Results**

In order to build further understanding of the development process, a detailed explanation will be given for the ethnic origins census variable. Only the final results will be given for mother tongue and religion census characteristics. This detailed look into the construction of the variables will serve to highlight the innovations in the development of culture indicators and their complexity.

**Ethnic Origins**

As seen in Table 4-1, the original census sub-categories for the category of ethnic origins were grouped into meaningful geographical regions. PCA was performed upon this conceptually grouped set of ethnic origins census variables. The results of this exploration would serve as the basis for grouping the dissemination areas. As mentioned earlier, the scree plot, the proportion of variance and the cumulate proportion of variance were used to determine the number of principal components used to group the dissemination areas. The scree plot (Figure 4-2) suggests that between five and eight principal components are reasonable. There is no clear elbow in the curve; it is clear that the curve levels off after then eighth principal component. Since the goal is to reduce the dimensionality of the problem, only a few principal components are desired. From Table 4-4, it can be seen that there are large differences between the first three components. The differences in the proportion of variance explained begin to decline rapidly after the fifth or sixth principal component. As well, over 91% of the variation in the data is explained by the first five components. Either 5 or 6 components appear to be reasonable and as the objective was to minimize the dimensionality of the problem, five
components were taken. The loadings for these five components represent the coefficients of the linear combination of the conceptually grouped census variables. Based upon these linear combinations, new values were calculated for the dissemination areas.

Table 4-4: PCA proportion of variance and cumulative proportion of variance for the conceptually grouped ethnic origins census variables

<table>
<thead>
<tr>
<th>Principal Component</th>
<th>Proportion of Variance</th>
<th>Cumulative Proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Component 1</td>
<td>0.412</td>
<td>0.412</td>
</tr>
<tr>
<td>Component 2</td>
<td>0.293</td>
<td>0.705</td>
</tr>
<tr>
<td>Component 3</td>
<td>0.103</td>
<td>0.808</td>
</tr>
<tr>
<td>Component 4</td>
<td>0.067</td>
<td>0.875</td>
</tr>
<tr>
<td>Component 5</td>
<td>0.039</td>
<td>0.914</td>
</tr>
<tr>
<td>Component 6</td>
<td>0.030</td>
<td>0.944</td>
</tr>
<tr>
<td>Component 7</td>
<td>0.019</td>
<td>0.963</td>
</tr>
<tr>
<td>Component 8</td>
<td>0.014</td>
<td>0.977</td>
</tr>
<tr>
<td>Component 9</td>
<td>0.011</td>
<td>0.988</td>
</tr>
<tr>
<td>Component 10</td>
<td>0.004</td>
<td>0.992</td>
</tr>
</tbody>
</table>

K-means was used to group the dissemination areas in the five-dimensional principal component space. The univariate analysis suggested that Canadian, British Isles, Chinese, Aboriginal, and East Indian communities may exist. The plot of the total within-cluster sum of squares versus the index for 2 through 25 clusters begins to level off around the eighth cluster (Figure 4-3). Significant reductions in the total within-cluster sum of squares are seen between two and three clusters and between three and four clusters.
The in-depth exploration with respect to the within-cluster sum of squares and the ability to interpret the clusters was restricted to investigations utilizing four through seven clusters.

Six was selected as the best number of clusters because the gains in the reduction of the total within-cluster sum of squares was minimal and the ability to interpret the clusters was considerably more difficult. Six clusters yielded a total within-cluster sum of squares of 99.5 whereas five clusters had 118.7 and seven clusters had 90.7. Although the clusters were defined in the principal component space, it was possible to identify the groups in the original conceptual variable space. The multivariate nature of the groups was summarized by the univariate measures for each of the eighteen conceptual groups (Table 4-5). Based upon these descriptions, an evocative label was given to each of the six groups. The “Heterogeneous” group has pockets of the various ethnic origins but no one ethnic origin appears to have a strong overall representation within this group. The “Aboriginal Presence” cluster has a strong aboriginal presence as the proportion of aboriginals in these dissemination areas ranges from 44% to 100%. One exceptional feature of this group is the noticeable absence of East Indian, Indian Sub-Continent and Caribbean origins. Pockets of European, Asian and Canadian origins can be seen. The dissemination areas in the “Chinese Presence” cluster have moderate to strong Chinese

**Figure 4-3:** Total within-cluster sum of squares (SS) versus the number of clusters
ethnic origin presence with proportions ranging from just over 28% to just fewer than 87%. The presence of other ethnic origins is more diverse than with the Aboriginal Presence cluster. The “British Isles Presence” cluster has a strong presence of ethnic origins from British Isles ranging from 37% to just fewer than 88%. As with the previous clusters, there are visible pockets of ethnic origins other than British Isles with Canadian origins exhibiting the strongest minority presence. The “East Indian Presence” group has dissemination areas with East Indian ethnic origins ranging from about 17% to 100%. The “Americas and Europe” cluster has dissemination areas with strong Canadian, British Isles, and North-Western Europe presence. This cluster also contains the dissemination areas with higher proportions of Latin American origins.

Table 4-5: Statistical descriptions of the population proportions for the six ethnic origin groups expressed by only the minimum value, the median and the maximum value

<table>
<thead>
<tr>
<th>Conceptual Group</th>
<th>Brief Statistical Description</th>
<th>Heterogeneous (1)</th>
<th>Aboriginal Presence (2)</th>
<th>Chinese Presence (3)</th>
<th>British Isles Presence (4)</th>
<th>East Indian Presence (5)</th>
<th>Americas and Europe (6)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Canadian</td>
<td>Min</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Median</td>
<td>0.100</td>
<td>0.000</td>
<td>0.060</td>
<td>0.138</td>
<td>0.079</td>
<td>0.155</td>
</tr>
<tr>
<td></td>
<td>Max</td>
<td>0.267</td>
<td>0.400</td>
<td>0.247</td>
<td>0.550</td>
<td>0.265</td>
<td>0.765</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>0.105</td>
<td>0.025</td>
<td>0.065</td>
<td>0.140</td>
<td>0.083</td>
<td>0.161</td>
</tr>
<tr>
<td>French</td>
<td>Min</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Median</td>
<td>0.031</td>
<td>0.000</td>
<td>0.013</td>
<td>0.049</td>
<td>0.022</td>
<td>0.054</td>
</tr>
<tr>
<td></td>
<td>Max</td>
<td>0.122</td>
<td>0.222</td>
<td>0.103</td>
<td>0.294</td>
<td>0.116</td>
<td>0.286</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>0.034</td>
<td>0.021</td>
<td>0.016</td>
<td>0.051</td>
<td>0.026</td>
<td>0.056</td>
</tr>
<tr>
<td>British Isles</td>
<td>Min</td>
<td>0.057</td>
<td>0.000</td>
<td>0.000</td>
<td>0.373</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Median</td>
<td>0.262</td>
<td>0.068</td>
<td>0.131</td>
<td>0.475</td>
<td>0.160</td>
<td>0.362</td>
</tr>
<tr>
<td></td>
<td>Max</td>
<td>0.484</td>
<td>0.375</td>
<td>0.375</td>
<td>0.875</td>
<td>0.333</td>
<td>0.486</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>0.263</td>
<td>0.080</td>
<td>0.134</td>
<td>0.486</td>
<td>0.155</td>
<td>0.352</td>
</tr>
<tr>
<td>North-Western Europe</td>
<td>Min</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Median</td>
<td>0.092</td>
<td>0.000</td>
<td>0.043</td>
<td>0.147</td>
<td>0.074</td>
<td>0.189</td>
</tr>
<tr>
<td></td>
<td>Max</td>
<td>0.278</td>
<td>0.333</td>
<td>0.186</td>
<td>0.375</td>
<td>0.299</td>
<td>0.790</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>0.096</td>
<td>0.033</td>
<td>0.048</td>
<td>0.148</td>
<td>0.081</td>
<td>0.197</td>
</tr>
<tr>
<td>Southern Europe and Mediterranean</td>
<td>Min</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Median</td>
<td>0.039</td>
<td>0.000</td>
<td>0.033</td>
<td>0.021</td>
<td>0.017</td>
<td>0.023</td>
</tr>
<tr>
<td></td>
<td>Max</td>
<td>0.325</td>
<td>0.200</td>
<td>0.303</td>
<td>0.196</td>
<td>0.179</td>
<td>0.232</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>0.051</td>
<td>0.007</td>
<td>0.049</td>
<td>0.026</td>
<td>0.025</td>
<td>0.030</td>
</tr>
<tr>
<td>Eastern European</td>
<td>Min</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Median</td>
<td>0.065</td>
<td>0.000</td>
<td>0.033</td>
<td>0.063</td>
<td>0.037</td>
<td>0.081</td>
</tr>
<tr>
<td></td>
<td>Max</td>
<td>0.382</td>
<td>0.182</td>
<td>0.159</td>
<td>0.385</td>
<td>0.207</td>
<td>0.400</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>0.073</td>
<td>0.008</td>
<td>0.038</td>
<td>0.067</td>
<td>0.043</td>
<td>0.087</td>
</tr>
<tr>
<td>East Asian</td>
<td>Min</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Median</td>
<td>0.060</td>
<td>0.000</td>
<td>0.077</td>
<td>0.003</td>
<td>0.049</td>
<td>0.007</td>
</tr>
<tr>
<td></td>
<td>Max</td>
<td>0.452</td>
<td>0.118</td>
<td>0.349</td>
<td>0.176</td>
<td>0.395</td>
<td>0.277</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>0.079</td>
<td>0.003</td>
<td>0.088</td>
<td>0.012</td>
<td>0.066</td>
<td>0.019</td>
</tr>
<tr>
<td>United States</td>
<td>Min</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Median</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.007</td>
<td>0.000</td>
<td>0.006</td>
</tr>
<tr>
<td></td>
<td>Max</td>
<td>0.073</td>
<td>0.071</td>
<td>0.083</td>
<td>0.104</td>
<td>0.047</td>
<td>0.212</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>0.006</td>
<td>0.002</td>
<td>0.004</td>
<td>0.009</td>
<td>0.003</td>
<td>0.010</td>
</tr>
</tbody>
</table>
From the statistical descriptions of the clusters, it would be expected that Aboriginal Presence, Chinese Presence and the East Indian Presence clusters would exhibit a degree of separation from the other groups when plotted. In Figure 4-4, the Heterogeneous, British Isles Presence and the Americas and Europe groups are close together in each plot. The Aboriginal Presence exhibits clear separation in plots 4-4a, e, f, and g. It shows minor separation in plots 4-4b, c, and d. This level of separation can be seen in the low proportions of non-Aboriginal conceptual groups. The Chinese Presence cluster
is far from the other clusters in plot 4-4a and it has minor separation in plots 4-4b and 4-4e. This cluster is close to the Heterogeneous, British Isles Presence and the Americas and Europe clusters in the remaining plots. This would validate the higher proportions of the Western European and Canadian conceptual ethnic origin categories in this cluster. The East Indian Presence cluster exhibits high and moderate levels of separation in plots 4-4b, c, e, f, h, and i.

![Figure 4-4: Scatter plots of the first five principal components with the centres for the six ethnic origins clusters: Heterogeneous (1), Aboriginal Presence (2), Chinese Presence (3) British Isles Presence (4), East Indian Presence (5) and the Americas and Europe (6)](image)

The culture indicator was initially assessed by investigating the proportion of the six ethnic origin groups in a census sub-division (Table 4-6). Surrey and Abbotsford have the highest proportion of East Indian Presence dissemination areas, 23.4% and 12.8%
respectively. Richmond, Vancouver, and Burnaby have the highest proportion of Chinese Presence dissemination areas, 29.4%, 30.4%, and 11% respectively. Victoria, West Vancouver, and Nelson have the highest proportion of British Isles dissemination areas, 84.7%, 79.2%, and 72.7% respectively. These regional attributes agree with the general profile of the various cities.

Table 4-6: Percentage of the six ethnic origins clusters in 22 British Columbian municipalities

<table>
<thead>
<tr>
<th>Census Subdivision</th>
<th>Heterogeneous</th>
<th>Aboriginal Presence</th>
<th>Chinese Presence</th>
<th>British Isles Presence</th>
<th>East Indian Presence</th>
<th>Americas and Europe</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abbotsford</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>12.8%</td>
<td>12.8%</td>
<td>74.4%</td>
</tr>
<tr>
<td>Burnaby</td>
<td>67.6%</td>
<td>0.0%</td>
<td>11.0%</td>
<td>6.3%</td>
<td>2.2%</td>
<td>12.9%</td>
</tr>
<tr>
<td>Chilliwack</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>21.4%</td>
<td>0.0%</td>
<td>78.6%</td>
</tr>
<tr>
<td>Coquitlam</td>
<td>35.2%</td>
<td>0.0%</td>
<td>3.4%</td>
<td>25.7%</td>
<td>0.0%</td>
<td>35.8%</td>
</tr>
<tr>
<td>Cranbrook</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>25.0%</td>
<td>0.0%</td>
<td>75.0%</td>
</tr>
<tr>
<td>Delta</td>
<td>4.8%</td>
<td>0.0%</td>
<td>0.6%</td>
<td>47.9%</td>
<td>8.4%</td>
<td>38.3%</td>
</tr>
<tr>
<td>Fort St. John</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>100.0%</td>
</tr>
<tr>
<td>Kamloops</td>
<td>0.8%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>48.5%</td>
<td>0.0%</td>
<td>50.8%</td>
</tr>
<tr>
<td>Kamloops 1</td>
<td>0.0%</td>
<td>40.0%</td>
<td>0.0%</td>
<td>40.0%</td>
<td>0.0%</td>
<td>20.0%</td>
</tr>
<tr>
<td>Kelowna</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>21.3%</td>
<td>0.0%</td>
<td>78.8%</td>
</tr>
<tr>
<td>Langley</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>44.0%</td>
<td>0.0%</td>
<td>56.0%</td>
</tr>
<tr>
<td>Maple Ridge</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>55.3%</td>
<td>0.0%</td>
<td>44.7%</td>
</tr>
<tr>
<td>Nelson</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>72.7%</td>
<td>0.0%</td>
<td>27.3%</td>
</tr>
<tr>
<td>New Westminster</td>
<td>4.3%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>34.8%</td>
<td>3.3%</td>
<td>57.6%</td>
</tr>
<tr>
<td>North Vancouver</td>
<td>4.2%</td>
<td>0.5%</td>
<td>0.0%</td>
<td>66.5%</td>
<td>0.0%</td>
<td>28.8%</td>
</tr>
<tr>
<td>Prince George</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>23.1%</td>
<td>0.0%</td>
<td>76.9%</td>
</tr>
<tr>
<td>Prince Rupert</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>6.3%</td>
<td>0.0%</td>
<td>93.8%</td>
</tr>
<tr>
<td>Richmond</td>
<td>54.5%</td>
<td>0.0%</td>
<td>29.4%</td>
<td>7.2%</td>
<td>1.3%</td>
<td>7.7%</td>
</tr>
<tr>
<td>Surrey</td>
<td>9.3%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>22.8%</td>
<td>23.4%</td>
<td>44.5%</td>
</tr>
<tr>
<td>Vancouver</td>
<td>30.5%</td>
<td>0.0%</td>
<td>30.4%</td>
<td>22.0%</td>
<td>3.4%</td>
<td>13.7%</td>
</tr>
<tr>
<td>Victoria</td>
<td>0.7%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>84.7%</td>
<td>0.0%</td>
<td>14.6%</td>
</tr>
<tr>
<td>West Vancouver</td>
<td>13.0%</td>
<td>0.0%</td>
<td>1.3%</td>
<td>79.2%</td>
<td>0.0%</td>
<td>6.5%</td>
</tr>
</tbody>
</table>
**Mother Tongue**

The same steps were taken for the conceptually grouped mother tongue variables (Table 4-2). The first three principal components, which explained 91% of the variation in the data, were taken. In this reduced variable space, various numbers of clusters were explored for low total within-cluster sum of squares and ease of interpretation. Five clusters were chosen with a total within-cluster sum of squares of 71.54. The five centres were labelled as:

- Punjabi Presence,
- Dominant English,
- Heterogeneous with Dominant English,
- Chinese Presence, and
- Strong English Presence.

The distributions for each group are given in the Appendix (Table 4-12).

When the five centres are visualised in the principal component space, it can be seen that there is good separation between the clusters in most of the pair-wise plots (Figure 4-5). In figure 4-5, a high degree of separation exists between cluster 1 and the other four. Cluster 4 has good separation in all three plots as well. Clusters 2, 3, and 5 have moderate separation in plots 4-5a and b. In plot 4-5c, clusters 5 and 3 are very close together with cluster 2 exhibiting only moderate separation from them.

![Figure 4-5](image-url)

*Figure 4-5: Scatter plots of the first three principal components with the centres for the five mother tongue clusters: Punjabi Presence (1), Dominant English (2), Heterogeneous with Dominant English (3), Chinese Presence (4), and Strong English Presence (5)*
With such close proximity of clusters 3 and 5 in plot 4-5c, it is expected that these clusters will share similar features. The labels suggest the commonality of English as mother tongue between cluster 3, “Heterogeneous with Dominant English”, and cluster 5, “Strong English Presence”. In cluster 3, the mother tongue profile is well mixed with many different language pockets emerging, but there is a noticeable presence of English which ranges from 12% to 68%. As well, there is a minor Chinese presence throughout these dissemination areas. Ranging from 0% to 44%, this minor presence rises quickly from 0% to 16% by the first quartile and reaches 31% by the third quartile. Cluster 5 has a Strong English Presence as it ranges from 48% to 81% within the dissemination areas. By the first quartile, this presence becomes a majority (65%). As with the third cluster, there are some strong non-English representations. These are fewer and not as large, for example the Chinese presence ranges from 0% to 28% with the third quartile at 8%.

Cluster 2, “Dominant English”, has the strongest English as the mother tongue presence, ranging from 78% to 100%, thus the slight separation of this cluster from that of 3 and 5. A number of languages have maximums between 15% and 22% such as Aboriginal, German, French, North-Western European, Indian sub-continental, East Asian, Spanish-Portuguese, Punjabi, and Eastern European, yet none of the third quartiles exceed 2%. From this, the proximity of clusters 3 and 5 to cluster 2 is reasonable.

Clusters 1 and 4 should represent distinct mother tongue groups. Cluster 1 has been identified as the “Punjabi Presence” cluster. Here the presence of this language ranges from 16% to 90%. Although it does not have such a distinct presence across all dissemination areas in this cluster, the first quartile is 32%, thus 75% of the dissemination areas in this cluster had more than a third of the people with Punjabi as their mother tongue. There is an Dominant English in this cluster, ranging from 0% to 61%. The first quartile is 24%, thus 75% of the dissemination areas have at least a quarter of the people with English as their mother tongue.

Cluster 4 has been identified as the “Chinese Presence” cluster. Here the presence of Chinese languages ranges from 26% to 87%. In all dissemination areas in this cluster at least a quarter of all the people have a Chinese language as their mother tongue. The first quartile is 42%, thus this minority quickly rises to be a strong presence. Again English does have a presence, but it is not as strong as with the Punjabi cluster, ranging from 4% to 49% with the third quartile being 34%.

When considering the distribution of these clusters with census sub-divisions, trends similar to that of ethnic origins emerge (Table 4-7). Surrey and Abbotsford have the largest Punjabi percentage of dissemination areas in the Punjabi Presence cluster. Richmond, Vancouver and Burnaby have the highest proportion of Chinese Presence dissemination areas. Nelson, Cranbrook and Kamloops 1 have all dissemination areas identified with the Dominant English cluster.
### Table 4-7: Percentage of the five mother tongue clusters in 22 British Columbian municipalities

<table>
<thead>
<tr>
<th>Census Subdivision</th>
<th>Punjabi Presence</th>
<th>Dominant English</th>
<th>Heterogeneous with Dominant English</th>
<th>Chinese Presence</th>
<th>Strong English Presence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abbotsford</td>
<td>15.1%</td>
<td>48.3%</td>
<td>1.2%</td>
<td>0.0%</td>
<td>35.5%</td>
</tr>
<tr>
<td>Burnaby</td>
<td>0.9%</td>
<td>1.6%</td>
<td>66.0%</td>
<td>12.3%</td>
<td>19.2%</td>
</tr>
<tr>
<td>Chilliwack</td>
<td>0.0%</td>
<td>95.7%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>4.3%</td>
</tr>
<tr>
<td>Coquitlam</td>
<td>0.0%</td>
<td>17.3%</td>
<td>29.6%</td>
<td>3.4%</td>
<td>49.7%</td>
</tr>
<tr>
<td>Cranbrook</td>
<td>0.0%</td>
<td>100.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>Delta</td>
<td>11.4%</td>
<td>47.9%</td>
<td>3.6%</td>
<td>1.8%</td>
<td>35.3%</td>
</tr>
<tr>
<td>Fort St. John</td>
<td>0.0%</td>
<td>100.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>Kamloops</td>
<td>0.0%</td>
<td>89.4%</td>
<td>0.8%</td>
<td>0.0%</td>
<td>9.8%</td>
</tr>
<tr>
<td>Kamloops 1</td>
<td>0.0%</td>
<td>100.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>Kelowna</td>
<td>0.0%</td>
<td>78.1%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>21.9%</td>
</tr>
<tr>
<td>Langley</td>
<td>0.0%</td>
<td>86.9%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>13.1%</td>
</tr>
<tr>
<td>Maple Ridge</td>
<td>0.0%</td>
<td>89.3%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>10.7%</td>
</tr>
<tr>
<td>Nelson</td>
<td>0.0%</td>
<td>100.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>New Westminster</td>
<td>4.3%</td>
<td>33.7%</td>
<td>3.3%</td>
<td>0.0%</td>
<td>58.7%</td>
</tr>
<tr>
<td>North Vancouver</td>
<td>0.0%</td>
<td>33.0%</td>
<td>5.1%</td>
<td>0.0%</td>
<td>61.9%</td>
</tr>
<tr>
<td>Prince George</td>
<td>0.7%</td>
<td>85.1%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>14.2%</td>
</tr>
<tr>
<td>Prince Rupert</td>
<td>0.0%</td>
<td>75.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>25.0%</td>
</tr>
<tr>
<td>Richmond</td>
<td>0.9%</td>
<td>3.8%</td>
<td>45.1%</td>
<td>36.2%</td>
<td>14.0%</td>
</tr>
<tr>
<td>Surrey</td>
<td>23.7%</td>
<td>24.7%</td>
<td>7.6%</td>
<td>0.0%</td>
<td>44.0%</td>
</tr>
<tr>
<td>Vancouver</td>
<td>3.1%</td>
<td>8.9%</td>
<td>28.5%</td>
<td>32.1%</td>
<td>27.5%</td>
</tr>
<tr>
<td>Victoria</td>
<td>0.0%</td>
<td>77.4%</td>
<td>0.7%</td>
<td>0.0%</td>
<td>21.9%</td>
</tr>
<tr>
<td>West Vancouver</td>
<td>0.0%</td>
<td>50.6%</td>
<td>15.6%</td>
<td>1.3%</td>
<td>32.5%</td>
</tr>
</tbody>
</table>

An interesting feature revealed by this exploration is the contrast between Abbotsford, Chilliwack and Langley. Abbotsford is situated between Chilliwack and Langley on the Trans-Canada highway and are both approximately 30 kilometres from Abbotsford. Interestingly, Langley and Chilliwack have more in common with each other than with Abbotsford with respect to the distribution of mother tongue languages. Both Chilliwack and Langley have a high proportion of dissemination areas identified with the Dominant English cluster, 95.7% and 86.9% respectively. Abbotsford has only 48.3%. The balance of dissemination areas, for Chilliwack and Langley fall in the Strong English Presence cluster. Abbotsford, unlike these two has 35.5% in that cluster and has 15.1% of the dissemination areas identified with the Punjabi Presence cluster. From a mother tongue perspective, Chilliwack and Langley have more in common with cities such as Nelson, Cranbrook and Prince George than with Abbotsford. These regional attributes agree with the general profile of the various cities.
Religion
From the PCA, four principal components, which explained just less than 97% of the variation in the data, were taken. Six clusters were chosen as they had a low total within-cluster sum of squares (145.3) and they could be easily described. Figure 4-6 reveals that two clusters, 2 and 4, have good separation from the other four clusters. Clusters 1, 3, 5, and 6 have reasonable separation in most scatter plots, but common themes are expected between them.

Figure 4-6: Scatter plots of the first four principal components with the centres for the six religion clusters

Clusters 1, 3, 5, and 6 are generally close together. This proximity is due to the subtle differences with respect to the presence of Protestants and those who have no religious affiliation. Clusters 1 and 6 are both heterogeneous with respect to religious affiliation. Cluster 1 has a moderate presence of Protestants (25% to 58%) and people with no religious affiliation (25% to 62%). Protestants in cluster 6 range from 0% to 45% and no
religions, ranging from 0% to 47%. Also, the Catholic presence in this cluster is similar to that of the Protestants and those with no religious affiliation, ranging from 0% to 45% with the first quintile just over 23%. These similarities can be seen in the close proximity of clusters 1 and 6 in figure 4-6a, 4-6b, and 4-6c. The differences between the two clusters are mainly experienced by the remaining religious identities.

Cluster 2 is labelled “Catholic Presence” (Appendix, Table 4-13). The proportion of Catholics, both Roman and Ukrainian, ranges from 42% to 100%. The first quartile is just less than 52%, thus 75% of all the dissemination areas in this cluster have a majority of Catholics. As usual, there are other religions represented in this cluster, but none exhibit the consistent presence that is presented by those who identify with the Catholic religion.

Cluster 3 has been labelled “Protestant Presence” (Appendix, Table 4-13). It ranges from just less than 37% to 100% with a median of approximately 53%. Catholic and no religious affiliation groups are present but do not exceed 45% as their maximum. They have medians less than or equal to 25%. The other religious groups have low representations in this cluster.

Cluster 4 is labelled “Sikh Presence” (Appendix, Table 4-13). The proportion of Sikhs ranges from just fewer than 18% to just over 92%. The first quartile is 28.6% with a median of 36.5%. This indicates that there is a moderate to strong Sikh presence in these dissemination areas. There is a Protestant presence within this cluster with ranges from 0% to 44% with a third quintile of 24%. Again, other religions are represented in this cluster to varying degrees.

Cluster 5 has been called the “No Religion Presence” cluster (Appendix, Table 4-13). Here, dissemination areas range from just over 33% to 100% in the proportion of people identifying no religious affiliation. The first quartile reaches 47%. The Protestant presence does not exceed 37% and the Catholic presence does not exceed 44%. Again, there are pockets of other religious affiliations, but none exhibit the densities of the no religion, Protestant, or Catholic groups.

When considering the distribution of the religion clusters within census sub-divisions, familiar results occur (Table 4-8). Surrey and Abbotsford have the highest proportion of dissemination areas defined as having a Sikh Presence. Richmond, Vancouver, and Burnaby, which had high proportions of “Chinese” ethnic origins and “Chinese” mother tongue, have varying results. Richmond has a high proportion of dissemination areas associated with the Heterogeneous cluster (45.1%) and has a moderately high proportion of dissemination areas associated with the No Religion Presence cluster (30.2%). Vancouver has the highest proportion of dissemination areas associated with the No Religion Presence cluster (47.0%) and it has a moderately high proportion in the Heterogeneous cluster (31.6%). Burnaby has a high proportion of dissemination areas associated with the Heterogeneous cluster (57%) and a moderate proportion with the No Religion cluster (21.1%).
Table 4-8: Percentage of the six religion clusters in 22 British Columbian municipalities

<table>
<thead>
<tr>
<th>Census Subdivision</th>
<th>Heterogeneous with Protestant and No Religion Presence</th>
<th>Catholic Presence</th>
<th>Protestant Presence</th>
<th>Sikh Presence</th>
<th>No Religion Presence</th>
<th>Heterogeneous</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abbotsford</td>
<td>12.8%</td>
<td>0.0%</td>
<td>65.1%</td>
<td>17.4%</td>
<td>0.0%</td>
<td>47.7%</td>
</tr>
<tr>
<td>Burnaby</td>
<td>14.2%</td>
<td>0.9%</td>
<td>3.8%</td>
<td>2.2%</td>
<td>21.1%</td>
<td>57.9%</td>
</tr>
<tr>
<td>Chilliwack</td>
<td>31.4%</td>
<td>0.0%</td>
<td>64.3%</td>
<td>0.0%</td>
<td>1.4%</td>
<td>29.9%</td>
</tr>
<tr>
<td>Coquitlam</td>
<td>19.6%</td>
<td>1.7%</td>
<td>10.6%</td>
<td>0.0%</td>
<td>12.3%</td>
<td>55.9%</td>
</tr>
<tr>
<td>Cranbrook</td>
<td>35.0%</td>
<td>0.0%</td>
<td>25.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>40.0%</td>
</tr>
<tr>
<td>Delta</td>
<td>28.1%</td>
<td>0.0%</td>
<td>28.7%</td>
<td>15.0%</td>
<td>6.0%</td>
<td>22.2%</td>
</tr>
<tr>
<td>Fort St. John</td>
<td>40.0%</td>
<td>0.0%</td>
<td>20.0%</td>
<td>0.0%</td>
<td>13.3%</td>
<td>26.7%</td>
</tr>
<tr>
<td>Kamloops</td>
<td>43.9%</td>
<td>0.0%</td>
<td>17.4%</td>
<td>1.5%</td>
<td>10.6%</td>
<td>26.5%</td>
</tr>
<tr>
<td>Kamloops 1</td>
<td>0.0%</td>
<td>0.0%</td>
<td>20.0%</td>
<td>0.0%</td>
<td>40.0%</td>
<td>40.0%</td>
</tr>
<tr>
<td>Kelowna</td>
<td>30.0%</td>
<td>0.0%</td>
<td>46.3%</td>
<td>0.0%</td>
<td>6.9%</td>
<td>16.9%</td>
</tr>
<tr>
<td>Langley</td>
<td>42.3%</td>
<td>0.0%</td>
<td>37.5%</td>
<td>0.6%</td>
<td>7.1%</td>
<td>12.5%</td>
</tr>
<tr>
<td>Maple Ridge</td>
<td>38.8%</td>
<td>0.0%</td>
<td>23.3%</td>
<td>0.0%</td>
<td>19.4%</td>
<td>18.4%</td>
</tr>
<tr>
<td>Nelson</td>
<td>88.9%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>11.1%</td>
<td>0.0%</td>
</tr>
<tr>
<td>New Westminster</td>
<td>23.9%</td>
<td>0.0%</td>
<td>8.7%</td>
<td>6.5%</td>
<td>8.7%</td>
<td>52.2%</td>
</tr>
<tr>
<td>North Vancouver</td>
<td>34.9%</td>
<td>0.5%</td>
<td>17.7%</td>
<td>0.0%</td>
<td>8.4%</td>
<td>38.6%</td>
</tr>
<tr>
<td>Prince George</td>
<td>30.6%</td>
<td>0.7%</td>
<td>11.2%</td>
<td>0.7%</td>
<td>18.7%</td>
<td>38.1%</td>
</tr>
<tr>
<td>Prince Rupert</td>
<td>25.0%</td>
<td>0.0%</td>
<td>31.3%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>43.8%</td>
</tr>
<tr>
<td>Richmond</td>
<td>19.6%</td>
<td>0.0%</td>
<td>3.4%</td>
<td>1.7%</td>
<td>30.2%</td>
<td>45.1%</td>
</tr>
<tr>
<td>Surrey</td>
<td>16.7%</td>
<td>0.6%</td>
<td>20.1%</td>
<td>28.3%</td>
<td>4.6%</td>
<td>29.8%</td>
</tr>
<tr>
<td>Vancouver</td>
<td>13.3%</td>
<td>1.2%</td>
<td>3.3%</td>
<td>3.7%</td>
<td>47.0%</td>
<td>31.6%</td>
</tr>
<tr>
<td>Victoria</td>
<td>46.0%</td>
<td>0.0%</td>
<td>14.6%</td>
<td>0.0%</td>
<td>28.5%</td>
<td>10.9%</td>
</tr>
<tr>
<td>West Vancouver</td>
<td>35.1%</td>
<td>0.0%</td>
<td>48.1%</td>
<td>0.0%</td>
<td>1.3%</td>
<td>15.6%</td>
</tr>
</tbody>
</table>

Abbotsford and Chilliwack have the highest proportion of Protestant Presence clusters from this list of municipalities, 65.1% and 64.3% respectively. These communities differ in that Abbotsford has a Sikh presence of 17.4% whereas Chilliwack has none. Chilliwack, on the other hand, has 31.4% of the dissemination areas associated with the Heterogeneous with Protestant and No Religion Presence cluster and Abbotsford has only 12.8%. These regional attributes agree with the general profile of the various cities.

**Indicator Validation**

The prevalence of the various clusters within census sub-divisions was used as a first step in assessing the reasonableness of the culture indicators described above. The next step was to check the consistency of the indicators by cross-tabulation. The two-way tables, not presented here:

- Ethnic origins and mother tongue,
- Ethnic origins and religion, and
- Mother tongue and religion.
were constructed and assessed for independence. A Chi-squared test for independence found that the ethnic origins clusters and mother tongue clusters were dependent (p-value <<0.001). Similarly, ethnic origin clusters and religion clusters were found to be dependent (p-value <<0.001) as were the mother tongue clusters and the religion clusters (p-value <<0.001).

These associations are highlighted with their conditional multinomial distributions. Table 4-9 is the distribution of mother tongue clusters given the cluster of ethnic origin. If the ethnic origin was from the Chinese Presence cluster, 92% of these dissemination areas were also members of the mother tongue Chinese Presence cluster. For the East Indian Presence ethnic origin cluster, 89.3% of the dissemination areas were from the Punjabi Presence mother tongue cluster. The Heterogeneous ethnic origins cluster had 77.4% of its dissemination areas associated with the Heterogeneous with Dominant English mother tongue cluster. The British Isles ethnic origin cluster which has 79.2% of its dissemination areas associated with the Dominant English cluster and 19.8% associated with the Strong Dominant English cluster.

<table>
<thead>
<tr>
<th>MT</th>
<th>Origins</th>
<th>Punjabi Presence</th>
<th>Dominant English</th>
<th>Heterogeneous with Dominant English</th>
<th>Chinese Presence</th>
<th>Strong Dominant English</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Heterogeneous</td>
<td>0.9%</td>
<td>0.0%</td>
<td>77.4%</td>
<td>8.7%</td>
<td>13.1%</td>
</tr>
<tr>
<td></td>
<td>Aboriginal Presence</td>
<td>0.0%</td>
<td>80.2%</td>
<td>2.5%</td>
<td>0.0%</td>
<td>17.3%</td>
</tr>
<tr>
<td></td>
<td>Chinese Presence</td>
<td>0.0%</td>
<td>0.0%</td>
<td>8.0%</td>
<td>92.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td></td>
<td>British Isles Presence</td>
<td>0.0%</td>
<td>79.2%</td>
<td>1.1%</td>
<td>0.0%</td>
<td>19.8%</td>
</tr>
<tr>
<td></td>
<td>East Indian Presence</td>
<td>89.3%</td>
<td>0.5%</td>
<td>3.4%</td>
<td>1.5%</td>
<td>5.3%</td>
</tr>
<tr>
<td></td>
<td>Americas and Europe</td>
<td>0.9%</td>
<td>62.0%</td>
<td>2.3%</td>
<td>0.0%</td>
<td>34.8%</td>
</tr>
</tbody>
</table>

The table of religious clusters given the ethnic origins is less well defined than the previous, but still exhibits associations desired of a well defined indicator function (Table 4-10). The cluster for East Indian ethnic origins has 94.2% of its dissemination areas associated with the Sikh Presence cluster. The Chinese Presence ethnic origin cluster has 64.9% of its dissemination areas associated with the No Religion Presence cluster. The Aboriginal Presence cluster has 48.9% of its dissemination areas associated with Catholic Presence clusters and almost 20% associated with Protestant Presence clusters.
Table 4-10: Percentage of clusters for religion groups given the ethnic origin group

<table>
<thead>
<tr>
<th>Religion</th>
<th>Origins</th>
<th>Heterogeneous with Protestant and No Religion Presence</th>
<th>Catholic Presence</th>
<th>Protestant Presence</th>
<th>Sikh Presence</th>
<th>No Religion Presence</th>
<th>Heterogeneous</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Heterogeneous</td>
<td>13.2%</td>
<td>1.7%</td>
<td>3.0%</td>
<td>2.2%</td>
<td>27.7%</td>
<td>52.1%</td>
</tr>
<tr>
<td></td>
<td>Aboriginal Presence</td>
<td>7.6%</td>
<td>48.9%</td>
<td>19.8%</td>
<td>0.0%</td>
<td>15.2%</td>
<td>8.4%</td>
</tr>
<tr>
<td></td>
<td>Chinese Presence</td>
<td>6.6%</td>
<td>0.7%</td>
<td>0.5%</td>
<td>1.2%</td>
<td>64.9%</td>
<td>26.1%</td>
</tr>
<tr>
<td></td>
<td>British Isles Presence</td>
<td>42.9%</td>
<td>0.2%</td>
<td>27.7%</td>
<td>0.1%</td>
<td>14.4%</td>
<td>14.8%</td>
</tr>
<tr>
<td></td>
<td>East Indian Presence</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>94.2%</td>
<td>1.0%</td>
<td>4.9%</td>
</tr>
<tr>
<td></td>
<td>Americas and Europe</td>
<td>34.0%</td>
<td>0.6%</td>
<td>23.0%</td>
<td>2.1%</td>
<td>14.3%</td>
<td>26.0%</td>
</tr>
</tbody>
</table>

When considering the religious cluster given the mother tongue cluster, patterns emerge that emulate those in the previous tables (Table 4-11). 95.8% of the dissemination areas were associated with the Sikh Presence cluster given they were from the Punjabi Presence cluster. 63.6% of the dissemination areas were associated with the No Religion Presence cluster given they were from the Chinese Presence cluster.

Table 4-11: Percentage of clusters for religion groups given the mother tongue group

<table>
<thead>
<tr>
<th>Religion</th>
<th>MT</th>
<th>Heterogeneous with Protestant and No Religion Presence</th>
<th>Catholic Presence</th>
<th>Protestant Presence</th>
<th>Sikh Presence</th>
<th>No Religion Presence</th>
<th>Heterogeneous</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Punjabi Presence</td>
<td>0.5%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>95.8%</td>
<td>1.4%</td>
<td>2.3%</td>
</tr>
<tr>
<td></td>
<td>Dominant English</td>
<td>42.7%</td>
<td>2.7%</td>
<td>27.3%</td>
<td>0.0%</td>
<td>14.7%</td>
<td>12.6%</td>
</tr>
<tr>
<td>Mother Tongue</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Heterogeneous with Dominant English</td>
<td>13.3%</td>
<td>2.0%</td>
<td>4.2%</td>
<td>1.2%</td>
<td>26.6%</td>
<td>52.7%</td>
</tr>
<tr>
<td></td>
<td>Chinese Presence</td>
<td>7.1%</td>
<td>1.1%</td>
<td>0.4%</td>
<td>2.0%</td>
<td>63.6%</td>
<td>25.8%</td>
</tr>
<tr>
<td></td>
<td>Strong English Presence</td>
<td>23.1%</td>
<td>2.1%</td>
<td>19.1%</td>
<td>3.1%</td>
<td>13.7%</td>
<td>38.9%</td>
</tr>
</tbody>
</table>
Discussion

This approach to developing an indicator for culture represents a departure from the current approach employed in Canadian research. By limiting the pre-processing of the data to the conceptual grouping of particular cultural features much of the original data structure is retained. Principal Component Analysis serves to summarise this data in a succinct way for the expressed purpose of grouping the dissemination areas based upon the multivariate structure resultant from PCA. As a result, the classification of dissemination areas is based upon all the available data.

By utilizing the multivariate distribution of the census data, this approach was able to identify cultural groups that are known to exist in British Columbia, such as the Indo-Canadian and Chinese cultural groups. As well, this approach was able to determine clusters that reflected the known geographic distributions of these cultural groups. Through the discovery of these groups, information was generated that can be mined for further insights and research.

The labelling of the groups with a meaningful title is another departure from the current methodology. Currently, the label precedes the formation of the groups. This suggests that the concept is imposed upon the groups a priori the construction. This new approach resists any labelling until the end of the process, thus labels are assigned based upon the composition of the groups. This distinction in when the label is assigned reflects more the investigator’s understanding of culture. The current methodology suggests that culture fits into a predetermined structure or identity whereas this new approach suggests that culture is defined by a complex web of relationships. Grouping dissemination areas based upon these webs of interconnectivity, a shift from modern to post-modern understandings of culture can emerge.

The Department of Canadian Heritage states that “Multiculturalism ensures that all citizens can keep their identities, can take pride in their ancestry and have a sense of belonging” 36. Furthermore, the Department of Canadian Heritage states that people are “free to choose for themselves, without penalty, whether they want to identify with their specific group or not” 36. This cultural pluralism suggests that the development of culture indicators must be premised upon the underlying multivariate distributions of the desired cultural constructs. In doing so, Canadian research will be poised to capture the complex interplay of cultures that constitutes the mosaic of Canadian society.

This new tool will be useful in understanding various regional structures. This exploration suggests that certain regions should not be considered homogeneous with respect to their cultural constructs. It was shown that there exist distinct differences within the Upper Fraser Valley with Abbotsford exhibiting different cultural characteristics than Chilliwack or Langley. These understandings can be used to specify better the services that may be needed in a region.
Limitations

Although this indicator does provide a weight of evidence for the validity of the constructs being identified which is essential for evidentiary support for the development of policy, it does suffer from several limitations. As the indicators are developed from the underlying multivariate distributions of a census-based cultural construct, they are limited to the measures determined by Statistics Canada. The interpretation of these indicators should be done in dialogue with sociologists and ethno-geographers in order to prevent disparaging ethnic profiling within the research community, thus interdisciplinary collaborative teams would be beneficial for the continued development of multivariate culture indicators. As well, the census cultural constructs should be combined in evocative ways to group dissemination areas upon more complex interplays of culture. This would entail the combining of several cultural features, such as ethnic origins, mother tongue and religion, into one data set and then group the dissemination areas based upon a more complete and complex representation of culture. Finally, a comprehensive assessment would need to be developed.

Conclusions

In order to identify Canadian populations that are systematically marginalized from health services due to non-economic sociodemographic constructs, further work in developing multivariate-based culture indicators must be done. This novel work represents a first step towards a new understanding of how to quantify culture in population-based research which uses census-based measures to describe culture constructs.
### Table 4-12: Statistical descriptions of the population proportions for the five mother tongue groups expressed by only the minimum value, the median and the maximum value

<table>
<thead>
<tr>
<th>Conceptual Group</th>
<th>Brief Statistical Description</th>
<th>Punjabi Presence (1)</th>
<th>Dominant English (2)</th>
<th>Heterogeneous with Dominant English (3)</th>
<th>Chinese Presence (4)</th>
<th>Strong English Presence (5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chinese</td>
<td>Min</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.262</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Median</td>
<td>0.000</td>
<td>0.000</td>
<td>0.241</td>
<td>0.500</td>
<td>0.0321</td>
</tr>
<tr>
<td></td>
<td>Max</td>
<td>0.343</td>
<td>0.147</td>
<td>0.443</td>
<td>0.8692</td>
<td>0.282</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>0.036</td>
<td>0.007</td>
<td>0.233</td>
<td>0.513</td>
<td>0.051</td>
</tr>
<tr>
<td>Aboriginal</td>
<td>Min</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Median</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Max</td>
<td>0.017</td>
<td>0.200</td>
<td>0.071</td>
<td>0.017</td>
<td>0.500</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>0.000</td>
<td>0.001</td>
<td>0.001</td>
<td>0.000</td>
<td>0.001</td>
</tr>
<tr>
<td>Southern European</td>
<td>Min</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Median</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Max</td>
<td>0.054</td>
<td>0.159</td>
<td>0.222</td>
<td>0.238</td>
<td>0.194</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>0.004</td>
<td>0.005</td>
<td>0.018</td>
<td>0.020</td>
<td>0.011</td>
</tr>
<tr>
<td>Eastern European</td>
<td>Min</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Median</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Max</td>
<td>0.150</td>
<td>0.171</td>
<td>0.355</td>
<td>0.184</td>
<td>0.397</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>0.013</td>
<td>0.012</td>
<td>0.420</td>
<td>0.017</td>
<td>0.035</td>
</tr>
<tr>
<td>North-West European</td>
<td>Min</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Median</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Max</td>
<td>0.103</td>
<td>0.200</td>
<td>0.319</td>
<td>0.056</td>
<td>0.261</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>0.005</td>
<td>0.013</td>
<td>0.008</td>
<td>0.003</td>
<td>0.015</td>
</tr>
<tr>
<td>Baltic</td>
<td>Min</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Median</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Max</td>
<td>0.035</td>
<td>0.061</td>
<td>0.089</td>
<td>0.080</td>
<td>0.073</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>0.000</td>
<td>0.000</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td>Middle East</td>
<td>Min</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Median</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Max</td>
<td>0.101</td>
<td>0.089</td>
<td>0.454</td>
<td>0.157</td>
<td>0.301</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>0.006</td>
<td>0.002</td>
<td>0.022</td>
<td>0.008</td>
<td>0.015</td>
</tr>
<tr>
<td>Indian Sub-Continent</td>
<td>Min</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Median</td>
<td>0.046</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Max</td>
<td>0.328</td>
<td>0.188</td>
<td>0.257</td>
<td>0.257</td>
<td>0.254</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>0.063</td>
<td>0.001</td>
<td>0.019</td>
<td>0.018</td>
<td>0.011</td>
</tr>
<tr>
<td>East Asia</td>
<td>Min</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Median</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Max</td>
<td>0.311</td>
<td>0.182</td>
<td>0.500</td>
<td>0.379</td>
<td>0.310</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>0.045</td>
<td>0.007</td>
<td>0.079</td>
<td>0.072</td>
<td>0.038</td>
</tr>
<tr>
<td>Spanish-Portuguese</td>
<td>Min</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Median</td>
<td>0.000</td>
<td>0.000</td>
<td>0.014</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Max</td>
<td>0.173</td>
<td>0.182</td>
<td>0.270</td>
<td>0.172</td>
<td>0.283</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>0.016</td>
<td>0.004</td>
<td>0.023</td>
<td>0.018</td>
<td>0.017</td>
</tr>
<tr>
<td>English</td>
<td>Min</td>
<td>0.000</td>
<td>0.784</td>
<td>0.128</td>
<td>0.040</td>
<td>0.478</td>
</tr>
<tr>
<td></td>
<td>Median</td>
<td>0.366</td>
<td>0.896</td>
<td>0.491</td>
<td>0.284</td>
<td>0.712</td>
</tr>
<tr>
<td></td>
<td>Max</td>
<td>0.615</td>
<td>1.000</td>
<td>0.687</td>
<td>0.494</td>
<td>0.811</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>0.356</td>
<td>0.896</td>
<td>0.484</td>
<td>0.280</td>
<td>0.703</td>
</tr>
<tr>
<td>French</td>
<td>Min</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Language</td>
<td>Median</td>
<td>0.000</td>
<td>0.009</td>
<td>0.000</td>
<td>0.000</td>
<td>0.014</td>
</tr>
<tr>
<td>----------</td>
<td>--------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
</tr>
<tr>
<td></td>
<td>Max</td>
<td>0.177</td>
<td>0.212</td>
<td>0.154</td>
<td>0.109</td>
<td>0.214</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>0.008</td>
<td>0.014</td>
<td>0.012</td>
<td>0.005</td>
<td>0.019</td>
</tr>
<tr>
<td>German</td>
<td>Min</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Median</td>
<td>0.000</td>
<td>0.018</td>
<td>0.000</td>
<td>0.000</td>
<td>0.020</td>
</tr>
<tr>
<td></td>
<td>Max</td>
<td>0.200</td>
<td>0.200</td>
<td>0.826</td>
<td>0.067</td>
<td>0.452</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>0.013</td>
<td>0.023</td>
<td>0.018</td>
<td>0.007</td>
<td>0.029</td>
</tr>
<tr>
<td>Punjabi</td>
<td>Min</td>
<td>0.161</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Median</td>
<td>0.3836</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Max</td>
<td>0.9097</td>
<td>0.174</td>
<td>0.235</td>
<td>0.316</td>
<td>0.297</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>0.423</td>
<td>0.004</td>
<td>0.022</td>
<td>0.026</td>
<td>0.034</td>
</tr>
<tr>
<td>Other</td>
<td>Min</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Median</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Max</td>
<td>0.159</td>
<td>0.200</td>
<td>0.778</td>
<td>0.104</td>
<td>0.500</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>0.011</td>
<td>0.007</td>
<td>0.018</td>
<td>0.010</td>
<td>0.017</td>
</tr>
<tr>
<td>Cluster Size</td>
<td>214</td>
<td>3660</td>
<td>753</td>
<td>450</td>
<td>1495</td>
<td></td>
</tr>
</tbody>
</table>
Table 4-13: Statistical descriptions of the population proportions for the six five mother tongue groups expressed by only the minimum value, the median and the maximum value

<table>
<thead>
<tr>
<th>Conceptual Group</th>
<th>Brief Statistical Description</th>
<th>Heterogeneous with Protestant and No Religion Presence (1)</th>
<th>Catholic Presence (2)</th>
<th>Protestant Presence (3)</th>
<th>Sikh Presence (4)</th>
<th>No Religion Presence(5)</th>
<th>Heterogeneous (6)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Catholic</td>
<td>Min</td>
<td>0.000</td>
<td>0.4167</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Median</td>
<td>0.139</td>
<td>0.6340</td>
<td>0.146</td>
<td>0.108</td>
<td>0.158</td>
<td>0.259</td>
</tr>
<tr>
<td></td>
<td>Max</td>
<td>0.258</td>
<td>1.000</td>
<td>0.441</td>
<td>0.381</td>
<td>0.433</td>
<td>0.4868</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>0.138</td>
<td>0.657</td>
<td>0.146</td>
<td>0.119</td>
<td>0.159</td>
<td>0.263</td>
</tr>
<tr>
<td>Orthodox</td>
<td>Min</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Median</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Max</td>
<td>0.179</td>
<td>0.118</td>
<td>0.190</td>
<td>0.125</td>
<td>0.244</td>
<td>0.235</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>0.006</td>
<td>0.004</td>
<td>0.006</td>
<td>0.006</td>
<td>0.009</td>
<td>0.014</td>
</tr>
<tr>
<td>Protestant</td>
<td>Min</td>
<td>0.253</td>
<td>0.000</td>
<td>0.367</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Median</td>
<td>0.396</td>
<td>0.076</td>
<td>0.533</td>
<td>0.167</td>
<td>0.207</td>
<td>0.287</td>
</tr>
<tr>
<td></td>
<td>Max</td>
<td>0.575</td>
<td>0.429</td>
<td>1.000</td>
<td>0.438</td>
<td>0.366</td>
<td>0.450</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>0.396</td>
<td>0.088</td>
<td>0.552</td>
<td>0.176</td>
<td>0.196</td>
<td>0.280</td>
</tr>
<tr>
<td>None</td>
<td>Min</td>
<td>0.252</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.337</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Median</td>
<td>0.406</td>
<td>0.197</td>
<td>0.247</td>
<td>0.1706</td>
<td>0.5234</td>
<td>0.321</td>
</tr>
<tr>
<td></td>
<td>Max</td>
<td>0.615</td>
<td>0.500</td>
<td>0.421</td>
<td>0.387</td>
<td>1.000</td>
<td>0.471</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>0.408</td>
<td>0.200</td>
<td>0.247</td>
<td>0.168</td>
<td>0.537</td>
<td>0.314</td>
</tr>
<tr>
<td>Muslim</td>
<td>Min</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Median</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.018</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Max</td>
<td>0.264</td>
<td>0.209</td>
<td>0.308</td>
<td>0.409</td>
<td>0.217</td>
<td>0.600</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>0.008</td>
<td>0.004</td>
<td>0.007</td>
<td>0.340</td>
<td>0.011</td>
<td>0.026</td>
</tr>
<tr>
<td>Sikh</td>
<td>Min</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.179</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Median</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.365</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Max</td>
<td>0.253</td>
<td>0.175</td>
<td>0.301</td>
<td>0.929</td>
<td>0.342</td>
<td>0.254</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>0.009</td>
<td>0.004</td>
<td>0.012</td>
<td>0.405</td>
<td>0.012</td>
<td>0.029</td>
</tr>
<tr>
<td>Jewish</td>
<td>Min</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Median</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Max</td>
<td>0.185</td>
<td>0.057</td>
<td>0.170</td>
<td>0.136</td>
<td>0.200</td>
<td>0.321</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>0.005</td>
<td>0.001</td>
<td>0.004</td>
<td>0.002</td>
<td>0.008</td>
<td>0.007</td>
</tr>
<tr>
<td>Other</td>
<td>Min</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Median</td>
<td>0.021</td>
<td>0.000</td>
<td>0.015</td>
<td>0.074</td>
<td>0.0482</td>
<td>0.047</td>
</tr>
<tr>
<td></td>
<td>Max</td>
<td>0.303</td>
<td>0.375</td>
<td>0.320</td>
<td>0.398</td>
<td>0.435</td>
<td>0.667</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>0.030</td>
<td>0.042</td>
<td>0.025</td>
<td>0.085</td>
<td>0.068</td>
<td>0.066</td>
</tr>
<tr>
<td>Cluster Size</td>
<td></td>
<td>2040</td>
<td>152</td>
<td>1317</td>
<td>270</td>
<td>1233</td>
<td>1560</td>
</tr>
</tbody>
</table>
References


88


Introduction

Current research into the location of death of cancer patients in Canada suggests that demographic and clinical attributes may be useful in predicting where a cancer patient will die. This notion of the utility of socioeconomic, demographic and clinical assessment for predicting place of death for cancer patients is supported by non-Canadian research. In order to come to this conclusion, univariate, multivariate, non-parametric, and regression techniques have been employed.

In determining the place of death of cancer patients it is desirable to obtain a set of socioeconomic, demographic and clinical predictors that are useful in determining the location of end of life care for cancer patients. A corollary to this is to determine which modelling technique allows for the best predictive accuracy. Interestingly, of all the classification techniques, the research reveals that logistic regression is the method of choice for this area of investigation. As well, there is a noticeable absence of location-of-death studies which compare the utility of different classifiers. Unfortunately preference does not determine the usefulness and appropriateness of a technique.

As logistic regression is a familiar modelling technique and has been implemented in a variety of statistical packages, it is not surprising that this classifier is frequently chosen. Even though it a popular technique, it is one of several technique that can be used for classification. Classifiers are often encountered within the field of Data Mining.

The field of data mining can be broken into two sections: unsupervised and supervised learning. If there is a known outcome, such as the location of death, then supervised learning techniques (classifiers) are used. If there is no expressed outcome in the data set, then unsupervised learning techniques are used. Logistic regression is one type of supervised learning technique along with linear discriminant analysis, neural networks, recursive partitioning, and nearest neighbours among others. Recognizing that logistic regression is one possible way to determine the predictors for the location of death, the exploration of data mining techniques is an appropriate endeavour for this area of investigation.

Cancer research is rapidly exploring the usefulness of these models. With applications in radiology, survival, diagnosis and prediction, these models are being used in two general ways. A portion of the research simply uses the classifiers as a statistical tool while the other portion is concerned about the usefulness and appropriateness of these classifiers within cancer research. Within this activity of application and assessment, place of death research is silent.

In July 2004, the Canadian Institutes of Health Research (CIHR) approved the grant application entitled “Palliative Care in cross-cultural context: A NET for equitable and quality cancer care for ethnically diverse populations”. This five year grant allows researchers in British Columbia, Nova Scotia, and Saskatchewan to focus upon three themes: Access to health services, caregivers and complementary and alternative therapies. The Access theme seeks to understand the cultural constructs that influence
the use of health services by cancer patients at the end-of-life. Not only does this involve
the construction and interpretation of predictive models, but it is also concerned with the
construction of culture indicators to be used in the models and in the quality of the data
used for the modelling process.

In order to gain a voice within the emerging research, two approaches can be taken. The
first is to merely infer the utility of data mining based upon the current body of research.
The second is to apply the insights gained from the current body of research and engage
in area specific research comparing the usefulness of the techniques. Spurious inference
of the usefulness of data mining for predicting the location of death for cancer patients is
thus avoided. Following the second approach, the objectives of this research are to
compare the performance of five classifiers, comment upon the clinical utility of the
classifiers and to understand which factors were predictive of dying out of hospital.

Methodology

Administrative Databases
For this retrospective population-based study, a secondary data set was created from the
linkage of two administrative databases: The British Columbia Vital Statistics Agency
(VS) and the British Columbia Cancer Registry (BCCR). The Statistics Canada Postal
Code Conversion File program version 4D (PCCF+4D) was used to associate subjects, at
the dissemination area level, with economic and demographic information derived from
the 2001 census. Linkage was performed by the VS for the BCCR and VS databases.
The census information was merged to the VS and BCCR database using SAS. The
merging of the census information and the study database was based upon the
dissemination area allocated to each person by the PCCF program.

The BC Cancer Registry, having been in existence from 1969, and maintained by the BC
Cancer Agency (BCCA) since 1980, collects and generates cancer statistics on the BC
population. This database is used for cancer surveillance and research which provides
information on the magnitude of the cancer problem within BC, assists in the
development and monitoring of programs for reduced mortality and morbidity, and
provides information for future planning. By providing a registry, the BCCA has
access to a source of information which avoids bias due to non-representative
participation. This translates into better quality data than that used in non-population-
based sources. The BC Cancer Registry has attained an overall silver certification
status from the North American Association of Central Cancer Registries (NAACCR).
The registry has reached gold standard achievement for specific items such as timeliness
and death certificate only cases.

Although there is much information contained in these data sources, the data has been
collected for purposes other than the understanding of the location of death for cancer
patients, thus making this investigation an observational study. From a classical
statistical perspective, observational studies are the most problematic type of study as
there is, in general, no reference to any experimental design. As a result, is has been
suggested that observational, non-randomized studies carry lower evidential weight than
randomized ones. It is because of this that a well-planned and credible method for analyzing data sets of this nature is needed.

Data mining has emerged in response to the desire to make sense of the vast amounts of data housed in administrative and corporate databases. These databases represent an efficient means, in terms of cost and time, to sketch a rough picture for population-based studies. These sketches are useful for hypothesis generation.

Often, the information desired from the databases does not naturally extend from the underlying motivation for collecting the information. For the current research, the BCCR database was not developed for research into the location of death. It is large in terms of the number of patients represented and in the number of potential variables. The use of this database and the VS database allows for the unique opportunity to use data mining techniques to generate understanding of population-based end-of-life health outcomes.

These attributes of the BCCR database indicate that the current research is an observational one and would be subject to the aforementioned critiques. Fortunately, data mining addresses such situations and allows researchers credibly to investigate such databases.

**Subjects**
The subjects were all adults, aged 20 years and older, who died of cancer, in British Columbia, as identified from the death certificates, from 1999 to 2003. Death Certificate Only (DCO) patients were excluded from the formation of the data set. These patients were those who had a cancer diagnosis at the time of death, although no prior information relating to a diagnosis of cancer was found after following up the diagnosis at the time of death (Appendix A). The postal code for the usual place of residence was obtained from British Columbia Vital Statistics. If this information was unavailable, then the postal code for the place of residence was obtained from the BCCR. Vital Statistics provided the location of death.

**Variables**

**Data Sources**
The response, location of death, was obtained from the VS database. The predictors were obtained from the BCCA database. The cause of death was taken from the BCCA. It should be noted that the BCCA database obtained this information from BC Vital Statistics.

**Response**
The response variable for the place of death was obtained from VS. The location of death was derived from a field which indicated hospital death. If a person died in hospital, this was identified with the hospital code. If a person died out of the hospital, there was no code. Place of death was dichotomized as in hospital or out of hospital.
Predictors

As this research is situated within a collaborative investigation between Nova Scotia (NS) and British Columbia (BC), the predictors chosen for this stage represented a modest expansion upon the published set of predictors\(^1\). To the set of predictors,

- Year of Death,
- Sex,
- Age (years),
- Income,
- Aggregated Tumour Group,
- Region, and
- Survival Time (days)

the following were added:

- City Size,
- Ethnic Origin Context,
- Mother Tongue Context, and
- Religious Context.

The year of death and sex were given by the BCCR. The age was derived from the BCCR database as the difference between the date of death and the date of birth and was not categorized as in the NS research\(^1\). The income per person equivalent was obtained from the Postal Code Conversion File (PCCF) program output and expressed as quintiles\(^5^1\).

The aggregate tumour group was derived from the ICD-O. Cancer causes of death were initially grouped according to the Canadian Cancer Society groupings (Appendix B)\(^5^6\). These groups were further aggregated as many of the groups had small numbers of patients. Canadian Cancer Society groups which had fewer than 1500 patients from the study group were grouped into the “Other” category. The final aggregate tumour groups were:

- breast,
- colorectal,
- haematological,
- lung,
- pancreas,
- prostate,
- unknown primary, and
- other.

The region was the BC Health Authority of usual residence as derived from the VS report. The survival time was derived from the difference between the date of death and the diagnosis date and was not categorized as in the NS research\(^1\). The city size was obtained from the PCCF program.
The cultural indicators were derived using unsupervised learning techniques. The Ethnic Origin context has the categories:
- Heterogeneous,
- Aboriginal Presence,
- Chinese Presence,
- British Isles Presence,
- East Indian Presence, and
- Americas and Europe.

The Mother Tongue context indicator is broken down into:
- Punjabi Presence,
- Dominant English,
- Heterogeneous with Dominant English,
- Chinese Presence, and
- Strong English Presence.

The Religious context indicator is broken into:
- Heterogeneous with Protestant and No Religion Presence
- Catholic Presence,
- Protestant Presence,
- Sikh Presence,
- No Religion Presence, and
- Heterogeneous.

The postal code definition of rural/urban as defined by the Nova Scotia research was not used. The zero in the second position of the postal code indicates delivery from a rural post office. This indicates the point of service, but not the rural or urban feature of the area. Some rural areas are urbanised and suburbanised, yet they have a zero in the second position, thus it may not be the best way to define the indicator. As well, rural route services and sub-urban route services are provided from urban post offices. For example, the entire province of New Brunswick is considered urban as no postal code has a zero in the second position. A measure based upon population size or on population density and distance to regional center may serve as starting points for such a definition.

The re-definition of rural and urban needs further work. The process of redefining rural and urban for this research was not in the scope of this investigation. Instead the city size was used. As the PCCF program uses the postal codes of usual residence, a city size could be associated with each patient. The five groups were:
- greater than 1,250,000 people,
- 500,000 - 1,249,999 people,
- 100,000 - 499,999 people,
- 10,000 - 99,999 people, and
- less than 10,000 people.
The PCCF support material suggested a definition of rural as those locations where the population was less than 10,000 people. The city size was chosen over the coarser population definition of rural/urban.

As this research is in collaboration with a research team in Nova Scotia the categories used for age and survival were based upon work previously done by the Nova Scotia members of research team.

**Data Quality**

Assessing the quality of the data is an important step when using administrative databases. The data quality assessment has been limited to variable completeness and the concordance of the diagnosis and the cancer cause of death due to database constraints.

Under the above definition, the study data set has a high degree of variable completeness. Only three variables are not complete and each of them is well over 99% complete (Figure 5-1). The agreement between cancer diagnosis and cancer cause of death suggests a high degree of data quality (Appendix D).

![Variable Completeness For Data Mining Data Set](image)

**Figure 5-1:** Graph for study set variable completeness.
Data Mining Methodologies

By investigating the health services that are used at the end of a patient’s life, we are using information from a variety of predictor variables to allocate patients into one health service category. For this research, the location of death is being used as the health outcome. A patient either used the hospital services at death or did not.

The outcome categories have been determined by the research question and we are able to determine them in the data set, thus it is natural to use supervised learning techniques to investigate the relationship, by building models, between the explanatory variables and the binary outcome for the location of death.

As previous investigations into the location of death in Canada have used logistic regression researchers have already used a supervised learning, data mining technique. We will introduce linear discriminant analysis (LDA), neural networks (NN), classification trees (CT), and K-nearest neighbours (KNN) to end-of-life research and compare the performance of these classifiers to that of logistic regression.

As the outcome variable is binary (death in a hospital versus death out of a hospital), the presentation of the classification methodology will be within the two class framework. For purposes of notation, “H” will be used to designate deaths in a hospital and “O” will be used for deaths out of a hospital. The outcome to be modelled will be out of hospital deaths unless otherwise expressed.

Linear Discriminant Analysis

Linear discriminant analysis is one of the oldest classification techniques. Introduced by Fisher in 1936 \(^1\), it is a classification technique that is used to distinguish between two or more groups based upon a set of characteristics (predictor variables).

With two classes, linear discriminant analysis (LDA) seeks to find a linear boundary between the two groups. The linear boundary is found by obtaining the function which best separates the two groups.

It does this by using a discriminant function,

\[
DF = \beta_1 X_1 + \cdots + \beta_p X_p = \beta^T X \tag{0.3}
\]

where there are \(p\) predictor variables, \(\beta_i\) is the \(i^{th}\) coefficient, and \(X_i\) is the \(i^{th}\) predictor variable. The coefficients or weights are estimated to give the best separation of the two groups.

In order to determine the coefficients that yield the best separation, two assumptions will be made. The first is that each group has a different mean. This suggests that the means of \(X_1, \ldots, X_p\) is different for patients who died in a hospital and for those who died out of a hospital. The second assumption is that the variance-covariance structure is the same for both groups \(^41,71\).

With this setup, the assumptions, conditioned upon each class, about the data are:
$X_1, \ldots, X_d \mid \text{class } H \sim (\mu^{(H)}, \Sigma)$, and
$X_1, \ldots, X_d \mid \text{class } O \sim (\mu^{(O)}, \Sigma)$,
where $\mu^{(H)}$ and $\mu^{(O)}$ are the $p$-dimensional mean vectors for deaths out of a hospital and hospital deaths respectively. $\Sigma$ is the $p \times p$ common covariance matrix. Interestingly, no distributional assumption has been made. LDA can be developed with a multivariate normal assumption, but this approach will not be presented in this paper. Hastie, et al. do discuss the distributional approach.

The statistical properties of the DF distributions, conditioned upon class $k$, where $k \in \{H,O\}$, can be determined. The expectation is,

$$E(DF \mid \text{Class } k) = E(\beta^T X \mid \text{Class } k) = \beta^T E(X \mid \text{Class } k) = \beta^T \mu^{(k)}.$$

The variance of the discrimination function is,

$$\text{Var}(DF) = \text{Var}(\beta^T X) = \beta^T \text{Var}(X) \beta = \beta^T \Sigma \beta.$$

As the goal is to find the best separation of the two groups, we want a DF that makes a good distinction between the two groups. This difference can be measured by the separation between the two means, $\beta^T \mu^{(H)} - \beta^T \mu^{(O)}$. An estimate of this distance is,

$$\beta^T \bar{X}^{(H)} - \beta^T \bar{X}^{(O)} = \beta^T \left( \bar{X}^{(H)} - \bar{X}^{(O)} \right)$$

where $\bar{X}^{(k)}$ is a random vector for $\bar{X}^{(k)}$, the realisation. If the means of the two groups are far apart, then the classification technique can adequately distinguish the two groups. This estimator has a variance of,

$$\text{Var}(\beta^T \left( \bar{X}^{(H)} - \bar{X}^{(O)} \right)) = \beta^T \text{Var}(\bar{X}^{(H)} - \bar{X}^{(O)}) \beta = \beta^T \left( \frac{\Sigma}{n_H} + \frac{\Sigma}{n_O} \right) \beta = \left( \frac{1}{n_H} + \frac{1}{n_O} \right) \beta^T \Sigma \beta.$$
As $\Sigma$ is the common variance-covariance between the two classes, this can be estimated by the pooling the sample variance-covariance matrices $S^H$, $S^O$:

$$S = \frac{(n_H - 1)S^{(H)} + (n_O - 1)S^{(O)}}{n_H + n_O - 2}.$$ 

With an estimate of the difference between the means of the two groups and an estimate of the variation in the estimate of the difference between the two means, it is possible to construct a statistic that resembles the two-sample t test statistic with a pooled sample variance:

$$t' = \frac{\beta^T \left( \bar{x}^{(H)} - \bar{x}^{(O)} \right)}{\sqrt{\left( \frac{1}{n_H} + \frac{1}{n_O} \right) \beta^T S \beta}}$$

the notation $t'$ has been used to denote that this is not a t statistic. No distributional assumption has been made, so it is only similar to the t statistic. The utility of $t'$ is that a very large value indicates good separation. This is similar to having a large t statistic for the two-sample t test in the context of hypothesis testing.

The coefficients of DF are chosen so that the absolute value of $t'$ is maximized. The square of $t'$, $t'^2$ is used as it is easier. The maximum of $t'^2$ is,

$$\beta \propto S^{-1} \left( \bar{x}^{(H)} - \bar{x}^{(O)} \right).$$

Fisher's Discriminant Function is, then,

$$\beta^T x = \left( x^{(H)} - x^{(O)} \right)^T S^{-1} x^{(H)}.$$  

From this, a new observation $x$ is classified based upon its proximity to the class means. the new patient is classified as dying out of a hospital, O, if,

$$|w^T x - w^T x^{(O)}| < |w^T x - w^T x^{(H)}|,$$

otherwise the patient is classified as dying in a hospital.

**Logistic Regression**

As logistic regression is a familiar epidemiological modelling technique, only a brief overview will be given. Logistic regression is a special case of a generalized linear model. The initial formulation of generalized linear models was by Nelder and Wedderburn in 1972.
In logistic regression, we model the probability of belonging to a certain class. As we are considering only the two class situation, the formulation of logistic regression is greatly simplified. We need only to model the probability of belonging to one of the classes. As deaths out of a hospital are of interest, we will model the conditional probability of dying out of a hospital given the explanatory variables \( x \), \( \Pr(\text{Class}=0 \mid x) \). This probability has immediate appeal as it is easy to make a decision as to how to classify the new observation.

If we wanted to know the odds of dying out of hospital when compared to dying in hospital, given the explanatory variables, then the odds in favour of dying out of a hospital would be,

\[
\text{odds} = \frac{\Pr(\text{Class}=0 \mid x)}{\Pr(\text{Class}=1 \mid x)} = \frac{\Pr(\text{Class}=0 \mid x)}{1 - \Pr(\text{Class}=0 \mid x)}
\]

where \( \text{odds} \in [0, \infty) \). Since the \( \text{odds} \) takes any value between zero and infinity, the logarithm of the odds will take any value in the real numbers, \( \log(\text{odds}) \in (-\infty, \infty) \). As the \( \log(\text{odds}) \) takes any real value, it is reasonable to model the \( \log(\text{odds}) \) with a linear function of the predictor variables \(^{43}\). For more information about the technical reasons for using the \( \log(\text{odds}) \) with the linear model see Dobson \(^{43}\) or Venables et al. \(^{42}\).

The \( \log(\text{odds}) \) is commonly referred to as the logit function. If \( \pi_0 = \Pr(\text{Class}=0 \mid x) \) and \( \beta^T \) is the row vector of coefficients \(^{43}\), then a linear model for the logit function is

\[
\log\left(\frac{\pi_0}{1 - \pi_0}\right) = \beta^T x. \tag{1.1}
\]

The model is linear in terms of the unknown coefficients. Here we have a linear model for the logarithm of the odds of dying out of a hospital when compared to dying in a hospital. What is wanted is \( \pi_0 = \Pr(\text{Class}=0 \mid x) \). This can be rewritten as,

\[
\pi_0 = \frac{e^{\beta^T x}}{1 + e^{\beta^T x}}. \tag{1.2}
\]

In this form, \( \pi_0 \) will take on values between zero and one, \( \pi_0 \in [0,1] \). If the probability of dying in a hospital is desired, the complement is taken.

**Neural Networks**

The human brain and neural networks are conceptual analogs. Information goes into a neuron. Information leaves the neuron and subsequently enters another neuron. Each neuron has inputs from other neurons and outputs to other neurons \(^{71}\). The human brain is extremely complicated, so neural networks attempts to emulate this process through a
very simple structure. This idea, having been around since the 1940s, was popularized when Rosenblatt introduced his idea of a perceptron, the mathematical version of the neuron. Widespread interest in neural networks was gained in the 1980s with improvements in the training algorithm.

For neural networks (NN), the fundamental idea is to create linear combinations of the explanatory variables as inputs to each of the nodes. The outputs of these nodes are, in turn, inputs to other nodes. This continues until the final node(s) are reached. At the final node(s), the model is a complex non-linear function of the original explanatory variables. As with logistic regression, we will model the probability of dying out of a hospital given the explanatory variables \( x \), \( \Pr(\text{Class}=0 \mid x) \).

To use neural networks, two particulars must be addressed. The first is the relationship between the nodes (perceptrons). By establishing this relationship, the architecture for the network is determined. The most frequently used model is a single-layer neural network which has only one hidden layer of nodes between the inputs and the output node (Figure 5-2). With the binary response being used in this research only a single output node is needed (Figure 5-2). The input variables are inputs to the hidden layer nodes. The hidden layer nodes output a linear combination of the input variables. The outputs from the hidden layer become inputs for the output node. If a skip is allowed, the constant input term is allowed to be an input to the output node.

![Figure 5-2: Example of a one layer architecture for a neural network with three input variables. The input variables are \( x_1, x_2, \) and \( x_3 \) with a constant input \( x_0 \). The transformation functions (perceptrons) for the hidden layer are \( h_1 \) and \( h_2 \). The output layer has a transformation function \( f \). A skip allows \( x_0 \) to skip over the hidden layer and be one of the weighted inputs to the output layer.](image)
The second item to determine is the function that is used in the nodes. For classification, the softmax function,

\[ h(z) = \frac{e^z}{1 + e^z}, \tag{2.1} \]

is used \(^{40,42}\). This function is similar to (1.2), where \( z = \beta^T x \). Here \( z \) is a linear combination of the inputs. Recall that the inputs can be either the original explanatory variables or the outputs from another node.

For Figure 5-2, the hidden node, \( h_1 \), would be

\[ h_1(\beta^T x) = \frac{e^{\beta^T x}}{1 + e^{\beta^T x}}, \]

where \( \beta_1 \) is the vector of coefficients for hidden node \( h_1 \). The hidden node, \( h_2 \), would have an analogous formulation. For the output node, with no skip, it would be,

\[ g_0(y^T h) = \pi_0 = \frac{e^{y^T h}}{1 + e^{y^T h}}, \]

where \( h \) is a column vector of the outputs from the hidden nodes and \( y^T \) is the vector of coefficients for the output node. This process is similar to logistic regression at each of the hidden layer nodes and then taking the outputs from these nodes and using them as inputs for a logistic regression at the output node. For reasons analogous to those expressed with logistic regression, the probability of dying out of a hospital given the explanatory variables \( x \), \( \Pr(\text{Class}=0 \mid x) \), is modelled. It is interesting to observe that if there is no hidden layer and the softmax function is used, then neural networks reduces to logistic regression. Logistic regression is a special case of neural networks.

For completeness, several technical issues will be briefly presented. Further information can be found in Hastie et al.\(^{40}\) and Venables et al.\(^{42}\). The single-layer architecture in Figure 5-2 with the softmax function represents a feed-forward neural network as the inputs feed forward through the network \(^{42,71}\). The values for the weights are those which minimize a measure of error, such as least squares or deviance \(^{40,42,71}\). In order to find these values, back propagation is used which is a generic approach to minimizing the chosen measure of error by gradient descent \(^{40,71}\). When the model has been fit, then classification occurs according to some predetermined cut-off value. If \( \pi_0 = \Pr(\text{Class}=0 \mid x) > c \), for \( c \in [0,1] \), then the neural network classifies the new patient as dying out of a hospital. Otherwise the patient is classified as dying in a hospital. By adding the product of a weight-decay parameter, \( \lambda \), and the squared sum of weights, \( w \), to the error function, \( E \), smoothness can be ensured:

\[ \text{Fitting criterion} = E + \lambda \sum w. \tag{40} \]
Classification Trees

The idea of tree-based methods can be traced back to at least Morgan and Sonquist in 1963 [7]. Breiman et al. popularized tree-based methods in the mid-1980s [7]. Classification trees divide the explanatory variable space into a set of disjoint hyper-rectangles. The hyper-rectangles are created so that they are as homogeneous as possible with respect to the outcome variable. When considering the location of death, we desire hyper-rectangles that are homogeneous with respect to deaths out of a hospital or deaths in a hospital.

The most common implementation is to make the hyper-rectangles as homogeneous as possible in the response variable through binary splits of the explanatory variables. For simplicity, consider the case where there are only two explanatory variables, \( x_1 \) and \( x_2 \). Figure 5-3 represents the explanatory variable space and each of the patients is identified by letter in that space. “H” refers to those patients that died in a hospital and “O” are those that died out of hospital.

![Figure 5-3: Example of how the classification tree algorithm partitions the explanatory variable space. There are two explanatory variables, \( x_1 \) and \( x_2 \). The binary splits are represented by a, b, c, and d.](image)

In the two-dimensional explanatory variable space, the goal is to create rectangles that are pure in one of the two outcomes. The classification algorithm will search \( x_1 \) and \( x_2 \) for the best place to split the explanatory variable into two parts (binary split). In Figure 5-3, the first binary split occurred along \( x_2 \) at point a. The decision rule is given for the root node (Figure 5-4). The upper portion of the variable space only has Hs, thus all patients in this portion of the explanatory variable space died in a hospital. In Figure 5-4 this is represented by the terminal node labelled H. The lower portion is a mix of Hs and Os so further splits will be needed to reach our goal of pure nodes.
Figure 5-4: The classification tree for Figure 5-3. If the condition is true, go down the left branch. If the condition is false, go down the left branch. "H" indicates death in a hospital. "O" indicates death out of a hospital. The terminal nodes are considered pure. There are only members of one class in each terminal node.

The second binary split was made along $x_1$ at point $c$. This split the lower rectangle in Figure 5-3 into two parts. The left hand side has only patients that died out of a hospital. The terminal node is labelled as O in Figure 5-4. The right hand side is a mix of Hs and Os. Further splits will be needed to reach our goal of pure nodes. This process continues until the nodes are pure or some predetermined stopping criteria is reached. In reality, of course, pure nodes are rare.

For each region, a probability of belonging to a particular class is the proportion of that class in the region. Translating this to the tree graph (Figure 5-4), the probability of belonging to a particular class in a terminal node is the proportion of patients of that class associated with the terminal node. For us, we can determine the probability of dying out of a hospital given the explanatory variables $x$, $\Pr(\text{Class}=O \mid x)$ by the proportion of patients who died out of a hospital for the terminal node associated with $x$. 
To make this precise, the probability of dying out of a hospital given that the explanatory variables, $x$, puts the patient in the $j$th node then the estimated value of $\Pr(\text{Class}=0 \mid x)$ is,

$$
\hat{p}_j = \frac{1}{N_j} \sum_{x_r \in R_j} I(\text{Class} = 0),
$$

(3.1)

where $x_r \in R_j$ denotes all patients in the $R_j$th node (or region) and $N_j$ is the number of observations in that node (or region). This is the proportion of class 0 in the $j$th node (or region) and this proportion is uniform over the entire region$^{40}$.

For completeness, the four methods used to determine the optimal split will be briefly presented. For further information see Hastie et al.$^{40}$ or Venables et al.$^{42}$. There are four methods for determining the optimal split. The simplest of them is the misclassification rate. The split is made for the smallest misclassification rate$^{42}$. This is generally not used$^{71}$. Secondly, using the idea of likelihood, a deviance can be defined as

$$
D = -2 \sum_i \sum_k n_{ik} \log p_{ik},
$$

(3.2)

where $n_{ik}$ is the number of observations in the $i$th node and the $k$th class, $p_{ik}$ is the proportion of class $k$ in the $i$th node. The split is made where the reduction in deviance is maximized$^{42}$. The final two methods use the idea of impurity in the node distributions. Entropy,

$$
\sum p_{ik} \log p_{ik},
$$

and the Gini index,

$$
\sum p_{ij} p_{ik} = 1 - \sum_k p_{ik}^2,
$$

select the split that reduces the average impurity$^{42}$. In addition, a cost function can be utilized to penalize for overly complex trees$^{40}$.

**K Nearest Neighbours**

K nearest neighbours (KNN) utilizes the idea that similar things are often found together. The approach emerge out of the electrical engineering and information systems fields of study. It can be traced back to Cover and Hart (1967)$^{71}$. KNN is premised upon the idea that the outcome of a patient will be similar to patients who have similar explanatory variables. If the explanatory variables of a patient put it in a cluster where the majority of patients have died out of a hospital, then we would expect that the new patient would have an outcome like that of the majority.
With k nearest neighbours (KNN), decisions are based upon the majority vote among the 
k nearest neighbours. This *vox populi* approach to classification depends upon the 
definition of a neighbourhood. The neighbourhood is defined by a measure of distance. 
This can be a measure such as Euclidean distance (straight line between two points). The 
probability of dying out of a hospital given the explanatory variables x, $Pr(Class=O | x)$ 
is taken as the proportion of neighbours that are from class O.

For completeness, a general metric for the distance between two observations, x and x', is 
defined as,

$$\sqrt{(x-x')^T A (x-x')}$$  \hspace{1cm} (4.1)

This allows for a variety of measures to be used. For instance, if A is the identity matrix, 
then the distance is the p-dimensional Euclidean distance; if it is the covariance matrix, 
then the Mahalanobis distance is used. Also the number of neighbours is generally 
experimentally determined.

**Advantages and Disadvantages**

Every technique has both advantages and disadvantages. For supervised learning 
techniques, there are some common advantages.

- They can be highly adaptable allowing for complex relationships such as 
  interactions and higher order terms.
- They are useful when analysing large data sets.
- They are useful for classification.

The most common disadvantages shared by all classifiers are:

- overly optimistic assessments of performance,
- over-fitting, and
- under-fitting.
These disadvantages will be addressed in a subsequent section.

Each particular classifier has its own set of advantages and disadvantages. For the 
remainder of this section, each of the five classifiers previously discussed will have their 
advantages and disadvantages presented. Table 5-1 is a summary of the advantages and 
disadvantages.
<table>
<thead>
<tr>
<th>Classifier</th>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
</table>
| Linear Discriminant Analysis | • Adaptable for non-linear patterns (Quadratic Linear Discriminant Analysis)  
• Distributional and non-distributional approaches | • Requires assumption of equal covariance structures  
• Parametric approach requires the assumption of a multivariate normal distribution |
| Logistic Regression        | • Can have interactions and higher power terms in the model  
• Widely available in statistical packages  
• Significance testing is possible  
• Odds-ratio and relative risk are feasible | • Limited by model specification |
| Neural Networks            | • Can find complex patterns  
• Flexible architecture  
• Many parameters available to enhance the performance of the classifier  
• Not limited to the softmax function | • Significance testing not available  
• Black-box  
• May not find global minimum  
• Large computation resources required |
| Classification Trees       | • Can find complex patterns  
• Results transfer easily to clinical setting  
• Interactions can be implied | • Significance testing not available  
• Does not optimize over the entire tree |
| Nearest Neighbours         | • Can find complex patterns  
• Memory-based  
• Can identify pockets of activity  
• Alternate measures of distance | • Significance testing not available  
• Black box |
Linear discriminant analysis has both parametric and non-parametric methods. Only the non-parametric approach was previously presented. Having a parametric method for the development of LDA does allow for statistical inference, but a multivariate normal distribution needs to be assumed. LDA does assume a linear relationship but a non-linear relationship can be explored by relaxing the assumption of equal variance-covariance matrices.

Logistic regression has many appealing properties which include:
- wide-spread implementation in statistical software packages,
- the use relative risk and the odds ratio, and
- hypothesis testing and inferential procedures.

Although logistic regression is a linear modelling technique, interactions and higher order terms can be included. Logistic regression does not automatically seek out these complex relationships, so it is limited to the model specified by the researcher.

As with logistic regression, neural networks has many appealing properties. It can find complex relationships, has a flexible architecture, and has many parameters that help fine tune its performance (e.g. decay parameter). As well, it is not restricted to a single type of function for the nodes. There are some drawbacks to this flexible classifier. It is a black box procedure in that it is challenging to extract information about the utility of the explanatory variables, but research into the extraction of important variables is emerging. A minimum will be found for the measure of error used, but there is no guarantee that the method has found the global minimum or the best solution.

Classification trees, in partitioning the variable space into hyper-rectangles, yield results in the form of classification rules. The obvious advantage of this approach is that it is easily transferred into a clinical setting as classification trees can be easily explained. These rules can be presented as written rules or pictorially as a decision tree. As well, interaction effects are suggested through a particular branch of the tree. The partitioning of the explanatory variable space into hyper-rectangles means that the classifier will perform poorly if the class probabilities have gradual transitions rather than abrupt ones. In this situation, linear discriminant analysis, logistic regression, or neural networks may out perform classification trees. Also, there is no attempt to optimize over the entire tree.

Nearest Neighbours is a memory-based method. It can produce demarcations for very complicated boundaries as well as identify pockets of peculiarity which may be overlooked by other methods. Another benefit of the procedure is that the notion of distance can be changed, thus the researcher is not limited to the Euclidean measure alone. Nearest neighbours is a black box approach and there is not an efficient means to determine which explanatory variables contribute meaningful information. Also, as the variable space increases in its dimensions, the space itself becomes more sparse which may degrade the performance.
As the data under analysis is not simulated or simple in its construction, it is difficult to make a priori decisions about the underlying structure of the data. Given that the structure of the data will be, in general, unknown, the perspectives given by each classifier will help to uncover the underlying story in the data. Each classifier provides insights into the data that may not be revealed by other classifiers, the strengths of one compensating for the weaknesses of another. For example, neural networks can be used to see if complex relationships between the explanatory variables is better than simple linear relationships, and classification trees can give insights into the types of interactions that should be explored. By using the strengths of each classifier, a more complete picture of the data can be obtained.

**Evaluating the Performance of Data Mining Methodologies**

**Performance Criteria**

Recall that each classifier will produce a conditional probability of dying out of a hospital, \( \pi_0 = \Pr(\text{Class}=0|x) \). This probability will be used to determine the class to which a patient belongs based upon a threshold value or it will be used to prioritize patients. The subsequent methods for assessing the performance of a classifier use the conditional probability of dying out of a hospital in one of these two ways.

When a model is fit to a data set, a natural thing to wonder about is its performance. The following techniques can be applied to any classification method. In assessing the performance of a classifier, two main issues emerge: the criterion for measuring the performance and the means by which the criterion is estimated. Four criteria will be considered: the misclassification rate, the receiver operating characteristic curve, the area under the receiver operating characteristic curve and the hit curve. Cross-validation will be considered for the estimation of the performance criterion.

**Misclassification, Loss Function and the Misclassification Curve**

The misclassification rate is the proportion of patients that have been misclassified. Frequently this is the first measure considered when assessing the performance of a classifier. It is easy to communicate and easy to calculate. For the two-class problem for predicting the location of death for cancer patients a general misclassification table is given by Table 5-2. The overall misclassification is

\[
\text{Miss} = \frac{B+C}{A+B+C+D}.
\]  

(5.1)

This formulation of the misclassification rate gives equal weight to all errors. This implies that a patient is classified as dying out of a hospital if \( \Pr(\text{Class}=0|x) > 0.5 \).
In reality, one type of misclassification may be worse than another. For example, it may be worse to classify a patient who died in a hospital as having died out of a hospital than to classify a patient who died out of a hospital as having died in a hospital. One method of accounting for differences in the costs of misclassifying a patient is the 0-1-c loss function.

In the two-class situation the loss function allows the researcher to specify the loss associated with a certain type of misclassification. Recall that the outcome (response) \( Y \) is \( H \) if a patient died in a hospital and \( O \) if a patient died out of a hospital. The loss function is given as

\[
L(Y, \hat{Y}) = \begin{cases} 
0 & \text{if } Y = H \text{ and } \hat{Y} = H \\
0 & \text{if } Y = O \text{ and } \hat{Y} = O \\
1 & \text{if } Y = H \text{ and } \hat{Y} = O \\
c & \text{if } Y = O \text{ and } \hat{Y} = H 
\end{cases} 
\]

(5.2)

where \( \hat{Y} \) is the predicted class. Table 5-3 presents this in tabular form.

Table 5-3: 0-1-c loss function; the truth is the real state (what was observed) and the decision is the decision made by the classifier. "H" indicates death in a hospital. "O" indicates death out of a hospital.
Since deaths out of a hospital are of primary interest in this research, we want to minimize the risk of misclassifying patients who died out of a hospital. The risk of misclassification given that the patient did die out of the hospital is

\[ R(Y_0, \hat{Y}) = cP(Y = H) + 0P(Y = O), \]  

(5.3)

where \( Y_0 \) is actual out of hospital death. The risk of misclassification given that the patient did die in the hospital is

\[ R(Y_H, \hat{Y}) = 0P(Y = H) + 1P(Y = O), \]  

(5.4)

where \( Y_H \) is actual in hospital death. To minimize the risk of misclassifying the patients who died out of a hospital, we want (5.3) to be less than (5.4),

\[ R(Y_0, \hat{Y}) < R(Y_H, \hat{Y}) \]

\[ cP(Y = H) < P(Y = O) \]

\[ \frac{c}{1+c} < P(Y = O) \]

\[ k < \pi_0, \]  

(5.5)

where \( P(Y = O) = \Pr(\text{Class}=O \mid x)=\pi_0 \) and cut point \( k = \frac{c}{1+c} \). If \( P(\hat{Y} = O) \) is greater than the value \( k \), then the patient is classified as having died out of the hospital. If it is less than \( k \), then the patient is classified as having died in a hospital.

This yields a decision rule based upon some cost, \( c \), associated with the misclassification. If \( c>1 \) then the cost of misclassifying an observation from class 1 as being a member of class 0 is greater than the cost of misclassifying a member of class 0 as belonging to class 1. The reverse is true if \( c<1 \) and the costs are equal for \( c=1 \).

In certain situations, the cost associated with misclassification is unknown. In these situations, the assumption of equal cost may be too naïve, thus an experimental approach can be utilized. A sequence of decision rules, or cut points, can be used. This will result in a sequence of misclassification tables; one for each cut point. The table with the lowest misclassification rate has the optimal cut point. From this cut point the cost can be determined.

More technically, consider classification to be a function, \( \hat{Y} \), of some value \( k \), where \( k \in [0,1] \). The function can be defined as

\[ \hat{Y}(k) = \begin{cases} O \text{ if } P(Y = O) > k \\ H \text{ otherwise} \end{cases}. \]  

(5.6)
This function is the predicted value for the patient (Table 5-2). The misclassification error is a function of \( Y \), the true outcome and \( \hat{Y}(k) \), the predicted outcome. The misclassification error, \( Miss \), becomes,

\[
Miss_k (Y, \hat{Y}(k)) = \frac{B_k + C_k}{A_k + B_k + C_k + D_k}.
\]

A misclassification table can be constructed for each value of \( k \), thus a misclassification rate can be obtained for each value of \( k \). A set of misclassification errors results,

\[
MISS = \{Miss_k (Y, \hat{Y}(k)) \mid k \in [0,1] \}.
\] (5.7)

The value of \( k \) chosen is the smallest misclassification rate,

\[
\hat{k} = \arg \min_k (MISS)
= \arg \min_k (Miss_k (Y, \hat{Y}(k))).
\]

With the optimal cut point determined, the value of \( c \) can be found,

\[
c = \frac{k}{1-k}.
\]

The value of \( c \) is the empirically derived cost that yields the minimum misclassification rate for the classifier. The value of \( c \) will offer an insight into the relative costs of misclassification for the particular data set as mentioned earlier. A plot of \( Miss_k \) by \( k \) yields a misclassification curve which graphically shows the minimum misclassification rate and the optimal cut point.

**Receiver Operating Characteristic Curve and Area Under the Curve**

Building upon the information obtained by constructing a misclassification table, the sensitivity and the specificity can be obtained. In this study, the sensitivity (true positive fraction) is the probability of predicting a death out of a hospital given the true state was death out of hospital,

\[
\text{Sensitivity} = \frac{A}{A+B}.
\] (6.1)

The specificity (true negative fraction) is the probability of predicting death in a hospital given the true state was death in a hospital,

\[
\text{Specificity} = \frac{D}{C+D}.
\] (6.2)
As with the misclassification rate, the sensitivity and specificity are subject to the cut point \( k \). Different values of \( k \) will result in different sensitivity and specificity values. A plot, similar to the misclassification plot can be constructed for these two measures. This plot is known as the receiver operating characteristic (ROC) curve.

The ROC curve is a plot of sensitivity versus 1-specificity (false positive fraction) for all possible cut points, \( k \). Emulating the process by which (5.7) was obtained, (6.1) and (6.2) are derived for each \( k \). This results in the pairs (1-specificity\(_k\), sensitivity\(_k\)).

A summary measure of the ROC curve is the area under the ROC curve commonly known as just the area under the curve (AUC). A straight line ROC curve from (0, 0) to (1, 1) will have an AUC of 0.5 and represents a classifier that is no better than random allocation to the classes (Figure 5-5)\(^{53}\). This suggests that the classes are not well separated by the classifier. A ROC curve that rises very quickly to a sensitivity of 1.0 will have an AUC close to 1.0. This suggests that the classes are well separated by the classifier (Figure 5-5)\(^{53}\).

\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{ROC_curve.png}
\caption{The ROC curve graphically represents the relationship between the sensitivity and the specificity. Line (A) represents a classifier that performs no better than flipping a coin. The associated AUC would be 0.5. Line (B) represents a classifier that performs well and has the ability to separate the two classes. The AUC would be near 1.0.}
\end{figure}
The ROC curve is the graphical representation of the AUC. If the ROC is a straight line from the bottom left corner to the top right corner, then the classifier does no better than a random allocation of patients to each group (death in a hospital, death out of a hospital). The associated AUC will be close to 0.5. If the ROC rises very quickly towards the top left corner and then levels off as it approaches the top right corner, then the classifier is able to do a very good job distinguishing between the two groups. The AUC will be close to 1.0.

**Hit Curve (Lift Curve)**

The above measures of performance assess how well the classifier classifies patients. Another means by which performance can be assessed is to determine how quickly it finds the patients of interest. In our situation, we are interested in finding those people who died out of a hospital.

The classifier assigns to each patient a probability of dying out of a hospital, \( \pi_0 = \Pr(\text{Class}=0 | x) \). These probabilities are sorted from largest to smallest. The patients with the highest probabilities are chosen first and a hit occurs when the chosen patient actually died out of a hospital (Table 5-4). A good classifier will quickly find the patients of interest. An effective way of presenting this is with a hit curve.

Table 5-4: Tabular representation of the hit curve where “O” is out of hospital death and “H” is hospital death.

| Rank | Study ID   | \( \pi_0 = \Pr(\text{Class}=0 | x) \) | \( y \) | Expected Number of hits | Hit/Miss |
|------|------------|--------------------------------------|-------|-------------------------|----------|
| 1    | 19991985   | 0.98                                 | O     | 0.98                    | Hit      |
| 2    | 20004563   | 0.96                                 | O     | 1.94                    | Hit      |
| 3    | 20016466   | 0.91                                 | O     | 2.85                    | Hit      |
| 4    | 20007651   | 0.82                                 | H     | 3.67                    | Miss     |
| 5    | 19970293   | 0.81                                 | O     | 4.48                    | Hit      |

More formally, if \( S \) patients are chosen and \( S_h \) of them are hits the coordinate pair \((S, S_h)\) results. For each value a coordinate pair is obtained. A plot can be constructed with the number of patients selected, \( S \), along the horizontal axis and the number of hits, \( S_h \), along the vertical axis. A classifier which quickly finds the cases of interest will have a hit curve that rises quickly (Figure 5-6).
Figure 5-6: The ideal hit curve (A). The curve rises at a 45 degree angle to the number of selected patients axis until it reaches the total number of patients who died out of hospital (n1). After this the curve is horizontal. The curve (B) is what is commonly seen with hit curves. The curve rises more slowly and needs more than n1 patients selected to find the n1 patients who died out of the hospital.

In reality, the ideal curve (A) in Figure 5-6 is rarely seen, thus the performance is compared to the expected hit curve. The expected hit curve is derived from plotting the expected number of hits against the number of patients selected. The expected number of hits can be taken as the sum of probabilities for the first S ordered patients (Table 5-4) \(^{55}\). It is desirable to see the hit curve perform at least as well as expected.

**Estimating the Performance Criterion**

**Out-of-Sample Prediction**

Four criteria for classifier performance have been presented. These four criteria will be used to assess the performance of the classifiers, but we still need to obtain an estimate for each of the performance criteria. What we need is an honest estimate of how the fitted classification model performs \(^{71}\). This needs to reflect how the fitted model will perform if new patients are classified in the future \(^{71}\). This is the problem of assessing how a fitted classifier performs with out-of-sample prediction.

When estimating the performance of a fitted classification model, three approaches can be taken. The first entails using the same data for both the construction of the model and the assessment of the model. The second entails the use of an independent set of data to
assess the fitted model. The third entails a sophisticated reuse of data so that the same
data is not used for fitting the classifier and estimating the performance of the fitted
classifier.

**Estimation by reusing the same data**

A naïve approach to estimating the performance is to use the same data to fit the model
and to assess the performance of the fitted classifier. Two related problems emerge when
data is used in this way: overly optimistic estimation and over-fitting.71

Any classifier will attempt to fit the data as well as it possibly can. So logistic regression
with only linear explanatory terms will fit the data as well as the model will allow. The
model will to some extent adjust to the random noise in the data71. If the same data used
to fit the classifier is used, then the estimation of performance will be too good because
the model has adapted to some degree to the data. The estimation will be an overly
optimistic evaluation of how the fitted classifier will perform on an independent data set.

The overly optimistic estimation of how a classifier performs can be made worse if the
model is too complex. A model can be made to adapt to the data. For example, if
interactions and higher powers of explanatory variables were included in the
specification of a logistic regression function, then the logistic regression function would
fit the data even better. In this situation, the classifier could be tailored to the data set,
thus it has been tailored to the noise in the data set. This complex model could perform
worse than a simpler model on an independent data set. This is known as over-fitting.

In order to overcome this naïve use of data, a more sophisticated view of how to use the
data is needed.

**Using data to prevent overly optimistic estimates and over-fitting**

If the data is used in such a way that overly optimistic estimates of performance are
avoided, then a significant step has been taken to also address the problem of over-fitting.
The failing of the previous approach to performance assessment was that the data was
reused and the estimate would not give an good picture of how the fit classifier would
perform on an independent data set. A simple solution is to split the data into two
parts40. The first part is used to fit the classifier and is called the “training set”. The
second part is used to assess the performance of the fit classifier and is called the “test
set”.

In most cases, more work needs to be done than just simply fit the model and assess it.
Often, the fitted model goes through a process of refinement. For logistic regression, this
would be the process by which a parsimonious model is found. For nearest neighbours,
it would be the process by which the best number of neighbours is selected. In these
cases, either the training data set or the test data set would have to be reused to refine the
model. This reuse of data would result in the same two problems that we are trying to
avoid.
In order to avoid these problems when fine-tuning the fit of the model, we could split the data into three parts (Figure 5-7)\(^{40}\). The "training set" would be used to fit the classifier. The "validation set" would be used to estimate the performance of the fitted model for model selection. When fine-tuning the fitted model, the validation set would be used as the independent data set for assessing the ability of the model to generalize to an independent data set. The "test set" would be held until all the fine-tuning of the fitted model was complete. This data has not been used for fitting the model or for selecting the best version of that model. The test set will yield an honest estimate of the performance of the fitted classifier's ability to generalize to an independent data set. In doing this, we will avoid the problems of overly optimistic estimates of a classifiers performance and over-fitting.

| Training | Validation | Testing |

**Figure 5-7:** The partitioning of data into three sets: Training set for model construction, validation set for model selection, and testing set for error estimation (generalization)

Reality rarely allows for such a partitioning of the data. It must be used in an efficient manner as the data may be difficult to obtain and expensive, in terms of time, money or both. To overcome this limitation on the use of data, sophisticated methods of data reuse have emerged. The solution utilized in this research was to efficiently reuse the data through cross-validation.

**Bias-Variance Trade-Off**

Before cross-validation is discussed, two other aspect of fitting a model will be addressed: bias and variance. These two measures are related to the complexity of the fitted model and thus to the problem of over-fitting.

When fitting a classifier, there are four items of interest\(^{40}\). There is the observed outcome, \(y\). There is the true model for the data, \(f(x)\). There is the fitted classifier used to approximate the true model, \(\hat{f}(x)\). Finally, there is the expected value of the fitted classifier, \(E\hat{f}(x)\), or simply the mean of the fitted classifier.

It is believed that the data follows some true function, \(f(x)\). The training data, \(T\), is chosen and can be defined as the set of all pairs of predictor vectors and outcomes for the training set, \((x_i, y_i)\) \(i=1,\ldots,N\), where the outcome follows the true model with some added noise, \(y_i = f(x_i) + \varepsilon_i\). Mathematically, the training set is defined as \(T=\{(x_i,\)}
y_i)_{i=1,\ldots,N} and y_i = f(x_i) + \varepsilon_i$. Furthermore, we assume that noise has mean zero, $E(\varepsilon)=0$, and constant variance, $\text{Var}(\varepsilon)=\sigma^2$.

A classifier is fit to the data. The resulting fitted model, $\hat{f}(x)$, is an approximation of the true model, $f(x)$. As before, we want to assess the performance of the classifier. For simplicity squared error loss is used. This is the squared difference between the observed value, $y_i$, and the fitted value, $\hat{f}(x_i)$. The expected prediction error ($\text{PErr}$) at $x_i$, based upon the squared error is,

$$
\text{PErr}(x_i) = E(y_i - \hat{f}(x_i))^2 \mid X = x_i)
$$

$$= E\left[y_i - f(x_i) + f(x_i) - \hat{f}(x_i)\right]^2
$$

$$= \sigma^2 + \text{MSE}(\hat{f}(x_i))
$$

$$= \sigma^2 + E\left[\hat{f}(x_i) - f(x_i)\right]^2
$$

$$= \sigma^2 + E\left[f(x_i) - E[\hat{f}(x_i)] + E[\hat{f}(x_i)] - f(x_i)\right]^2
$$

$$= \sigma^2 + [E\hat{f}(x_i) - f(x_i)]^2 + E\left[\hat{f}(x_i) - E\hat{f}(x_i)\right]^2
$$

$$= \sigma^2 + \text{Bias}^2(\hat{f}(x_i)) + \text{Var}(\hat{f}(x_i))
$$

$$= \text{Irreducible Error} + \text{Bias}^2 + \text{Variance}.$$

It is good to observe that the prediction error can be broken into three parts. The first portion, the irreducible error cannot be changed. It is the variance of the observed value. This is the squared difference between the observed value, $y_i$, and the true functional value, $f(x)$. This is the variance of the noise in the data, $\text{Var}(\varepsilon)=\sigma^2$.

The second and third parts are under our control and they comprise the mean squared error. The mean squared error is the squared difference between fitted value and the true value. This is composed of the bias and variance.

The bias term is the squared difference between the expected value of the estimate, $\hat{f}(x_i)$ and the true value, $f(x_i)$. The bias can be eliminated by constructing a complex model that is tailored to the true one. When this occurs, the fitted value is the same as the observed value, $\hat{f}(x_i) = y_i$. When the bias is eliminated the mean of the fitted classifier equals the true value:

$$E\left[\hat{f}(x_i)\right] = E[y_i]
$$

$$= E\left[f(x_i) + \varepsilon_i\right]
$$

$$= f(x_i).$$

118
It may be believed that this would be good, but the variance term becomes $\sigma_e^2$ which increases the overall variance. This translates to poor out-of-sample prediction.

The variance term is the squared difference between the estimated value and the expectation of that value. If the variance is minimized by having a simple model, then the bias will be much larger than desired. The model will be under-fit. Again this translates to poor out-of-sample prediction.

The challenge is to find a balance between the variance and the bias. A simple solution is efficiently to re-use the data with a procedure such as cross-validation.

**Cross-Validation**

Cross-validation appears to have been introduced in 1951 by Moiser as a means by which the predictive validity of linear regression equations could be evaluated. This was utilized to prevent an overly optimistic assessment of how the fitted linear regression would perform with independent data. The data was divided into two parts, the training set and the test set. It was found that a better estimate of the model's performance was obtained.

Cross-validation is a non-analytic method used to give an honest estimate of the performance of the fitted classification model on an independent set of data. This is known as test error or the generalization error.

In cross-validation, the data is split into $V$, roughly equal, parts or folds (Figure 5-8). We chose $V=10$. In its simplest form, one part is removed from the data set. The other $V-1$ parts are used to fit a classification model. The fit model is assessed with the data that was removed. As this data was not used to train the classifier, it is an independent set of data. This approach allows for the training of a classifier and the assessment of its performance with independent set of data. This is repeated and each of the $V$ parts of the data are removed in turn while the other $V-1$ parts of the data are used to fit a classifier. This is called $V$-fold cross-validation.
Figure 5-8: An example of cross-validation illustrated by 10-fold cross-validation

When each part has been held out once, then there will be $V$ assessments of the performance of the classifier, each of which is the result of a data set that is independent of the data used to fit the model. For example, if the AUC is the measure of performance then there will be ten measures of the AUC for a classifier. Each AUC came from a model fitted with nine parts of the data and assessed by the one part that was withheld from the process of training the classifier.

The cross-validation assessment of how a classifier will perform on an independent set of data is an estimate of a method’s test error. It is the average of the test errors from the $V$ folds,

$$Err^{CV} = \frac{1}{V} \sum_{i=1}^{V} Err_i,$$

where $Err_i$ is the test error associated with the $i^{th}$ fold, $Err^{CV}$ is the cross-validation estimate of the performance of the classifier. The $Err$ is the estimate from one of the four methods of assessing the performance of the fitted classifier: misclassification, ROC, AUC, or hit curve.

Cross-Validation and model refinement
A classifier can have parameters which can help tune the classifier and enhance its performance. For linear discriminant analysis and logistic regression, we can think of a model being “fine-tuned” in terms of the number of variables in the model, number of
interactions, number of higher order terms, or the transformation of the variables\textsuperscript{75}. For neural networks, there is the weighted-decay, the presence of the skip layer, and the specification of the architecture. Classification trees has a penalty for the size of the tree, the minimum number of observations that must be in a node in order to split the node, the minimum number of observations in a terminal node, and surrogate splits for example. Nearest neighbours has the size of the neighbourhood and the measure of distance.

Having these types of parameter adds an extra level of complexity. We need more than just the training and test data sets. A third, intermediate data set is need to validate the model. As indicated above, portioning the data set into three parts allows for the model to be fine tuned or selected with a data set that is independent of the data used to fit the model. Also, it allows for an independent set of data for the assessment of model performance. Hastie et al.\textsuperscript{40} give insight into a general approach.

Consider a classifier that has a parameter that can be used to fine tune the model. Let $\alpha$ be this parameter. Then the model can be expressed as a function of both the data and the parameter, $f(x, \alpha)$. For any method of assessing the performance of a trained classifier (misclassification, ROC, AUC, or hit curve), the measure of performance is $Err_\alpha$. For each value of $\alpha$ an estimate of how the classifier performs can be obtained. To prevent over-fitting, under-fitting, and overly optimistic assessments of the performance, we use cross-validation. In this situation, the estimate of how the classifier performs for each value of $\alpha$ needs to be obtained by cross-validation. With this, the cross-validation estimate of classifier performance is a function of $\alpha$, thus (8.1) is modified resulting in

$$Err_{\text{CV}}(\alpha) = \frac{1}{V} \sum_{i=1}^{V} Err_{i, \alpha}, \quad (8.2)$$

where $Err_{i, \alpha}$ is the estimate of performance associated with the $i^{th}$ fold and $\alpha$, and $Err_{\text{CV}}$ is the cross-validated measure of performance. This is an average of the $V$ folds as in (8.1), but it is for a particular value of $\alpha$. The desired parameter is the value of $\alpha$ which minimizes the generalization error (measure of performance). With this chosen, all the data is then fit\textsuperscript{40}.

**Two-layer Cross-Validation**

The method proposed above for model refinement finds the value of $\alpha$ that minimizes the estimate of model performance. Although this estimate is derived from cross-validation, the selection of the best estimate of model performance is not necessarily an honest use of the data. The cross-validation process was used to select the best model based upon the optimal value of $\alpha$. The estimate of the performance has not been done with data independent of the selection process. As a result, the measure of performance will be overly optimistic as the same data has been used for model selection (finding $\alpha$) and for assessing the performance of the model.
To address this concern, we used a two-layer cross-validation process. The two-layer cross-validation is where $V$-fold cross-validation is applied to the training set (Figure 5-9). From the training set, the cross-validation process is used to determine the optimal value of a parameter, $\alpha$. The process of model selection is now done with only the original training data set. Once the model is selected, it is assessed with the original test data set (Figure 5-9).

Figure 5-9: An example of two-layer cross-validation illustrated by 10-fold two layer cross-validation
We will call the original split of the data the outer layer and the split of the training data the inner layer. In more detail, the outer layer is a V-fold cross-validation applied to the entire data set. The inner layer is a V-fold cross-validation applied to the V-1 outer layer portions of data that comprise the training set (Figure 5-9). The inner layer uses V-fold cross-validation to find the optimal value of $\alpha$. The outer layer test data, which is an independent data set, is used to obtain an estimate of the fitted classifier based upon the optimal parameter $\alpha$. An overall estimate of optimal classifier performance is,

$$Err_{CV}^{outer} = \frac{1}{V_{outer}} \min_{\alpha} \sum_{j=1}^{V_{outer}} \left[ Err_{inner}^{CV}(\alpha) = \frac{1}{V_{inner}} \sum_{i=1}^{V_{inner}} Err_{i,\alpha} \right].$$  (8.3)

This is an average over all the best values of $\alpha$, where the best value of $\alpha$ was chosen using (8.2).

**Fitting the classifiers with cross-validation**

The cross-validation approach utilized was determined by the needs of the classifier. 10-fold cross-validation was used for all classifiers. When model selection was necessary, AUC was used to estimate the performance of the fitted classifier. As the AUC was used (8.3) was modified to

$$AUC_{CV}^{outer} = \frac{1}{V_{outer}} \max_{\alpha} \sum_{j=1}^{V_{outer}} \left[ AUC_{inner}^{CV}(\alpha) = \frac{1}{V_{inner}} \sum_{i=1}^{V_{inner}} AUC_{i,\alpha} \right],$$  (8.4)

as a larger AUC indicates better performance.

LDA used (8.1) as there were no tuning parameters to optimize. Logistic regression used (8.1) in two stages. The first use of was for the manual stepwise selection process by which a parsimonious model would be selected. The deviance was used as the measure of performance for this selection process. Equation (8.1) was then applied to the resultant model using all the data.

Neural networks has several parameters that can be adjusted to increase the performance of the model: decay parameter, number of nodes in the hidden layer, and the use of a skip layer. To understand how these parameters affect the performance of the classifier a $2^3$ factorial design was used. The response was taken as the average of three AUC estimates of performance. Each AUC was estimated using 10-fold cross-validation and each AUC was an estimate of a neural network fit based upon a different random starting point for the back-propagation optimizing algorithm (Table 5-5). It was found that the decay was not significant ($p > 0.05$) and that performance increased with no skip ($p < 0.0001$) and a higher number of nodes in the hidden layer ($p < 0.01$). The model was fit using 10-fold cross-validation, no skip and 20 nodes in the hidden layer.
Table 5-5: $2^3$ design for determining the best set of parameters for a neural network

<table>
<thead>
<tr>
<th>Run</th>
<th>Number of Nodes</th>
<th>Decay</th>
<th>Skip</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5</td>
<td>0</td>
<td>T</td>
<td>0.5393</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>0</td>
<td>F</td>
<td>0.5812</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>0.0001</td>
<td>T</td>
<td>0.5474</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>0.0001</td>
<td>F</td>
<td>0.5785</td>
</tr>
<tr>
<td>5</td>
<td>20</td>
<td>0</td>
<td>T</td>
<td>0.5573</td>
</tr>
<tr>
<td>6</td>
<td>20</td>
<td>0</td>
<td>F</td>
<td>0.6094</td>
</tr>
<tr>
<td>7</td>
<td>20</td>
<td>0.0001</td>
<td>T</td>
<td>0.5526</td>
</tr>
<tr>
<td>8</td>
<td>20</td>
<td>0.0001</td>
<td>F</td>
<td>0.6149</td>
</tr>
</tbody>
</table>

Classification trees were built using 10-fold two-layer cross-validation. The inner layer was used to construct the tree and the outer layer was used to test the resulting tree. The overall performance of classification trees is given by (8.4).

The optimal tree was selected by modifying the two-layer cross-validation approach so that,

$$
\text{optimal classifier} = \max \left[ AUC(\hat{f}_{\min}(x)) \right],
$$

where $\hat{f}_{\min}(x)$ is the 10-fold cross-validated classification tree that minimizes the entropy. Here the entropy is minimized in the inner layer while constructing the classification tree, $\hat{f}_{\min}(x)$. The tree was tested with the test data of the outer layer. The best tree is selected based upon the maximum AUC from the outer layer.

Nearest neighbours used the approach given by (8.4). The inner layer was used to select the optimal number of neighbours and the outer layer tested it. The cross-validated AUC is a measure of the overall performance of nearest neighbours.

An estimate of the standard error of the cross-validated AUC was found by using the 10 AUC estimates from the cross-validation procedure. The standard error was found using,

$$
SE(AUC_{cv}) = \frac{SD(AUC)}{10} = \frac{1}{10} \sqrt{\frac{1}{9} \sum_{i=1}^{10} (AUC_i - AUC_{cv})^2}.
$$

(8.5)
Normal approximation was used to construct the 95% confidence intervals for the AUC\textsuperscript{CV} of each classifier.

**Analysis**

We used cross-tabulation (chi-squared test for homogeneity) to see if each variable was homogeneous with respect to the location of death. This follows the methodology as established in Nova Scotia\textsuperscript{1}. If the variable was not homogeneous with respect to the place of death then it was retained as an explanatory variable for each of the classifiers.

Linear discriminant analysis, logistic regression, neural networks, classification trees, and nearest neighbours were the classifiers used in this investigation. A parsimonious model was found for logistic regression. The odds ratios for both the full model and the reduced model were found. Cross-validated stepwise elimination was used to find the parsimonious model with a decision p-value of 0.05.

10-fold cross-validation was used to estimate the performance of the classifiers. 10-fold cross-validation was maintained for each layer when two-layer cross-validation was used. Model selection was based upon the AUC. SAS was used for the cross-tabulations and the statistical package R was used for all other analysis. The AUC was calculated using the implementation in the statistical analysis software R\textsuperscript{54}. The hit curve function was implemented in R as developed by Welch\textsuperscript{12}.

**Results**

A total of 38,050 British Columbian adults died of cancer in B.C. between 1999 and 2003. We excluded the data for 1618 people because of missing information pertaining to the income quintile or the region of residence. Of the remaining 36,432 people, 23,385 (64.2%) died in the hospital. The characteristics of the two groups are presented in Table 5-6. All the characteristics were found to be associated with the location of death (p < 0.0001).

<table>
<thead>
<tr>
<th>Year</th>
<th>In Hospital</th>
<th>Percent in Hospital</th>
<th>Out of Hospital</th>
<th>Percent out of Hospital</th>
</tr>
</thead>
<tbody>
<tr>
<td>1999</td>
<td>4232</td>
<td>61.86%</td>
<td>2609</td>
<td>38.14%</td>
</tr>
<tr>
<td>2000</td>
<td>4515</td>
<td>66.41%</td>
<td>2284</td>
<td>33.59%</td>
</tr>
<tr>
<td>2001</td>
<td>4566</td>
<td>64.73%</td>
<td>2488</td>
<td>35.27%</td>
</tr>
<tr>
<td>2002</td>
<td>4897</td>
<td>64.49%</td>
<td>2696</td>
<td>35.51%</td>
</tr>
<tr>
<td>2003</td>
<td>5175</td>
<td>63.54%</td>
<td>2970</td>
<td>36.46%</td>
</tr>
</tbody>
</table>

Table 5-6: Characteristics of adults in BC who died of cancer from 1999 to 2003, by place of death

\[125\]
<table>
<thead>
<tr>
<th>Sex</th>
<th>12624</th>
<th>65.21%</th>
<th>6736</th>
<th>34.79%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>10761</td>
<td>63.03%</td>
<td>6311</td>
<td>36.97%</td>
</tr>
<tr>
<td>Age, yr</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20-44</td>
<td>791</td>
<td>66.53%</td>
<td>398</td>
<td>33.47%</td>
</tr>
<tr>
<td>45-64</td>
<td>5792</td>
<td>66.13%</td>
<td>2967</td>
<td>33.87%</td>
</tr>
<tr>
<td>65-74</td>
<td>6555</td>
<td>66.59%</td>
<td>3289</td>
<td>33.41%</td>
</tr>
<tr>
<td>75-84</td>
<td>7322</td>
<td>63.71%</td>
<td>4170</td>
<td>36.29%</td>
</tr>
<tr>
<td>≥85</td>
<td>2925</td>
<td>56.82%</td>
<td>2223</td>
<td>43.18%</td>
</tr>
<tr>
<td>Neighbourhood IPPE</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td>5880</td>
<td>65.46%</td>
<td>3103</td>
<td>34.54%</td>
</tr>
<tr>
<td>Lower Middle</td>
<td>4908</td>
<td>65.25%</td>
<td>2614</td>
<td>34.75%</td>
</tr>
<tr>
<td>Middle</td>
<td>4598</td>
<td>63.97%</td>
<td>2590</td>
<td>36.03%</td>
</tr>
<tr>
<td>Upper Middle</td>
<td>4104</td>
<td>63.51%</td>
<td>2358</td>
<td>36.49%</td>
</tr>
<tr>
<td>Upper</td>
<td>3895</td>
<td>62.05%</td>
<td>2382</td>
<td>37.95%</td>
</tr>
<tr>
<td>Missing</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aggregate Tumour Group</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Breast</td>
<td>1726</td>
<td>60.14%</td>
<td>1144</td>
<td>39.86%</td>
</tr>
<tr>
<td>Colorectal</td>
<td>2182</td>
<td>61.67%</td>
<td>1356</td>
<td>38.33%</td>
</tr>
<tr>
<td>Hematologic</td>
<td>2466</td>
<td>71.52%</td>
<td>982</td>
<td>28.48%</td>
</tr>
<tr>
<td>Lung</td>
<td>6152</td>
<td>65.20%</td>
<td>3283</td>
<td>34.80%</td>
</tr>
<tr>
<td>Other</td>
<td>6888</td>
<td>63.56%</td>
<td>3949</td>
<td>36.44%</td>
</tr>
<tr>
<td>Pancreas</td>
<td>1238</td>
<td>62.56%</td>
<td>741</td>
<td>37.44%</td>
</tr>
<tr>
<td>Prostate</td>
<td>1451</td>
<td>58.32%</td>
<td>1037</td>
<td>41.68%</td>
</tr>
<tr>
<td>Unknown Primary</td>
<td>1282</td>
<td>69.79%</td>
<td>555</td>
<td>30.21%</td>
</tr>
<tr>
<td>Region of Death</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interior</td>
<td>4338</td>
<td>59.69%</td>
<td>2930</td>
<td>40.31%</td>
</tr>
<tr>
<td>Fraser</td>
<td>7988</td>
<td>70.45%</td>
<td>3350</td>
<td>29.55%</td>
</tr>
<tr>
<td>Vancouver Coastal</td>
<td>5241</td>
<td>65.54%</td>
<td>2756</td>
<td>34.46%</td>
</tr>
<tr>
<td>Vancouver Island</td>
<td>4730</td>
<td>59.13%</td>
<td>3269</td>
<td>40.87%</td>
</tr>
<tr>
<td>Northern</td>
<td>1088</td>
<td>59.45%</td>
<td>742</td>
<td>40.55%</td>
</tr>
<tr>
<td>Length of Survival, days</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>≤60</td>
<td>5902</td>
<td>74.75%</td>
<td>1994</td>
<td>25.25%</td>
</tr>
<tr>
<td>61-120</td>
<td>2354</td>
<td>63.45%</td>
<td>1356</td>
<td>36.55%</td>
</tr>
<tr>
<td>≥121</td>
<td>15129</td>
<td>60.94%</td>
<td>9697</td>
<td>39.06%</td>
</tr>
<tr>
<td>City Size, people</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>≥1,250,000</td>
<td>11270</td>
<td>69.06%</td>
<td>5050</td>
<td>30.94%</td>
</tr>
<tr>
<td>500,000 - 1,249,999</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>100,000 - 499,999</td>
<td>3923</td>
<td>55.71%</td>
<td>3119</td>
<td>44.29%</td>
</tr>
<tr>
<td>10,000 - 99,999</td>
<td>4966</td>
<td>61.48%</td>
<td>3111</td>
<td>38.52%</td>
</tr>
<tr>
<td>&lt; 10,000</td>
<td>3226</td>
<td>64.61%</td>
<td>1767</td>
<td>35.39%</td>
</tr>
<tr>
<td>Origins</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Heterogeneous</td>
<td>2477</td>
<td>63.20%</td>
<td>1442</td>
<td>36.80%</td>
</tr>
<tr>
<td>Aboriginal Presence</td>
<td>82</td>
<td>62.12%</td>
<td>50</td>
<td>37.88%</td>
</tr>
<tr>
<td>Chinese Presence</td>
<td>1286</td>
<td>66.63%</td>
<td>644</td>
<td>33.37%</td>
</tr>
<tr>
<td>Great Britain Presence</td>
<td>10157</td>
<td>63.19%</td>
<td>5916</td>
<td>36.81%</td>
</tr>
<tr>
<td>East Indian Presence</td>
<td>473</td>
<td>72.43%</td>
<td>180</td>
<td>27.57%</td>
</tr>
<tr>
<td>Americas and Europe</td>
<td>8910</td>
<td>64.92%</td>
<td>4815</td>
<td>35.08%</td>
</tr>
<tr>
<td>Mother Tongue</td>
<td>Count</td>
<td>Median</td>
<td>Mean</td>
<td>SD</td>
</tr>
<tr>
<td>-------------------------------------</td>
<td>-------</td>
<td>---------</td>
<td>-------</td>
<td>-------</td>
</tr>
<tr>
<td>Punjabi Presence</td>
<td>500</td>
<td>73.21%</td>
<td>183</td>
<td>26.79%</td>
</tr>
<tr>
<td>Dominant English</td>
<td>2280</td>
<td>63.07%</td>
<td>1335</td>
<td>36.93%</td>
</tr>
<tr>
<td>Heterogeneous with Dominant English</td>
<td>5399</td>
<td>66.76%</td>
<td>2688</td>
<td>33.24%</td>
</tr>
<tr>
<td>Chinese Presence</td>
<td>13797</td>
<td>62.86%</td>
<td>8152</td>
<td>37.14%</td>
</tr>
<tr>
<td>Strong English Presence</td>
<td>243</td>
<td>69.38%</td>
<td>183</td>
<td>26.79%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Religion</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Heterogeneous with Protestant and</td>
<td>7862</td>
<td>62.91%</td>
<td>4635</td>
<td>37.09%</td>
<td>7862</td>
<td>62.91%</td>
<td>4635</td>
<td>37.09%</td>
</tr>
<tr>
<td>No Religion Presence</td>
<td>104</td>
<td>68.87%</td>
<td>47</td>
<td>31.13%</td>
<td>104</td>
<td>68.87%</td>
<td>47</td>
<td>31.13%</td>
</tr>
<tr>
<td>Catholic Presence</td>
<td>6290</td>
<td>63.53%</td>
<td>3611</td>
<td>36.47%</td>
<td>6290</td>
<td>63.53%</td>
<td>3611</td>
<td>36.47%</td>
</tr>
<tr>
<td>Protestant Presence</td>
<td>658</td>
<td>73.52%</td>
<td>237</td>
<td>26.48%</td>
<td>658</td>
<td>73.52%</td>
<td>237</td>
<td>26.48%</td>
</tr>
<tr>
<td>Sikh Presence</td>
<td>3248</td>
<td>63.64%</td>
<td>1856</td>
<td>36.36%</td>
<td>3248</td>
<td>63.64%</td>
<td>1856</td>
<td>36.36%</td>
</tr>
<tr>
<td>No Religion Presence</td>
<td>5223</td>
<td>66.25%</td>
<td>2661</td>
<td>33.75%</td>
<td>5223</td>
<td>66.25%</td>
<td>2661</td>
<td>33.75%</td>
</tr>
</tbody>
</table>

Note: Ethnic Origins, Mother Tongue, and Religion are cultural indicators developed for this research. The names given are a description of the dominant features of the particular cluster. The clusters were developed using principal components and k-means clustering algorithm.

Table 5-7 reveals that the average age for patients in each group is in the mid-70s and that the mean survival times are greater than 2 years. There is a statistically significant difference between the mean ages for those who died out of hospital and those who died in hospital (p < 0.0001). Those who died in hospital and those who died out of hospital were found to have a statistically significant difference in survival time (p < 0.0001).

Table 5-7: Characteristics for age and survival as continuous variables

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>In Hospital</th>
<th>Out of Hospital</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (years)</td>
<td>23,385</td>
<td>13,047</td>
</tr>
<tr>
<td>Survival (days)</td>
<td>23,385</td>
<td>13,047</td>
</tr>
</tbody>
</table>

Note: SD = Standard Deviation

There appears to be no noticeable overall trend in the location of death through the study period (Figure 5-10). There is a decrease in the proportion of subjects dying out of hospital when comparing 1999 to 2000, but from 1999 to 2000 the trend is towards an increase in the proportion of patients who are dying out of hospital (Table 5-6).
Figure 5-10: Proportion of death out of hospital and in hospital among adults in British Columbia who died of cancer, 1999-2003.

Classifier Comparison Results

Receiver Operating Characteristic Curve Comparison

In Figure 5-5, curve B is the shape of ROC curve that is desired. None of the classifiers perform as well as curve B in Figure 5-5 (Figure 5-11). Neural networks produces the best ROC curve of all the classifiers and is not intersected by any of the other ROC curves. Nearest neighbours performs the worst of all the classifiers. Linear discriminant analysis, logistic regression and classification trees have similar ROC curves.
**Figure 5-11**: Receiver Operating Characteristic Curves for linear discriminant analysis, logistic regression, neural networks, classification trees, and nearest neighbours. Neural network has the best ROC curve and nearest neighbours has the worst. Linear discriminant analysis, logistic regression, and classification trees are similar.

**Area Under the Receiver Operating Characteristic Curve**

Each of the classifiers performs better than random ($p < 0.0001$). Although the AUC are not as large as we would have liked, all the classifiers are better at allocating patients to the appropriate group than a random allocation of patients to each group. This suggests that this measure has utility for measuring classifier performance for this data set.
The standard errors for the area under the receiver operating characteristic curve, as calculated in (8.5), are small. This indicates that for each classifier there was little variation amongst the measures of AUC over the partitions of the data. The 95% confidence intervals suggests three groups of classifiers (Table 5-8). The first group, with the worst performance is comprised only of nearest neighbours. The second group is comprised of linear discriminant analysis the full model for logistic regression and the parsimonious logistic regression model. The third group has neural networks as its only member. Classification trees can be assigned to both the second and third groups because it has such a large confidence interval. It performs better on average than the second group, but the variation in its performance is large.

### Table 5-8: Cross-validated Area Under the Curve and cross-validated estimate of the AUC standard error

<table>
<thead>
<tr>
<th>Classifier</th>
<th>AUC&lt;sup&gt;CV&lt;/sup&gt;</th>
<th>SE(AUC&lt;sup&gt;CV&lt;/sup&gt;)</th>
<th>CI (95%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Lower</td>
</tr>
<tr>
<td>Linear Discriminant Analysis</td>
<td>0.5967</td>
<td>0.0083</td>
<td>.5916</td>
</tr>
<tr>
<td>Logistic Regression (Full Model)</td>
<td>0.5968</td>
<td>0.0082</td>
<td>.5917</td>
</tr>
<tr>
<td>Logistic Regression (Parsimonious Model)</td>
<td>0.5968</td>
<td>0.0073</td>
<td>.5923</td>
</tr>
<tr>
<td>Neural Networks</td>
<td>0.6181</td>
<td>0.0103</td>
<td>.6117</td>
</tr>
<tr>
<td>Classification Tree</td>
<td>0.5988</td>
<td>0.0235</td>
<td>.5842</td>
</tr>
<tr>
<td>Nearest Neighbours</td>
<td>0.5620</td>
<td>0.0062</td>
<td>.5581</td>
</tr>
</tbody>
</table>

Note: CV=Cross-validation, SE=Standard Error, CI=Confidence Interval

Interestingly, even though the rounded values for the AUC for the full and parsimonious logistic models are identical, there is a statistically significant difference between them (p < 0.001). Between the two versions of logistic regression there is an AUC difference of 3.86E-5 with a correlation of $r = 0.993412$ between them. In terms of the AUC, the full model slightly outperforms the reduced model. This suggests that added information in the full model (Religion) makes a meaningful contribution even though a stepwise regression based upon the deviance would suggest the reduced model.

### Hit Curve

Recall that the hit curve is a graphical representation of how quickly a classifier finds the patients who died out of hospital. The desired hit curve would be one that at least performs as well as expected. As there are 13,047 patients who died out of a hospital in this data study, the first 5,000 cases were used to give an indication of how the hit curve performed.
Overall, linear discriminant analysis performs as expected. It does perform worse than expected at first, but after the first few thousand patients, it begins to find those patients who died out of hospital as quickly as expected (Figure 5-12).

Figure 5-12: Linear discriminant analysis hit curve
The full logistic regression model performs better than LDA. As with LDA, it finds the patients who died out of hospital more slowly than expected at first, but it reaches expectation much more quickly than LDA (Figure 5-13). The parsimonious model reaches expectation more quickly than the previous classifiers (Figure 5-14).

Figure 5-13: Logistic regression hit curve for the full model.
Figure 5-14: Logistic regression hit curve for the parsimonious model.

The neural network classifier performs poorly when compared to the expectation of its performance. It never reaches expectation (Figure 5-15). The distance between the expectation curve and the actual curve widens as more patients are selected. Even though neural networks does not perform as well as expected, it is able to find a similar number of patients as logistic regression and LDA over the 5,000 patients.
Figure 5-15: Neural network hit curve

The classification tree performs poorly when compared to its expected performance (Figure 5-16). Even though it never reaches the expected number of hits, it is able to find a similar number of patients as the previous classifiers over the 5,000 patients.
As with the neural network and classification tree, the nearest neighbour classifier does not perform as well as expected (Figure 5-17). Even though it never reaches the expected number of hits, it is able to find a similar number of patients as the previous classifiers over the 5,000 patients.
The expected hit curves are not the same for all the classifiers. LDA, logistic regression and classification trees have similar expectation curves. Over the 5,000 patients selected, approximately 2,500 hits are expected. For neural networks and nearest neighbours, 3,000 hits are expected over the 5,000 patients.
Misclassification Rate and Empirically Derived Cost of Misclassification

For completeness the results pertaining to the misclassification rate will be presented. As all of the classifiers performed only slightly better than classifying all the patients as dying in a hospital (0.358), this measure provides little insight into the overall performance of the classifiers (Table 5-9). The misclassification curves are presented in Appendix E.

Table 5-9: Minimum misclassification rates and empirically derived costs for each classifier. The misclassification rate for classifying all patients as dying in a hospital is given at the end of the table.

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Minimum Misclassification Rate</th>
<th>Cut Point</th>
<th>Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear Discriminant Analysis</td>
<td>0.3559</td>
<td>0.5</td>
<td>1</td>
</tr>
<tr>
<td>Logistic Regression (Full Model)</td>
<td>0.3560</td>
<td>0.5</td>
<td>1</td>
</tr>
<tr>
<td>Logistic Regression (Parsimonious Model)</td>
<td>0.3567</td>
<td>0.475</td>
<td>.905</td>
</tr>
<tr>
<td>Neural Networks</td>
<td>0.3531</td>
<td>0.525</td>
<td>1.105</td>
</tr>
<tr>
<td>Classification Tree</td>
<td>0.3531</td>
<td>0.525</td>
<td>1.105</td>
</tr>
<tr>
<td>Nearest Neighbours</td>
<td>0.3578</td>
<td>0.65</td>
<td>1.857</td>
</tr>
<tr>
<td>Classify all patients as dying in a hospital</td>
<td>0.358</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

The misclassification rate does not appear to be a good estimator for the performance of the classifiers. It suggests that classifying all the patients as having died in a hospital is just as effective as using a sophisticated classifier. By indicating that classifying all the patients as having died in a hospital is a reasonable approach, this measure of performance does little to aid the identification of the predictors for the location-of-death of cancer patients. As the utility of this estimate of classifier performance is low, no subsequent discussion for this estimate will occur.

There appears to be no consensus on the empirically derived cost associated with misclassification. Linear discriminant analysis and the full logistic model indicate no differentiation in the cost for either type of misclassification. The parsimonious logistic model suggests that the cost of misclassifying a patient who died out of hospital as a patient who died in the hospital is less than the cost associated with classifying a patient who died in the hospital as a patient who died out of the hospital. The reverse is true for neural networks, classification trees and nearest neighbours.
Health Results

In order to put the classification results into the desired health context, information concerning the utility of the explanatory variables is necessary. For linear discriminant analysis, neural networks, and nearest neighbours, we can say that the chosen set of predictors are useful in determining the location of death of British Columbian cancer patients. Within the constraints of this investigation, further expounding of these classifiers is difficult.

Logistic regression and classification trees will be considered in this section. Logistic regression has recourse to inferential methods to determine a parsimonious model. As well, the results can be expressed in terms of the odds ratio. Classification trees does find a parsimonious model, thus the relative utility of explanatory variables can be presented. These features enable us to determine which explanatory variables are useful in uncovering the patterns of health service use at the end of life.

Logistic Regression: Full and Parsimonious Models

The only difference between the full and parsimonious models is the indicator of religion. Recall that the AUC comparison of the full and parsimonious models suggested that the indicator of religion added useful information. Due to the fact that evidence exists for both the inclusion and exclusion of the indicator of religion, both logistic regression models are presented in Table 5-10 to highlight the differences that occur due to the presence of the indicator of religion.
Table 5-10: Adjusted odds of dying out of the hospital, by predictors, for British Columbia for both the full model and the parsimonious model

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Full Model</th>
<th></th>
<th>Parsimonious Model</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Adjusted OR† 95% CI</td>
<td></td>
<td>Adjusted OR‡ 95% CI</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Lower Bound</td>
<td>Upper Bound</td>
<td>Lower Bound</td>
</tr>
<tr>
<td><strong>Year of Death</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1999*</td>
<td>1.00</td>
<td></td>
<td></td>
<td>1.00</td>
</tr>
<tr>
<td>2000</td>
<td>0.81</td>
<td>0.79</td>
<td>0.83</td>
<td>0.81</td>
</tr>
<tr>
<td>2001</td>
<td>0.87</td>
<td>0.84</td>
<td>0.89</td>
<td>0.87</td>
</tr>
<tr>
<td>2002</td>
<td>0.87</td>
<td>0.85</td>
<td>0.88</td>
<td>0.87</td>
</tr>
<tr>
<td>2003</td>
<td>0.90</td>
<td>0.88</td>
<td>0.92</td>
<td>0.90</td>
</tr>
<tr>
<td><strong>Sex</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>0.90</td>
<td>0.88</td>
<td>0.92</td>
<td>0.90</td>
</tr>
<tr>
<td>Female*</td>
<td>1.00</td>
<td></td>
<td></td>
<td>1.00</td>
</tr>
<tr>
<td><strong>Age (10 year increment)</strong></td>
<td>1.08</td>
<td>1.07</td>
<td>1.09</td>
<td>1.08</td>
</tr>
<tr>
<td><strong>Aggregate Tumour Group</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Breast</td>
<td>1.11</td>
<td>1.06</td>
<td>1.52</td>
<td>1.11</td>
</tr>
<tr>
<td>Colorectal</td>
<td>1.14</td>
<td>1.06</td>
<td>1.21</td>
<td>1.14</td>
</tr>
<tr>
<td>Hematological</td>
<td>0.71</td>
<td>0.67</td>
<td>0.74</td>
<td>0.71</td>
</tr>
<tr>
<td>Lung*</td>
<td>1.00</td>
<td></td>
<td></td>
<td>1.00</td>
</tr>
<tr>
<td>Other</td>
<td>1.05</td>
<td>0.99</td>
<td>1.12</td>
<td>1.05</td>
</tr>
<tr>
<td>Pancreas</td>
<td>1.09</td>
<td>1.02</td>
<td>1.16</td>
<td>1.09</td>
</tr>
<tr>
<td>Prostate</td>
<td>1.21</td>
<td>1.14</td>
<td>1.29</td>
<td>1.21</td>
</tr>
<tr>
<td>Unknown Primary</td>
<td>0.78</td>
<td>0.73</td>
<td>0.84</td>
<td>0.78</td>
</tr>
<tr>
<td><strong>Income Quintile</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low*</td>
<td>1.00</td>
<td></td>
<td></td>
<td>1.00</td>
</tr>
<tr>
<td>Lower Middle</td>
<td>1.03</td>
<td>1.01</td>
<td>1.05</td>
<td>1.03</td>
</tr>
<tr>
<td>Middle</td>
<td>1.10</td>
<td>1.07</td>
<td>1.12</td>
<td>1.10</td>
</tr>
<tr>
<td>Upper Middle</td>
<td>1.15</td>
<td>1.12</td>
<td>1.18</td>
<td>1.15</td>
</tr>
<tr>
<td>Upper</td>
<td>1.21</td>
<td>1.17</td>
<td>1.24</td>
<td>1.20</td>
</tr>
<tr>
<td><strong>Region</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interior</td>
<td>1.19</td>
<td>1.16</td>
<td>1.23</td>
<td>1.18</td>
</tr>
<tr>
<td>Fraser</td>
<td>0.83</td>
<td>0.80</td>
<td>0.87</td>
<td>0.82</td>
</tr>
<tr>
<td>Vancouver Coastal*</td>
<td>1.00</td>
<td></td>
<td></td>
<td>1.00</td>
</tr>
<tr>
<td>Vancouver Island</td>
<td>1.04</td>
<td>0.99</td>
<td>1.08</td>
<td>1.03</td>
</tr>
<tr>
<td>Northern</td>
<td>1.28</td>
<td>1.23</td>
<td>1.33</td>
<td>1.26</td>
</tr>
<tr>
<td><strong>Survival (2 year increment)</strong></td>
<td>1.03</td>
<td>1.03</td>
<td>1.03</td>
<td>1.03</td>
</tr>
<tr>
<td><strong>Community Size</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt; 10,000</td>
<td>1.10</td>
<td>1.06</td>
<td>1.14</td>
<td>1.10</td>
</tr>
<tr>
<td>10,000 - 99,999</td>
<td>1.32</td>
<td>1.28</td>
<td>1.35</td>
<td>1.31</td>
</tr>
<tr>
<td>100,000 - 499,999</td>
<td>1.74</td>
<td>1.69</td>
<td>1.80</td>
<td>1.74</td>
</tr>
<tr>
<td>500,000 - 1,249,999 **</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>≥1,250,000*</td>
<td>1.00</td>
<td></td>
<td></td>
<td>1.00</td>
</tr>
<tr>
<td><strong>Ethnic Origins</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Heterogeneous*</td>
<td>1.00</td>
<td></td>
<td></td>
<td>1.00</td>
</tr>
</tbody>
</table>
Aboriginal Presence  | 0.93 | 0.79 | 1.06 | 0.86 | 0.74 | 0.98
Chinese Presence   | 1.00 | 0.93 | 1.07 | 1.02 | 0.95 | 1.09
British Isles Presence | 0.77 | 0.72 | 0.82 | 0.77 | 0.72 | 0.82
East Indian Presence | 1.00 | 0.88 | 1.13 | 0.92 | 0.81 | 1.02
Americas and Europe | 0.74 | 0.70 | 0.78 | 0.74 | 0.70 | 0.78

Mother Tongue

| Punjabi Presence | 0.95 | 0.86 | 1.04 | 0.77 | 0.71 | 0.83
| Dominant English* | 1.00 | 1.00 |
| Heterogeneous with Dominant English | 1.12 | 0.95 | 1.29 | 1.12 | 0.98 | 1.26
| Chinese Presence | 0.86 | 0.73 | 0.99 | 0.87 | 0.76 | 0.97
| Strong English Presence | 1.02 | 0.88 | 1.16 | 1.00 | 0.89 | 1.12

Religion

| Heterogeneous with Protestant and No Religion Presence* | 1.00 |
| Catholic Presence | 0.80 | 0.63 | 0.97 |
| Protestant Presence | 0.97 | 0.95 | 0.99 |
| Sikh Presence | 0.73 | 0.69 | 0.78 |
| No Religion Presence | 1.06 | 1.05 | 1.08 |
| Heterogeneous | 0.96 | 0.94 | 0.97 |

Note: OR = Odds Ratio, CI=Confidence Interval
†Adjusted for year of death, sex, age, aggregate tumour group, income, region, survival, community size, ethnic origins, mother tongue, and religion
‡Adjusted for year of death, sex, age, aggregate tumour group, income, region, survival, community size, ethnic origins, and mother tongue
* Reference Category
** There were no BC cities in this category

It is difficult to observe an overall trend in Figure 5-10, but patients were more likely to die in hospital from 2000 to 2003 when compared to 1999 (Table 5-10). This substantiates the observation that a change in the location of death occurred from 61.8% in hospital in 1999 to 66.4% in hospital in 2000 (Table 5-6 and Figure 5-10).

Table 5-10 shows the adjusted odds ratio (OR) for both the full and parsimonious logistic regression models. After all other significant factors have been taken into consideration; people who were more likely to die out of hospital were females, those who live in neighbourhoods in a higher income per person equivalent quintile, those who live in either the Interior or Northern Health Authorities, and those who live in communities with less than 500,000 people. These people also died from breast, colorectal, pancreatic or prostate cancer. It was found that older people and those with longer survival times were more likely to die out of the hospital.

The parsimonious model included two of the three indicators of culture. Both ethnic origins and mother tongue remained. Although the indicator for religion did not remain in the model, it did contribute information when considering the AUC.
In the full model, it was found that people associated with the British Isles Presence ethnic origin cluster (OR 0.77, 95% CI 0.72-0.82) and the Americas and Europe cluster (OR 0.74, 95% CI 0.70-0.78) were less likely to die out of hospital when compared against the Heterogeneous cluster. All other ethnic origin clusters were similar to the Heterogeneous cluster. The parsimonious model found that people associated with the Aboriginal Presence cluster (OR 0.86, 95% CI 0.74-0.98), the British Isles Presence ethnic origin cluster (OR 0.77, 95% CI 0.72-0.82) and the Americas and Europe cluster (OR 0.74, 95% CI 0.70-0.78) were less likely to die out of hospital when compared against the Heterogeneous cluster. The Chinese Presence and the East Indian Presence clusters were similar to the Heterogeneous cluster.

When considering the mother tongue, the full model found that people associated with the Chinese Presence (OR 0.86, 95% CI 0.73-0.99) cluster were less likely to die out of hospital than the reference cluster, Dominant English. When the parsimonious model was considered, both the Chinese Presence cluster (OR 0.87, 95% CI 0.76-0.97) and the Punjabi Presence (OR 0.77, 95% CI 0.71-0.83) cluster were less likely to die out of hospital than the reference cluster.

The consideration of religion is restricted to the full model. People who were associated with the Catholic Presence (OR 0.80, 95% CI 0.63-0.97), Protestant Presence (OR 0.97, 95% CI 0.95-0.99), Sikh Presence (OR 0.73, 95% CI 0.69-0.78) and the Heterogeneous clusters (OR 0.96, 95% CI 0.94-0.97) were less likely to die out of hospital than the reference group Heterogeneous with Protestant and No Religion Presence.

**Classification Tree**

By utilizing (8.3), the optimal classification tree was fit to the data. The realized model is interesting in that all predictors are used as inputs, yet a parsimonious model results without the selection procedures so often used in more familiar regression models. Appendix F presents the graph of the tree. The variables which are useful for classification are:

- Survival,
- Region of usual residence,
- Age,
- City size,
- Mother Tongue,
- Aggregate Tumour Group,
- Year of death, and
- Sex.

The first split is based upon the length of survival. A person who has a survival of less than 31.5 days is classified as dying in the hospital. If the contrary is true, then further decisions are necessary for deciding the location of death. The next decision is based upon the region of usual residence. If the person lives in the Fraser Health Authority or Vancouver Coastal Health Authority then the next split is based upon the age. If the person lives in the Vancouver Island, Interior, or Northern Health Authorities, then the next split is based upon the city size. The tree continues in this fashion.
Although it does not have the immediate clinical appeal that logistic regression has, it can provide interesting insights into the structure of the data and the utility of a sub-set of predictors. Interactions which are omitted from the logistic regression are suggested in the classification tree. An example of a potential interaction is that the combined effect of survival time, region of residence, age, and mother tongue are influential in the location of death. The age cut point is consistently over 70 years of age. People younger than 70 tend to die in the hospital as all but one age cut point sends all people less than the expressed age to a terminal node of death in the hospital. A pure terminal node has all members of one class, for example a pure out of hospital terminal node would have all members dying out of hospital. There are no such nodes in the tree. The purest node has 69% of the members dying in the hospital. The purest terminal node for people dying out of the hospital is 65%. This indicates that the optimal classification tree is unable to find hyper-rectangles in the predictor variable space that are homogeneous with respect to the response.

Discussion
Two main points of discussion emerge from this research. The first pertains to the performance of the classifiers. The second pertains to the health-related insights gained from this investigation.

Classifier Performance
Linear discriminant analysis, logistic regression, neural networks, classification trees, and nearest neighbours were explored for their utility in end-of-life research. In terms of overall performance, the neural network performed the best (AUC=.6180) and nearest neighbours performed the worst (AUC=.5620) with the remaining having similar results (Table 5-8). The minimum misclassification rates were only slightly lower than 0.385 which would occur if all the study subjects were classified as dying in the hospital. Due to this, the misclassification rate is not a good measure of performance.

In terms of the ability of a classifier quickly to find the people who died out of the hospital, all the classifiers found approximately the same number of patients who died out of a hospital, but they did not all reach the expected hit curve.

The hit curve results suggest that all the classifiers find patients who died out of a hospital with similar speed. The difference is in how they perform with respect to their expected performance. Neural networks and nearest neighbours had an expectation of approximately 3,000 hits over the first 5,000 patients. The other classifiers had expectations of approximately 2,500 hits over the first 5,000 patients. It appears as if NN and KNN may have the potential for better performance. This may be due to the flexibility of these classifiers.

The neural network performed the best in terms of the AUC and performed as well as the other classifiers in its ability to find patients who died out of a hospital. Although it represents a type of black box methodology in that it is difficult to extract information
pertaining to the utility of the predictor variables (compared against logistic regression), it does yield important insights.

Neural networks is a flexible classifier in that it can model complex relationships among the explanatory variables. The strength of its performance suggests that there are important non-linear relationships involved in the prediction of the location of death. If a classifier such as logistic regression is chosen for its clinical utility, then interaction terms and higher order terms must be included in the model. The exclusion of these relationships could impair a classifier's performance. Even by selecting a set of interactions and higher order terms for logistic regression, the best set of these complex relationships may not be found. Neural networks has the ability to find these complex relationships and exploit them. It is this natural ability of neural networks that makes it worthy of future developments to increase its utility for health research.

Nearest neighbours (KNN), another black box approach was the worst performing classifier. With the lowest AUC and misclassification rate is was the least useful in accurately classifying the patients, yet it does have the ability to find patients as quickly as the other classifiers. This performance may be due to not having enough information. If extra information was included, this new information would provide good performance. Another possibility is that there is too much information. There could be variables included in the analysis that add extra noise. This noise causes the poor performance. A smaller, well-defined set of explanatory variables is needed. If the best set of predictors was found that substantially improved the performance of KNN, then valuable information would be gained as to which explanatory variables are important for understanding where cancer patients die.

Linear discriminant analysis, logistic regression and classification trees performed similarly with respect to the AUC and the hit curves. Although there were differences for each classifier with respect to how it performed against its expected performance, there was little difference in the actual performance of the hit curves.

Logistic regression, although not yielding the best performance with respect to the AUC has the advantage that a parsimonious model is not difficult to find and it can be interpreted with relative ease. Although the parsimonious model is typically the preferred model, it has a statistically significant lower AUC than the full model. The parsimonious model suggests that the cultural indicator of religion is not significant, yet the AUC implies that the cultural indicator of religion does add enough information to improve the overall performance of the classifier.

This ambiguity as to the utility of the indicator of religion may suggest that the indicator itself is measuring more than just religion. There may be redundant information in the indicator of religion, thus it is excluded from the parsimonious logistic regression model, but there is still useful information in the indicator hence the higher AUC for the full logistic model. Investigation into the relationship between the derived indicator of religion and each of the explanatory variables would be needed to ferret out the peculiarities of this measure.
By looking at the similarities of the two logistic regression models immediate insights can be gained. The utility of the predictor variables of year of death, age, sex, region, income, aggregate tumour group, city size, survival, ethnic origins, and mother tongue as predictors of the location of death is reinforced.

Although the classification tree did not perform as well as logistic regression with respect to the hit curve, it did have the second highest AUC. A parsimonious set of predictors was identified by the algorithm: year of death, age, sex, region, aggregate tumour group, city size, survival, and mother tongue. This is a subset of the predictors shared by the logistic models. Only income and ethnic origins were excluded. This validates the relevance of these predictors for understanding where people die.

With the need for good classification and the ability to extract a set of meaningful predictors, both the logistic regression and classification tree classifiers are reasonable choices for end-of-life research at this point in time. Logistic regression is not in the same performance group as neural networks (Table 5-8), but it does have the clinical appeal that the neural network currently lacks.

The classification tree bridges the group to which logistic regression belongs and the group to which neural networks belongs and therefore presents an intriguing alternative to logistic regression. It served to validate the core set of predictors for the location of death in end-of-life research. It yields a useful graph which can help the overall understanding of how the predictors may relate to one another. Also, it gives insight into the complex interplay between demographic, cultural and geographic predictors, suggesting interactions that the researcher may choose to include in a logistic regression model.

The use of data mining techniques has been insightful. The different techniques have provided insights that are valuable in uncovering the predictors for the location of death. Neural networks emphasises that necessity for a more complex relationship between the explanatory variables. Classification trees supports this by implying certain interactions. Nearest neighbours suggests that the best set of predictors has not yet been found. By incorporating these findings into future models we may be able to further uncover the story behind the use of health services at the end of a patient’s life.

**Health Related Insights**

We found a shift in the place of death among the people with cancer over the study period, from out of hospital to in hospital. This shift in the location of death may be related to legislative policy or in a change in how health services are delivered in British Columbia. From 1999 to 2000 the proportion of deaths out of the hospital decreased from 38.1% to 33.6%. The proportion of deaths out of hospital increased to 36.5%, but still remained below the levels in 1999.

It was found that women, people who lived in neighbourhoods in a higher income per person equivalent quintile, people in the Interior or Northern Health Authorities, and people in communities with less than 500,000 inhabitants were more likely to die out of
the hospital. As well, it was found that people with breast, colorectal, pancreatic or prostate cancer were more likely to die out of the hospital. With an increase of age by ten years, the odds of dying out of the hospital were found to be 1.08 and an increase of survival by two years had an associated odds ratio of 1.03.

People associated with the ethnic origins clusters of Aboriginal Presence, the British Isles Presence and the Americas and Europe were less likely to die out of hospital. People associated with the mother tongue clusters of Chinese Presence and Punjabi Presence were less likely to die out of the hospital.

This result between the indicator for ethnic origins and the indicator for mother tongue was unexpected. This could be a result of how people self-report their ethnic origin in the census. A person who had just come to Canada from mainland China and a person who was the grandchild of a person who came from mainland China can both report Chinese as their ethnic origin, but they may report different mother tongues. If the grandchild reports English as the mother tongue, then some adaptation to the surrounding British Columbia context has occurred. A person’s language may be a better measure than ethnicity to identify vulnerable populations.

It was found that women are more likely than men to die out of hospital. This is congruent with Canadian and Italian research, yet there is conflicting evidence as some studies find that either sex is not included in the final model or women are less likely to die out of hospital. The particular results concerning gender may be related to the regional socio-demographic and economic structure, regional emphasis upon sex-specific cancers, geographic differences or even differences in how countries finance health care and related services.

With a ten year increase in age, the odds ratio was found to be 1.08, thus indicating a slight, yet significant increase in the likelihood of dying out of the hospital with increasing age. In Nova Scotia, it was found that patients 75 years and older were more likely to die out of the hospital. In Manitoba, it is suggested that individuals aged 45 and older who have died in the hospital were more likely to have died due to cancer and respiratory diseases. These findings concur with those of Constantini et al. (2000) and Bruera et al. (2002/2003), yet Tang (2001) finds that the relationship between age and the location of death is as yet unsettled.

Focusing upon the Canadian context, BC is reported to have one of the oldest populations. It has a median age of 38.4 years. Nova Scotia also has an older population as it has a median age of 38.8 years. In contrast, the Prairie Provinces were the youngest. Manitoba had a median age of 36.8, yet it had the second highest proportion of seniors aged 65 and over (14%) following Saskatchewan (15%). Manitoba saw the largest increase in median age over the last decade due to substantial increases in the number of elderly people.

In BC, the aging population is not a trivial matter. BC saw a 49% increase in the 45 to 64 age group, a 27% increase in the 70 to 79 age group, and a 54% increase in the 80 and over group between the 1996 and 2001 censuses. Of the 25 oldest municipalities in...
Canada, 13 were found in BC. The oldest BC community was Qualicum Beach (Median age: 58.1 years) followed by White Rock (Median age: 50.9 years). Victoria was the second oldest census metropolitan area with a median age of 41.0 years. Vancouver's median age was 37.4 years. The Fraser Valley and Vancouver Island saw their median age increase from 36.0 years in 1996 to 38.1 years in 2001.

With an aging population a better understanding of how older people use health services is needed. If these people are dying out of the hospital by choice, then health service delivery must evolve to meet the needs of a growing population. With such an evolution of service, congruence between the preference for the place of death and the actual place of death can be achieved.

Income was found to have a statistically significant contribution to the prediction of the location of death. This stands in contrast to the Nova Scotia research where income did not contribute additional information when all the other predictors were in the logistic regression equation. This contrast of the role of income in Canadian health outcomes is a familiar one. Research, utilizing income as a socioeconomic measure, in Toronto and Nova Scotia suggests that there is no connection between income and health outcomes. Roos (2004) and Veugelers (2001) imply that the lack of association is due to Canada's universal health care system. Yet research in Vancouver implies connections between socioeconomic measures such as income and health outcomes. This inconsistency could be related to differences in the delivery of health care services or in the differences in coverage for additional services between provinces.

In contrast to the Nova Scotia results, the aggregate tumour group was found to be predictive of the location of death. Those patients with breast, colorectal, pancreatic or prostate cancer were more likely to die out of the hospital compared with patients who had lung cancer. In Italy, the cancer site was found to contribute to the prediction of the location of death. People with cancer in the upper area of the digestive tract, in the female genital organs, haemopoietic system and of a non-specific nature were more likely to die in the hospital.

Patients who reside in the Fraser Health Authority are less likely to die out of hospital when compared with the region comprising the Vancouver Coastal Health Authority. The Fraser Health Authority and the Vancouver Coastal Health Authority capture much of the BC population (61.4%) as well as many of the major urban centres. Both of these health authorities have the lowest number of acute care days (Fraser: 503 days/1000 pop, Vancouver Coastal: 496 days/1000 pop).

Patients who reside in the Northern and Interior Health Authorities are more likely to die out of the hospital. The Northern Health Authority has few urban centers and is the least urbanized. The Interior has mid-range urbanization rates with approximately 700,000 people covering a quarter of British Columbia. Interestingly, the Northern Health Authority has the highest number of acute care days (734 days/1000 pop), followed by the Interior (609 days/1000 pop).
In BC, the regions with higher odds of dying out of hospital are those regions where the length of stay in acute care facilities is longest. This may be related to decisions made by people when required to travel large distances for treatment. For patients outside the Fraser and Vancouver Health Authorities, hospitalization may occur when other local options have been exhausted. This unexpected relationship poses a challenge for BC researchers as this relationship may be intimately connected to utilization, geographic distribution and the administration of health services as well as the influence of other support systems.

It was found that people who lived in communities with less than 500,000 individuals were more likely to die out of hospital. A gradient was observed in that smaller cities had a smaller likelihood of this event than larger cities.

The ethnic origin and mother tongue cultural indicators were relevant to the classification of the location of death. People associated with the ethnic origins clusters of Aboriginal Presence, the British Isles Presence, and the Americas and Europe were more likely to die in a hospital than the other ethnic origin clusters. Bruera et al. (2002) found that Black patients in the Houston area were more likely to die in a hospital than Hispanics and non-Hispanic whites. It is reasonable to surmise that immigrants would have a different ethnic origin than people who identify themselves as being from the country in which they live. In Italy, it was found that immigrants were more likely to die in a hospital than non-immigrants. In Nova Scotia, it was found that immigrants were more likely to die out of the hospital. Unfortunately, these results do not clarify the matter. Factors such as the underlying structure of the health care system may be influencing the outcomes. As well, the immigrant or ethnic composition of the country may contribute to the outcome. Even by narrowing the perspective on the Nova Scotia research, the cultural indicators of ethnicity were not included in the parsimonious model leaving only immigrant status. If the above assumption about immigrants is correct, then there is some level of congruence between the BC and NS results and an implication that there may be populations that are marginalized from the health services network.

People associated with the mother tongue clusters of Chinese Presence and Punjabi Presence were less likely to die out of the hospital than the Dominant English Cluster. In Nova Scotia, language did contribute to the prediction of location of death, but the odds ratio did not suggest any differences between the non-official languages community and the official languages community. By identifying particular language communities where the likelihood of dying out of a hospital is greater than in the community associated with the lingua franca of the country, vulnerable communities may be identified.

Our data was restricted to the administrative database at the British Columbia Cancer Agency and the British Columbia Vital Statistics. In the future, it is hoped that more administrative databases pertaining to palliative care, hospital separations and hospice care will be added to the existing database. It is believed that with this information a better understanding of how palliative care and primary care services are used by cancer patients at the end of their life may be obtained.
Limitations

Although the use of administrative databases in health research represents an efficient means by which research can be executed, it does bring limitations to the study. The number of fields is limited as the primary intent of the databases was not for research. The VS and BCCA databases do not collect socioeconomic and demographic information. Due to this, the income, ethnic origin, mother tongue, and religion measures used in this study were obtained from dissemination area census data. The use of such data does have ramifications upon the utility of the model in the form of the ecological fallacy. Researchers are confined to using the coding system established by the database providers. These codes may change over the years thus further confounding their utility. Even with these and the aforementioned limitations, administrative databases do represent an efficient means, in terms of cost and time by which population-based health data can be obtained. This allows for a rough sketch of population health to be made with relative ease.

The response, death out of the hospital may be too vague. These deaths are not a hospital death, but they may occur in an institution such as a nursing home or a palliative care facility. Although technically not a hospital, locations such as these represent a mixed environment as medical services and place of usual residence are intermingled. What is needed is clearer information on the location of death so that pseudo-residential locations can be identified.

The problem of re-used postal codes presents a limitation in that only those patients who had a recorded address could undergo postal code verification. When a postal code was identified as correctly identifying an institution, then all instances of that postal code were taken as correct. Unfortunately patients with an old postal code that was not associated with an institution could not be followed up if they did not have a recorded address. Also, the address was taken from the BCCR and may not represent the place of usual residence. Although a limitation, this was restricted to 207 patients in the data set.

This research only describes where people have died. As administrative databases were used, no comment can be made on the congruence between the location of death and the patient’s preferred place of death. Understanding the reasons for dying in a certain location are also eclipsed.

As this paper represents the first departure from the research begun by the Nova Scotia team, only minor changes were made to the original methodology and the original set of variables. This departure has allowed for the BC team to consider methodological approaches and variables that may not be conducive to the research in Nova Scotia. In doing so, the limitations of working within the Nova Scotia methodological framework have been lessened.
Conclusion

Classification Techniques
Using administrative databases for population health research requires an alternate perspective from that commonly used. The studies which use such data are observational and should use statistical techniques specifically developed for such purposes. Data mining provides a conceptual framework which is appropriate for such research.

As this investigation was a retrospective study utilizing health-related administrative databases, the use of data mining techniques was appropriate. Linear discriminant analysis, logistic regression, neural networks, classification trees and nearest neighbours were used to investigate the utility of a data mining perspective for end-of-life research. Neural networks and classification trees suggests that complex relationships may be necessary to better understand where people die. The inclusion of interactions and higher order explanatory terms in logistic regression may uncover pieces of a complex story. Nearest neighbours suggests that the set of predictors is not the best set. Logistic regression confirms this as the role of religion in classification is uncertain as it is not in the parsimonious model, yet the full model has a higher AUC. Further investigation into the set of predictors is warranted.

As different classifiers have contributed to the overall understanding, we can say that these techniques are valuable to end-of-life research. As future research allows higher epidemiological utility for these classifiers, it is anticipated that the relevance to this type of research will increase.

Population Health
Given the current limitations for the clinical utility for many of the classifiers used in this study, many of the health-related observations were based upon the results of the logistic regression. Even though this is the case, insights can be gained.

Socioeconomic, demographic and geographic gradients can be seen. Income, language, ethnicity, region, and sex were predictors for the place of death. The inclusion of community size and cultural indicators presents novel insights that can be used to gain an understanding of the location and constitution of potential vulnerable populations. The set of predictors from the logistic regression model in conjunction with the insights from the classification tree suggest that a complex interplay of factors may underlie marginalization from health services. As this investigation only used the location of death, the inclusion of palliative care information and hospital separations may yield further insights into the constitution of populations who are not accessing health services in a manner that could enhance their end-of-life experience. These insights will be necessary to develop effective health policy and service delivery programs for all.

General Conclusion
The year of death, sex, age, region of residence, tumour group, survival time, city size and mother tongue are predictive in classification trees and logistic regression. These predictors constitute the core set of relevant predictors for understanding where cancer
patients die. This set of predictors is derived from the classification tree and logistic regression results as they both yield parsimonious models. Although logistic regression was not the best performing classifier it does provide some immediate advantages for epidemiologists. As it allows researchers to find a parsimonious set of predictors and determine risk, it has a clear advantage from an epidemiological point of view. Even though it has some evocative advantages, these benefits should not eclipse further investigations into the utility of other classification methods. Further research along this path should continue to assess the performance of logistic regression against other classification methods, and should seek further theoretical development of the other classifiers so that they too share in the epidemiological utility of logistic regression. For now, ease of use, widespread implementation in statistical software, and epidemiological utility make logistic regression the preferred classifier, yet these things should not be the sole determinants for choosing a statistical methodology. The choice of modelling technique should not be predicated upon ease of use, but upon the performance of the classifier to meet both statistical and clinical objectives.
Appendices
Appendix A: The DCI/DCO process in British Columbia

Abbreviations:
VS: British Columbia Vital Statistics
DCI: Death Certificate Identified
DCO: Death Certificate Only

This flow chart represents the current process for identifying which cases are classified as death certificate identified (DCI) or death certificate only (DCO).
Method of diagnosis is DCO and death date = diagnosis date.
1. Prior information can include:
   a. pathology report
   b. doctor assuming that the patient did have cancer but chose not to confirm through invasive testing
## Appendix B: ICD Groupings

Table 5-11: Tumour groups and associated ICD-9, ICD-10, and ICD-0 codes.

<table>
<thead>
<tr>
<th>Tumour Group</th>
<th>ICD-9</th>
<th>ICD-10</th>
<th>ICD-0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bladder</td>
<td>188</td>
<td>C67</td>
<td>C67</td>
</tr>
<tr>
<td>Brain</td>
<td>191-192</td>
<td>C70-C72</td>
<td>C70-C72</td>
</tr>
<tr>
<td>Breast</td>
<td>174-175</td>
<td>C50</td>
<td>C50</td>
</tr>
<tr>
<td>Cervix</td>
<td>180</td>
<td>C53</td>
<td>C53</td>
</tr>
<tr>
<td>Colorectal</td>
<td>153-154</td>
<td>C18-C21</td>
<td>C18-C21</td>
</tr>
<tr>
<td>Endocrine Glands</td>
<td>193-194</td>
<td>C73-C75, C37</td>
<td>C37-C73-C75</td>
</tr>
<tr>
<td>Esophagus</td>
<td>150</td>
<td>C15</td>
<td>C15</td>
</tr>
<tr>
<td>Hematologic</td>
<td>200-208</td>
<td>C81-C85, C88, C90-C95</td>
<td>C42, C77</td>
</tr>
<tr>
<td>Kidney</td>
<td>189</td>
<td>C68, C64-C66</td>
<td>C64-C66, C68</td>
</tr>
<tr>
<td>Larynx</td>
<td>161</td>
<td>C32</td>
<td>C32</td>
</tr>
<tr>
<td>Liver</td>
<td>155</td>
<td>C22</td>
<td>C22</td>
</tr>
<tr>
<td>Lung</td>
<td>162</td>
<td>C33-C34</td>
<td>C33-C34</td>
</tr>
<tr>
<td>Oral</td>
<td>141-149</td>
<td>C00-C14</td>
<td>C00-C14</td>
</tr>
<tr>
<td>Other</td>
<td>195, 190, 170-173</td>
<td>C96-97, C76, C69, C49, C46-C47, C43-C44, C40-C41</td>
<td>C40-C41, C44, C47, C49, C69, C76</td>
</tr>
<tr>
<td>Other Digestive</td>
<td>158-159, 156</td>
<td>C48, C26, C23-C24, C17</td>
<td>C17, C23-C24, C26, C48</td>
</tr>
<tr>
<td>Other Genitourinary</td>
<td>187</td>
<td>C63, C60</td>
<td>C60, C63</td>
</tr>
<tr>
<td>Other Gynecological</td>
<td>183-184, 181</td>
<td>C56-C58, C51-C52</td>
<td>C51-C52, C56-C58</td>
</tr>
<tr>
<td>Other Thoracic and Neck</td>
<td>163-165, 160</td>
<td>C45, C38-C39, C30-C31</td>
<td>C30-C31, C38-C39</td>
</tr>
<tr>
<td>Pancreas</td>
<td>157</td>
<td>C25</td>
<td>C25</td>
</tr>
<tr>
<td>Prostate</td>
<td>185</td>
<td>C61</td>
<td>C61</td>
</tr>
<tr>
<td>Stomach</td>
<td>151</td>
<td>C16</td>
<td>C16</td>
</tr>
<tr>
<td>Testis</td>
<td>186</td>
<td>C62</td>
<td>C62</td>
</tr>
<tr>
<td>Unknown Primary</td>
<td>196-199</td>
<td>C77-C80</td>
<td>C80</td>
</tr>
<tr>
<td>Uterus</td>
<td>182, 179</td>
<td>C54-C55</td>
<td>C54-C55</td>
</tr>
</tbody>
</table>
Appendix C: British Columbia Health Authorities

There are five geographic health authorities in British Columbia.

1. Interior
2. Fraser
3. Vancouver Coastal
4. Vancouver Island
5. Northern

The sixth health authority is provincial. The Provincial Health Services Authority’s (PHSA) primary role is to ensure access to network of specialized health care services. It operates agencies such as BC Children's Hospital, the BC Transplant Society, and Riverview Hospital. It is responsible for specialized services such as chest surgery and cardiac services.

Figure 5-18: British Columbia Health Authorities
Appendix D: Data Quality

The agreement between the diagnosis of cancer and the cancer cause of death has been proposed as one measure of data quality.39

Figure 5-19: Agreement between the diagnosis of cancer and the cancer cause of death.
Appendix E: Misclassification Curves

Figure 5-20: Misclassification curve for linear discriminant analysis model.
Figure 5-21: Misclassification curve for the full logistic regression model
Figure 5-22: Misclassification curve for the parsimonious logistic regression model
Misclassification Curve

Figure 5-23: Misclassification curve for neural networks model
Figure 5-24: Misclassification curve for classification trees model
Figure 5-25: Misclassification curve for the nearest neighbours model
Appendix F: Classification Tree

How to read the classification tree:

- If the condition is true then proceed to the left. If it is false proceed to the right.
- Terminal nodes are indicated by an oval
- The probability under the terminal nodes is the probability of dying out of hospital given the data (1=death out of the hospital)
- Survival and age were taken as continuous variables
Aggregate Tumour Group: Haematological, Unknown Primary

Age < 87.5 years

In Hospital: $p(l|x) = 0.37$
Out of Hospital: $p(l|x) = 0.61$

Survival < 63.5 days

In Hospital: $p(l|x) = 0.39$


Sex: Male

Out of Hospital: $p(l|x) = 0.60$

Aggregate Tumour Group: Breast, Colorectal, Lung, Other, Prostate

Age < 72.5 years (Continued on following page)

In Hospital: $p(l|x) = 0.47$
Out of Hospital: $p(l|x) = 0.62$
Age < 72.5 years

Aggregate Tumour Group: Breast, Lung, Other, Pancreas

Survival ≥116 days

In Hospital

Out of Hospital

Out of Hospital

p(l|e) = 0.60

Survival >116 days

p(l|e) = 0.65

p(l|e) = 0.45

p(l|e) = 0.60

In Hospital

Out of Hospital
References


172


70. Martin, S. Canada Post [Personal communication 2004 April 20]


75. Zamar, R. [Personal Communication 2005 August 15]


Chapter 6: Conclusion

Conclusion
Administrative databases represent an efficient means, in terms of cost and time, by which to gain a general understanding of how cancer patients in British Columbia used end-of-life services. Although efficient, the British Columbia Cancer Registry (BCCR) and the British Columbia Vital Statistics (VS) databases do not contain socioeconomic (SES) and cultural information. In order to have such information census data were used. Dissemination area or enumeration area level data were obtained from Census Canada and associated with each patient in the study. Thus, by design we had an observational, retrospective, ecological study based upon a very large dataset. In addition, this study aimed at replicating the methodology used by the NS team on their study about trends in place of death and developing a BC population-based database for further research into care service use at EOL. Although the path uncovered many challenges, this thesis served to:

- assess the necessity and appropriateness of ecological research in end-of-life studies,
- characterize the trends in the place of death from 1997 to 2003 for adult British Columbians who died of cancer,
- explore a novel multivariate approach for the development of census-based cultural indicators, and
- assess and compare the utility of five data mining techniques for end-of-life research.

At this stage of the research, the ecological model is both necessary and appropriate. From the databases, individual level information was obtained. This included predictor variables such as the year of death, age, survival time from diagnosis, and the sex of the patient. To these predictors, aggregate census data were associated. The outcome was the location of death. This was a binary outcome variable indicating either “death in the hospital” or “out of the hospital”. Each patient had both individual-level and aggregate-level predictors and an individual-level outcome. This means that this is a mixed-level study and is inherently ecological (Chapter 2).

In using census-based indicators of culture as predictor variables, it is necessary to have a clear understanding how the variables have been developed and are to be used. Currently, the culture indicators are prescribed by the researcher, based upon cut-points which have been determined by a univariate exploration of the data. This approach presupposes a cultural definition that makes it difficult to determine how this variable fits into the Susser-Macintyre-Duncan classification scheme (Chapter 2).

A descriptive, data-driven approach was proposed as an alternative means of constructing a derived measure for the underlying cultural construct. By utilizing unsupervised learning techniques, such as principal components and k-means, the data were organized into groups. The goal of this organization is to have small within-group variation and larger between-group variation. The advantage is that a probabilistic
distribution can be associated with each group and is the best description of each group. An evocative label is given in order to facilitate communication. The groups are based upon similarities, thus the groups represent some underlying cultural construct, but they are clearly derived. This approach yields Susser-type contextual variables that are understood and used as such. In terms of Susser-Macintyre-Duncan labels, it is anticipated that these will function as contextual-compositional variables, but a contextual-contextual realisation is not precluded.

What is sought is an understanding of how underlying cultural constructs affect health services utilization at the end of life. By utilizing a descriptive development of the ecological indicators of culture, the use of census-based SES and cultural data is appropriate. As well, we are obligated to use such data as it is not available in the BCCR and VS databases. The ecological model is both appropriate and necessary for this thesis.

Having established the appropriateness and necessity of an ecological approach, an initial point of reference for end-of-life health service utilization by cancer patients in British Columbia was needed. As this research was part of a collaborative research, the seminal Nova Scotia research was reproduced. The exercise of reproducing this work enabled us to use the Nova Scotia methodology, develop a critique of it, and then develop means by which the critiques could be addressed.

The study itself (Chapter 3) yielded much needed insights about how cancer patients in British Columbia utilize health services. A shift has occurred in the place of death for BC cancer patients between 1997 and 2003, with the peak of “death our of hospital” occurring in 1999. While the peak may reflect an implementation of a commitment for home related health care delivery systems, it warrants further investigation into the health policies being implemented in the 1998/1999 time period. We found that the proportion of out-of-hospital deaths among adults with cancer increased over the study period. The proportion rose from 25.9% in 1997 to 36.5% in 2003. It was found that women, patients over 45 years of age, patients in neighbourhoods in the Middle to Upper IPPE quintiles, patients with colorectal cancer, patients who lived outside the Fraser Valley, and those who have survival times greater than 60 days were more likely to die outside of a hospital.

Having established a point of reference, the development of the indicators of culture was necessary for the next stage of research. By integrating the critique levied in the first manuscript with the Nova Scotia methodology used in establishing our first look at how health services are used in British Columbia, a new indicator was developed.

Descriptive indicators of culture were developed for the 2001 census categories of ethnic origin, mother tongue and religion (Chapter 4). The pre-processing of the data was limited to the conceptual grouping of particular cultural features. Principal Component Analysis (PCA) was used to further organize the data and reduce the amount of noise in the data. The dissemination areas were grouped based upon the principal component variables. This approach was the first step in developing a novel approach for the construction of culture indicators.
By utilizing the multivariate distribution of the census data, this approach was able to identify cultural groups that are known to exist in British Columbia, such as the Indo-Canadian, Chinese, and Aboriginal cultural groups. As well, this approach was able to determine clusters that reflected the known geographic distributions of these cultural groups.

The labelling of the groups with meaningful titles was another departure from the current methodology. Currently, the label precedes the formation of the groups. This suggests that the concept is imposed upon the groups a priori the construction. This new approach resists any labelling until the end of the process, thus labels are assigned based upon the composition of the groups resulting in contextual-compositional variables, contextual-contextual variables, or both. The distinction in when the label is assigned reflects more the investigator's understanding of culture. The current methodology suggests that culture fits into a predetermined structure or identity whereas this new approach suggests that culture is defined by a complex web of relationships. Grouping dissemination areas based upon these webs of interconnectivity, a shift from modern to post-modern understandings of culture can emerge.

This construction of indicators of culture seeks to appropriate the Canadian ideal of multiculturalism. This approach allows for the identities chosen by people to be integrated into health research while minimizing the philosophical ramifications of imposing a cultural definition upon geographically-grouped people. By seeking innovative ways to quantify culture, researchers will move towards a preservation of the cultural pluralism of Canada.

The new indicators of culture were integrated into a study which compared the utility of five data mining classifiers for end-of-life health research (Chapter 5). Linear discriminant analysis, logistic regression, neural networks, classification trees, and nearest neighbours were applied to the problem of determining the location of death for British Columbian cancer patients.

Neural networks and classification trees comprised the group of best performing classifiers. Logistic regression and linear discriminant analysis were in the group with the next best performance and nearest neighbours performed the worst. The classification tree and logistic regression were able to provide parsimonious models and insights relevant to epidemiologic research. Given the current limitations for neural networks, it was unable to give valuable insights into the relative utility of the predictors. Although it performed better than logistic regression, the clinical utility (i.e. the interpretability of the results in terms of odds ratio) is the key disadvantage to its current use in end-of-life research. This is a common disadvantage for many of the classifiers which must be overcome.

By considering a data mining perspective, end-of-life research has been introduced to an unfamiliar set of statistical methods. This exposure has yielded useful insights. Logistic regression, as with all the classifiers, performed more poorly than was hoped. Its
inability to make clear distinctions between patients dying in a hospital and those dying out of the hospital has not been considered in the current body of work. This impediment must be taken into consideration when assessing the relevance of the predictors for classification. Also, the inability of logistic regression to outperform the other classifiers suggests that research should be directed into exploring ways to make other data mining techniques more easily used within an epidemiological context. Based on the comparison of results the use of logistic regression may appear as driven by the availability of software and methodological preference. Evidential support must be supplied for any classification technique and if the best classifier for the data is not selected as the modelling strategy then the reasons for such a selection should be clearly expressed. Although logistic regression was not the best classifier, insights could be obtained as it was the most useful for understanding the relevant predictors for determining the location of death for BC cancer patients. It was found that women, people who lived in neighbourhoods in a higher income per person equivalent quintile, people in the Interior or Northern Health Authorities, and people in communities with less than 500,000 inhabitants were more likely to die out of the hospital. As well, it was found that people with breast, colorectal, pancreatic or prostate cancer were more likely to die out of the hospital. With an increase of age by ten years, the odds of dying out of the hospital were found to be 1.08 and an increase of survival by two years had an associated odds ratio of 1.03. People associated with the ethnic origins clusters of Aboriginal Presence, the British Isles Presence and the Americas and Europe were less likely to die out of hospital. People associated with the mother tongue clusters of Chinese Presence and Punjabi Presence were less likely to die out of the hospital.

Through understanding the nature of an ecological study, a novel methodology for the construction of culture indicators was developed. Through the collaboration with a research team in Nova Scotia, a foundation for future research was established. By integrating these two advances for cancer research in BC, a comparison of classification techniques was performed. Although logistic regression was not the best performing classifier, it was the most useful from the perspective of a health researcher. The health related findings from this investigation are evocative as reflecting distinct health service utilization patterns among BC cancer patients.

Strengths
The strengths of this research can be summarized in the following achievements: 1) the development of a BC population based database for EOL research; 2) the characterization of trends in place of death for BC cancer patients over a seven-year period; 3) the use of a novel approach for developing culture indicators; and 4) the assessment and comparison of performance of different data mining techniques into EOL research.

An end-of-life database of British Columbia cancer patients was created as part of this investigation. The creation of this database represents an efficient use of resources in order to gain a understanding of how health services are utilized at the end of life. From this database, study sets are obtained to further investigate research questions relating to end-of-life health service utilization. As well, the constructed database will be used to
link to other health databases such as Pharmacare or Medical Service Plan. This database is central to the continued investigations into issues of access and utilization of health services at the end of life.

This research has established a foundation for further research. Chapter 3 is the first time trends in the place of death for cancer patients in British Columbia have been considered. Chapter 5 expands upon it by including variables such as the city size, ethnic origins, mother tongue and religion. These two papers have given a rough sketch of where British Columbian cancer patients have died in relation to clinical, demographic, socioeconomic and cultural explanatory variables.

A novel approach to the creation of culture indicators was developed. Multivariate techniques were used to create indicators of culture for the census categories of ethnic origin, mother tongue, and religion. The groups that were obtained for each census category were statistically described and then given evocative labels. This approach is a departure from the current methodology which imposes a cultural label upon the census data. By discovering the cultural groups within a census category, this methodology serves as a first step towards addressing concerns about cultural safety.

Chapter 5 gives a sketch of how various classifiers perform with this data. The clinical and statistical utility of logistic regression was readily seen in this context. Although neural networks outperformed logistic regression, logistic regression has a much higher utility for end-of-life health service utilization. By understanding the advantages of logistic regression for health research, new areas of potential classifier development were identified. Also, it was determined that until other classifiers have the same utility for end-of-life research, the clinical utility of logistic regression is evident.

**Limitations**

A couple of limitations apply to this study. First, by design this is an ecological mixed-level study so interpretation of the results should consider the constrains of the design. The predictor variables are individual-level and group-level. The outcome variable is still individual-level.

Second, the use of administrative databases in health research represents an efficient means by which research can be executed, but it does bring limitations to the study. The number of fields is limited as the primary intent of the databases was not for research. The VS and BCCA databases do not collect socioeconomic and demographic information. Due to this, the income, ethnic origin, mother tongue, and religion measures used in this study were obtained from dissemination area census data. The use of such data does have ramifications upon the utility of the model in the form of the ecological fallacy. In addition, researchers are confined to using the coding system established by the database providers. These codes may change over the years thus further confounding their utility (e.g. changes in Place of Death codes used by VS over the study period). However, even with these and the aforementioned limitations, administrative databases do represent an efficient means, in terms of cost and time by which population-based health research can be done.
Third, in terms of variable definitions several limitations are acknowledged. For instance, the outcome, “death out of the hospital” may be too vague. These deaths are not a hospital death, but they may occur in an institution such as a nursing home or a palliative care facility. Although technically not a hospital, locations such as these represent a mixed environment as medical services and place of usual residence are intermingled. What is needed is clearer information on the location of death to improve the accuracy in designating the place of death.

The use of the postal code for an indicator of rural versus urban residence is questionable. The zero in the second position of the postal code indicates delivery from a rural post office. This indicates the point of service, but not the rural or urban feature of the area. Some rural areas are urbanised and suburbanised, yet they have a zero in the second position, thus it may not be the best way to define the indicator. As well, rural route services and sub-urban route services are provided from urban post offices. For example, the entire province of New Brunswick is considered urban as no postal code has a zero in the second position. A definition must be established in order to use this type of indicator with precision and clarity. A measure based upon population size or on population density and distance to regional center may serve as a useful alternative for differentiating geographic areas and merits further exploration.

As this paper is a first step in an inter-provincial comparison, the categories for age and survival from the Nova Scotia research were used. As these categorizations may not be appropriate for the BC population, this research is limited by the conceptual framework implemented for the population in Nova Scotia. A common framework must be established in order to facilitate lucid comparisons.

The problem of re-used postal codes presents a limitation in that only those patients who had a recorded address could undergo postal code verification. When a postal code was identified as correctly identifying an institution, then all instances of that postal code were taken as correct. Unfortunately patients with an old postal code that was not associated with an institution could not be followed up if they did not have a recorded address.

The derived indicator of culture does provide a weight of evidence for the validity of the constructs being identified which is essential for evidentiary support for the development of policy, but it does suffer from limitations. As the indicators are developed from the underlying multivariate distributions of a census-based cultural construct, they are limited to the measures determined by Statistics Canada. As well, the census categories of ethnic origin, mother tongue and religion are derived from a 20% sample of the Canadian population.

The development of the indicators of culture only considers the multinomial distribution of one census category. Census cultural constructs should be combined in evocative ways to group dissemination areas upon more complex interplays of culture. This would
entail the combining of several cultural features, such as ethnic origins, mother tongue and religion, into one data set and then group the dissemination areas based upon a more complete and complex representation of culture.

To fully capture the complexity of each identified cultural group. The interpretations should be done in dialogue with sociologists and ethno-geographers in order to prevent disparaging ethnic profiling within the research community. Thus interdisciplinary collaborative teams would be beneficial for the continued development of multivariate cultural indicators.

Lastly, this research only describes where people have died. As administrative databases were used, no comment can be made on the congruence between the location of death and the patient’s preferred place of death. Understanding the reasons for dying in a certain location needs more in depth research. Qualitative and quantitative approaches can be set into an analytic dialogue to understand the complex interactions among the factors underlying the reasons for access and service use at the end of a patient’s life.

**Significance of this Work**

Prior to this thesis, there was no research into the trends in location of death for British Columbian cancer patients. Based upon the seminal work done in Nova Scotia, this work represents the establishment of a foundation of research upon which further investigations in end-of-life research can occur. By providing the first sketch of how British Columbian cancer patients are using health services at the end of their life, researchers are beginning to uncover how end-of-life health care utilization varies with respect to demographics, income and culture.

Through the investigation into the appropriateness and necessity of an ecological study for the data to be utilised in this research, a novel approach for the development of cultural indicators was determined. This methodology used principal component analysis for data and noise reduction. K-means was used for grouping the dissemination areas based upon the principal components. This approach represents a significant departure form the current methodology used in end-of-life health service utilization research.

The use of administrative databases as the source for the study sets represented an opportunity to investigate the utility of supervised learning techniques in end-of-life research. This thesis introduces data mining to this area of study. By comparing five different types of classifiers, it was shown that the standard method of logistic regression is not the best classifier. The counterpoint to this was that the best classifier, neural networks, has little utility for epidemiological research into end-of-life health service utilization. Logistic regression is preferred as it delivers the type of information that is most useful. A tension between performance and utility has been identified. The resolution of this tension will need further statistical development for many of the classifiers.

In summary, there are two benefits of this work. The first is the generation of new knowledge in terms of health service use at the end of life in BC. The second follows
from the first in that this knowledge has practical implications for policy makers in terms of service planning given the projected increase in chronic disease burden and the aging of our population.

Future Work and Recommendations

As this research represents seminal work for illustrative the use of end-of-life services in British Columbia, the following are recommendations for future work.

As the explanatory variables are at the individual and group levels, a mixed-level ecological study may be avoided by using a hierarchical model. It is recommended that a two-stage model be developed. The first stage models the aggregate-level data. The dissemination area is taken as the observational unit. The response would be the percentage of people in the dissemination area with a particular outcome, for instance death out of the hospital. The predictors would be aggregate-level variables, such as the indicators of culture. The second stage would model either higher or lower levels of data conditioned upon the first stage.

As the indicator of culture is novel in this thesis, further development needs to occur. The number of predictor variables can be expanded. More census categories can be used as well as more database variables. The indicators of culture should be developed for the 1997 census and validation procedures need to be developed and implemented to ensure sound research results.

The development of the culture indicators used principal components. The principal components were derived from explanatory variables that were proportions. These proportions will be constrained to sum to one. This constraint may have implications upon how principal components will perform. Research into the effects of using constrained explanatory variables should occur to ensure a sound methodology for the development of culture indicators.

The utility of logistic regression for health research emphasises the need for investigations that develop the same utility in other classifiers. Measures such as the odds ratio for classifiers such as classification trees and neural networks are recommended.

As the first step in a program of research on end-of-life care, this research only addresses the location of death. However, future work will include linkage to other population bases health databases such as hospital separations and continuing care information. This will allow researchers to obtain a detailed illustration of what and how health services are used and by whom at the end of a patient’s life. Also, special investigations should occur to determine the congruency between a patient’s desired location of death and the actual location of death.

Further data quality assessment and validation needs to occur. The database developed for this thesis was restricted in the means by which data quality could be assessed. With the addition of more databases, cross-database measures of data quality can be obtained.
References

1. Johnston, G.M., Boyd, C. J., MacIssac, M.A, Rhodes, J. W., Grimshaw, R. N. Effectiveness of letters to Cape Breton women who have not had a recent Pap smear. Chronic Diseases in Canada. 2003; 24(2/3).


5. Zamar, Ruben. [Personal Communication July 2005]