Applying the Apriori and FP-Growth Association Algorithms to Liver Cancer Data

by

Fabiola M. R. Pinheiro

B.Comp.Sc., Concordia University, 2007

A Thesis Submitted in Partial Fulfillment

of the Requirements for the Degree of

MASTER OF SCIENCE

in the School of Health Information Science

University of Victoria

SUPERVISORY COMMITTEE

Applying the Apriori and FP-Growth Association Algorithms to Liver Cancer Data

by

Fabiola M. R. Pinheiro

B.Comp.Sc., Concordia University, 2007

Supervisory Committee

Dr. Alex M. H. Kuo, Supervisor
(School of Health Information Science, University of Victoria)

Mr. Jeff Barnett, Departmental Member
(School of Health Information Science, University of Victoria)

Supervisory Committee

Dr. Alex M. H. Kuo, Supervisor
(School of Health Information Science, University of Victoria)
Mr. Jeff Barnett, Departmental Member
(School of Health Information Science, University of Victoria)

ABSTRACT

Cancer is the leading cause of deaths globally. Although liver cancer ranks only fourth in incidence worldwide among all types of cancer, its survivability rate is the lowest. Liver cancer is often diagnosed at an advanced stage, because in the early stages of the disease patients usually do not have signs or symptoms. After initial diagnosis, therapeutic options are limited and tend to be effective only for small size tumors with limited spread and minimal vascular invasion. As a result, long-term patient survival remains minimal, and has not improved in the past three decades. In order to reduce morbidity and mortality from liver cancer, improvement in early diagnosis and the evaluation of current treatments are essential.

This study tested the applicability of the Apriori and FP-Growth association data mining algorithms to liver cancer patient data, obtained from the British Columbia Cancer Agency. The data was used to develop association rules which indicate what combinations of factors are most commonly observed with liver cancer incidence as well as with increased or decreased rates of mortality.

Ideally, these association rules will be applied in future studies using liver cancer data extracted from other Electronic Health Record (EHR) systems. The main objective of making these rules available is to facilitate early detection guidelines for liver cancer and to evaluate current treatment options.

ACKNOWLEDGEMENTS

I would like to express my deepest appreciation to my committee members for all the valuable feedback and encouragement and especially:

- Dr. Alex M.H. Kuo, my supervisor, for suggesting this study idea and for all the guidance and revisions of this manuscript and other publications,
- Mr. Jeff Barnett, from the British Columbia Cancer Agency (BCCA), Adjunct Professor at the School of Health Information Science, for assisting with everything related to the BCCA, including the ethics approval process and the acquisition of the data set,
- Dr. Alex Thomo, from the Department of Computer Science, for all the guidance related to the association analysis, including the algorithm implementations used.

TABLE OF CONTENTS

LIST OF FIGURES

1. INTRODUCTION

1.1. Study background

According to the World Health Organization, cancer is the leading cause of deaths globally, having been responsible for 7.6 million (around 13%) of all deaths in 2008 (WHO, 2012). Although liver cancer ranks only fourth in incidence among all types of cancer (Pellegrino, 2006; Cardenes & Lasley, 2012), with 747,000 new cases estimated worldwide in 2008 (Lee et al., 2010), its survivability rate is the lowest.

The overall five-year survival rate for liver cancer observed in the Surveillance, Epidemiology, and End Results (SEER) Program database of the National Cancer Institute from 2003 to 2007 is a mere 14% (Jou & Fisher, 2010). Cancer in the liver causes approximately 1 million deaths each year (Pellegrino, 2006), and its incidence is projected to continue rising, with 12.7 million deaths expected in the year 2030 (Ferlay et al., 2010). According to the Public Health Agency of Canada, the incidence rate of liver cancer is significantly rising in both men and women (PHAC, 2013).

Incidence of liver cancer is associated with several risk factors and varies by geographic region. For this reason, researchers believe that environmental factors, in addition to patient characteristics, play a significant role in the development of the disease (Barber & Nelson, 2000; Chang et al., 2006).

The highest estimated incidence of liver cancer occurs in East Asia (Lee et al., 2010). In Mongolia, where liver cancer is the most common cause of death in both males and females (Sangdagdorj et al., 2010), there are on average 116 cases per 100,000 people (Lee et al., 2010). Moreover, Asian countries, which represented more than 60% of the world population in 2008 (United Nations, 2009), are seeing an increase in the impact of cancer as a health issue, due to an aging population, and to changes in lifestyles

associated with economic development (Shin et al., 2010; Yoo, 2010). According to the GLOBOCAN estimates in 2002, 45% (4.9 million) of all new cancer cases diagnosed in the world and 50% (3.4 million) of cancer deaths occurred in Asia (Shin et al., 2010).

Although with a much lower rate than in Asia, liver cancer in several developed countries in Europe and North America have been rising in the past decade, with possible links to the hepatitis C virus and diabetes (El-Serag, 2002; Llovet, Burroughs, & Bruix, 2003; Giovanucci et al., 2010; Cardenes & Lasley, 2012). These increased rates are accompanied by a shift in incidence from typically elderly patients to relatively younger patients between the ages of forty and sixty years.

In the United States, liver cancer incidence has approximately doubled over the past thirty years, and its rate is currently about 7 cases per 100,000 people (Lee et al., 2010). Incidence is higher in people of Asian origin, Pacific Islanders, and Native Americans, being approximately double or triple that of African Americans, who, in turn, are two to three times less often affected than Caucasians (El-Serag, 2002). Men are up to four times more often affected than women (El-Serag, 2002; Pellegrino, 2006).

For the past few decades, cancer registries such as that of the BC Cancer Agency have helped plan and evaluate the care of individual cancer patients and their survival. These registries also play a population-wide role in cancer control (Parkin, 2006; Sandagdorj et al., 2010) and can also extend the level of collaboration in cancer research, especially among countries (Valsecchi & Steliarova-Foucher, 2008; Curado, Voti, & Sortino-Rachou, 2009). In addition, these cancer registries constitute valuable sources of information on the incidence and characteristics of specific cancers.

1.2. Motivation

As a result of its high and increasing incidence and mortality rates, liver cancer has emerged as an area in urgent need of epidemiologic and outcomes research (Jou & Fisher, 2010). The BC Cancer Agency data set could constitute a good source of information about factors most often associated with liver cancer incidence, as well as the outcome of treatments in terms of survival. In addition, understanding cancer and what to expect can help patients make better decisions with regards to treatment options and costs, lifestyle changes, quality of life (Delen, 2009).

1.3. Research objectives

The objectives of this study include:

(1) To explore the associations between liver cancer and patient characteristics.

(2) To develop association rules that indicate what combination of patient characteristics (demographics, geographic locations, disease stage, treatment, and clinical factors) are frequently observed in cases of liver cancer.

(3) To test the applicability of the Apriori and FP-Growth association data mining algorithms in analyzing liver cancer data. Also to potentially their applicability for future studies using liver cancer data extracted from the EHR systems of the National Taiwan University Hospital and the Mongolian University of Science and Technology Hospital.

(4) To test the suitability of the Apriori and FP-Growth association data mining algorithms as a reliable means of detecting cases of liver cancer from new patient data,

taking into account patient demographics and clinical characteristics, and possibly providing an early alert to those patients whose characteristics categorize them as having a high risk of liver cancer.

(5) To test the association of different courses of treatment (such as radiotherapy, surgery, and chemotherapy) with survivability. Survivability is in this case defined as the period of time the patient survives after being diagnosed with liver cancer.

1.4. Research constraints/limitations

This study was carried out on anonymized data obtained from the BC Cancer Agency. Anonymized data are data that have the patient identification information permanently removed subsequent to collection (Delen, 2009) due to privacy and ethics concerns. Such anonymization was carried out by the BC Cancer Agency before the data was provided and consisted of removing the patient's name (first, middle, and last), BC personal health number (PHN), BC Cancer Agency unique identifier (agency ID), home address, city, postal code, and phone number. Patient identification is not considered important for the purposes of this research, so the data and potential study outcome will not suffer any loss from the anonymization procedure.

Clinical and geographic information were used as provided by the BC Cancer Agency. Some variables needed to be cleaned, either due to missing values, or through categorization of values, as will be explained later.

2. LITERATURE REVIEW

2.1. Incidence of Liver Cancer

Tumors of the liver either arise from native hepatic cells (primary nature) or due to metastasis or the direct spread of neoplasia from adjacent organs (secondary nature) (McKillop & Schrum, 2005). The primary nature is usually related to presence of chronic inflammation, mainly triggered by exposure to infectious agents or to toxic compounds (Berasain et al., 2009). Hepatocellular carcinoma is the most common primary cancer of the liver (Greene at al., 2006). Cholangiocarcinoma is another, less common form (Yoo, 2010). As for the secondary nature, the liver is a common site for metastases for cancers such as colorectal, breast, and lung (Lock et al., 2012).

In terms of infectious agents, cirrhosis and the hepatitis (B and C) virus are the two most common factors contributing to liver cancer worldwide (Barber & Nelson, 2000; El- Serga, 2002; Llovet, Burroughs, & Bruix, 2003; McKillop & Schrum, 2005; Chang et al., 2006; Greene et al., 2006; Pellegrino, 2006; Luk et al., 2007; Hainaut & Boyle, 2008; Valsecchi & Steliarova-Foucher, 2008; Berasain et al., 2009). Although most patients with liver cancer have underlying cirrhosis, liver cancer does not develop in all patients with cirrhosis (McKillop & Schrum, 2005). Still, approximately 20% of all patients who die of cirrhosis have liver cancer (Barber & Nelson, 2000).

Case-control studies demonstrate that carriers of hepatitis B or C virus are fifteen- or seventeen times respectively more likely to have liver cancer than healthy individuals (El-Serag, 2002). Approximately 60% to 90% of cases of liver cancer are believed to have been caused by chronic infection with hepatitis B (Barber & Nelson, 2000). Hepatitis C became a major cause of hepatic cancer in the United States in the 1960s

when an epidemic started due to injected drug use and blood transfusions using unscreened blood (Pellegrino, 2006). An even higher risk is reported in hepatitis B endemic areas, where transmission is mostly through sexual and parenteral routes (El-Serag, 2002; Hainaut & Boyle, 2008). Progressive hepatitis infection causes liver inflammation, eventually leading to fibrosis and cirrhosis (Pellegrino, 2006).

Liver flukes, such as *Clonorchis* sinensis (endemic in Korea) and *Opistorchis viverini* (endemic in Thailand) may also be factors associated with liver cancer, based on evidence that they cause cholangio-carcinoma, whose incidence is high in Korea and Thailand (Yoo, 2010). Another condition, although not infectious, that is frequently associated with liver cancer is hemochromatosis. This is a hereditary metabolic disease which causes excess iron to accumulate in the liver. This accumulation eventually leads to inflammation and cirrhosis (Pellegrino, 2006).

Diabetes may also be a factor (Chang et al., 2006). Type 2 diabetes, which accounts for 95% of prevalent diabetes cases, is also often observed in liver cancer patients. Liver cancer and diabetes are diagnosed within the same individual more frequently than would be expected by chance, even after adjusting for age. However, how the two diseases are connected has not been discovered yet. Diabetes-related factors including steatosis, non-alcoholic fatty liver disease and cirrhosis may enhance susceptibility to liver cancer (Pellegrino, 2006; Giovanucci et al., 2010). Moreover, evidence from observational studies suggests that some medications used to treat hyperglycemia in diabetics may be associated with an increased incidence of cancer. Some studies also suggest diabetes may increase mortality in patients with cancer (Giovanucci et al., 2010).

In terms of toxic compounds, exposure to arsenic and polyvinyl chloride and to aflatoxins (particularly aflatoxin B1) have been highly correlated with increased risk of liver cancer. Aflatoxins are toxic compounds produced by *Aspergillus flavus* and *A. parasiticus* molds which are often found in improperly stored grains and nuts (Barber & Nelson, 2000; El-Serag, 2002; Barrett, 2005; McKillop & Schrum, 2005; Pellegrino, 2006; Hainaut & Boyle, 2008; Valsecchi & Steliarova-Foucher, 2008; Shils, 2008; Delen, 2009). Other products suspected of increasing risk of liver cancer include tobacco, androgenic steroids, and oral contraceptives (Chang et al., 2006; Pellegrino, 2006).

Greater alcohol intake is another factor normally associated with increased liver cancer (Barber & Nelson, 2000; El-Serag, 2002; McKillop & Schrum, 2005; Chang et al., 2006; Greene et al., 2006; Pellegrino, 2006; Delen, 2009; Berasain et al., 2009; Giovanucci et al., 2010). The liver is the major site of metabolism of ethanol alcohol ingested, producing acetaldehyde and free radicals that bind rapidly to numerous cellular targets. In addition to direct DNA damage, acetaldehyde depletes glutathione, an antioxidant involved in detoxification.

Chronic ethanol abuse also leads to induction of hepatocyte microsomal cytochrome P450 2E1, an enzyme that metabolizes ethanol to acetaldehyde and is also associated with activation of procarcinogens, changes in cell cycle, nutritional deficiencies, and altered immune system responses. Competing factors of ethanol consumption, such as the high incidence of cigarette smoking in ethanol-dependent patients, may also play a significant role in the development of cirrhosis, cell transformation, and tumor progression (McKillop & Schrum, 2005).

Other patient characteristics that are associated with higher liver cancer risk include age, ethnicity and diet, similarly with other types of cancer. In terms of age, cases of liver cancer are rarely seen before age forty, with exceptions including mostly patients who acquired hepatitis B at birth or during childhood (El-Serag, 2002). As for ethnicity, as mentioned previously, liver cancer affects people of Asian origin in greater proportions, with African Americans coming second and Caucasians being the least affected population (El-Serag, 2002; Pellegrino, 2006). In addition, incidence in males is much higher than in females, again as with other types of cancer, with male to female ratios of between 2:1 to 4:1 (El-Serag, 2002; Pellegrino, 2006; Berasain et al., 2009; Giovanucci et al., 2010).

Studies suggest that diets low in red and processed meats and higher in vegetables, fruits, and whole grains are associated with a lower risk of many types of cancer (Giovanucci et al., 2010). Coffee consumption has also been suggested to decrease the risk of liver cancer, even in heavy alcohol drinkers (La Vecchia, 2005).

2.2. Screening and Detection of Liver Cancer

As in other types of cancer, early detection increases the chance of cure (Pendharkar et al., 1999; Lareinjam & Wasan, 2009). In the case of liver cancer, this is even more so the case, as effective treatment is limited to the early stages of the disease (Richards et al., 2001). Ando et al. (2006) reported that early detection significantly prolonged the five-year survival rate in Japan.

Early diagnosis of liver cancer is hampered by the absence of symptoms and markers (Osl et al., 2008). In that context, several countries have implemented screening programs for liver cancer with positive results in reducing cancer mortality (Lee et al., 2010; Yoo, 2010). Korea, for instance, started such a program in 2003, identifying the population at high risk by means of testing for hepatitis C virus antibodies and hepatitis B antigens in patients over 40 years of age. Common screening tests include ultrasound and measurements of alfa-fetoprotein levels in serum (Lee et al., 2010).

The American Association for the Study of Liver Diseases recommends liver cancer screening at six-month intervals for high risk individuals (Jou & Fisher, 2010; Lee et al., 2010), with varying figures given in the literature for the cost of surveillance per year of life saved (or quality adjusted life year), ranging from $26,000 to $55,000 (Lee et al., 2010). In addition to screening programs, nationwide hepatitis B vaccination efforts have proved successful in various Asian countries, such as Taiwan, Malaysia, and Mongolia (Yoo, 2010).

Screening and diagnosis determine the incidence of liver cancer. As such, no drop in incidence will result immediately from early detection screening programs. In fact, the introduction of a screening program should be followed by a rise in incidence. However, the incidence of advanced cases should fall, as well as the mortality rates (Parkin, 2006).

2.3. Staging Systems

Although the prognosis of solid tumors is generally related to tumor stage, in liver cancer it is greatly influenced as well by the underlying liver dysfunction and needs to

take into account liver function and performance status, as well as the impact of treatment. Historically, the TNM or Okuda staging systems have been used, despite neither being useful in determining a prognosis with the most adequate forms of therapy. This is particularly true in patients with early or intermediate stages of liver cancer (Cardenes & Lasley, 2012).

The TNM Classification of Malignant Tumours (TNM) is a cancer staging system that describes the extent of cancer in a patient's body:

i. **T** describes local **tumor** size and its growth and spread to nearby tissue,
ii. **N** describes regional lymph **nodes** that are involved,
iii. **M** describes distant **metastasis** (spread of cancer from one body part to another).

Spread to regional lymph nodes and/or distant metastasis occur before they are discernible by clinical examination. As a result, pathologic classification and staging based on the examination of a surgically resected specimen with sufficient tissue may differ from clinical staging and is recorded as well. Both clinical and surgical classification should be maintained in the patient's permanent medical record. The clinical stage is used as a guide to the selection of primary therapy, whereas the surgical stage can be used as a guide to the need for adjuvant therapy (Greene et al., 2006).

2.4. Treatment of Liver Cancer

As mentioned above, staging of liver cancer is required for treatment selection. The factors used to determine treatment are mainly tumor location and size, vascular

invasion, lymph node involvement, the presence and extent of metastases, and the patient's liver function (Barber & Nelson, 2000; Pellegrino, 2006). With current treatment practices, survival rate is low if the tumor size is over 5 cm (Richards et al., 2001). This emphasizes the importance of early detection, which can not only save patients' lives but also avoid more drastic treatment procedures (Pendharkar et al., 1999).

There are four main categories of treatment for patients diagnosed with liver cancer:

(1) surgical interventions, including tumor resection and liver transplantation;

(2) percutaneous interventions, including ethanol injection and radio frequency thermal ablation;

(3) transarterial interventions, including embolization and chemoembolization;

(4) other treatments, such as drugs, gene and immune therapies (Blum, 2005).

The main goals of these therapies are to prolong survival time and to maintain quality of life (Barber & Nelson, 2000).

2.4.1. Surgical Interventions

Some authors consider surgery, either liver resection or transplant, the only chance of recovery (Pellegrino, 2006). Lock et at. (2012) indicate a general lack of type I evidence for any local liver treatment other than surgery. Type I evidence indicates cause of an outcome and includes factors such as its magnitude and severity (Rabin & Brownson, 2012).

Although surgical resection remains the most common treatment, only 40% to 50% of patients diagnosed with long-term (5 to 7 years) liver cancer are eligible. Survival rates typically fall between 10% and 25% (Okuda, 2002; Sitzmann & Abrams, 1993).

Partial hepatectomy consists of removing the diseased portion of the liver. Liver function determines the remaining liver's ability to regenerate. The 5-year survival rate is about 30% (Pellegrino, 2006). Partial hepatectomy is recommended for patients with:

(1) an encapsulated tumor less than 2 to 5 cm confined to one lobe,

(2) minimal cirrhosis,

(3) adequate liver function,

(4) absence of extra-hepatic metastasis,

(5) no vascular invasion by the tumor,

(6) no portal hypertension (Barber & Nelson, 2000; Pellegrino, 2006).

Total hepatectomy consists of removing the tumor and cirrhotic liver and is performed in conjunction with a cadaveric liver transplant (Pellegrino, 2006). Liver transplantation, although greatly limited by organ donor availability and the absence of metastasis before transplant, benefits patients who have decompensated cirrhosis and one tumor smaller than 5 cm or three nodules smaller than 3 cm (Llovet, Burroughs, & Bruix, 2003; McKillop & Schrum, 2005). The 4-year survival rate can be as high as 85% (Pellegrino, 2006). In theory, transplantation might simultaneously cure the tumor and the underlying cirrhosis, but only has benefits if the waiting time is 6 months or less (Llovet, Burroughs, & Bruix, 2003).

2.4.2. Percutaneous Interventions

Percutaneous treatments provide good results, with a 5-year survival rate between 40 and 50%. The response rates and outcomes, however, are not comparable to surgical treatments (Llovet, Burroughs, & Bruix, 2003; McKillop & Schrum, 2005).

Percutaneous ethanol injection achieves responses of 90–100% in liver tumors smaller than 2 cm, 70% in those of 3 cm, and 50% in liver tumors of 5 cm in diameter (Llovet, Burroughs, & Bruix, 2003). The ethanol injected diffuses within the neoplastic cells, inducing cellular dehydration, coagulation necrosis, and vascular thrombosis, eventually destroying the cancerous tissue by ischemia (Barber & Nelson, 2000).

Radiofrequency ablation is a minimally invasive image-guided technique for destroying tumor cells using thermal energy. It is a type of Radiotherapy and, as practiced today, involves the insertion of a needle electrode percutaneously into the liver tumor under imaging guidance (Barber & Nelson, 2000; Pellegrino, 2006; Munneke, 2008), but it can also be done laparoscopically, or during laparotomy (Llovet, Burroughs, & Bruix, 2003). Studies have reported 5-year survival rates of 38-64% for radiofrequency-ablated patients (Munneke, 2008). Alternatives to radiofrequency ablation are laser ablation and microwave ablation (Barber & Nelson, 2000). A new technique of stereotactic body radiotherapy, using “targeted, highly conformal, hypofractionated ablative radiotherapy” has shown positive results (Cardenes & Lasley, 2012).

2.4.3. Transarterial Interventions

Other therapies, including intra-arterial embolism and cryoablation, have been used with limited success (McKillop & Schrum, 2005). Intra-arterial embolysm consists

of inserting a catheter through a femoral artery to the desired portion of the hepatic artery, which is then plugged with small particles injected. The goal is to cause cell death by stopping blood supply to the tumor.

Chemo-embolization consists of embolizing the artery, then infusing a chemotherapy agent into the blocked artery. This exposes the tumor to concentrated chemotherapy without causing systemic effects (Barber & Nelson, 2000; Pellegrino, 2006). Radioembolization with hepatic arterial yttrium-90 appears to be a new alternative (Cardenes & Lasley, 2012; Lock et al., 2012). Cryosurgery involves using an intra-tumoral circulating liquid nitrogen probe to freeze and destroy liver tumors (Barber & Nelson, 2000; Pellegrino, 2006).

2.4.4. Other treatments, such as drugs, gene and immune therapies

Gene therapy involves manipulating cancer genes in order to make cancer cells vulnerable to treatment. This manipulation may be done by:

(1) correcting the underlying genetic defect that led to tumor development,

(2) altering the immunogenicity of a tumor,

(3) interjecting cytokinones into tumor cells,

(4) introducing a suicide gene that would increase the sensitivity of the tumor to chemicals (Barber & Nelson, 2000).

2.5. Knowledge Discovery Process

Data mining is an integral part of the field of knowledge discovery in databases (KDD), which is the overall process of converting raw data into previously unknown and potentially useful information (Pendharkar et al., 1999; Papageorgiou, Kotsioni & Linos, 2005; Kaur & Wasan, 2006; Tan, Steinbach, & Kumar, 2006). In other words, it is the process of discovering useful patterns and trends in large data repositories (e.g., data warehouses) (Clifton & Thuraisingham, 2001; Seeja, Alam & Jain, 2008; Sangster-Gormley et al., 2013). The KDD process consists of the following main phases:

(1) problem definition,

(2) data pre-processing,

(3) pattern recognition,

(4) pattern validation (Figure 1).

The problem definition phase (1) defines the questions to be answered (Sangster-Gormley et al., 2013), such as "what demographic factors lead to increased incidence of liver cancer?" or "should the course of treatment for a liver cancer patient include surgery alone or surgery plus chemotherapy or surgery plus radiation?" (Kaur & Wasan, 2006).

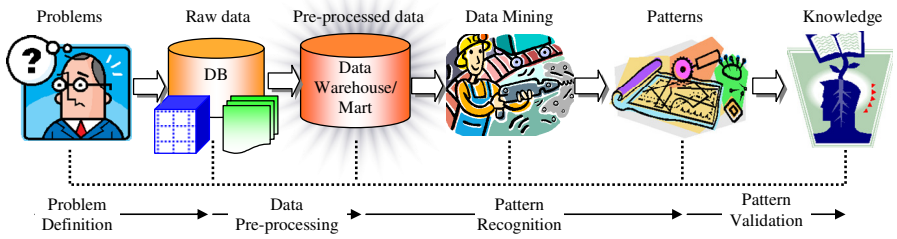

Figure 1. Generic process of knowledge discovery (from Sangster-Gormley et al., 2013)

The data pre-processing phase (2) cleans and transforms the raw data into a format that is appropriate for subsequent analysis. In the pattern recognition phase (3) a data mining algorithm is selected, such as the Apriori and FP-Growth algorithms used in this study. In the pattern validation phase (4) the performance of the data mining models is assessed (Sangster-Gormley et al., 2013) in terms of the knowledge produced.

Data mining applied to healthcare and medical areas is challenging yet successful (Richards et al., 2001; Kaur & Wasan, 2006; Concaro et al., 2009; Delen, 2009; Kirshners, Parshutin & Leja, 2012). The challenges derive from large, complex, heterogeneous, hierarchical healthcare data sets which may include time series and fragmented data of varying quality (Delen, 2009). According to Kaur and Wasan (2006), without data mining it is difficult to realize the full potential of healthcare data, which are "massive, highly dimensional, distributed, and uncertain".

When applied to cancer detection and treatment, data mining can be of great utility, as suggested by previous studies (Ho, Jee, Lee, & Park, 2004; Li et al., 2004; Chen et al., 2006; Henning et al., 2007; Luk et al., 2007; Chun et al., 2008; Delen, 2009; Lareinjam & Wasan, 2009; Lisboa et al., 2010; Agrawal & Choudhary, 2011).

In addition to cancer data, mining of association rules can be applied to other types of health care data as well. For instance, Mukhopadhyay et al. (2010) have used association rules for clarifying mechanisms of viral-human interactions in HIV patients. Also, Kuo et al. (2010) have positively assessed the suitability of the Apriori association algorithm for detecting adverse drug reactions.

Association rules have been widely used with bioinformatics and biomedical data, because they can manage large heterogeneous data sets and the rules produced are intuitively interpreted (Karabatak & Ince, 2009a) and humanly understandable. According to Li, Fu & Fahey (2009), data mining methods that use classification instead of association, on the other hand, are not well suited to medical data. For liver cancer, association rule data mining could provide better understanding of the cancer development mechanisms and help reveal which risk factors are associated with increased incidence. Eventually, the better understanding gained will enable the development of mechanisms for early detection.

2.6. Association Rule Algorithms

2.6.1. Overview of association analysis

Among the various data mining techniques, association analysis is a popular and well researched method for discovering interesting associations and/or relationships among variables in large databases (Tan, Steinbach, & Kumar, 2006; Karabatak & Ince, 2009a). For a pattern to be interesting, it should be logical and actionable (Delen, 2009). Association rule learning basically consists of revealing relationships among attribute values that occur frequently together in a dataset, and then representing them in the form of association rules (Tan, Steinbach, & Kumar, 2006; Karabatak & Ince, 2009a).

Association rules do not indicate causality, but instead suggest strong co-occurrence relationships that can be further investigated as associated factors. Causality

would require knowledge of causal and effect attributes, typically by observing relationships over time (Tan, Steinbach, & Kumar, 2006).

2.6.2. Support and Confidence

Two important metrics in association analysis are support and confidence. Support indicates how often a rule is applicable to a specific data set and can be used to eliminate uninteresting rules, such as those that occur simply by chance (Tan, Steinbach, & Kumar, 2006; Hu, 2010). Confidence measures the reliability of the inference made by a rule, for instance, X -> Y, measured as how frequently items or attributes in Y appear in transactions or patients that contain X.

Minimum support and confidence thresholds are selected for assessing the association rules extracted from the data. An itemset is frequent if its support is greater than or equal to that minimum support value (Tan, Steinbach, & Kumar, 2006). One important issue with mining association rules in large data sets is the fact that it can be computationally expensive depending on the algorithm used. A brute-force approach for discovering patterns from data would consist of computing the support and confidence for every possible rule. As the number of rules that can be obtained from a data set grows exponentially with the number of items in that set, this brute-force approach becomes prohibitively expensive. This approach also results in wasted transactions, as many of the rules would be discarded for falling below the minimum support and confidence levels selected (Tan, Steinbach, & Kumar, 2006).

Lift, also known as interest factor for binary variables, is the ratio between the rule's confidence and the support of the itemset in the rule's consequent (Tan, Steinbach,

& Kumar, 2006). It is the relative improvement of the frequency of a pattern against a baseline frequency or average of the target attribute computed across the database under the statistical independence assumption (Tan, Steinbach, & Kumar, 2006; Agrawal & Choudhary, 2011). For correlation analysis, Pearson's correlation coefficient can be used between a pair of continuous variables, and the fi-coefficient between a pair of binary variables (Tan, Steinbach, & Kumar, 2006).

2.6.3. Apriori and FP-Growth Algorithms

Many algorithms for generating association rules have been presented over time, such as the Apriori and FP-Growth algorithms (Tan, Steinbach, & Kumar, 2006; Hu, 2010). A common strategy that these algorithms implement, in terms of performance improvement, is to decompose the problem into two subtasks:

a) frequent itemset generation, accomplished by reducing either the number of:
    i. candidate itemsets based on the support measure, as in the Apriori algorithm, or
    ii. comparisons, as in the FP-Growth algorithm;

b) rule generation, which first excludes rules that have empty antecedents or consequents and then checks that, after splitting itemset Y into two non-empty subsets (X and Y – X), rule X -> Y – X satisfies the confidence threshold. Y – X in this case is known as the rule consequent. Rule generation does not require any additional passes over the dataset (Tan, Steinbach, & Kumar, 2006).

2.6.4. The Apriori Algorithm

The Apriori association data mining algorithm is probably the best known (Hu, 2010) association algorithm and was originally introduced for market basket data analysis. It analyzes the dataset to produce association rules which show attribute value conditions that occur frequently together. The goal is to capture association rules that are useful and important in explaining the presence of certain attributes according to the presence of other attributes (Agrawal, Imilienski, & Swami, 1993; Tan, Steinbach, & Kumar, 2006; Karabatak & Ince, 2009b; Kuo et al., 2010).

The Apriori algorithm assumes that "the support for an itemset never exceeds the support for its subsets". This guarantees that, if an itemset is frequent (i.e., its support measure is greater than or equal to the minimum support threshold), then all subsets of that itemset must also be frequent. On the other hand, it also guarantees that if an itemset is infrequent, then all of its supersets must be infrequent. This allows for the exponential search space to be trimmed by removing all supersets of an infrequent itemset, in what is known as support-based pruning. Although this significantly improves performance, the Apriori algorithm still incurs considerable input/ouput overhead since it makes several passes over the dataset. Performance may also be affected for dense datasets due the increasing width of transactions (Tan, Steinbach, & Kumar, 2006).

Historically, Apriori was the first association rule mining algorithm to systematically use support-based pruning to limit the number of candidate itemsets. The Apriori algorithm initially makes a single pass over the data in order to determine the support of each item. That way, the set of all frequent 1-itemsets is created, and the algorithm can then generate candidate 2-itemsets, using a function named apriorigen. The

Apriori algorithm then makes an additional pass over the data to count the support of the candidates and eliminate those that fall under the minimum threshold. This procedure is repeated iteratively in a level-wise mode, from 1-itemsets to k-itemsets. Each level corresponds to the number of items that belong to the rule consequent. In the rule generation step of the Apriori algorithm, the confidence of each rule is determined by using the support counts computed during frequent itemset generation (Tan, Steinbach, & Kumar, 2006).

2.6.5. The FP-Growth Algorithm

The FP-Growth algorithm encodes the input data set into a compact data structure known as an FP-tree. In certain data sets, the FP-Growth algorithm outperforms the standard Apriori algorithm by several orders of magnitude, depending on the compaction factor of the FP-tree.

The FP-tree is constructed by reading the data set one transaction at a time and mapping each to a path in the tree. Paths in the FP-tree overlap when different transactions share common items. The more paths overlap, the greater the compression achieved with the FP-tree structure. If the size of the FP-tree is small enough, it will fit in main memory, from which frequent itemsets can be directly extracted. Otherwise repeated passes need to be made over the data on disk storage.

In building the FP-tree, one pass is made over the data to determine support count for each item, discard infrequent items, and sort the frequent ones in order of decreasing support counts. The data set is then scanned once more to read each transaction and add its corresponding path to the initial tree, which consists of simply a root node represented

by the null symbol. For each transaction read, a new set of nodes is created, as long as the paths do not share a common prefix. When paths share a common prefix (same initial item), they overlap in the tree, and the support count for the shared node (prefix) is incremented by one. Once every transaction has been read and mapped on a path, the resulting FP-tree is ready. The size of the resulting tree depends on the ordering of the items (Tan, Steinbach, & Kumar, 2006).

The frequent itemset generation in the FP-Growth algorithm is done in a bottom-up fashion, starting with a particular ending item. Only the paths containing that node are examined, after ensuring that it is a frequent itemset itself. The support counts along the prefix paths are updated to count only transactions that include the node in question, as well as to truncate the paths by removing that node. Then the algorithm tries to find frequent itemsets ending in that node paired with each of the other nodes that immediately precede it in the FP-tree. This is done in a recursive fashion.

In order to avoid an unmanageable amount of rules created, it is important to clearly set criteria for evaluating the quality of association patterns. This can be done:

a) objectively, through measures that use statistics, such as support, confidence, lift (or interested factor), and correlation, to determine the interestingness of a pattern;

b) subjectively, by using domain expertise to determine whether the information or knowledge revealed about the data is interesting.

2.7. Previous related research

Several studies have reported using association rules in the discovery of cancer prevention factors and also in cancer detection and surveillance. For instance, Karabatak & Ince (2009b) have created an automatic diagnosis system for detecting breast cancer based on association rules and neural network. Mavaddat et al. (2010) have developed an association algorithm to predict the risk of breast cancer based on the probabilities of genetic mutations. Malpani et al. (2011) investigated association rules between the set of transcription factors and the set of genes in breast cancer.

Javier Lopez et al. (2009) have attempted to integrate the most widely used prognostic factors associated with breast cancer (primary tumor size, the lymph node status, the tumor histological grade, and tumor receptor status) with whole-genome microarray data to study the existing associations between these factors and gene expression profiles.

Osl et al. (2008) mined association rules with metabolic markers in prostate cancer. Agrawal & Choudhary (2011) performed association rule mining analysis on survival time in the lung cancer data from the Surveillance, Epidemiology, and End Results (SEER) Program.

Bener et al. (2010) determined association rules between family history of colorectal cancer in first-degree relatives, parental consanguinity, lifestyle and dietary factors, and risk of developing colorectal cancer. The study was performed with controls matched by age, gender, and race. The results indicated that parental consanguinity, family history of colorectal cancer, smoking, obesity, and dietary consumption of bakery items were more often associated with a colorectal cancer diagnosis.

Nahar et al. (2011) compared three association rule mining algorithms (Apriori, predictive Apriori and Tertius) for uncovering the most significant prevention factors for bladder, breast, cervical, lung, prostate, and skin cancers. Based on their experimental results, they concluded that Apriori is the most useful association rule mining algorithm for the discovery of prevention factors.

Several other studies have applied the Apriori association algorithm to the analysis of cancer data, focusing on treatment outcome as well as detection. Fan et al. (2010) have shown that Apriori association rules constitute an efficient method of exploring the relationships between breast cancer treatment options and survivability.

Hu (2010) studied the association rules between degree of malignance, number of invasion nodes, tumor size, tumor recurrence, and radiation treatment of breast cancer. That author proposed an improved Apriori association algorithm to reduce the size of candidate sets and better predict tumor recurrence in breast cancer patients.

Apriori association rule data mining has been applied to bioinformatics as well. Automonova et al. (2005) employed it in finding exceptions to the rules produced in protein annotation databases and reporting those as errors. This was developed into a methodology of error detection for improving the quality of the biological annotation data.

Additionally, Rodriguez, Carazo, & Trelles (2005) used association rules to discover homologies and evolutionary relationships between protein sequences that share short similar fragments with proteins of known function. The authors even developed their own association rule discovery algorithm, especially suitable for data in which

elements appear clustered in sparse but strongly related groups, where they say the Apriori and FP-Growth algorithms would be limited in performance.

Although plenty of studies have applied association rule mining to different types of cancer, liver cancer is not one of those. A classification algorithm has been proposed using alpha-fetoprotein and ultrasound results for liver cancer surveillance (Jou & Fisher, 2010; Richardson et al., 2010).

3. STUDY METHOD

3.1. The Data Mining Algorithm

This study will test the applicability of the Apriori and FP-Growth association rule mining algorithms and the generation of a group of association rules. The principal contribution is a group of association rules, developed, tested and validated on real liver cancer patient data. These association rules would ideally be used for future studies using liver cancer data extracted from other repositories and EHR systems. The overall goal is providing early detection guidelines for liver cancer.

3.2. Data Collection

3.2.1. British Columbia Cancer Agency (BCCA) liver cancer data set description

A data set consisting of liver cancer cases diagnosed in patients living in British Columbia (BC) and the Yukon Territory (YT) between January 1, 1970 and December 31, 2010 was obtained from the British Columbia Cancer Agency (BCCA). The data from the BC Cancer Agency Info System (CAIS) consisted of a total of 6,064 patients (6,047 from BC, 17 from the Yukon). One of those 6,064 patients (patient #4398) had a confirmed diagnosis of cancer in two separate liver sites. As a result of this, the total of liver cancer sites is 6,065, greater than the number of patients by one. Some of the 6,064 patients have had multiple radiation and/or surgical treatments. For that reason, the data file (in Excel 2003 format) contained 6,479 rows of data. The data attributes are described in Appendix A.

Unfortunately, the data do not contain any information regarding race or ethnic background or preexisting conditions such as hepatitis or diabetes. The data, however do mention:

a) the extent of the disease at the time of diagnosis, if completed on the Cancer Registration form by the patient's physician,
b) the stage, or histopathological degree of dedifferentiation of malignant neoplasms or the total number of histopathological features translated into a grade.

3.2.2. Case Selection Criteria

- BC - includes any cases where the patient had a BC postal code and /or Statistics Canada BC geographic code at the time of diagnosis. This also includes any records where either one or both of the above attributes was left blank at the time of diagnosis.

- Yukon - includes any cases where the patient had a YT postal code and/or Statistics Canada YT geographic code at the time of diagnosis.

- Includes ICD-O-3 cancer sites C220 (liver) and C221 (intrahepatic bile duct)

- Includes all histological behaviors - benign (0), borderline (1), insitu (2) and malignant (3)

- Includes all ICD-O-3 and SNOMED histologies

- Excludes pending cases, i.e., where the diagnosis is still being investigated and awaiting reports.

3.3. Data Pre-processing

Preliminary data processing and exploratory data analysis were performed using ACL Desktop version 9.3.0. This software was chosen due to my degree of familiarity with it, as, at the time of writing this thesis, I work with it on a daily basis.

3.3.1. Creation of additional attributes

The initial step of the preliminary data processing was to create a few additional attributes as computed fields. Some of these calculated fields were for the purpose of filling voids in the data set, by adding necessary additional attributes, such as calculating ages from years. Others were meant to categorize data or encode attribute values as numbers, the only type of data the FP-Growth algorithm application used can process. In order for these numbers to be told apart in the files that are output by the FP-Growth program, they were coded to include a unique prefix consisting of either one or two digits. Such data transformation is a typical step in the data mining process (Maimon & Rokach, 2010).

The additional attributes were created as described below:

- Number of years from diagnosis to death: calculated as the difference between the year of death and the year of diagnosis

- Age at time of death: calculated as the difference between the year of death and the year of birth for the deceased patients; “NA” (not applicable) for the patients who are alive
- Age group at time of diagnosis: obtained following the rules below:
    - "0 to 9" IF age at time of diagnosis < 10
    - "10 to 19" IF age at time of diagnosis < 20
    - "20 to 29" IF age at time of diagnosis < 30
    - "30 to 39" IF age at time of diagnosis < 40
    - "40 to 49" IF age at time of diagnosis < 50
    - "50 to 59" IF age at time of diagnosis < 60
    - "60 to 69" IF age at time of diagnosis < 70
    - "70 to 79" IF age at time of diagnosis < 80
    - "80 to 89" IF age at time of diagnosis < 90
    - "90 to 100" IF age at time of diagnosis < 100
    - "100 or older" IF age at time of diagnosis >= 100
- Age group at time of death: obtained following rules similar to those given above for age at time of diagnosis, but using the age at time of death instead of the age at time of diagnosis, and “NA” (not applicable) for the patients who are alive
- Age at time of diagnosis code: to each of the age groups at time of diagnosis a code was assigned, which is a number consisting of prefix 2 followed by two other digits in increasing order:
    - "201" IF age at time of diagnosis < 10
    - "202" IF age at time of diagnosis < 20

- "203" IF age at time of diagnosis < 30
- "204" IF age at time of diagnosis < 40
- "205" IF age at time of diagnosis < 50
- "206" IF age at time of diagnosis < 60
- "207" IF age at time of diagnosis < 70
- "208" IF age at time of diagnosis < 80
- "209" IF age at time of diagnosis < 90
- "210" IF age at time of diagnosis < 100
- "211" IF age at time of diagnosis >= 100
- “200” IF undefined

- Age at time of death code: to each of the age groups at time of death a code was assigned, which is a number consisting of prefix 3 followed by two other digits in increasing order, except for “NA” (not applicable), for the cases where the patient is not deceased:
  - “NA” IF age at time of death is “NA”
  - "301" IF age at time of death < 10
  - "302" IF age at time of death < 20
  - "303" IF age at time of death < 30
  - "304" IF age at time of death < 40
  - "305" IF age at time of death < 50
  - "306" IF age at time of death < 60
  - "307" IF age at time of death < 70
  - "308" IF age at time of death < 80

- "309" IF age at time of death < 90
- "310" IF age at time of death < 100
- "311" IF age at time of death >= 100
- "300" IF undefined

- Gender code:
  - “401” for female
  - “402” for male
  - “400” for not stated / not recorded
- Tumor group code: consisting of prefix 5 followed by digits 00 through 04:
  - "501" IF tumor group = " Gastro-intestinal
  - "502" IF tumor group = "Lymphoma"
  - "503" IF tumor group = "Melanoma"
  - "504" IF tumor group = "Sarcoma"
  - "500" IF undefined
- Tumor subgroup code: consisting of prefix 6 followed by digits 00 through 04:
  - "601" IF tumor subgroup = " Hodgkin’s disease
  - "602" IF tumor subgroup = "Liver"
  - "603" IF tumor subgroup = "Non Hodgkin’s"
  - "604" IF tumor subgroup = "Other"
  - "600" IF undefined
- Patient status code:
  - "701" IF patient status = "alive (A)"
  - "702" IF patient status= "deceased (D)"

    - "700" IF undefined
- Code for diagnosis health authority: applies only to BC patients and consists of prefix 8 followed by digits 00 through 05:
    - "801" IF "Interior"
    - "802" IF "Fraser"
    - "803" IF "Vancouver Coastal"
    - "804" IF "Vancouver Island"
    - "805" IF "Northern"
    - "800" IF undefined
- Code for the ECOG performance status: consisting of prefix 9 followed by digits 00 through 06:
    - "901" IF " Fully active (Karnofsky 90-100)"
    - "902" IF "Restricted in physically strenuous activity (Karnofsky 70-80)"
    - "903" IF "Ambulatory, unable to carry out any work activities (Karnofsky 50-60)"
    - "904" IF "Limited self-care, confined to bed/ chair > 50% of waking hours (Karnofsky 30-40)"
    - "905" IF "Completely disabled (Karnofsky 10-20)"
    - "906" IF "Unknown"
    - "900" IF undefined

The ECOG score, also called the WHO or Zubrod score, runs from 0 to 5, with 0 denoting perfect health and 5 denoting patient deceased (Oken et al., 1982). It is considered preferable over the Karnofsky scale due to its simplicity. The Karnofsky

performance status scale allows patients to be categorized according to their functional impairment and is included in Appendix B. A lower Karnofsky score indicates a lower chance of survival for most serious illnesses (Hospice Patients Alliance, 2011).

a) 0 – Asymptomatic (Fully active, able to carry on all predisease activities without restriction)
b) 1 – Symptomatic but completely ambulatory (Restricted in physically strenuous activity but ambulatory and able to carry out work of a light or sedentary nature. For example, light housework, office work)
c) 2 – Symptomatic, <50% in bed during the day (Ambulatory and capable of all self care but unable to carry out any work activities. Up and about more than 50% of waking hours)
d) 3 – Symptomatic, >50% in bed, but not bedbound (Capable of only limited self-care, confined to bed or chair 50% or more of waking hours)
e) 4 – Bedbound (Completely disabled. Cannot carry on any self-care. Totally confined to bed or chair)
f) 5 – Patient deceased

- Code for the clinical TNM system stage indicating the absence or presence of distant metastasis: consisting of prefix 10 followed by digits 00 through 04:
    - "1001" IF No distant metastasis
    - "1002" IF Distant metastasis

  - "1003" IF No classification is recommended in 6th Edition of the TNM system (2003)
  - "1004" IF Distant metastasis cannot be assessed
  - "1000" IF Not available
- Code for the clinical TNM system stage indicating the absence or presence and existence of regional lymph node metastasis. Staging data are not available for non-referred cases (i.e., cases not referred to a BC cancer centre). Consists of prefix 11 followed by digits 00 through 04 and categories include:
  - "1101" IF No regional lymph node metastasis exists
  - "1102" IF Regional lymph node metastasis exists
  - "1103" IF No classification is recommended in 6th Edition of the TNM system (2003)
  - "1104" IF Regional lymph nodes cannot be assessed
  - "1100" IF Not available
- Code for the clinical TNM system stage indicating the extent of the primary tumor: consisting of prefix 12 followed by digits 00 through 07:
  - "1201" IF No evidence of primary tumor
  - "1202" IF Solitary tumor, greatest dimension (GD) <= 2cm without vascular invasion
  - "1203" IF Solitary tumor with vascular invasion, or multiple tumors, all <= 5cm
  - "1204" IF Solitary tumor, GD >2cm with vascular invasion, or multiple tumors, same lobe, without vascular invasion

- “1205” IF Multiple tumors, more than one lobe, portal/hepatic veins, other
- “1206” IF No classification is recommended in 6th Edition of the TNM system (2003)
- “1207” IF Primary tumor cannot be assessed
- “1200” IF Undefined

- Code for whether the patient had chemotherapy up to 3 months post-BCCA admission, if the information is known: consisting of prefix 13 followed by digits 00 through 06:
  - "1301" IF "none"
  - "1302" IF "initial treatment"
  - "1303" IF "subsequent treatment"
  - "1304" IF "both initial and subsequent"
  - "1305" IF "exact treatment unknown either initial or subsequent"
  - "1306" IF "unknown"
  - "1300" IF undefined
- Code for whether the patient had hormone therapy up to 3 months post-BCCA admission, if the information is known: consisting of prefix 14 followed by digits 00 through 04:
  - "1401" IF "none"
  - "1402" IF "initial treatment"
  - "1403" IF "subsequent treatment"
  - "1404" IF "both initial and subsequent"
  - “1400” IF undefined

- Code for whether the patient had radiation therapy, including pre-admission non-BCCA radiation therapy, if the information is known:
    - “1501”IF “no radiation therapy”
    - “1502”IF “BCCA radiation therapy”
    - “1503” IF “exact information unknown either BCCA or not”
    - “1500” IF undefined
- Code for whether the patient had diagnostic or other surgery up to 3 months post-BCCA admission, if the information is known:
    - "1601" IF "none"
    - "1602" IF "not BCCA"
    - "1603" IF "diagnostic only"
    - "1604" IF " exact information unknown either BCCA or not "
    - "1605" IF "unknown"
    - "1600" IF undefined
- Code for status at referral: consisting of prefix 17 followed by digits 00 through 04:
    - "1701" IF "follow-up"
    - "1702" IF "new patient"
    - "1703" IF "recurrence"
    - "1704" IF "residual disease"
    - "1700" IF undefined
- Year of first treatment: obtained by extracting the year only from the date of first treatment, which when present had a month-day-year format.

- Years between diagnosis and first treatment: calculated by subtracting the year of diagnosis from the year of treatment for the patients who did undergo treatment.
- Grouped death cause: created by locating and grouping similar or related primary causes of death and assigned each group a code:
    - "1801" IF the primary cause of death contains "malignant neoplasm" as well as well as "liver"
    - "1832" IF the primary cause of death contains "malignant neoplasm" but does not contain "liver"
    - "1833" IF the primary cause of death contains "neoplasm" as well as “multiple sites" but does not contain "liver"
    - "1835" IF the primary cause of death contains "myeloma"
    - "1834" IF the primary cause of death contains "neoplasm" as well as either "biliary tract" or “bile duct" or "gallbladder"
    - "1836" IF the primary cause of death contains "liver" as well as either "disease" or "abscess" or "disorders"
    - "1837" IF the primary cause of death contains "hepatic failure"
    - "1838" IF the primary cause of death contains "hepatoblastoma"
    - "1839" IF the primary cause of death contains "BILIARY TRAC-CA" or "GALLBLADDER-CA" or "EXTRHPTC B.D-CA" or "bile duct carcinoma"
    - "1840" IF the primary cause of death contains "CHOLEL&OTH" or "biliary tract" or "bile duct"

- "1841" IF the primary cause of death contains "neoplasm", but does not contain either "multiple sites" or "liver"
- "1802" IF the primary cause of death contains either "angiosarcoma" or "sarcomas" as well as "liver"
- "1803" IF the primary cause of death contains "HIV" or "Human Immunodeficiency Virus"
- "1804" IF the primary cause of death contains "hepatitis"
- "1805" IF the primary cause of death contains "Non-Hodgkin's lymphoma"
- "1806" IF the primary cause of death contains "lymphoma" but does contain "Non-Hodgkin's"
- "1807" IF the primary cause of death contains "Hodgkin's disease"
- "1808" IF the primary cause of death contains "VSA modified"
- "1809" IF the primary cause of death contains "lymphosarcoma"
- "1811" IF the primary cause of death contains "cardiovascular" or "myocarditis" or "cardiomyopathies" or "cardiac" or "cardiomyopathy" "endocarditis" or "myocardial" or"myoc."
- "1811" IF the primary cause of death contains "heart"or "atrial" or "aorta" or "mitral"
- "1812" IF the primary cause of death contains "thrombosis" or "thrombophlebitis"
- "1810" IF the primary cause of death contains "aneurysm" or "valve"

- "1813" IF the primary cause of death contains "cerebral" or "intracranial" or "intracerebral" or "cerebrovascular" or "brain" or "subdural" or "stroke" or "CERE"
- "1814" IF the primary cause of death contains "pneumonia"
- "1815" IF the primary cause of death contains "pneumonitis"
- "1816" IF the primary cause of death contains "pulmonary" or "pleural" or "respiratory" or "lung", death_cause_orig_desc
- "1817" IF the primary cause of death contains "PNCREAS" or "pancreas"
- "1818" IF the primary cause of death contains "gastric ulcer" or "peptic ulcer" or "duodenal ulcer" "gastrointestinal" or"gastroenteritis" or "intestine" or"diverticula"
- "1819" IF the primary cause of death contains "INTRHPTC B.D-CA" or "Intrahepatic bile duct carcinoma"
- "1820" IF the primary cause of death contains "septicaemia" or "septicemia"
- "1821" IF the primary cause of death contains "renal" or "haematuria" or "uropathy" or"urinary" or "nephritic" or "kidney" or "KDNY" or "PYELONEPHRI" or "pyelonephritis"
- "1822" IF the primary cause of death contains "CIRR" or "cirrhosis"
- "1823" IF the primary cause of death contains "atherosclerosis"
- "1824" IF the primary cause of death contains "diabetes"
- "1825" IF the primary cause of death contains "dementia"

- "1826" IF the primary cause of death contains "Varicose veins of lower extremities"
- "1827" IF the primary cause of death contains "mesothelioma"
- "1828" IF the primary cause of death contains "Disorders of Iron Metabolism"
- "1829" IF the primary cause of death contains "alcohol" or "alcoholic"
- "1830" IF "accident" or "accidental" or "drowning" or "fall" or "fire"
- "1842" IF the last three letters are –CA and none of the previous apply
- "1831" IF "intentional" or "suicide"
- "1899" IF the primary cause of death is blank
- "1800"

4. DATA ANALYSIS

4.1. Preliminary Data Analysis

The pre-processing data steps listed below are designed to better understand the data set. Tables and graphs plotting the results are given in the Results section.

- Summarizing the BC data set on study ID in order to remove the effect of duplicate records representing the same patient. The table produced this way contains a total of 6,047 records, i.e., one record per patient.
- Summarizing the BC data table with one record per patient on patient status (deceased or alive) and gender, and then calculating a life-death ratio (ratio of the number of patients alive to the number of patients deceased).
- Summarizing the BC data table with one record per patient on current health authority and patient status (deceased or alive) and then calculating a life-death ratio (ratio of the number of patients alive to the number of patients deceased) for each BC health authority where the patient currently lives.
- Summarizing the BC data table with one record per patient on diagnosis health authority and patient status (deceased or alive) and then calculating a life-death ratio (ratio of the number of patients alive to the number of patients deceased) for each BC health authority where the patient lived at the time of diagnosis. Some records were blank for the diagnosis health authority, which indicates that either the location at diagnosis was blank or that the exact location at diagnosis is unknown but is located in BC.

- Summarizing the Yukon data set on patient status (deceased or alive) and gender and then calculating a life-death ratio (ratio of the number of patients alive to the number of patients deceased).
- Stratifying the BC data table with one record per patient on diagnosis age groups, i.e. calculating the number of patients and the percentage for each age group of patients who are still alive.
- Stratifying the Yukon data set on diagnosis age groups and calculating the percentage for each age group of patients who are still alive.
- Stratifying the BC data table with one record per patient on diagnosis age groups and calculating the percentage for each age group of patients who are deceased.
- Stratifying the Yukon data set on diagnosis age groups and calculating the percentage for each age group of patients who are deceased.
- Summarizing the BC data set by treatment start time (in years) after diagnosis and patient status (deceased or alive).
- Summarizing the Yukon data set by treatment start time (in years) after diagnosis and patient status (deceased or alive).
- Summarizing the BC data set by each of the four treatment types (chemotherapy, hormone therapy, radiotherapy, surgery), individually (one table and one graph for each treatment type), and by patient status (deceased or alive).
- Summarizing the Yukon dataset by each of the four treatment types (chemotherapy, hormone therapy, radiation therapy, surgery), individually (one table and graph for each treatment type), and patient status (deceased or alive).

- Classifying the BC data table with one record per patient on the highest level used to confirm the patient's diagnosis, namely:
    - Autopsy
    - Cytology
    - Histology
    - Surgery (without histology), or clinical diagnosis
    - Radiology or laboratory diagnosis other than specified above
    - Death certificate only
    - For date of diagnosis in 2004 and onward
    - Method of diagnosis unknown
- Classifying the Yukon data set on the highest level used to confirm the patient's diagnosis, namely:
    - Cytology
    - Histology
    - Surgery (without histology), or clinical diagnosis
    - Death certificate only
    - For date of diagnosis in 2004 and onward
- Classifying the BC data table with one record per patient on the highest ICD-O-3 (International Classification of diseases for Oncology, third revision) histology code of the patient's distinct primary disease, which included the same categories for the Yukon data set (item above) in addition to several others

- Classifying the Yukon data set on the highest ICD-O-3 (International Classification of diseases for Oncology, third revision) histology code of the patient's distinct primary disease, which included:
    - Hepatocellular carcinoma, Not Otherwise Specified (NOS)
    - Neoplasm, malignant
    - Hepatoblastoma
    - No evidence of malignancy (SNOMED)
    - Adenocarcinoma, Not Otherwise Specified (NOS)
    - Cholangiocarcinoma
- Classifying the BC data table with one record per patient on the tumor group assigned to the patient's primary disease, which included:
    - Gastro-intestinal
    - Lymphoma
    - Melanoma
    - Sarcoma
- Classifying the Yukon data set on the tumor group assigned to the patient's primary disease, which included only Gastro-intestinal
- Classifying the BC data table with one record per patient on the tumor subgroup assigned to the patient's primary disease, which included:
    - Liver
    - Hodgkin's disease
    - Non-Hodgkin's
    - Other

  - Unknown

- Classifying the Yukon data set on the tumor subgroup assigned to the patient's primary disease, which included only Liver
- Classifying the BC data on the clinical TNM system stage indicating the extent of the primary tumor. Staging data are not available for non-referred cases (i.e., cases not referred to a BC cancer centre). Categories include:
  - Solitary tumor, GD (greatest dimension) <= 2cm without vascular invasion
  - Solitary tumor with vascular invasion, or multiple tumors, all <= 5cm
  - Solitary tumor, GD >2cm with vascular invasion, or multiple tumors, same lobe, without vascular invasion
  - Multiple tumors, more than one lobe, portal/hepatic vein, other or…
  - Primary tumor cannot be assessed
  - No evidence of primary tumor
  - No classification is recommended in 6th Edition of the TNM system (2003)
  - Not available
- Classifying the Yukon data on the clinical TNM system stage indicating the extent of the primary tumor. Staging data are not available for non-referred cases (i.e., cases not referred to a BC cancer centre). Categories include:
  - Multiple tumors, more than one lobe or portal/hepatic vein invasion
  - Primary tumor cannot be assessed
  - Not available

- Classifying the BC data set on the clinical TNM system stage indicating the absence or presence of distant metastasis. Staging data are not available for non-referred cases (i.e., cases not referred to a BC cancer centre). Categories include:
    - Distant metastasis
    - No distant metastasis
    - Distant metastasis cannot be assessed
    - No classification is recommended in 6th Edition of the TNM system (2003)
    - Not available
- Classifying the Yukon data set on the clinical TNM system stage indicating the absence or presence of distant metastasis. Staging data are not available for non-referred cases (i.e., cases not referred to a cancer centre). Categories include:
    - Distant metastasis
    - Distant metastasis cannot be assessed
    - Not available
- Classifying the BC data set on the clinical TNM system stage indicating the absence or presence and existence of regional lymph node metastasis. Staging data are not available for non-referred cases (i.e., cases not referred to a BC cancer centre). Categories include:
    - Regional lymph node metastasis
    - No regional lymph node metastasis
    - Regional lymph nodes cannot be assessed

  - No classification is recommended in 6th Edition of the TNM system (2003)
  - Not available

- Classifying the Yukon data set on the clinical TNM system stage indicating the absence or presence and existence of regional lymph node metastasis. Staging data are not available for non-referred cases (i.e., cases not referred to a BC cancer centre). Categories include:
  - Regional lymph node metastasis
  - Regional lymph nodes cannot be assessed
  - Not available
- Classifying the BC data set on original cause of death group and calculating the percentage each group represents of the total cause of death

4.2. Selection of Attributes for Data Mining

Feature extraction, by excluding less informative attributes, was performed in order to improve the performance of the algorithm (Karabatak & Ince, 2009b; Kirsnhers, Parshutin & Leja, 2012). This was accomplished by selecting a final data set of lower dimension to work with, based on the results of the preliminary analysis. The list of selected attributes consisted of:

1. Sex_code: Indicates patient's gender.
2. Diagnosis_age_code: Indicates the age group of the patient when diagnosed.

3. Curr_hlth_auth: The health authority code of the patient's current BC home address postal code. Does not apply to Yukon patients.
4. Dx_h_auth_code: The health authority code of the patient's BC postal code at the time of diagnosis. Undefined indicates that either the location at diagnosis was blank or that the location at diagnosis is in BC but the exact location is unknown.
5. Death_age_code: Indicates the age group of the patient at the time of death.
6. Years_diag_to_death: The number of years between death and diagnosis.
7. Pat_status_code: Indicates the patient's status of 'Alive' or 'Deceased'.
8. Perform_status_code: The patient's ECOG performance status code.
9. Death_cause_original: The BC Vital Statistics primary cause of death.
10. Site: The ICD-O site code for the patient's distinct primary disease: C220 (liver) and C221 (intrahepatic bile duct).
11. Hist1: The highest ICD-O histology code of the patient's distinct primary disease. Includes all ICD-O-3 and SNOMED histologies.
12. Tumour_group_code: Code for the tumour group assigned to the patient's primary disease.
13. Tumor_subgrp_code: Code for the subgroup of the tumour group assigned to the patient's primary disease.
14. Tnm_clin_m_code: Code for the clinical TNM system stage indicating the absence or presence of distant metastasis.
15. Tnm_clin_n_code: Code for the clinical TNM system stage indicating the absence or presence and existence of regional lymph node metastasis.

16. Tnm_clin_t_code: Code for the clinical TNM system stage indicating the extent of the primary tumour.
17. Bcca_chemo_code: Code for the flag indicating the patient had chemotherapy up to 3 months post-BCCA admission.
18. Bcca_rad_code: Code indicating the patient had radiotherapy.
19. Bcca_surg_code: Code indicating the patient had surgery.

- Splitting the data set into the two subsets below:

A. Preliminary information:

This data set is meant to identify which factors lead to an increased incidence of liver cancer. It consists of the following attributes:

1. Sex_code
2. Diagnosis_age_code
3. Curr_hlth_auth (BC only)
4. Dx_h_auth_code (BC only)
5. Site
6. Hist1
7. Tumor_subgrp_code (BC only)

For the Yukon patients, the tumor subgroup code was not included as it consisted solely of liver cancer. The clinical TNM system stage attributes (Tnm_clin_m_code, Tnm_clin_n_code, Tnm_clin_t_code) and performance status had originally been included in the prior data set. After an initial run through the FPGrowth algorithm, they were removed as they generated a high level of meaningless associations. Any attributes

that had a large number of unknown or undefined values were also left out of the above data set.

B. Survivability:

This data set is meant to identify which factors are associated with a patient status of alive or deceased. It consists of the following attributes:

1. Sex_code
2. Diagnosis_age_code
3. Curr_hlth_auth (BC only)
4. Dx_h_auth_code (BC only)
5. Death_age_code
6. Pat_status_code
7. Site
8. Hist1
9. Tumor_subgrp_code (BC only)
10. Bcca_chemo_code (BC only)
11. Bcca_rad_code (BC only)
12. Bcca_surg_code (BC only)

For the BC data set, two data subsets were produced and processed, in order to verify whether feature selection can indeed increase efficiency in the system (Kirsnhers, Parshutin & Leja, 2012):

a) One removing all records where treatment (chemotherapy, radiotherapy, or surgery) was undefined. The resulting data set consisted of 1675 records.
b) Another removing all treatment attributes mentioned above: Bcca_chemo_code, Bcca_rad_code, Bcca_surg_code. This retained all 6047 patient records.

- Creating five comma-delimited text files, keeping only the data attributes (encoded numerically for the FP-Growth algorithm as previously described) required for each analysis:
    - BC preliminary
    - Yukon preliminary
    - BC survival without undefined treatment records
    - BC survival without treatment attributes
    - Yukon survival
- Submitting each of the comma-delimited text files with numerically encoded attributes through the FP-Growth algorithm using varying levels of support repeatedly from the command line until the best minimum level of support was determined. The support for each rule was not kept constant, in contrast to the implementation of the Apriori algorithm where it was.
- Creating another five comma-delimited text files, keeping only the data attributes required for each analysis, but this time in their original nominal form:
    - BC preliminary
    - Yukon preliminary
    - BC survival without undefined treatment records

  - BC survival without treatment attributes
  - Yukon survival

- Submitting each of the comma-delimited text files with nominal attributes through the Apriori algorithm using the WEKA data mining application with a confidence level of 0.85. WEKA's Apriori starts with minimum support set at 100% of the dataset items and decreases it in steps of 5% until it either obtains at least 10 rules with the minimum confidence specified or until support reaches a lower bound of 10% of the data set items, whichever occurs first (Witten & Frank, 2005). The support for all the rules is the same, since they are submitted in a batch and controlled by the software, in contrast to the FP-growth algorithm implementation, where everything is done manually from the command line and the support was not constant for all rules.

5. RESULTS:

In the Yukon, each patient had one single record. In BC, on the other hand, 235 patients had more than one record.

In BC, one of the 6,064 patients, patient #4398, had 3 records. This patient (#4398) had a confirmed diagnosis of cancer in two separate liver sites. The first diagnosis was in 1993, when the patient was 38, whereas the second was in 2004, when he was 49. In both occasions, the performance status of the patient was classified as fully active (90-100 on the Karnofsky scale). The highest ICD-O histology code at the time of each diagnosis was, respectively, Cholangiocarcinoma (C22.1, C24.0, 81603) and Hepatocellular carcinoma, NOS (81703). The treatment following the first diagnosis was not stated / recorded. The treatment after the second diagnosis consisted of both surgery and radiotherapy, although it is not known whether either of those was performed by BCCA or not. Two palliative surgical interventions were done, one in January 2005, one in February 2006. The latter one involved liver transplantation. Radiotherapy was done on the thoracic spine in February – March, 2005. The patient unfortunately passed away in 2006 and the primary cause of death was recorded as intrahepatic bile duct carcinoma (C221).

In BC, 4,173 liver cancer patients were men, compared to 1,874 women, which gives an incidence rate of liver cancer in men being 2.226 that of women. However, the life-death ratio, calculated as the number of living to deceased patients (Figure 2), is the same, 0.10, for both men and women.

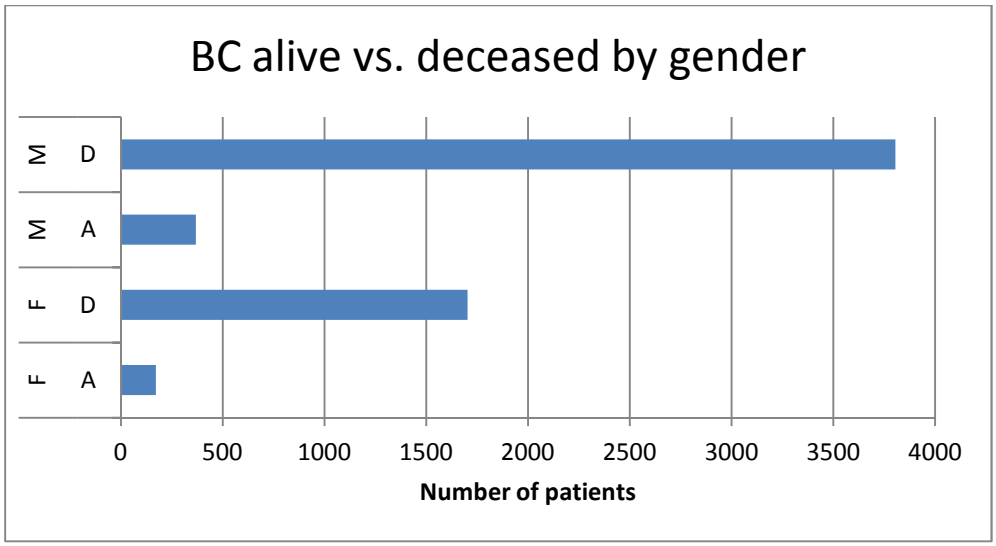

Figure 2. Patients living (A) vs. deceased (D) in BC by gender (F= female, M= male)

In the Yukon, incidence of liver cancer in men was about three times as much as in women. The life-death ratio is 1.00 for women (2 patients deceased, 2 patients alive) and 0.30 for men (10 patients deceased, 3 patients alive).

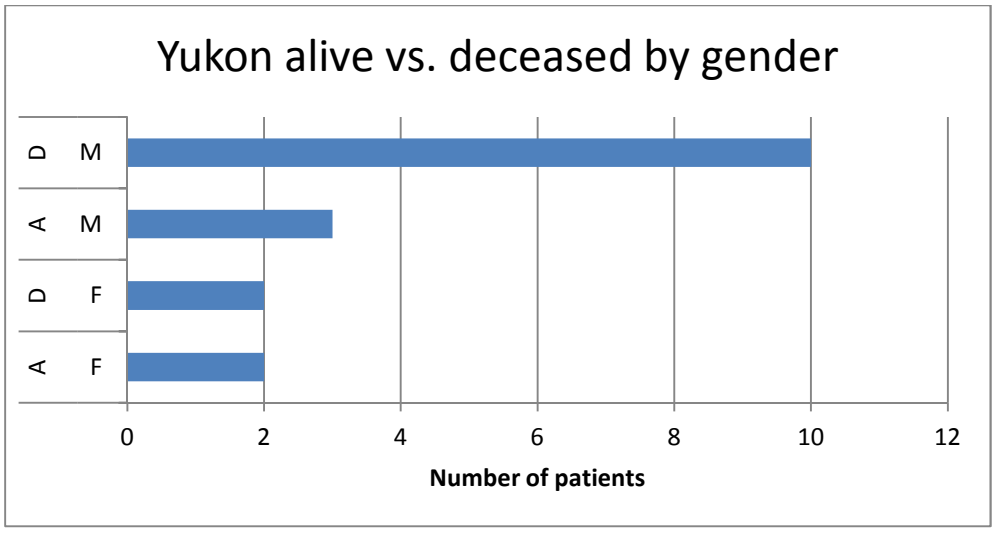

Figure 3. Yukon patients living (A) vs. deceased (D) by gender (F= female, M= male)

Liver cancer incidence in BC is highest for in the Vancouver Coastal and Vancouver Island Health Authorities (Figure 4), although the Fraser Health Authority actually has the largest population (Figure 5).

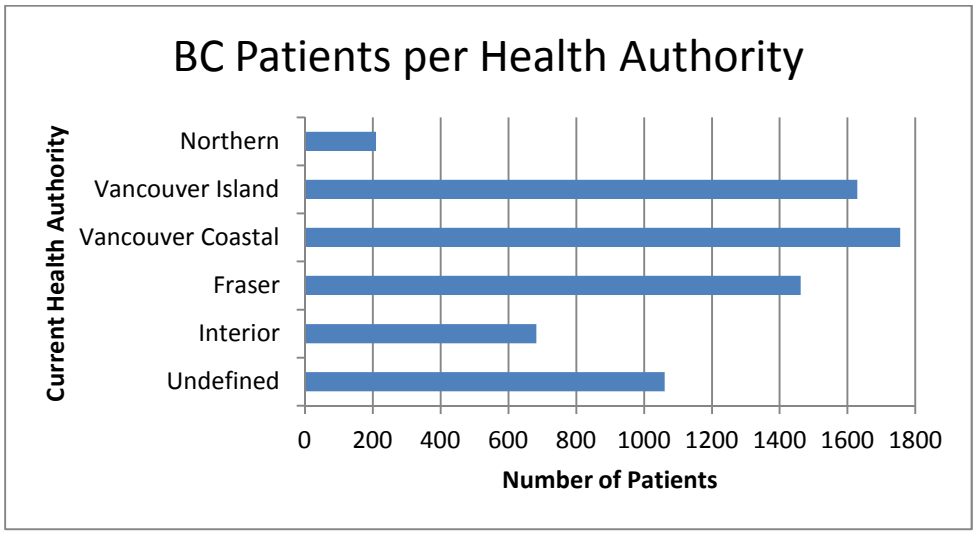

Figure 4. Number of patients per current Health Authority in BC

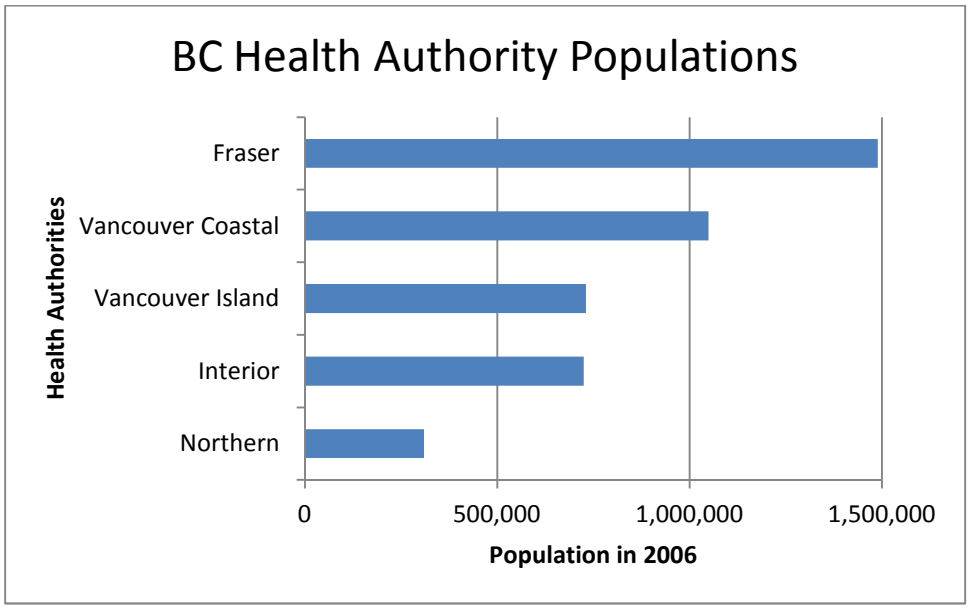

Figure 5. Health authority populations (from British Columbia Ministry of Health, 2006)

Life-death ratio in BC is markedly greater in the Greater Vancouver Area, which includes both Vancouver Coastal and Fraser Health Authorities (Figures 6 and 7).

The Yukon data was not organized in Health Authorities.

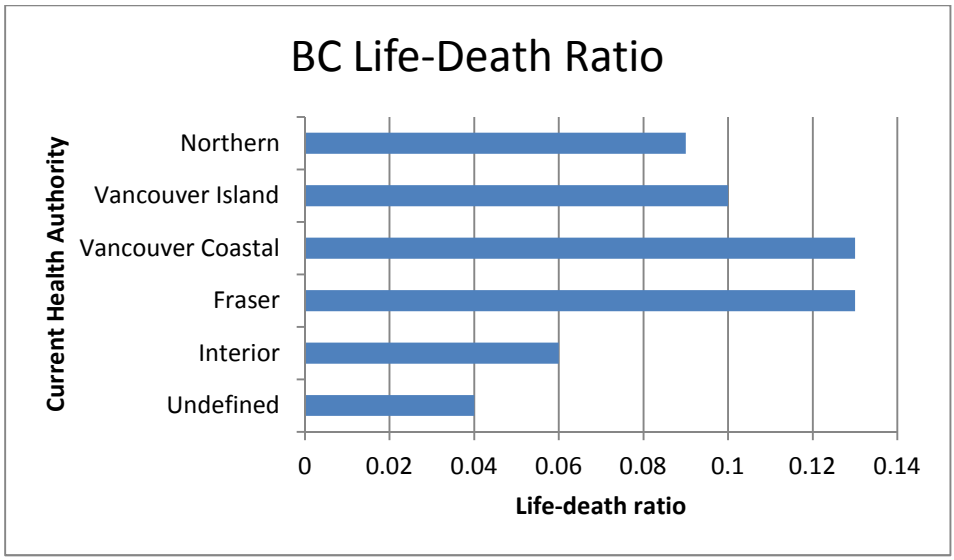

Figure 6. BC Life-death ratio by patient's current Health Authority.

Current Health Authority	Patient Status	Number of patients
Undefined	Alive	36
Undefined	Deceased	1025
Interior	Alive	39
Interior	Deceased	643
Fraser	Alive	167
Fraser	Deceased	1295
Vancouver Coastal	Alive	203
Vancouver Coastal	Deceased	1552
Vancouver Island	Alive	77
Vancouver Island	Deceased	800
Northern	Alive	17
Northern	Deceased	193

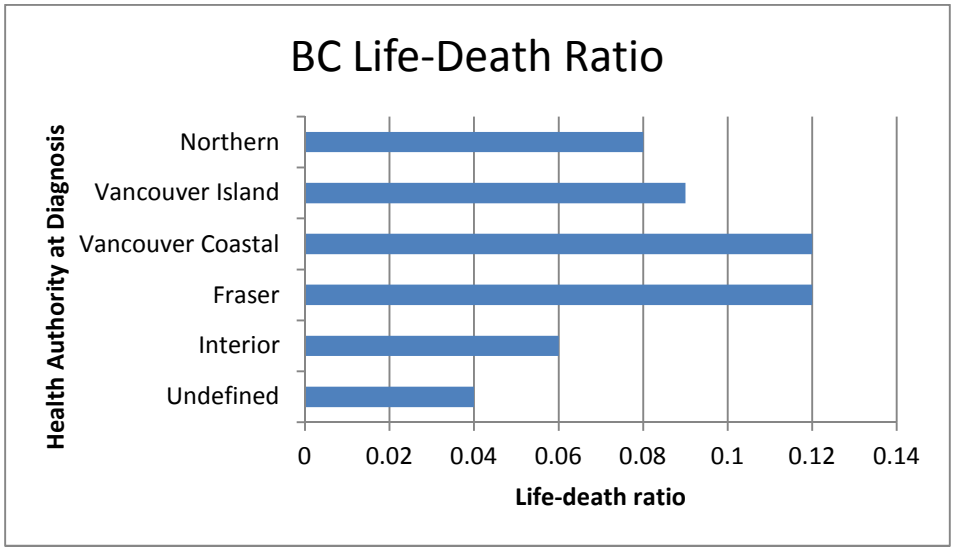

Figure 7. BC Life-death ratio by Health Authority of diagnosis.

In BC, the patients who are alive were diagnosed at a younger age than those who are deceased (Figure 8). In the Yukon, the trend is the opposite (Figure 10).

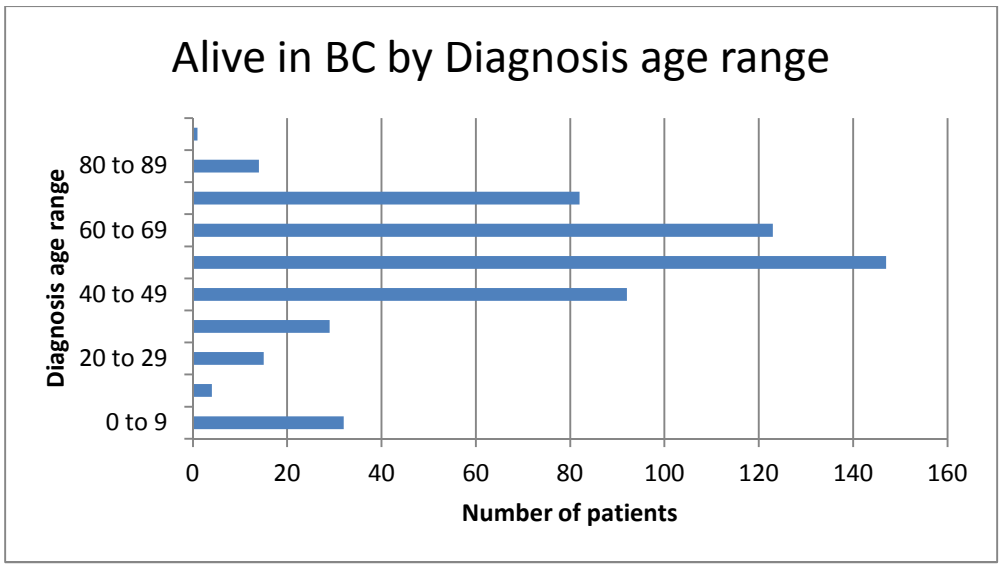

Figure 8. Number of living patients in BC by diagnosis age range

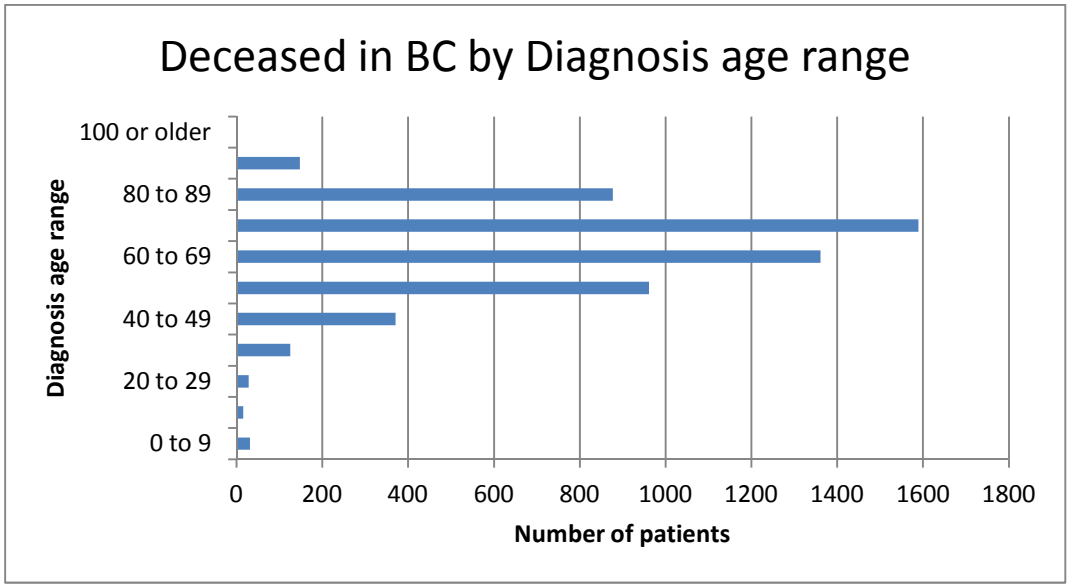

Figure 9. Number of deceased patients in BC by diagnosis age range

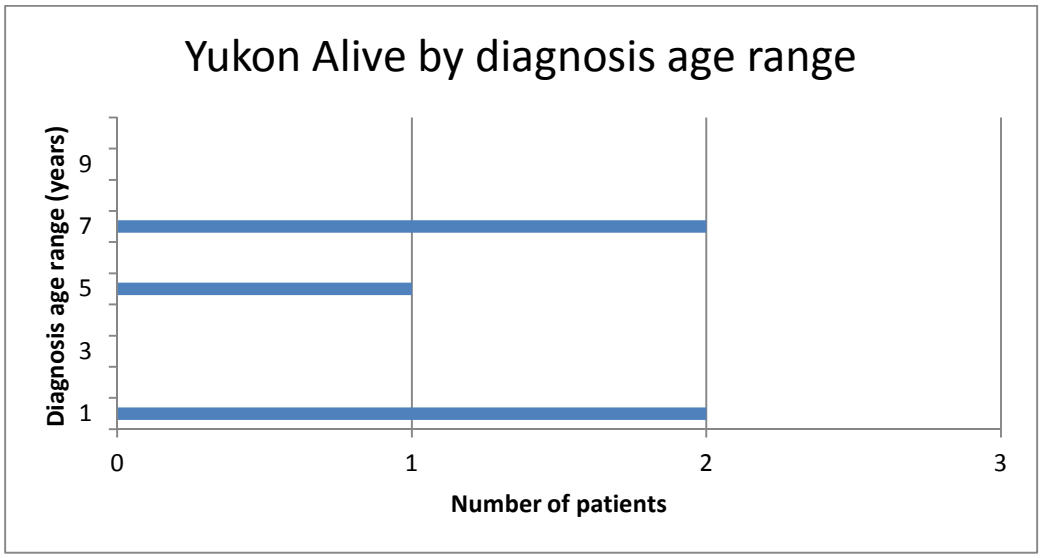

Figure 10. Number of living patients in the Yukon by diagnosis age range

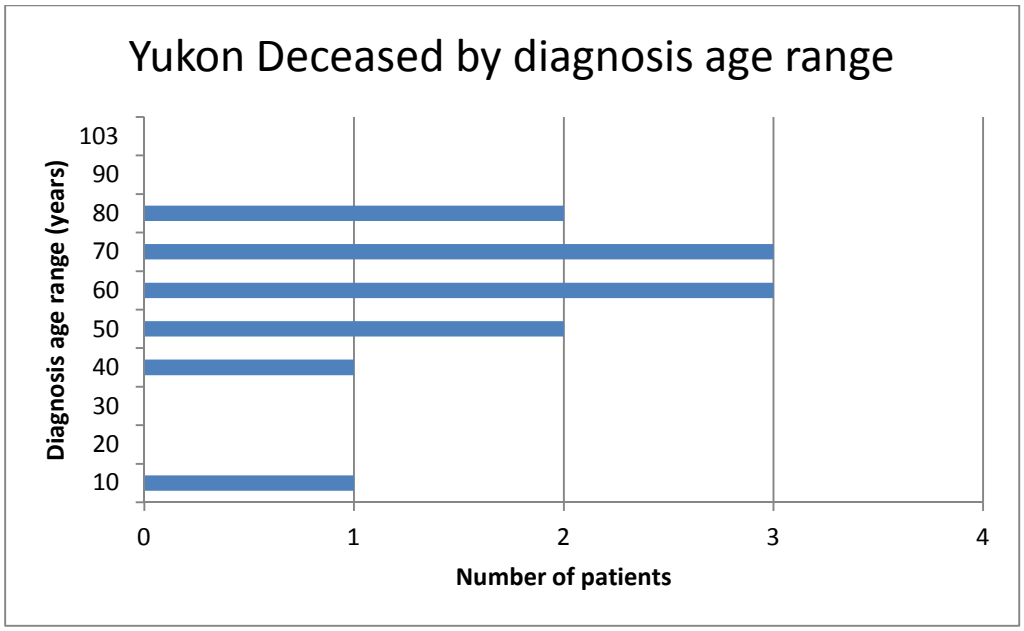

Figure 11. Number of deceased patients in the Yukon by diagnosis age range

In BC, 229 patients are known to have had radiotherapy, including pre-admission non-BCCA radiotherapy. Only 3 knowingly had radiotherapy through BCCA (Figure 12).

Undefined	U	4117
None	0	1701
BCCA	1	3
Exact info unknown either BCCA or not	7	226

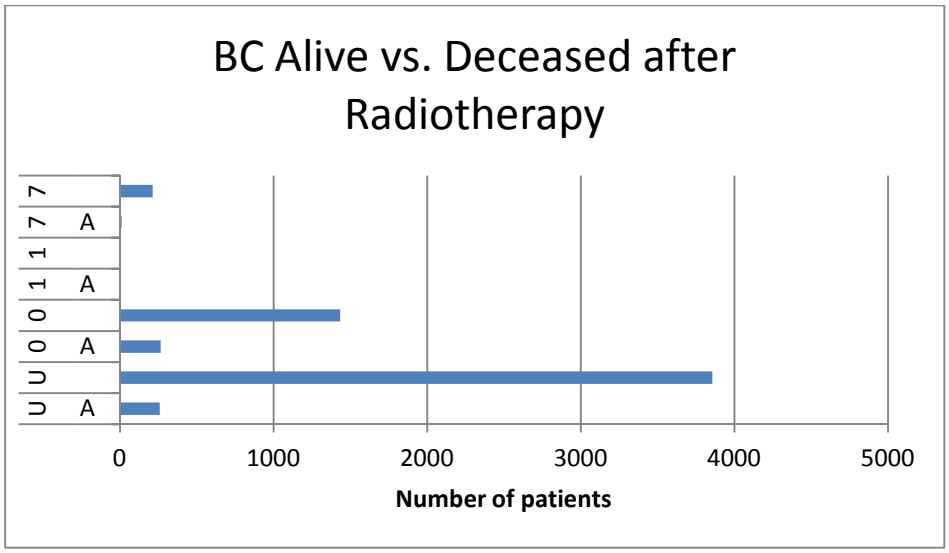

Figure 12. Number of patients living (A) vs. deceased (D) in BC after radiotherapy (0 = No Radiotherapy; 1 = BCCA Radiotherapy; 7 = Unknown whether BCCA Radiotherapy; U = Undefined)

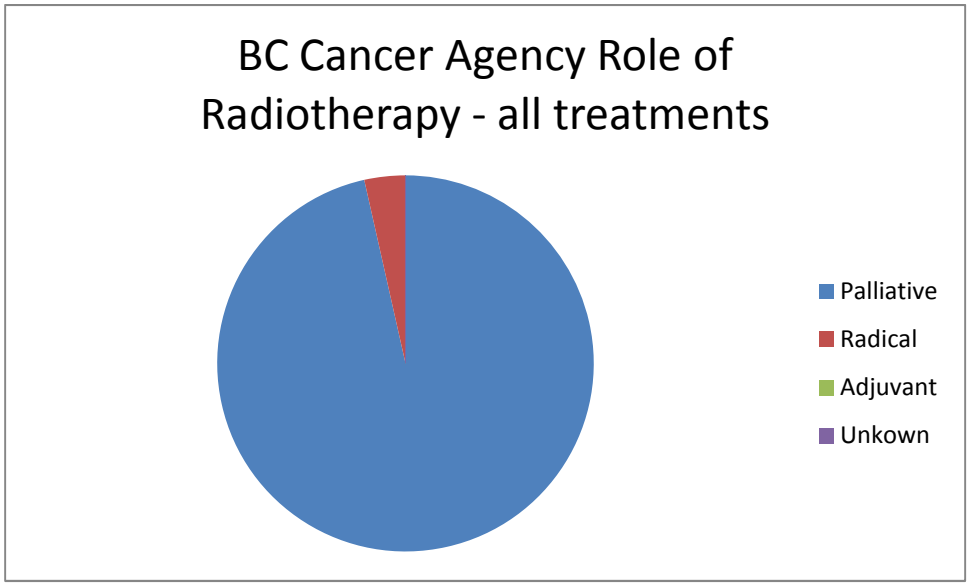

Figure 13. Role of radiotherapy, all treatments included, in BC

Of all BC patients who underwent radiotherapy, the intent of radiotherapy was palliative for 92.86% of those patients and radical for the remainder 7.14% (Figure 14).

Nowadays, palliative care is recommended in all chronic, potentially fatal illnesses, as symptoms not treated early on become difficult to manage as the disease progresses (Sepulveda et al., 2002). Palliative radiotherapy, however, is relevant only to patients not responsive to curative treatment (Sepulveda et al., 2002; Hung, 2007), with about 34 – 50% of patients treated with radiotherapy receiving that treatment with palliative intent (Hung, 2007).

Low-dose palliative radiotherapy is also beneficial to patients with symptomatic diffuse liver metastases, regardless of liver function (Lock et al., 2012). The main goals of palliative radiotherapy are to control the symptoms, to enhance the patient's quality of life, and, hopefully, to positively influence the course of illness and optimize the patient's remaining time (Sepulveda et al., 2002; Hung, 2007).

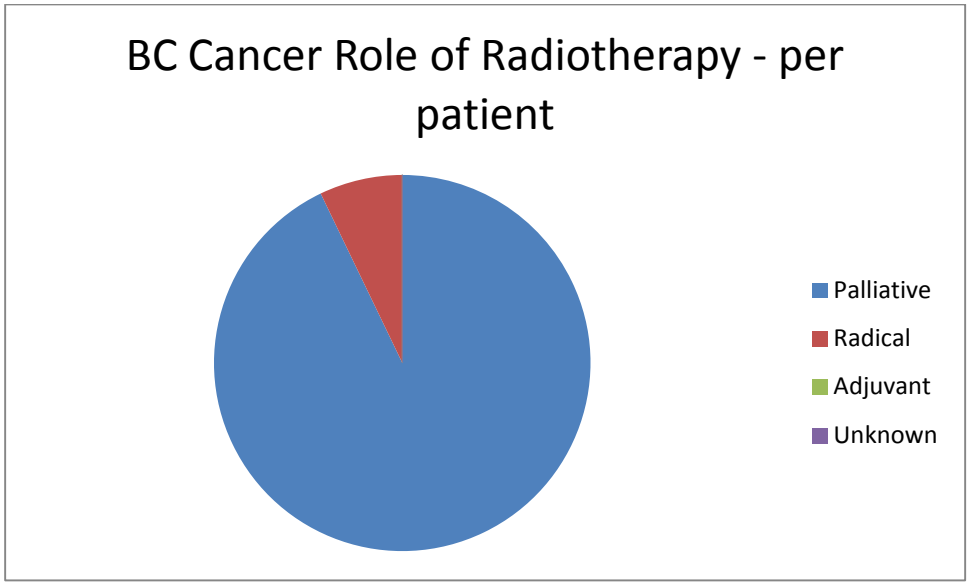

Figure 14. Role of radiotherapy, per patient, in BC

According to Lock et al. (2012), the use of radical intent radiotherapy varies globally between regions, from 90% in the United States and 55% in Europe to 42% in Canada and 9% in Australia and New Zealand.

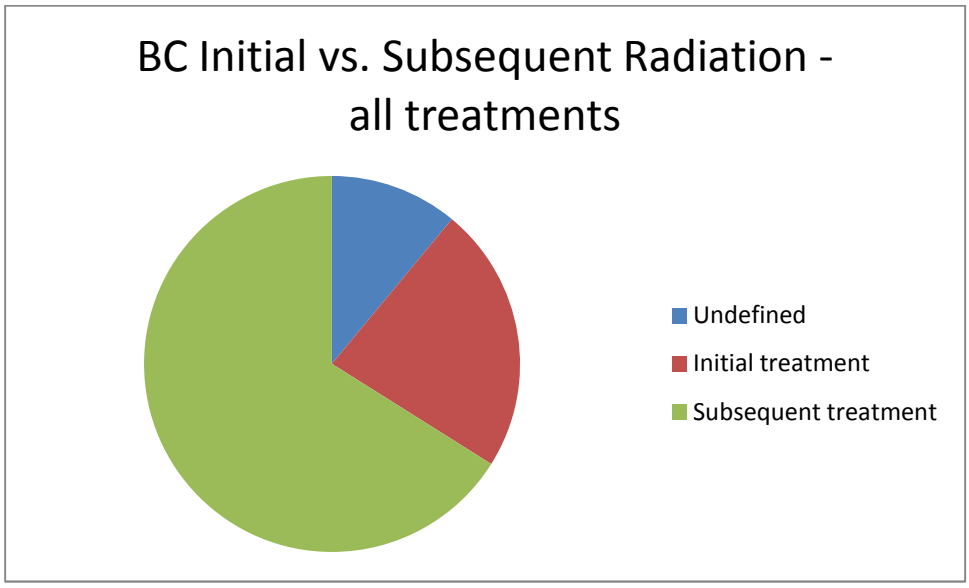

Figure 15. Initial vs. subsequent radiation, all treatments, in BC

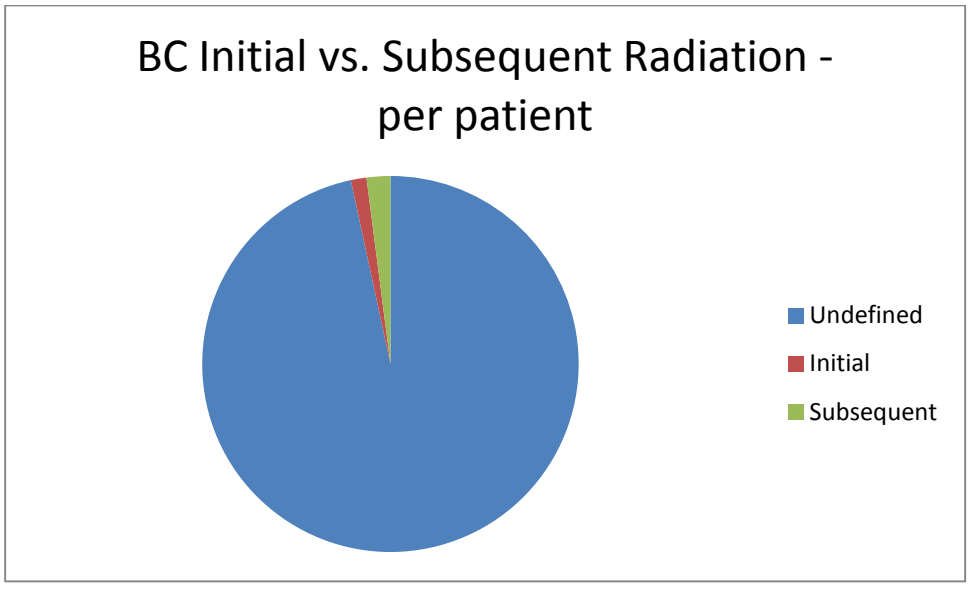

Figure 16. Initial vs. subsequent radiation, per patient, in BC

The radiation treatments ranged in length from a minimum of 1 day to a maximum of 34 days. Per patient, the number of radiation treatments ranged between a minimum of 1 and a maximum of 8, which corresponded to an accumulated length of treatment of 47 days.

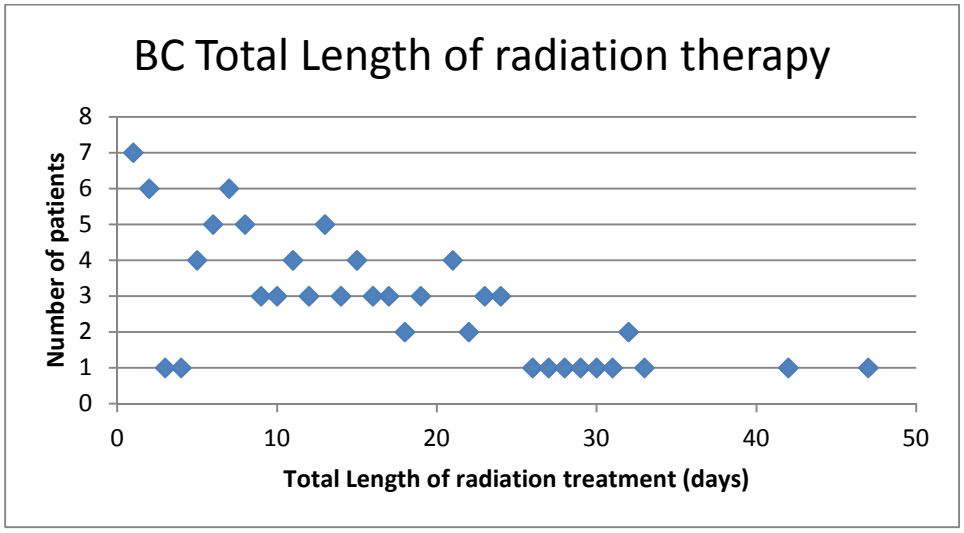

Figure 17. Total length of radiation therapy, in BC

Only one Yukon patient had radiotherapy and there are insufficient details about that treatment instance (Figure 18).

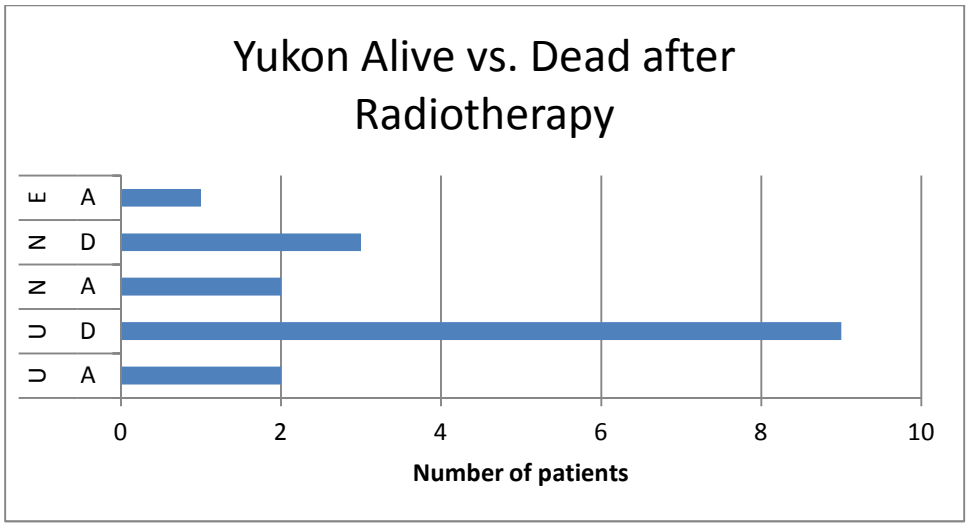

Figure 18. Yukon patients living (A) vs. deceased (D) after radiotherapy

Only 11 BC patients are known as having had hormone therapy up to three months after BCCA admission:

- Eight patients (522, 3435, 3948, 4303, 4512, 4555, 4757 and 4824) had initial hormone treatment only.
- Patients 554 and 4280 are recorded as having had subsequent hormone treatment only, with 4280 having two subsequent treatments on record.
- Patient 2754 is recorded as having had both initial and subsequent hormone treatments (Figure 19).

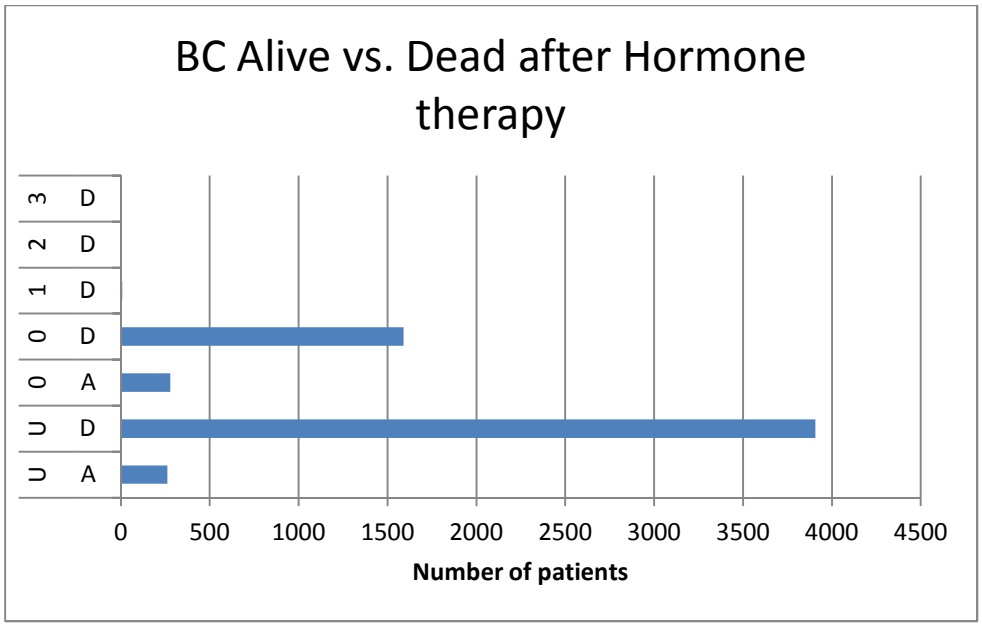

Figure 19. Number of patients living (A) vs. deceased (D) in BC after hormone therapy (0 = No Treatment; 1 = Initial Treatment; 2 = Subsequent Treatment; 3 = Both Initial and Subsequent Treatments; U = Undefined)

U	Undefined	4168
0	None	1868
1	Initial treatment	8
2	Subsequent treatment	2
3	Both Initial and subsequent	1

No Yukon patient is known to have had hormone therapy (Figure 20).

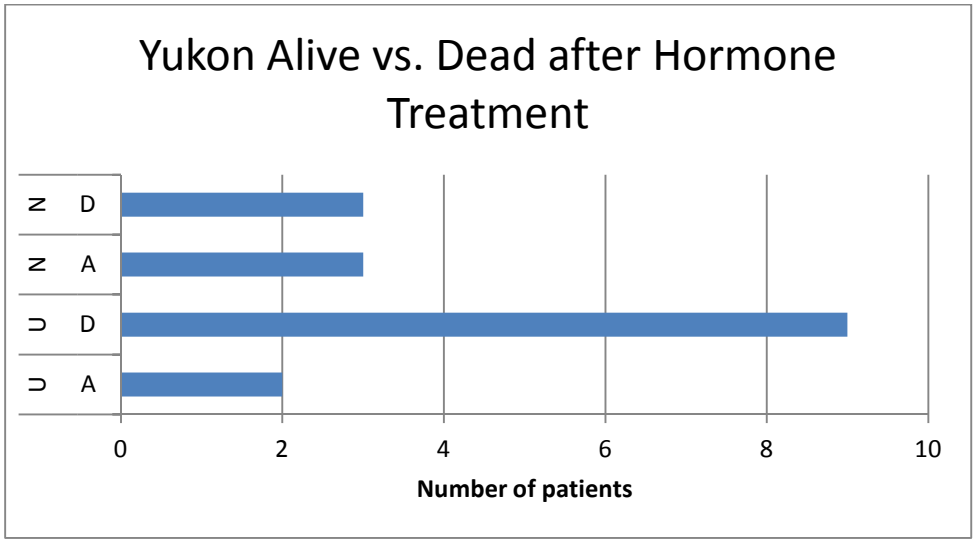

Figure 20. Number of patients living (A) vs. deceased (D) in the Yukon after hormone therapy (N = No hormone treatment; U = Undefined)

In BC, 672 liver cancer patients have been recorded as having had chemotherapy up to three months after BCCA admission. Of those, only 10 patients had both initial and subsequent treatments, whereas 320 had only initial treatment (Figure 21).

Undefined	U	3904
None	0	1467
Initial treatment	1	320
Subsequent treatment	2	42
Both Initial and subsequent	3	10
Unknown either initial treatment or subsequent	7	300
Unknown	99	4

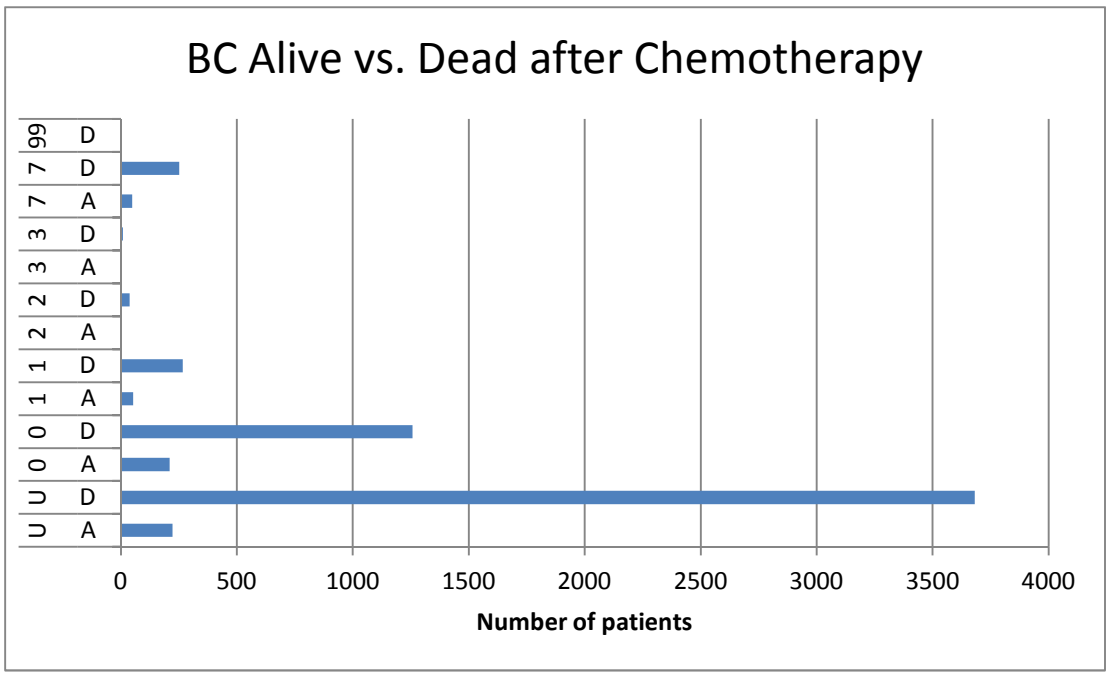

Figure 21. Number of patients living (A) vs. deceased (D) in BC after chemotherapy

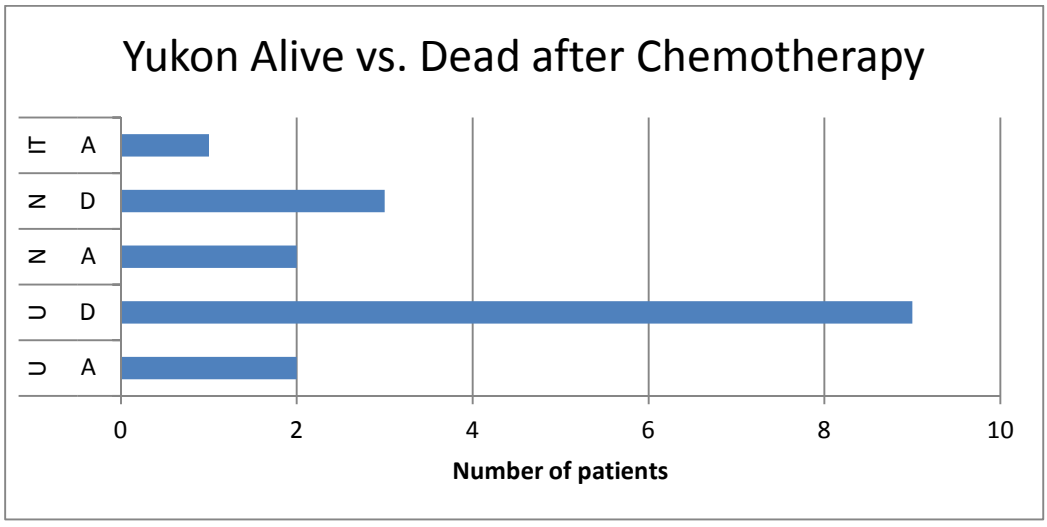

Figure 22. Number of patients living (A) vs. deceased (D) in the Yukon after chemotherapy (IT = Initial Treatment; N = No Treatment; U = Undefined)

Only one Yukon patient is known for having had chemotherapy (Figure 22).

In BC, 900 patients are known to have had surgery, but it was either only diagnostic surgery, or not known to have been performed by BCCA (Figure 23).

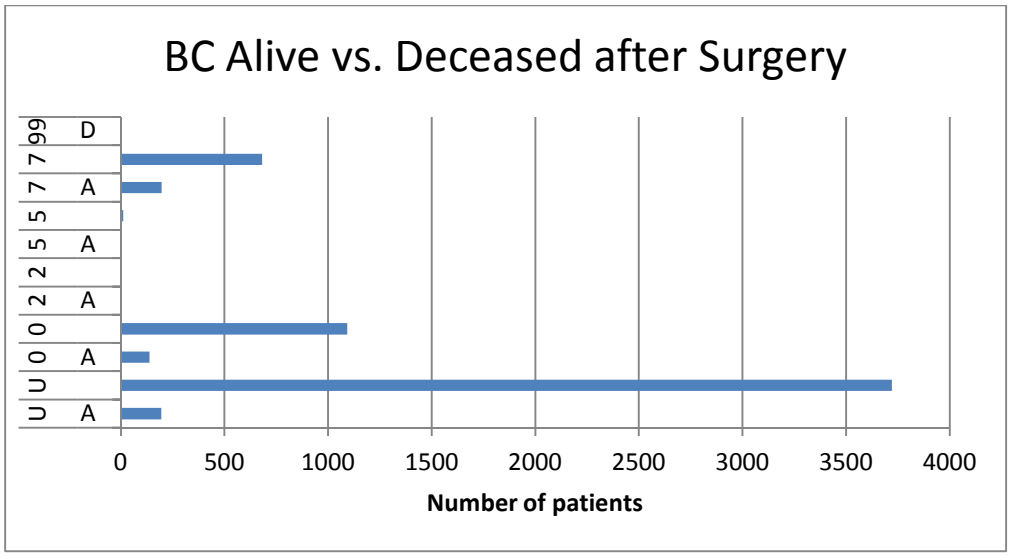

Figure 23. BC patients living (A) vs. deceased (D) after surgery (0 = No Surgery; 2 = Surgery but not with BCCA; 5 = Diagnostic surgery only; 7 = Surgery but unknown whether with BCCA or not; 99 = Unknown; U = Undefined)

Undefined	U	3916
None	0	1230
Not BCCA	2	6
Diagnostic only	5	16
Exact info unknown either BCCA or not	7	878
Unknown	99	1

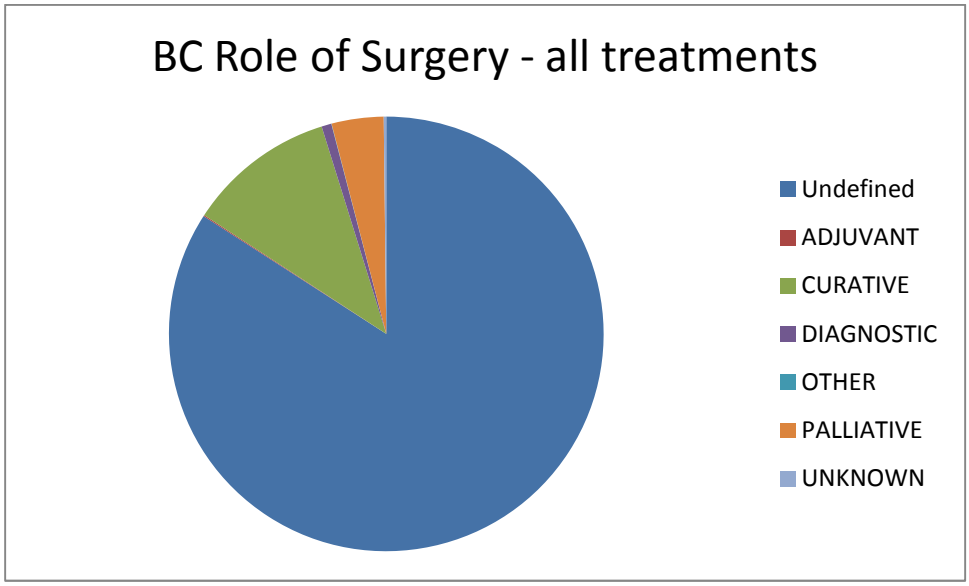

Figure 24. Role of surgery, all treatments included, in BC

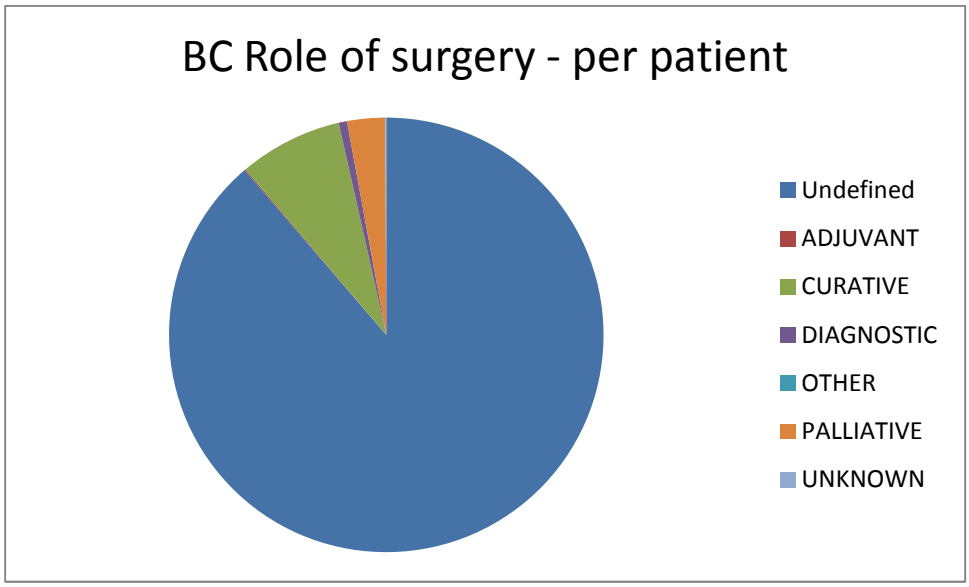

Figure 25. Role of surgery, per patient, in BC

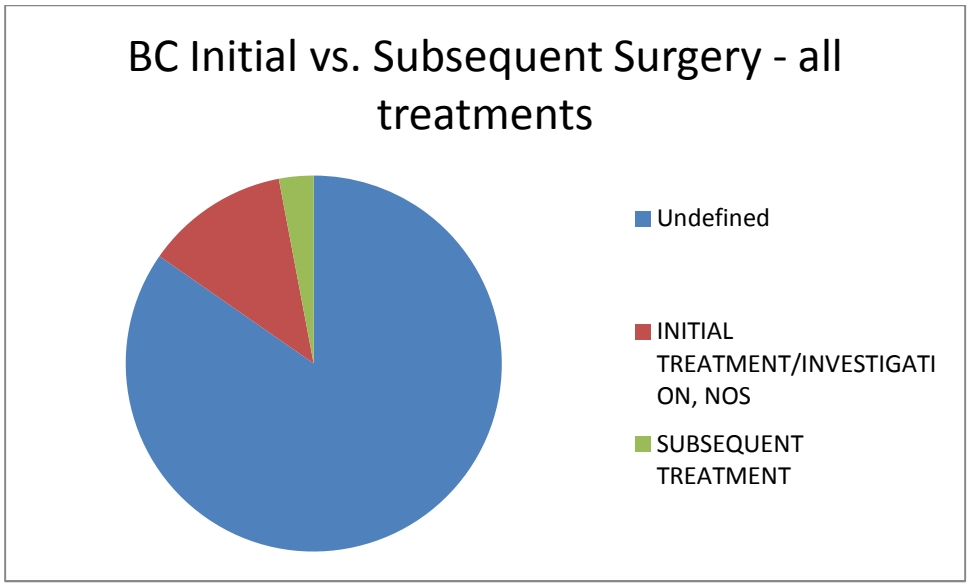

Figure 26. Initial vs. subsequent surgery, all treatments, in BC

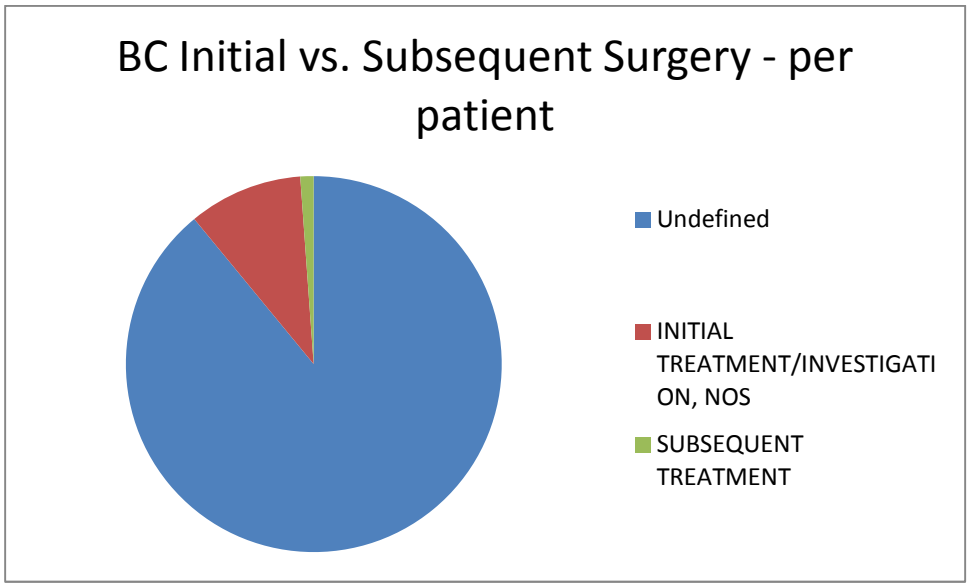

Figure 27. Initial vs. subsequent surgery, per patient, in BC

Only two Yukon patients had surgery and not enough details are known about it (Figure 28).

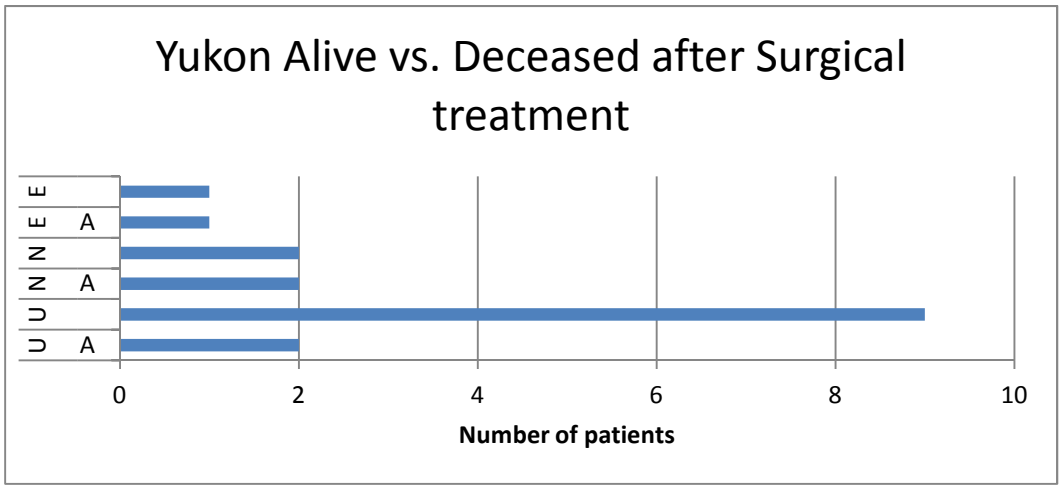

Figure 28. Number of patients living (A) vs. deceased (D) in BC after surgery (E = Unknown whether surgery was with BCCA; N = No Surgery; U = Undefined)

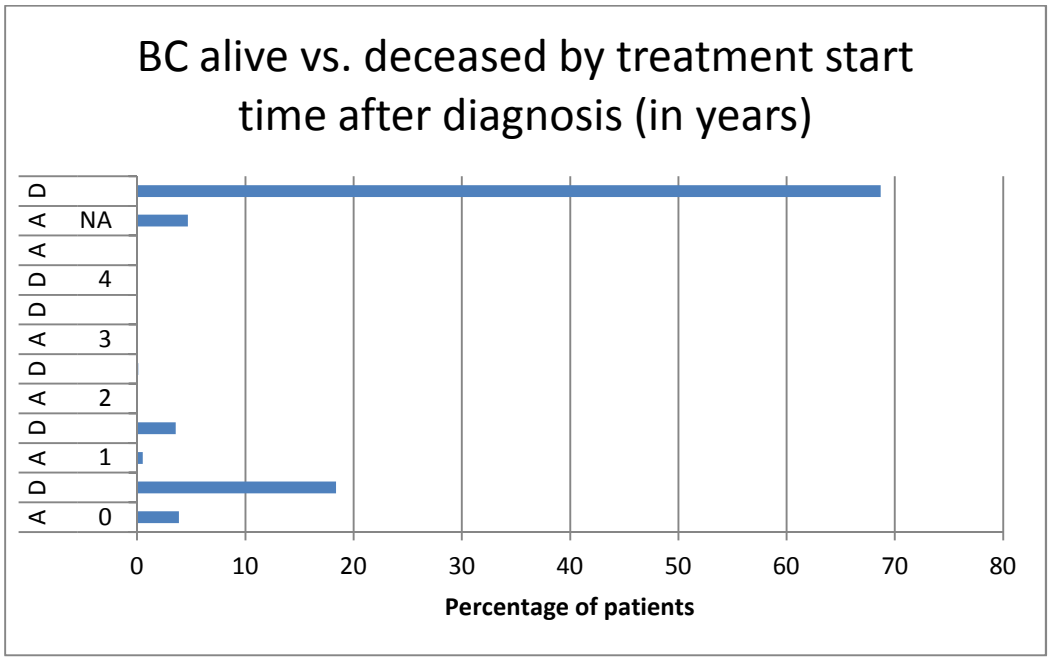

Figure 29. Number of patients living (A) vs. deceased (D) in BC by treatment start time after diagnosis (in years)

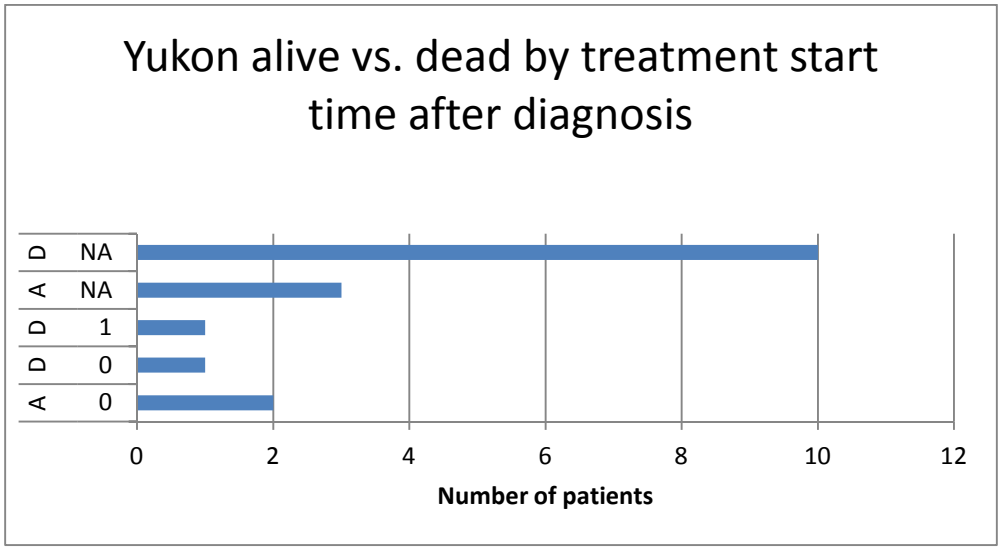

Figure 30. Number of patients living (A) vs. deceased (D) in the Yukon by treatment start time after diagnosis (in years)

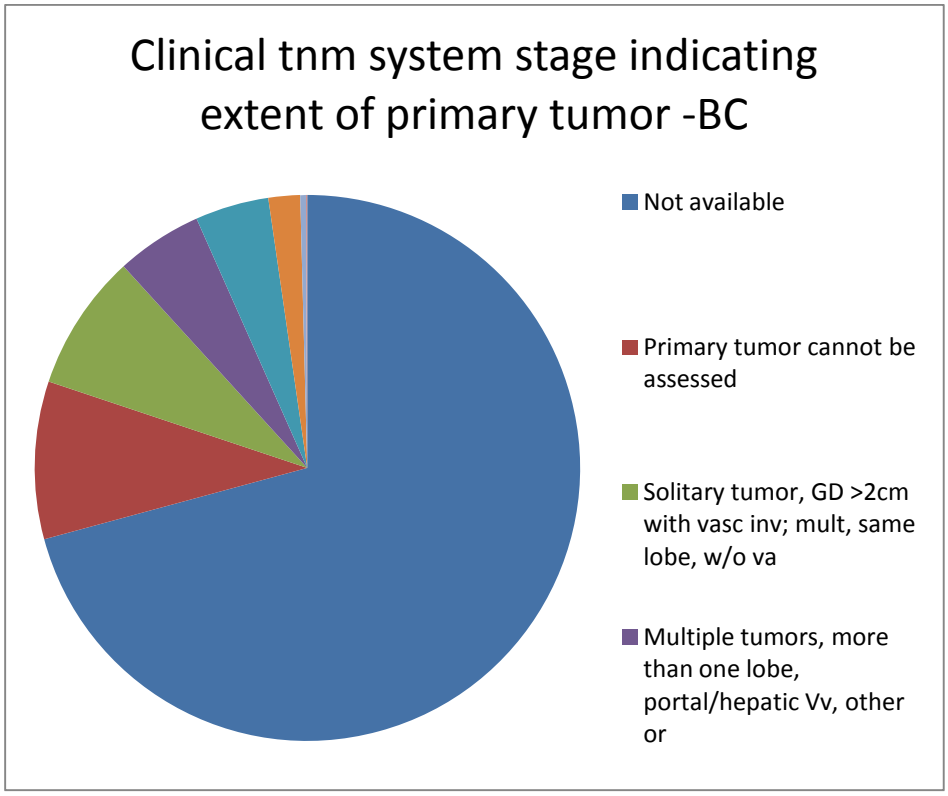

Figure 31. BC Clinical tnm system indicating the extent of the primary tumor

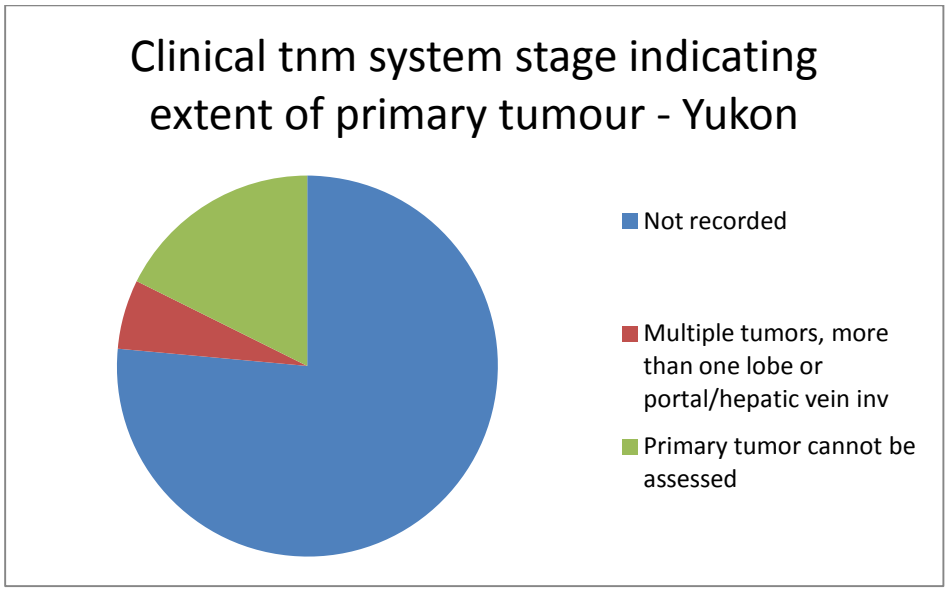

Figure 32. Yukon Clinical tnm system indicating extent of primary tumor

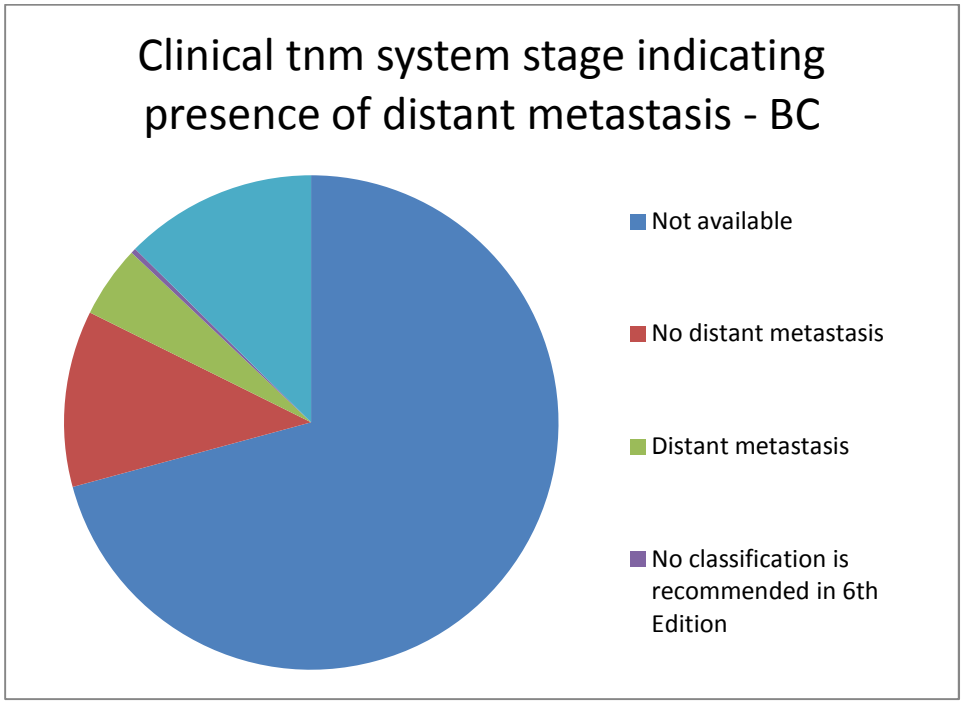

Figure 33. BC Clinical tnm system indicating the absence or presence of distant metastasis

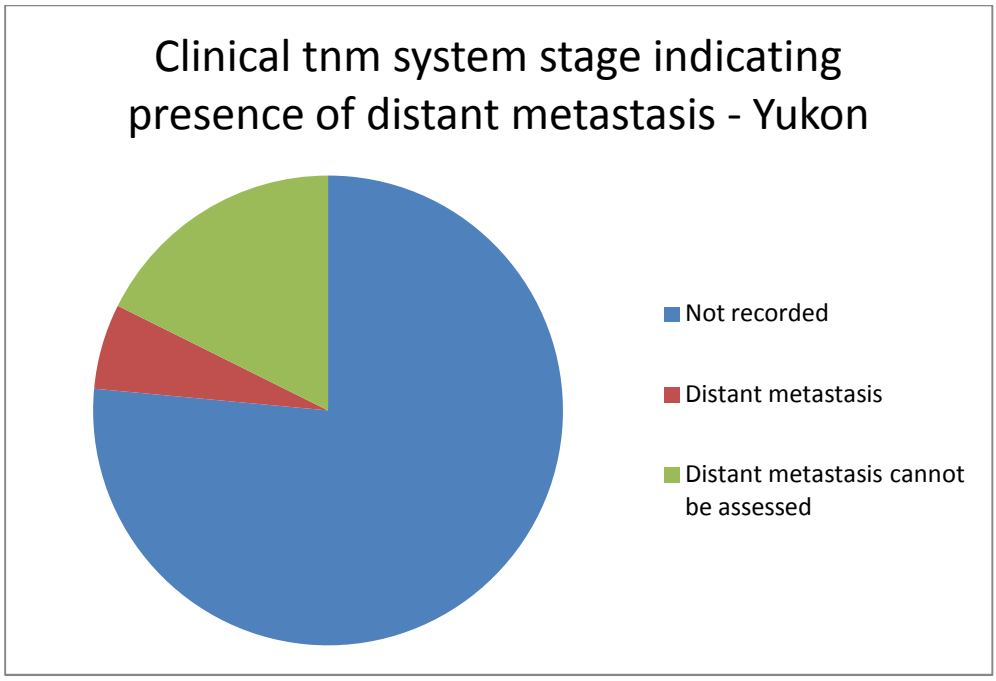

Figure 34. Yukon Clinical tnm system indicating the absence or presence of distant metastasis

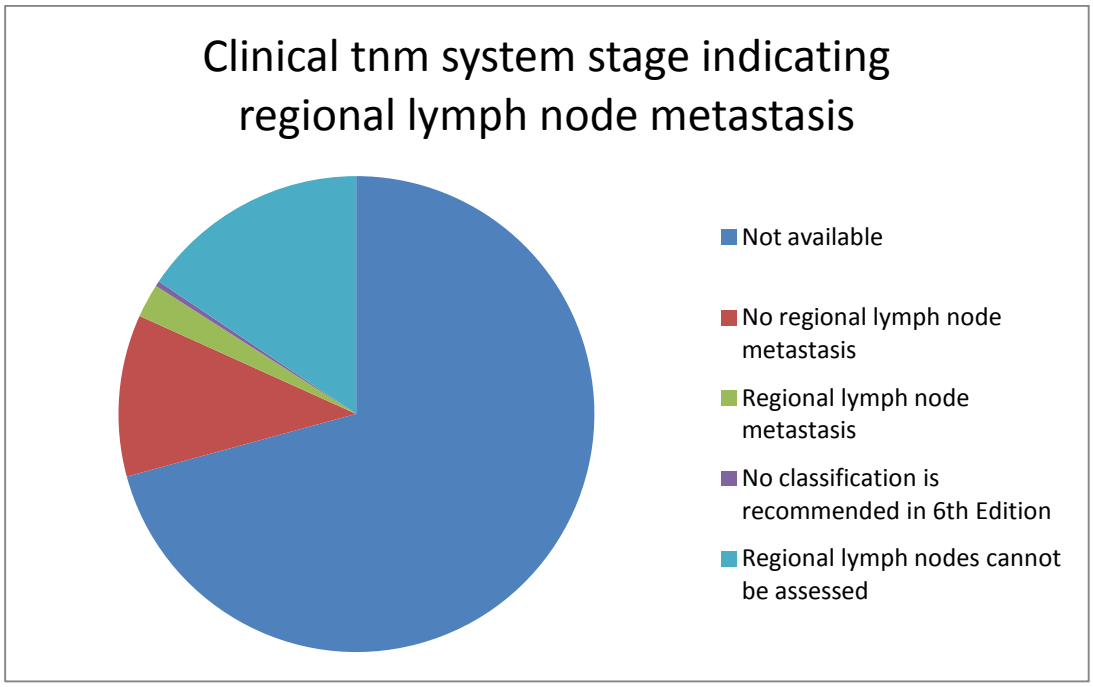

Figure 35. BC Clinical tnm system indicating the absence or presence and existence of regional lymph node metastasis

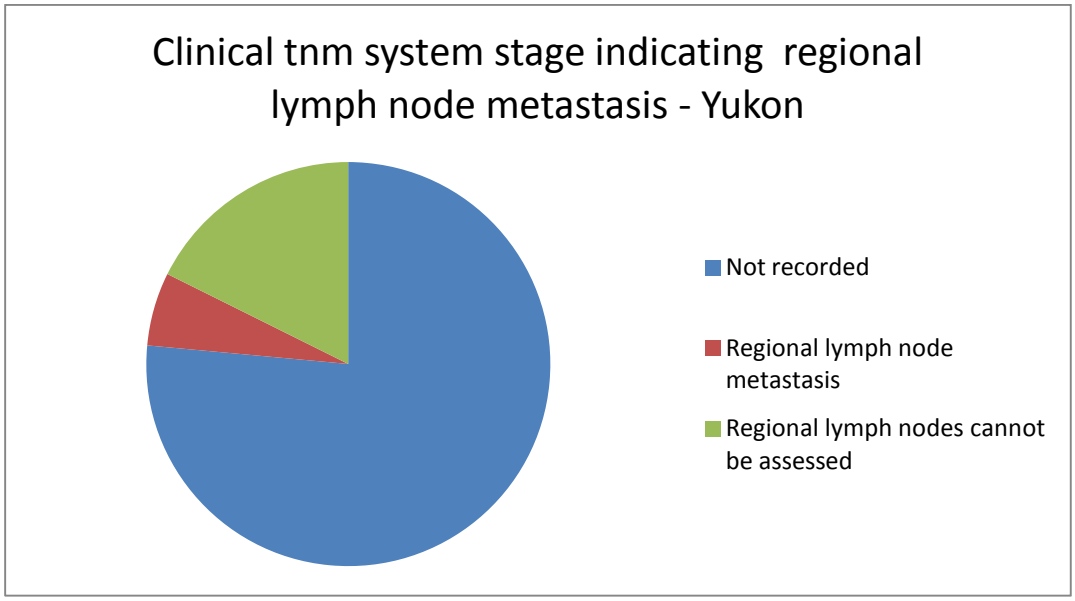

Figure 36. Yukon Clinical tnm system indicating the absence or presence of lymph node metastasis

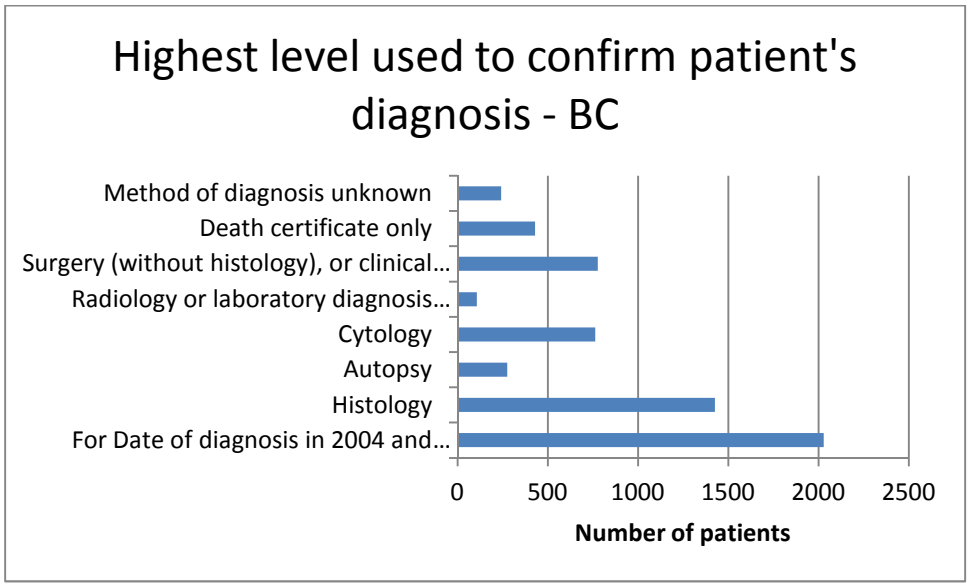

Figure 37. Highest level used to confirm patient's diagnosis in BC

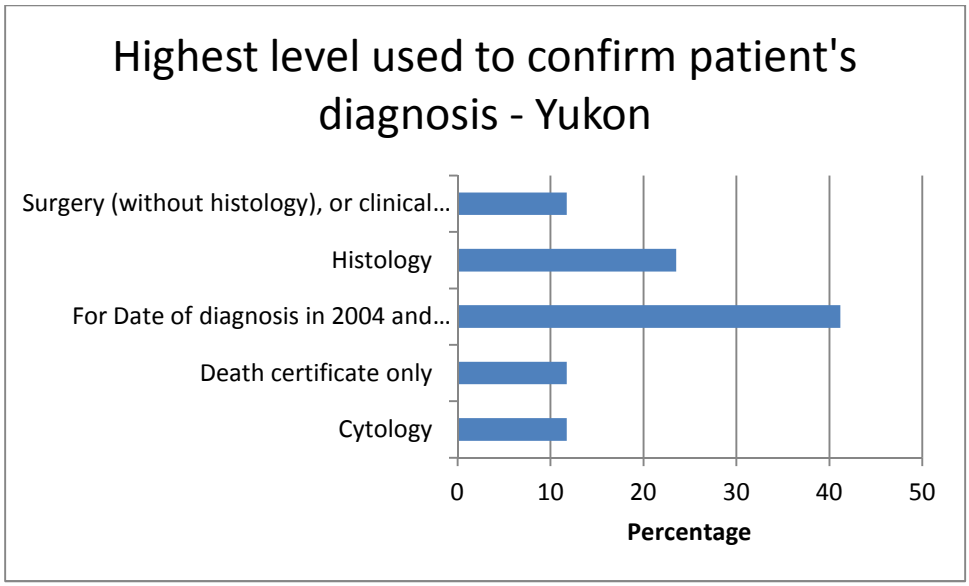

Figure 38. Highest level used to confirm patient's diagnosis in the Yukon

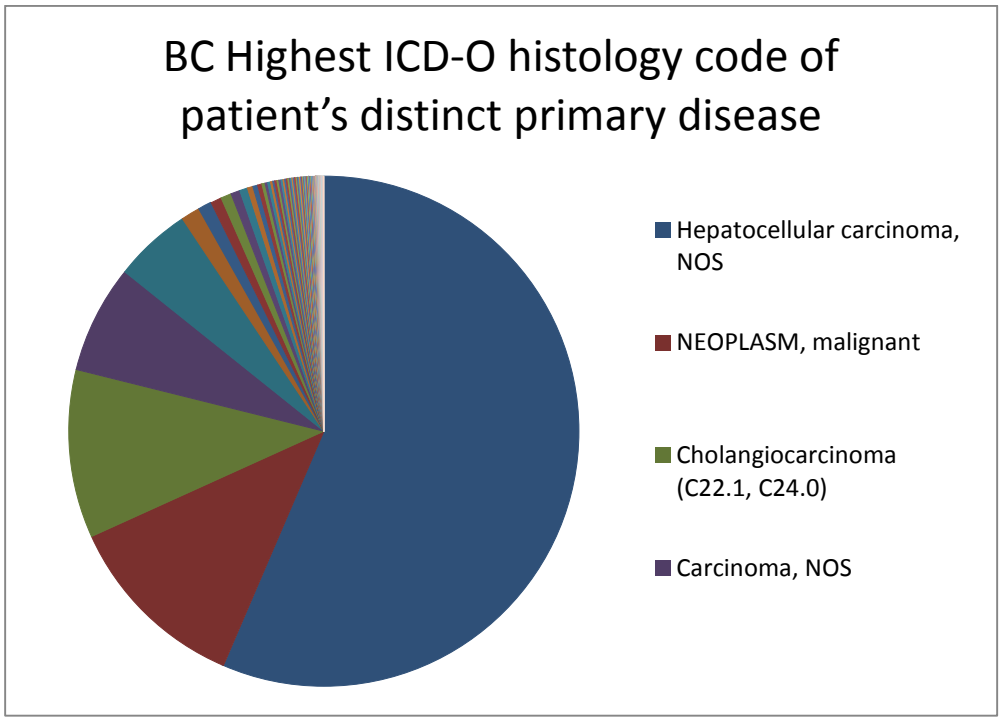

Figure 39. Highest ICD-O histology code of patient's distinct primary disease in BC

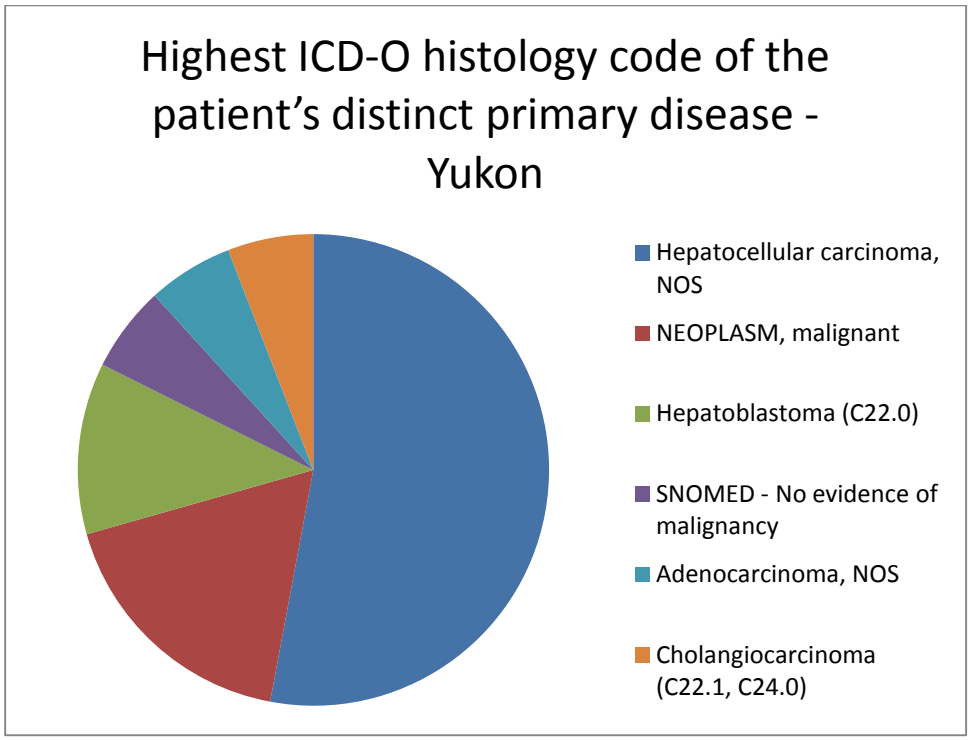

Figure 40. Yukon Highest ICD-O histology code of patient's distinct primary disease

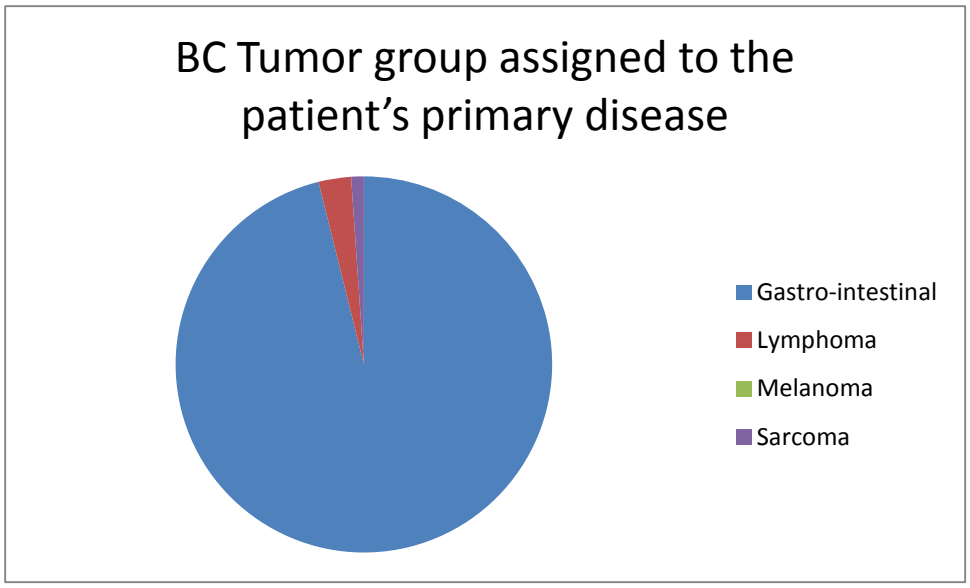

Figure 41. Tumor group assigned to the patient's primary disease in BC

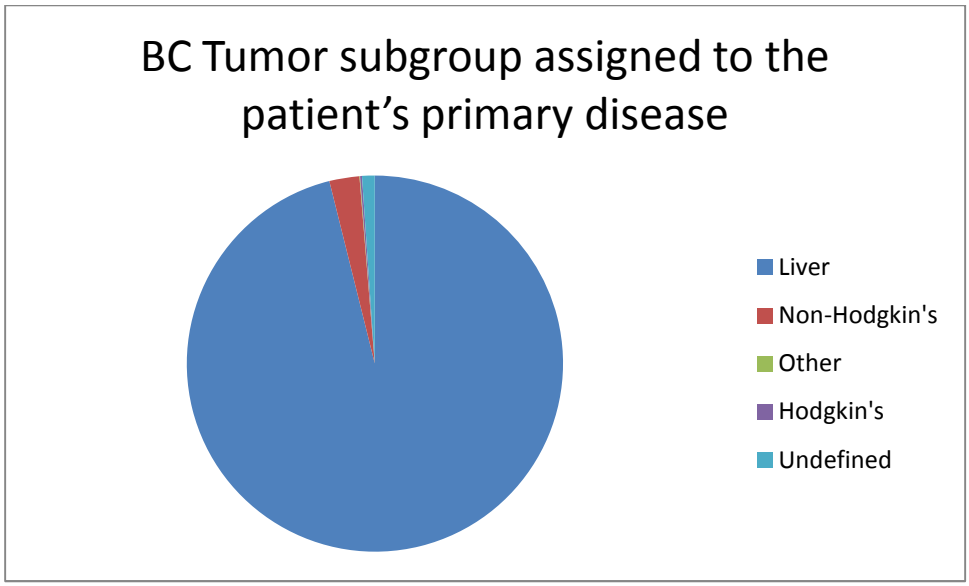

Figure 42. Tumor subgroup assigned to the patient's primary disease in BC

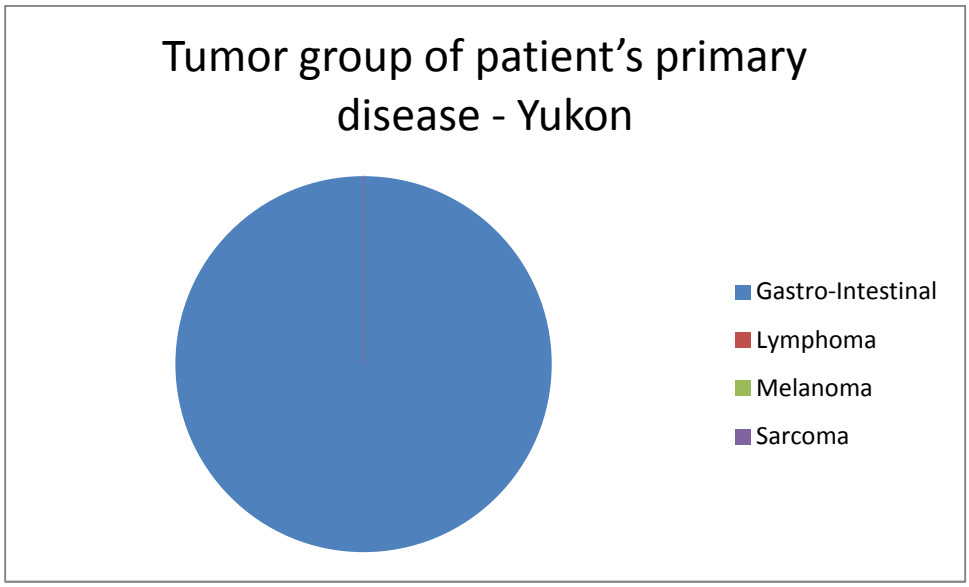

Figure 43. Tumor group assigned to the patient's primary disease in the Yukon

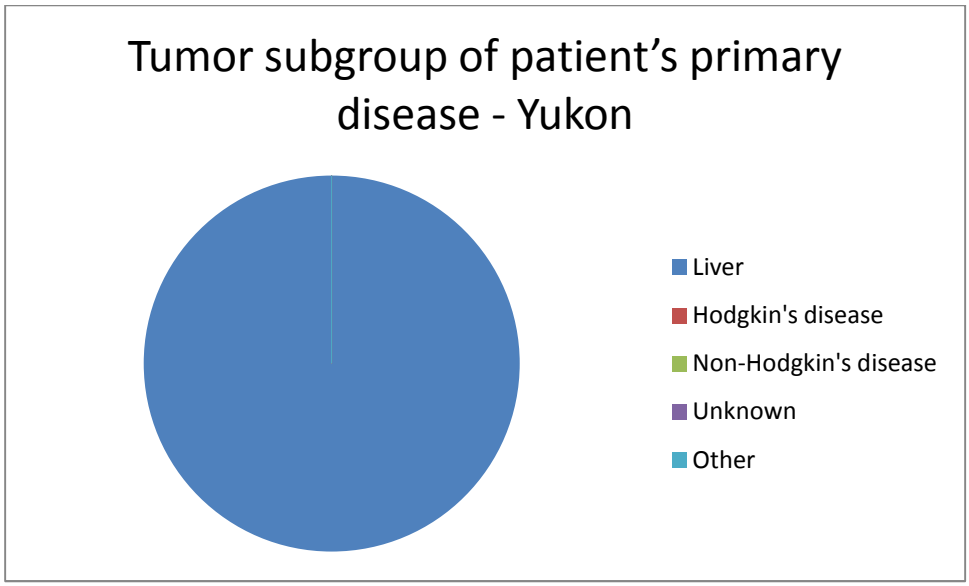

Figure 44. Tumor subgroup assigned to the patient's primary disease in the Yukon

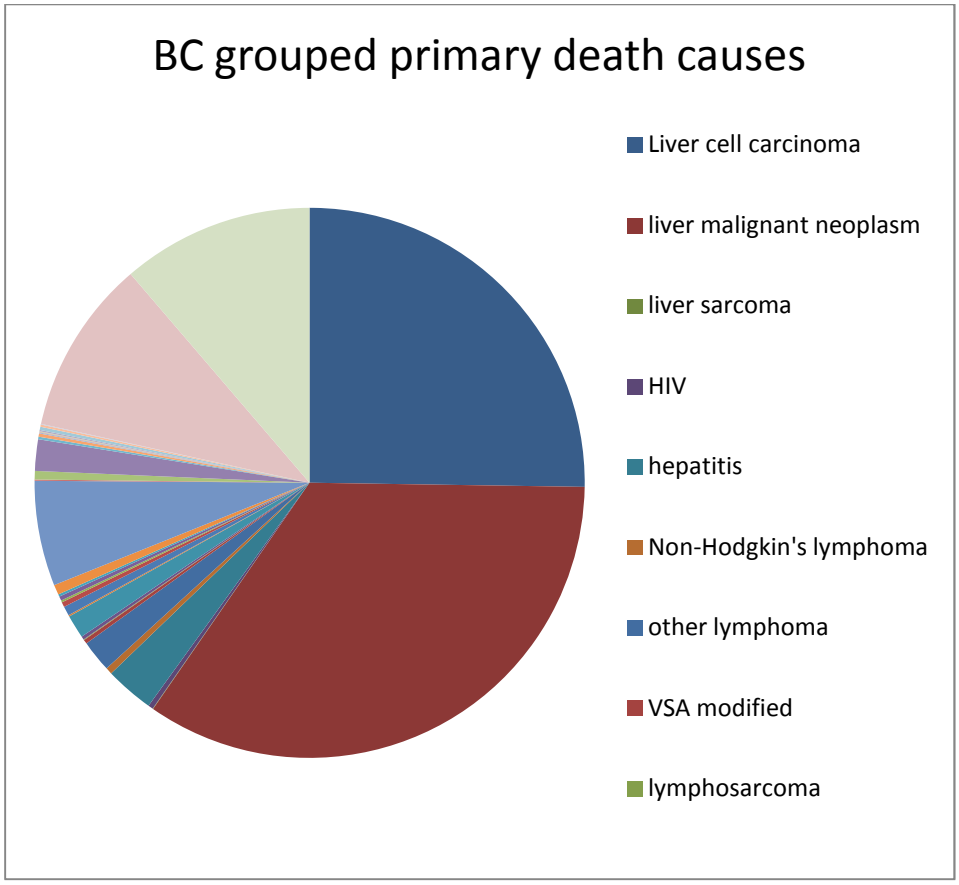

Figure 45. BC grouped primary death causes

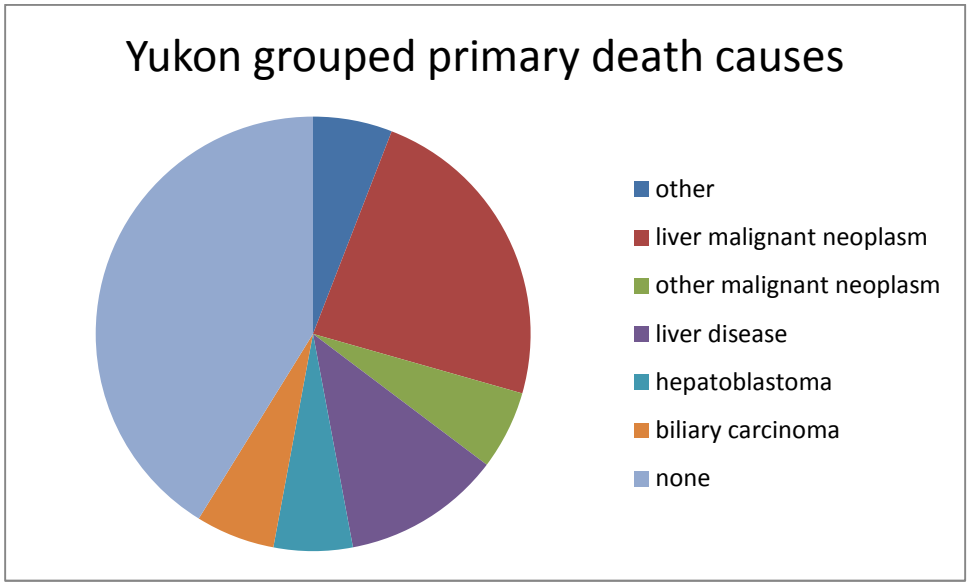

Figure 46. Yukon grouped primary death causes

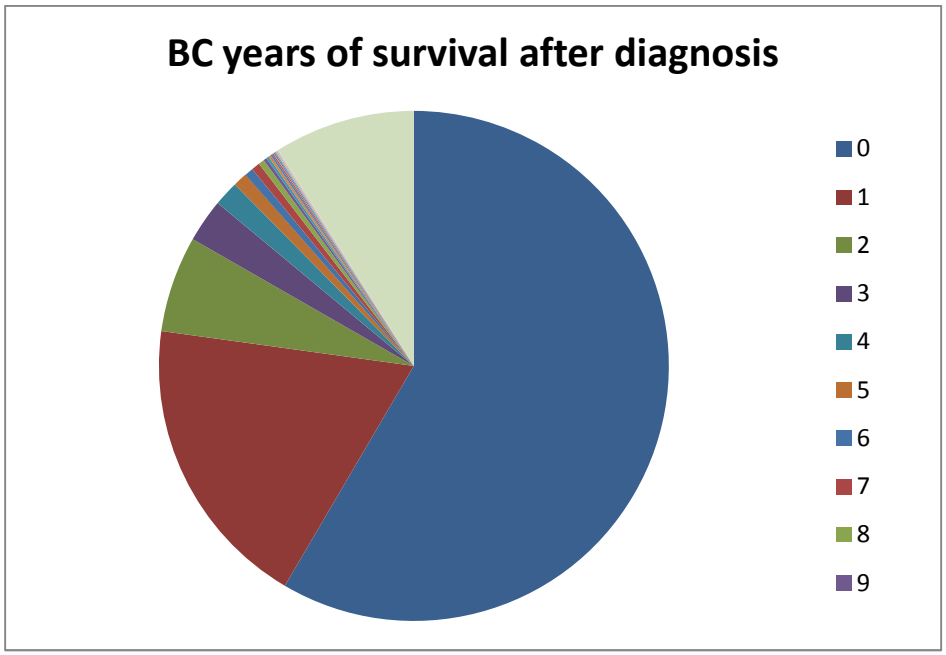

Figure 47. BC patient percentage per survival length

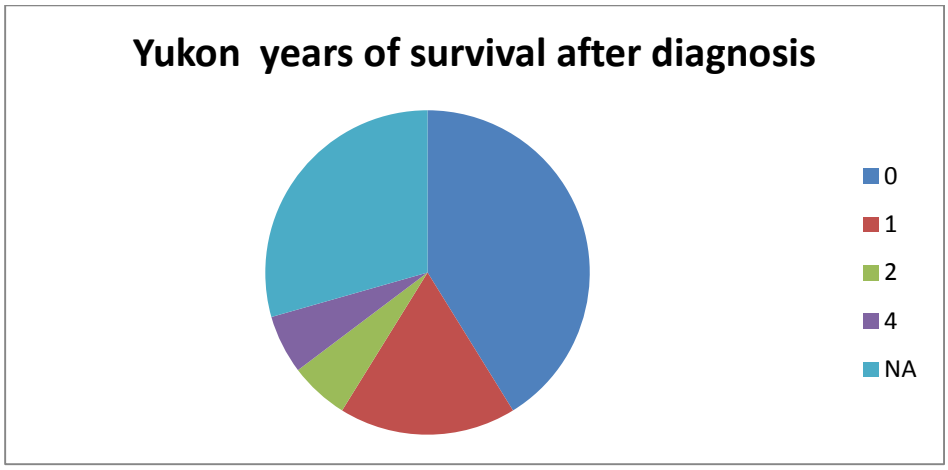

Figure 48. Yukon percentage of patients per length of survival in years

6. RULE VALIDATION AND DISCUSSION

Full results of the association rule mining, including the rules with their confidence and support calculations, are given in Appendix C.

6.1. BC Preliminary Information

6.1.1. Apriori algorithm

The main rules obtained from the Apriori algorithm applied to the BC Preliminary data include:

Rule	Confidence	Support
Diagnosis age range 70-79 → Liver Tumor Subgroup	0.97	0.25
Current Health Authority Vancouver Coastal → Liver Tumor Subgroup	0.96	0.25
Diagnosis Health Authority Vancouver Coastal → Liver Tumor Subgroup	0.96	0.25
Diagnosis Health Authority Vancouver Coastal → Liver Cancer	0.96	0.25
Current Health Authority Vancouver Coastal, Diagnosis Health Authority Vancouver Coastal → Liver Tumor Subgroup	0.96	0.25
Current Health Authority Vancouver Coastal, Diagnosis Health Authority Vancouver Coastal → Liver Cancer	0.97	0.25
Male → Liver Cancer	1.00	0.25
Male → Liver Tumor Subgroup	0.96	0.25

The confidence parameter describes the probability of the consequent to occur given that the antecedent has occurred. The support value indicates on what percentage of the subjects included in the study the relationship indicated by the rule holds (Concaro et al., 2009). For instance, for the first rule above, 97% of the patients diagnosed in age range 70 to 79 had a tumor of the liver subgroup, and that relationship holds for 25% of the patients in the population studied. This is probably the most interesting of the rules presented above, since age range 70 to 79 is also the diagnosis age range with the highest number of deceased patients (Figure 9).

The rule for 100% of the male patients included in the study having liver cancer is not interesting, since we know that all males included in the study did have liver cancer. However, 96% of those male patients had a tumor of the liver subgroup, whereas the remainder had non-Hodgkin's type tumor (Figure 42).

Also, 96% of the patients from the Vancouver Coastal Health Authority, both at diagnosis and currently, had a tumor of the liver subgroup, and, again, the relationship hold for 25% of the population studied. Vancouver Coastal was the Health Authority with the largest number of liver cancer cases (Figure 4), so this rule is expected. Considering the fact that Fraser, and not Vancouver Coastal, is the health authority with the largest population, this is not an artifact of the data.

6.1.2. FP-Growth algorithm

Unlike for the Apriori rules given above, in the FP-Growth rules the consequent was not differentiated into Liver Cancer and Liver Tumor Subgroup, since they would be redundant.

A new rule indicated by the FP-Growth algorithm is that 91 to 96% of the patients diagnosed in age range 60 to 69 have liver cancer. Again this is an interesting rule, since age range 60 to 69 is the diagnosis age range with the second highest number of deceased patients (Figure 9).

The main rules obtained from the FP-Growth algorithm applied to the BC Preliminary data include:

Rule	Confidence	Support
Diagnosis age range 60-69 → Liver Cancer	0.91 – 0.96	0.22 – 0.23
Diagnosis age range 70-79 → Liver Cancer	0.87 – 0.97	0.24 – 0.27
Current Health Authority Fraser → Liver Cancer	0.95	0.23
Diagnosis Health Authority Fraser → Liver Cancer	0.88 – 0.96	0.23 – 0.24
Current Health Authority Fraser, Diagnosis Health Authority Fraser → Liver Cancer	0.96	0.22
Current Health Authority Vancouver Coastal → Liver Cancer	0.88 – 0.96	0.25 – 0.28
Diagnosis Health Authority Vancouver Coastal → Liver Cancer	0.88 – 0.96	0.27 – 0.30
Current Health Authority Vancouver Coastal, Diagnosis Health Authority Vancouver Coastal → Liver Cancer	0.88 – 0.91	0.25 – 0.28
Female → Liver Cancer	0.96	0.30
Male → Liver Cancer	0.88 – 0.96	0.61 – 0.66

Although both males and females are included in the rules above, there was one rule with females and three rules with males (grouped into one rule in the tabulation above) with confidence ranging between 0.88 and 0.96. Also, the support for the rule with females is less than half (30%) the support for the rule with males (66%). As observed earlier (Figure 2), there were about twice as many male as female patients.

Other than that, the FP-Growth algorithm agreed with all the rules produced by the Apriori algorithm but suggested more rules, as summarized below in the form of factors frequently associated with liver cancer incidence:

Factor	Algorithm
Diagnosis age ranges 60 – 69	only FP-Growth
Diagnosis age ranges 70 – 79	both Apriori and FP-Growth
Diagnosis Health authority Vancouver Coastal	both Apriori and FP-Growth
Current Health authority Vancouver Coastal	both Apriori and FP-Growth
Diagnosis Health authority Fraser	only FP-Growth
Current Health authority Fraser	only FP-Growth
Male gender	both Apriori and FP-Growth
Female gender	only FP-Growth

6.2. Yukon Preliminary Information

6.2.1. A priori algorithm

The main rules obtained from the Apriori algorithm applied to Yukon Preliminary data include:

Rule	Confidence	Support
Diagnosis age range 60-69 → Liver Cancer	1.00	0.25
Male, Diagnosis age range 60-69 → Liver Cancer	1.00	0.25
Female → Liver Cancer	1.00	0.25
Male → Liver Cancer	1.00	0.25

The rules for 100% of the female patients and 100% of the male patients included in the study having liver cancer is not interesting, since we know that all females and all males included in the study did have liver cancer. We conclude gender is not an important factor in cancer incidence according to the Yukon data.

6.2.2. FP-Growth algorithm

The main rules obtained from the FP-Growth algorithm applied to the Yukon Preliminary data include:

Rule	Confidence	Support
Diagnosis age range 50-59 → Liver Cancer	1.00	0.18
Male, Diagnosis age range 50-59 → Liver Cancer	1.00	0.18
Diagnosis age range 0-9 → Liver Cancer	1.00	0.18
Diagnosis age range 40-49 → Liver Cancer	1.00	0.18
Diagnosis age range 60-69 → Liver Cancer	1.00	0.29
Male, Diagnosis age range 60-69 → Liver Cancer	1.00	0.24
Female → Liver Cancer	1.00	0.71
Male → Liver Cancer	0.92	0.24

Several diagnosis age ranges are represented in the rules above. However, with the exception of the rule for diagnosis age range 60 to 69, all of those rules have support levels less than 20%.

As for gender, the same comment made for the Apriori rules applies again, except that the confidence level for male patients with liver cancer is now 92% instead of 100%.

The FP-Growth and Apriori algorithms agreed on the rules produced, as summarized below as factors frequently associated with liver cancer incidence:

Factor	Algorithm
Diagnosis age ranges 60-69	both Apriori and FP-Growth
Diagnosis age ranges 60-69, Male	both Apriori and FP-Growth

6.3. BC Survivability Information

6.3.1. Apriori algorithm

a) Excluding records where treatment was undefined:

Rule	Confidence	Support
Male, Liver Cancer → Deceased	0.89	0.50
Liver Cancer → Deceased	0.88	0.50

b) Excluding all treatment attribute:

Rule	Confidence	Support
Male, Liver Cancer → Deceased	0.88 – 0.92	0.40
Liver Cancer → Deceased	0.88 – 0.90	0.40

6.3.2. FP-Growth algorithm

a) Excluding records where treatment was undefined:

Rule	Confidence	Support
Male, Liver Cancer → Deceased	0.87 – 0.89	0.53 – 0.52
Male, Liver Tumor Subgroup → Deceased	0.87	0.62
Liver Cancer → Deceased	0.86 – 0.88	0.67 – 0.77
Liver Tumor Subgroup → Deceased	0.86	0.82
Male → Deceased	1.00	0.29

b) Excluding all treatment attribute:

Rule	Confidence	Support
Male, Liver Cancer → Deceased	0.90 – 0.91	0.39 – 0.57
Male, Liver Tumor Subgroup → Deceased	0.92	0.61
Female, Liver Tumor Subgroup → Deceased	0.92	0.27
Liver Cancer → Deceased	0.90	0.51 – 0.80
Liver Tumor Subgroup → Deceased	0.92	0.88
Male → Deceased	9.1	0.63
Female → Deceased	0.91	0.28
Diagnosis health authority Vancouver Coastal → Deceased	0.89	0.28
Diagnosis health authority Vancouver Coastal, Liver Tumor Subgroup → Deceased	0.90	0.27
Death age range 70 – 79 → Deceased	1.00	0.27

How the treatments were pre-processed did not seem to affect the Apriori rules obtained. The only two Apriori rules that stand out indicate that 88 to 92% of male patients with liver cancer are likely to decease, and 88 to 90% of all patients with liver cancer are likely to decease. These rules hold for 40 to 50% of the population studied.

Excluding all treatment attributes produces more FP-Growth rules than excluding only records where treatment was undefined. The former FP-Growth rules, however, have support levels less than 30%, whereas rules from excluding records where treatment was undefined had support between 29 and 77%. This may demonstrate that feature selection can indeed increase efficiency in the system (Ribeiro et al., 2009; Kirsnhers, Parshutin & Leja, 2012).

The FP-Growth algorithm agreed with all the rules produced by the Apriori algorithm, but suggested more rules, as summarized below in the form of factors frequently associated with liver cancer incidence.

Factor	Algorithm
Age of death range 70 – 79	only FP-Growth
Diagnosis Health authority Vancouver Coastal	only FP-Growth
Male gender	both Apriori and FP-Growth
Female gender	only FP-Growth

6.4. Yukon Survivability Information

6.4.1. Apriori algorithm

The main rules obtained from the Apriori algorithm applied to the Yukon Survivability data include:

Rule	Confidence	Support
Death age range 60 - 69 → Deceased	1.00	0.30

The rule above seems to describe the age when most patients in the Yukon decease (Figure 11). The relationship holds for 30% of the population study.

6.4.2. FP-Growth algorithm

The main rules obtained from the FP-Growth algorithm applied to the Yukon Survivability data include:

Rule	Confidence	Support
Death age range 60 - 69 → Deceased	1.00	0.24
Male, Death age range 60 - 69 → Deceased	1.00	0.24
Liver Cancer, Death age range 60 - 69 → Deceased	1.00	0.24 – 0.29
Male, Liver Cancer, Death age range 60 - 69 → Deceased	1.00	0.24

The FP-Growth algorithm agrees with the rule produced by the Apriori algorithm, but adds male gender as another factor in decrease survivability. The factors frequently associated with patient death are summarized below:

Factor	Algorithm
Age of death range 60 – 69	both Apriori and FP-Growth
Male gender	only FP-Growth

7. CONCLUSIONS AND FUTURE WORK

The patient data attributes that were consistently recorded are the ones that manifest themselves in the association rules, primarily age, gender, and health authority. Age was calculated from the year of birth, since the full date of birth was not available.

Recommendations for future work would include stratifying the data from British Columbia by Health Authority and applying the association algorithms individually to each Health Authority, as that may reveal association rules and factors that were lost in the current study due to different numbers of patients in each Health Authority. Vancouver Health Authority (VIHA), for instance, has a much higher average age than the other Health Authorities and, as such, it may have specific rules that are not frequent when patient data from all Health Authorities are analyzed together.

Unfortunately the treatment data was too sparse to have an impact on the survivability results. Future work should also try and link with secondary treatment data to evaluate the impact of the findings on patient survivability and cost efficiency.

One possibility for validating the rules obtained in this study would be to apply the association algorithms to similar liver cancer patient repository data, such as from the Surveillance Epidemiology and End Results (SEER) program of the National Cancer Institute (Agrawal & Choudhary, 2011). Another possibility would be to engage health care professionals focused on dealing with liver cancer patients to help validate the rules.

It may also prove useful to gather data from the general population, possibly via surveys or focus groups, to validate the rules obtained from the Preliminary data sets.

8. REFERENCES

Agrawal, A.; Choudhary, A. 2011. Identifying HotSpots in Lung Cancer Data Using Association Rule Mining. 11th IEEE International Conference on Data Mining Workshops. Pp. 995 - 1002

Agrawal, R., Imielinski, T., Swami, A. 1993. Mining Association Rules between Sets of Items in Large Databases. Proceedings of the 1993 ACM SIGMOID Conference. Washington, DC, USA.

Ando, E., Kuromatsu, R., Tanaka, M., Takada, A., Fukushima, N., Sumie, S., Nagaoka, S., Akiyoshi, J., Inoue, K., Torimura, T., Kumashiro, R., Ueno, T., Sata, M. 2006. Surveillance program for early detection of hepatocellular carcinoma in Japan: results of specialized department of liver disease. J. Clin. Gastroenterol. 40 (10): 942-948.

Automonova, I.I., Frishman,G., Gelfand, M.S., & Frishman, D. 2005. Mining Sequence Annotation Databanks for Association Patterns. Bioinformatics 21 (3): iii49 – iii57

Barber, F.D., Nelson, J.P. 2000. Liver Cancer: Looking to the Future for Better Detection and Treatment. American Journal of Nursing 100: 4, Supplement: Oncology Nursing. Pp. 41-46. www.nursingcenter.com

Barrett, J. R. 2005. Liver Cancer and Aflatoxin: New Information from the Kenyan Outbreak Environmental Health Perspectives. Volume 113, NUMBER 12

Bener, A., Moore, A.M., Ali, R., El Ayoubi, H.R. 2010. Impacts of family history and lifestyle habits on colorectal cancer risk: a case-control study in Qatar. Asian Pac J. Cancer Prev. 11: 963-968

Berasain, C., Castillo, J. Perugorria, M.J., Latasa, M.U., Prieto, J., Avila, M.A. 2009. Inflammation and Liver Cancer: New Molecular Links. Steroid Enzymes and Cancer: Ann. N.Y. Acad. Sci. 1155: 206–221

Blum, H. E. 2005. Liver Cancer. Eur J Gastroenterol Hepatol 17:475–476

British Columbia Ministry of Health. 2006. Profiles of British Columbia's Six Health Authorities http://www.bcbudget.gov.bc.ca/2007/sp/hlth/default.aspx?hash=10

Cardenes, H. R. & Lasley, F. 2012. Primary Liver Cancer. In: S. S. Lo et al. (eds.), Stereotactic Body Radiation Therapy, Medical Radiology. Springer-Verlag, Berlin

Chang, C.K., Astrakianakis, G., Thomas, D.B., Seixas, N.S., Ray, R.M., Gao, D.G., Wernli, K.J., Fitzgibbons, E.D., Vaughan, T.L., Checkoway, H. 2006. Occupational exposures and risks of liver cancer among Shanghai female textile workers—a case–cohort study. International Journal of Epidemiology 35: 361–369

Chun, S.C., Kim, J., Hahm, K.B., Park, Y.J., Chun, S.H. 2008. Data mining technique for medical informatics: detecting gastric cancer using case-based reasoning and single nucleotide polymorphisms. Expert Systems 25(2)

Clifton, C. & Thuraisingham, B. 2001. Emerging Standards for Data Mining. Computer Standards and Interfaces 23: 187-193

Concaro, S., Sacchi, L., Cerra, C., Bellazzi, R. 2009. Mining Administrative and Clinical Diabetes Data with Temporal Association Rules. European Federation for Medical Informatics. In: K.P. Adlassnig et al. (eds.). Medical Informatics in a United and Healthy Europe. IOS Press. Pp. 574 - 578

Curado, M.P., Voti, L., & Sortino-Rachou, A.M. 2009. Cancer registration data and quality indicators in low and middle income countries: their interpretation and

potential use for the improvement of cancer care. Cancer Causes Control 20:751–756

Delen, D. 2009. Analysis of cancer data: a data mining approach. Expert Systems 26 (1)

El-Serag, H.B. 2002. Hepatocellular Carcinoma: An Epidemiologic View. Journal of Clinical Gastroenterology 35(Suppl. 2): S72–S78

Fan, Q., Zhu, C.H., Xiao, J.Y., Wang, B.H., Yin, L., Xu, X.L., & Rong, F. 2010. An application of the Apriori algorithm in SEER breast cancer data. 2010 International Conference on Artificial Intelligence and Computational Intelligence. IEEE Computer Society.

Ferlay, J., Shin, H.R., Bray, F., Forman, D., Mathers, C., & Parkin, D.M. 2010. Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008. Int. Journal Cancer 127: 2893–2917

Giovannucci, E., Harlan, D.M., Archer, MC, Bergenstal, R.M., Gapstur, S.M., Habel, L.A., Pollak, M, Regensteiner, J.G., Yee, D. 2010. Diabetes and Cancer: A Consensus Report. CA CANCER J CLIN 2010 60: 207–221

Greene, F.L., Compton, C.C., Fritz, A.G., Shah, J.P., Winchester, D.P. 2006. AJCC Cancer Staging Atlas. Springer, New York. Pp. 1 – 9, 127

Hainaut, P., Boyle, P. 2008. Curbing the liver cancer epidemic in Africa. The Lancet 371: 367-368 www.thelancet.com

Henning, J.S., Dusza, S.W., Wang, S.Q., Marghoob, A.A., Rabinowitz, H.S., Polsky, D., Kopf, A.W. 2007. The CASH (Color, Architecture, Symmetry, and Homogeneity) Algorithm for Dermoscopy. J. Am. Academy of Dermatology 56 (10): 45-52

Ho, S.H., Jee, S.H., Lee, J.E. Park, J.S. 2004. Analysis on risk factors for cervical cancer using induction technique, Expert Systems with Applications 27(1): 97-105

Hospice Patients Alliance. 2011. Karnofsky Performance Status Scale. http://www.hospicepatients.org/karnofsky.html

Hu, R. 2010. Medical Data Mining Based on Association Rules. Computer and Information Science 3 (4 F)

Hung, W.K. 2007. Palliative Radiotherapy and Palliative Chemotherapy. Fourth Hong Kong Palliative Care Symposium. HKSPM Newsletter Issues 1 & 2, April & August 2007. P 12.

Jou, J.H. & Fisher, D.A. 2010. Predictive Algorithms: Uses and Limitations. Dig. Dis. Sci. 55:3016–3017

Karabatak, M., & Ince, M.C. 2009a. A new feature selection method based on association rules for diagnosis of erythemato-squamous diseases. Expert Systems with Applications 36: 12500-12505

Karabatak, M., & Ince, M.C. 2009b. An expert system for detection of breast cancer based on association rules and neural network. Expert Systems with Applications 36: 3465–3469

Kaur, H. & Wasan, S.K. 2006. Empirical Study on Applications of Data Mining Techniques in Healthcare. J. Computer Science 2 (2): 194 - 200

Kirshners, A., Parshutin, S., Leja, M. 2012. Research on Application of Data Mining Methods to Diagnosing Gastric Cancer. 12th Industrial Conference on Data Mining, ICDM 2012, Volume LNAI 7377, Advances in Data Mining. Applications and Theoretical Aspects. Springer-Verlag, Berlin. pp 24-37

Kuo, M.K., Kushniruk, A.W., Borycki, E.M., & Greig, D. 2009. Application of the Apriori Algorithm for Adverse Drug Reaction Detection. Studies in Health Technology and Informatics 148:95–101

La Vecchia, C. 2005. Coffee, liver enzymes, cirrhosis and liver cancer. Journal of Hepatology 42: 444–446

Lareinjam, B. & Wasan, S.K. 2009. Neural Network with Classification based on Multiple Association Rula for Classifying Mammographic Data. In: E. Corchado & H. Yin (eds.). IDEAL 2009, LNCS 5788: 465-476

Lee, E.H., Han, M.A., Lee, H.Y., Jun, J.K., Choi, K.S., Park, E.C. 2010. Liver Cancer Screening in Korea: a report on the 2008 National Cancer Screening Programme. Research Communication. Asian Pacific J. Cancer Prev. 11:1305-1310

Li, J., Fu, A.W.C, Fahey, P. 2009. Efficient discovery of risk patterns in medical data. Artificial Intelligence in Medicine 45: 77-89

Li, L., Tang, H., Wu, Z., Gong, J., Gruidl, M., Zou, J., Tockman, M., Clark, R.A. 2004. Data mining techniques for cancer detection using serum proteomic profiling. Artificial Intelligence in Medicine 32, 71—83

Lisboa, J.G., Vellido, A., Tagliaferri, R., Napolitano, F., Ceccarelli, M., Martin-Guerrero, J.D., Biganzoli, E. 2010. Data Mining in Cancer Research. Ieee Computational Intelligence Magazine, February 2010

Llovet, J.P., Burroughs, A., Bruix, J. 2003. Hepatocellular Carcinoma. Lancet 362: 1907–1917

Lock, M.I., Morten, H., Bydder, S.A., Okunieff, P., Hahn, C.A., Vichare, A., Dawson, L.A. 2012. An International Survey on Liver Metastases Radiotherapy. Acta Oncologica 51: 568 – 574.

Lopez, F, Cuadros, M., Blanco, A., Concha, A. Unveiling Fuzzy Associations Between Breast Cancer Prognostic Factors and Gene Expression Data. Database and Expert Systems Application, 2009. 20th International Workshop on Database and Expert Systems Application. Pp 338 – 342

Luk, J.M., Lam, B.Y., Lee. N.P.Y., Ho, D.W., Sham, P.C., Chen, L., Peng, J., Leng, X., Day, P.J., Fan, S.T. 2007. Artificial neural networks and decision tree model analysis of liver cancer proteomes. Biochemical and Biophysical Research Communications 361 (2007) 68–73

Maimon, O. & Rokach, L. 2010. Introduction to Knowledge Discovery and Data Mining. In O. Maimon, L. Rokach (eds.), *Data Mining and Knowledge Discovery Handbook*, 2nd ed., Springer. 2010

Malpani, R., Lu, M., Zhang, D., Sung, W.K. 2011. Mining Transcriptional Association Rules from Breast Cancer Profile Data. IEEE IRI 2011, August 3-5, 2011, Las Vegas, Nevada, USA

Mavaddat, N., Rebbeck, T.R., Lakhani, S.R., Easton, D.F., & Antoniou, A.C. 2010. Incorporating tumour pathology information into breast cancer risk prediction algorithms. Breast Cancer Research, 12:R28

McKillop, I.H., and Schrum, L.W. 2005. Alcohol and Liver Cancer. Alcohol 35: 195–203

Mukhopadhyay, A., Maulik, U., Bandyopadhyay, S. & Eils, R. 2010. Mining Association Rulesfrom HIV-Human Protein Interactions. Proceedings of 2010 International

Conference on Systems in Medicine and Biology 16-18 December 2010, IIT Kharagpur, India

Munneke, G.J. 2008. Interventional radiology in liver cancer. Imaging*, 20:* 185-193

Nahar, J., Tickel, K.S., Shawkat Ali, A.B.M., & Chen, Y.P.P. 2011. Significant Cancer Prevention Factor Extraction: An Association Rule Discovery Approach. J Med Syst 35:353–367

Oken MM, Creech RH, Tormey DC, *et al.* (1982). "Toxicity and response criteria of the Eastern Cooperative Oncology Group". *Am. J. Clin. Oncol.* **5** (6): 649–55

Okuda, K. 2002. Hepatocellular carcinomadhistory, current status and perspectives. Dig Liver Dis 34, 613–616.

Osl, M., Dreiseitl, S., Pfeiffer, B., Weinberger, K., Klocker, H., Bartsch, G., Schafer, G., Tilg, B., Graber, A., & Baumgartner, C. 2008. A new rule-based algorithm for identifying metabolic markers in prostate cancer using tandem mass spectrometry. Bioinformatics 24 (24): 2908-2914

PHAC (Public Health Agency of Canada). 2013. Canadian Cancer Statistics 2012. http://www.phac-aspc.gc.ca/cd-mc/cancer/ccs-scc-2012-eng.php

Papagiorgiou, E., Kotsioni, I., Linos, A. 2005. Data Mining: a new technique in medical research. Hormones 4(4):210-212

Parkin, D.M. 2006. The evolution of the population-based cancer registry. *Nat* Rev Cancer 6: 603-612

Pellegrino, A. 2006. Looking at Liver Cancer. Nursing 36: 10. www.nursing2006.com

Pendharkar, P.C., Rodger, J.A., yaverbaum, G.J., Herman, N., Benner, M. 1999. Association, statistical, mathematical and neural approaches for mining breast cancer patterns. Expert Systems with Applications 17(3): 223-232

Rabin, B.A. & Brownson, R.C. 2012. Developing the Terminology for Dissemination and Implementation Research. In: R.C. Brownson, G.A. Colditz, E.K. Proctor (eds.), Dissemination and Implementation Research in Health: Translating Science to Practice. Oxford University Press. P. 25

Ribeiro, M.X., Bugatti, P.H., Traina Jr., C., Marques, P.M.A, Rosa, N.A., Traina, A.J.M. 2009. Supporting content-based image retrieval and computer-aided diagnosis systems with association rule-based techniques. Data and Knowledge Engineering 68: 1370-1382.

Richards, G., Rayward-Smith, V.J., Sonksen, P.H., Carey, S., Weng, C. 2001. Data mining for indicators of early mortality in a database of clinical records. Artificial Intelligence in Medicine 22: 215-231

Richardson, P., Henderson, L., Davila, J.A., Kramer, J.R., Fitton, C.P., Chen, G.J., & El-Serag, H.B. 2010. Surveillance for Hepatocellular Carcinoma: Development and Validation of an Algorithm to Classify Tests in Administrative and Laboratory Data. Digestive Diseases and Sciences 55 (11): 3241-3251

Rodriguez, A., Carazo, J.M., Trelles, O. 2005. Mining Association Rules from Biological Databases. J. AM. Soc. For Information Science and Technology 56(5): 493-504

Sandagdorj, T., Sanjaajamts, E., Tudev, U., Oyunchimeg, D., Ochir, C., Roder, D. 2010. Cancer Incidence and Mortality in Mongolia – National Registry Data. Asian Pacific J. Cancer Prev 11 (6): 1509 – 1514

Sangster-Gormley, E., Kuo, M.H., Borycki, E.M., Schreiber, R. 2013. Use of Knowledge Discovery Techniques to Understand Nurse Practitioner Practice Patterns and their Integration into a Healthcare System. Studies in Health Technology and Informatics 183: 111-115

Seeja, K.R., Alam, M.A., Jain, S.K. 2008. Identification of Co-regulated Signature Genes in Pancreas Cancer – A Data Mining Approach. In: D.S. Huang, D.C. Wunsch, D.S. Levine, K.H. Jo (eds.), Advanced Intelligent Computing Theories and Applications: With Aspects of Theoretical and Methodological Issues. Lecture Notes in Computer Science. Volume 5226: 138-145

Sepulveda, C., Marlin A., Yoshida, T., Ullrich, A. 2002. Palliative Care: The World Health Organization's Global Perspective. Journal of Pain and Symptom Management 24 (2): 91 – 96

Shils, M. E. 2008. Nutritional and Dietary Factors in Neoplastic Development**.** CA: A Cancer Journal for Clinicians 21 (6): 399-406

Shin, H.R., Masuyer, E., Ferlay, J., & Curado, M.P. 2010**.** Cancer in Asia - Incidence Rates Based on Data in Cancer Incidence in Five Continents IX (1998-2002) *Asian Pacific J Cancer Prev,* **11**, Asian Cancer Epidemiology Supplement, 11-16

Sitzmann, J. V., & Abrams, R. 1993. Improved survival for hepatocellular cancer with combination surgery and multimodality treatment. Ann Surg 217, 149–154.

Tan, P.N., Steinbach, M., Kumar, V. 2006. Introduction to Data Mining. Pearson, Addison-Wesley, Boston, MA.

United Nations. 2009. World Population Prospects: The 2008 Revision. http://www.un.org/esa/population/publications/wpp2008/wpp2008_highlights.pdf

Valsecchi M.G., & Steliarova-Foucher, E. 2008. Cancer registration in developing countries: luxury or necessity? *Lancet Oncol*, **9**, 159-67.

Witten, I.H. & Frank, E. 2005. Data Mining: Practical Machine Learning Tools and Techniques. Second Edition. Elsevier. Morgan Kaufmann Publishers, San Francisco

World Health Organization (WHO). Fact sheet No. 297. February 2012. http://www.who.int/mediacentre/factsheets/fs297/en/

Yoo, K.Y. 2010. Cancer Prevention in the Asia Pacific Region. Asian Pacific J. Cancer Prev 11: 839-844

APPENDIX A

Column Name	Description	Comments
DEMOGRAPHICS		
study_id	Non identifying study number	The BCCA identification number was replaced with a non-identifying study_id number. This number is random and is in no way associated with the BCCA identification number. If a patient has more than one disease site this is identified by a combination of the study_id and the site_num data fields. The cross-walk file will reside with Cancer Surveillance and Outcomes, BCCA.
site_num	Unique disease site number	

Column Name	Description	Comments
reg_group	Identifies whether the location at diagnosis is British Columbia (B) or Yukon Territory (Y).	BC (6047 pts) includes any cases with a BC location at diagnosis postal code, Statistics Canada BC geographic code and/or blank values. Yukon (17 pts) includes any cases with a YT location at diagnosis postal code and/or Statistics Canada YT geographic code.
birth_year	The patient's year of birth.	
sex	The patient's gender.	
curr_hlth_auth	The health authority code of the patient's current BC home address postal code	
curr_desc_hlth_auth	The health authority description of the patient's current BC home address postal code	
MORTALITY/FOLLOW-UP		
pat_status	The patient's status of 'Alive' or 'Deceased'.	
death_yr	The year of death.	

Column Name	Description	Comments
death_autopsy	A flag indicating if an autopsy was done, if known.	
bcca_cod	The ICD code assigned by the BCCA for the cause of death.	
bcca_cod_desc	The ICD site description for the bcca_cod	
death_cause_original	The BC Vital Statistics primary cause of death.	
death_cause_orig_desc	Description of the BC Vital Statistics primary cause of death.	
death_sec_cause	The BC Vital Statistics secondary cause of death.	
death_sec_cause_desc	Description of the BC Vital Statistics secondary cause of death.	
DISEASE		
age_at_diagnosis	The age of the patient at the time of diagnosis (system-generated '01's are not included in the calculation).	
age_at_diag_estimated	A flag to indicate age_at_diagnosis was calculated using a birth date and/or diagnosis date where the day or month were not entered.	

Column Name	Description	Comments
diag_type_desc	The level at which the patient's disease information has been entered into the system (Registry, Provisional, Final, Amended).	
diagnosis_yr	The year the patient's disease was diagnosed.	1970-2010
diagnosis_fuzz	A flag indicating the day of diagnosis is system-generated to '01' (D) or the day and month of diagnosis are system-generated to '01' (M).	
dx_hlth_auth	The health authority code of the patient's BC postal code at the time of diagnosis.	Any blanks here indicate that either the location at diagnosis was blank or that the location at diagnosis is in BC but the exact location is unknown.
dx_hlth_auth_desc	Description of the health authority of the patient's BC postal code at the time of diagnosis.	

Column Name	Description	Comments
dx_hsda_cc	The health service delivery area and cancer centre catchment of the patient's BC postal code at the time of diagnosis.	AC Abbotsford Centre CN Centre for the North SI Southern Interior VA Vancouver Centre VI Vancouver Island Centre ?? There are two local health areas within the FHA that cannot be mapped to any one specific cancer centre. These are programmed as '??'.
ext_of_disease	For non-referred cases the extent of the disease at the time of diagnosis, if filled in on the Cancer Registration form by the patient's physician.	
ext_of_disease_desc	For non-referred cases the extent of the disease at the time of diagnosis, if filled in on the Cancer Registration form by the patient's physician.	
site	The ICD-O (International Classification of Diseases for Oncology) site code for the patient's distinct primary disease.	ICD-O-3 sites C220 liver and C221 intrahepatic bile duct

Column Name	Description	Comments
site_desc	The English description for the ICDO2 site code	
laterality	The anatomical side of the patient's distinct primary disease.	
laterality_desc	The description of the anatomical side of the patient's distinct primary disease.	
hist1	The highest ICD-O (International Classification of Diseases for Oncology) histology code of the patient's distinct primary disease.	Includes all ICD-O-3 and SNOMED histologies
behavior	The fifth digit of the patient's ICD-O histology code (between 8000 and 9999) entered into the hist1 column.	
hist1_desc	The English description for the hist1 ICDO2 histology code	
hist2	The second highest ICD-O histology code of the patient's distinct primary disease, if applicable	
hist2_desc	The English description for the hist2 ICD-O histology code	
hist3	The third highest ICD-O histology code of the patient's distinct primary disease, if applicable.	
hist3_desc	The English description for the hist3 ICD-O histology code	

Column Name	Description	Comments
method_of_confirm	The highest level used to confirm the patient's diagnosis.	
method_of_confirm_desc	The description of the highest level used to confirm the patient's diagnosis.	
performance_status	The patient's ECOG performance status code.	
performance_status_desc	The description of the patient's ECOG performance status code.	
status_at_referra	The status of the patient's primary disease at referral to a cancer centre.	
status_at_referral_desc	The description of the status of the patient's primary disease at referral to a cancer centre.	
tumour_group	The tumour group assigned to the patient's primary disease	
tumour_group_desc	The description of the tumour group assigned to the patient's primary disease	
tumour_subgroup	The subgroup of the tumour group assigned to the patient's primary disease.	
tumour_subgroup_desc	The description of the subgroup of the tumour group assigned to the patient's primary disease.	
STAGE		
amended_stage	A flag indicating if the stage was amended.	

Column Name	Description	Comments
tnm_clin_t	The clinical tnm system stage indicating the extent of the primary tumour	Staging data are not available for nonreferred cases (ie. cases not referred to a BC cancer centre)
tnm_clin_t_desc	Description of the clinical tnm system stage indicating the extent of the primary tumour	
tnm_clin_n	The clinical tnm system stage indicating the absence or presence and existence of regional lymph node metastasis	
tnm_clin_n_desc	Description of the clinical tnm system stage indicating the absence or presence and existence of regional lymph node metastasis	
tnm_clin_m	The clinical tnm system stage indicating the absence or presence of distant metastasis.	
tnm_clin_m_desc	Description of the clinical tnm system stage indicating the absence or presence of distant metastasis.	
tnm_clin_yr	The revision year of the clinical tnm staging system being used.	
tnm_surg_t	The surgical tnm system stage indicating the extent of the primary tumour	

Column Name	Description	Comments
tnm_surg_t_desc	Description of the surgical tnm system stage indicating the extent of the primary tumour	
tnm_surg_n	The surgical tnm system stage indicating the absence or presence and existence of regional lymph node metastasis	
tnm_surg_n_desc	Description of the surgical tnm system stage indicating the absence or presence and existence of regional lymph node metastasis	
tnm_surg_m	The surgical tnm system stage indicating the absence or presence of distant metastasis.	
tnm_surg_m_desc	Description of the surgical tnm system stage indicating the absence or presence of distant metastasis.	
tnm_surg_yr	The revision year of the surgical tnm staging system being used.	
TREATMENT		
bcca_chemo	A flag indicating the patient had chemotherapy up to 3 months post-BCCA admission, if the information is known. For non-referred cases hormone therapy is included.	

Column Name	Description	Comments
bcca_chemo_desc	Description of the flag indicating the patient had chemotherapy up to 3 months post-BCCA admission, if the information is known. For non-referred cases hormone therapy is included.	
bcca_horm	A flag indicating the patient had hormone therapy up to 3 months post-BCCA admission, if the information is known. For non-referred cases see chemotherapy.	
bcca_horm_desc	Description of the flag indicating the patient had hormone therapy up to 3 months post-BCCA admission, if the information is known. For non-referred cases see chemotherapy.	
bcca_rad	A flag indicating the patient had radiation therapy, including pre-admission non-BCCA RT if the information is known.	
bcca_rad_desc	Description of the flag indicating the patient had radiation therapy, including pre-admission non-BCCA RT if the information is known.	
bcca_surg	A flag indicating the patient had other than diagnostic surgery up to 3 months post-BCCA admission, if the information is known.	

Column Name	Description	Comments
bcca_surg_desc	Description of the flag indicating the patient had other than diagnostic surgery up to 3 months post-BCCA admission, if the information is known.	
fst_treat_date	The date of the patient's first definitive treatment for a distinct primary disease.	
known_fst_treat_date	The fst_treat_date as entered (with '0's) without system-generated '01's.	
not_treated	A code indicating the reason a patient had no initial treatment.	
not_treated_desc	A description of the code indicating the reason a patient had no initial treatment.	
rt_start_date	The date radiation therapy treatment started.	Radiation treatment data are not available for nonreferred cases (ie. cases not referred to a BC cancer centre)
rt_end_date	The date radiation therapy treatment stopped.	

Column Name	Description	Comments
rt_treat_intent	A flag which describes how the radiation therapy fits into the treatment protocol. A = Adjuvant P = Palliative R = Radical X = Unknown	
rt_treat_plan	Describes how the radiotherapy fits into the treatment protocol. I = Initial Treatment U = Subsequent Treatment	
rt_treat_region	The anatomic site where the patient received radiotherapy treatment.	
tx_region_desc	The description of the anatomic site where the patient received radiotherapy treatment.	
surg_treat_date	The date the surgery was performed.	Surgery data are not available for nonreferred cases (ie. cases not referred to a BC cancer centre)
surg_treat_intent	A flag indicating the expected result of the surgical treatment.	
surg_treat_intent_desc	A description of the flag indicating the expected result of the surgical treatment.	

Column Name	Description	Comments
surg_treat_plan	A flag which indicating how the surgery fits into the treatment protocol	
surg_treat_plan_desc	A description of the flag which indicates how the surgery fits into the treatment protocol	
surg_code	A code (Canadian Classification of Diagnostic, Therapeutic, and Surgical Procedures) used to define the surgery performed	The surgery codes are listed in multiple fields if more than one surgery was performed on a single date ie. surg_code0, surg_code1.. up to surg_code10.

Source: CAIS Patient Information

Date Retrieved: November 26, 2012

Prepared by: Cancer Surveillance & Outcomes datareq@bccancer.bc.ca

APPENDIX B

KARNOFSKY PERFORMANCE STATUS SCALE DEFINITIONS RATING (%) CRITERIA

Able to carry on normal activity and to work; no special care needed.	100	Normal no complaints; no evidence of disease.
	90	Able to carry on normal activity; minor signs or symptoms of disease.
	80	Normal activity with effort; some signs or symptoms of disease.
Unable to work; able to live at home and care for most personal needs; varying amount of assistance needed.	70	Cares for self; unable to carry on normal activity or to do active work.
	60	Requires occasional assistance, but is able to care for most of his personal needs.
	50	Requires considerable assistance and frequent medical care.
Unable to care for self; requires equivalent of institutional or hospital care; disease may be	40	Disabled; requires special care and assistance.

progressing rapidly.	30	Severely disabled; hospital admission is indicated although death not imminent.
	20	Very sick; hospital admission necessary; active supportive treatment necessary.
	10	Moribund; fatal processes progressing rapidly.
	0	Deceased

Source: http://www.hospicepatients.org/karnofsky.html

APPENDIX C

C.1. FP-Growth Association Rules

C.1.1. BC Prior Information Data Set:

Support levels lower than 1300 reveal too much detail and are not clear enough in distinguishing the most frequent factors.

Candidate 1-itemsets:

Factors	Support Count	Support
Current health authority Fraser	1462	0.241773
Diagnosis health authority Fraser	1546	0.255664
Diagnosis Age range 60 through 69	1484	0.245411
Diagnosis Age range 70 through 79	1671	0.276335
Current health authority Vancouver Coastal	1755	0.290227
Female	1874	0.309906
Diagnosis health authority Vancouver Coastal	1883	0.311394
Male	4173	0.690094
Hepatocellular carcinoma	3413	0.564412
Liver site	5355	0.885563
Liver subgroup	5812	0.961138

Candidate 2-itemsets:

Factors	Support Count	Support
Current health authority Fraser, Liver subgroup	1394	0.230528
Current health authority Fraser, Diagnosis health authority Fraser	1409	0.233008
Diagnosis Age range 60 through 69, Liver site	1344	0.222259
Diagnosis Age range 60 through 69, Liver subgroup	1429	0.236316
Diagnosis health authority Fraser, Liver site	1360	0.224905
Diagnosis health authority Fraser, Liver subgroup	1478	0.244419
Diagnosis Age range 70 through 79, Liver site	1456	0.240781
Diagnosis Age range 70 through 79, Liver subgroup	1619	0.267736
Current health authority Vancouver Coastal, Liver site	1598	0.264263
Current health authority Vancouver Coastal, Liver subgroup	1692	0.279808
Current health authority Vancouver Coastal, Diagnosis health authority Vancouver Coastal	1706	0.282123
Female, Liver site	1535	0.253845
Female, Liver subgroup	1795	0.296841
Diagnosis health authority Vancouver Coastal, Male	1338	0.221267
Diagnosis health authority Vancouver Coastal, Liver site	1719	0.284273
Diagnosis health authority Vancouver Coastal, Liver subgroup	1817	0.30048
Hepatocellular carcinoma, Male	2642	0.436911
Hepatocellular carcinoma, Liver site	3413	0.564412
Hepatocellular carcinoma, Liver subgroup	3413	0.564412
Male, Liver site	3820	0.631718
Male, Liver subgroup	4017	0.664296
Liver site, Liver subgroup	5120	0.846701

Candidate 3-itemsets:

Factors	Support Count	Support
Current health authority Fraser, Liver subgroup, Diagnosis health authority Fraser	1347	0.222755
Diagnosis Age range 70 through 79, Liver site, Liver subgroup	1404	0.232181
Current health authority Vancouver Coastal, Liver site, Liver subgroup	1535	0.253845
Current health authority Vancouver Coastal, Liver site, Diagnosis health authority Vancouver Coastal	1553	0.256822
Current health authority Vancouver Coastal, Liver subgroup, Diagnosis health authority Vancouver Coastal	1645	0.272036
Female, Liver site, Liver subgroup	1456	0.240781
Diagnosis health authority Vancouver Coastal, Liver site, Liver subgroup	1653	0.273359
Hepatocellular carcinoma, Liver site, Male	2642	0.436911
Hepatocellular carcinoma, Liver subgroup, Male	2642	0.436911
Hepatocellular carcinoma, Liver site, Liver subgroup	3413	0.564412
Male, Liver site, Liver subgroup	3664	0.60592

Candidate 4-itemsets:

Factors	Support Count	Support
Current health auth. Vancouver Coastal, Liver site, Liver subgroup, Diagnosis health authority Vancouver Coastal	1492	0.246734
Hepatocellular carcinoma, Liver site, Liver subgroup, Male	2642	0.436911

After eliminating all the rules that are not relevant because they do not contain liver cancer, we are left with:

Candidate 2-itemsets:

Factors	Support Count	Support
Current health authority Fraser, Liver subgroup	1394	0.230527534
Diagnosis Age range 60 through 69, Liver site	1344	0.222258971
Diagnosis Age range 60 through 69, Liver subgroup	1429	0.236315528
Diagnosis health authority Fraser, Liver site	1360	0.224904912
Diagnosis health authority Fraser, Liver subgroup	1478	0.24441872
Diagnosis Age range 70 through 79, Liver site	1456	0.240780552
Diagnosis Age range 70 through 79, Liver subgroup	1619	0.267736067
Current health authority Vancouver Coastal, Liver site	1598	0.264263271
Current health authority Vancouver Coastal, Liver subgroup	1692	0.279808169
Female, Liver site	1535	0.253844882
Female, Liver subgroup	1795	0.296841409
Diagnosis health authority Vancouver Coastal, Liver site	1719	0.284273193
Diagnosis health authority Vancouver Coastal, Liver subgroup	1817	0.300479577
Hepatocellular carcinoma, Male	2642	0.436910865
Hepatocellular carcinoma, Liver site	3413	0.564412105
Hepatocellular carcinoma, Liver subgroup	3413	0.564412105
Male, Liver site	3820	0.631718207
Male, Liver subgroup	4017	0.664296345
Liver site, Liver subgroup	5120	0.846700843

Candidate 3-itemsets:

Factors	Support Count	Support
Current health authority Fraser, Liver subgroup, Diagnosis health authority Fraser	1347	0.222755085
Diagnosis Age range 70 through 79, Liver site, Liver subgroup	1404	0.232181247
Current health authority Vancouver Coastal, Liver site, Liver subgroup	1535	0.253844882
Current health authority Vancouver Coastal, Liver site, Diagnosis health authority Vancouver Coastal	1553	0.256821564
Current health authority Vancouver Coastal, Liver subgroup, Diagnosis health authority Vancouver Coastal	1645	0.27203572
Female, Liver site, Liver subgroup	1456	0.240780552
Diagnosis health authority Vancouver Coastal, Liver site, Liver subgroup	1653	0.27335869
Hepatocellular carcinoma, Liver site, Male	2642	0.436910865
Hepatocellular carcinoma, Liver subgroup, Male	2642	0.436910865
Hepatocellular carcinoma, Liver site, Liver subgroup	3413	0.564412105
Male, Liver site, Liver subgroup	3664	0.605920291

Candidate 4-itemsets:

Factors	Support Count	Support
Current health authority Vancouver Coastal, Liver site, Liver subgroup, Diagnosis health authority Vancouver Coastal	1492	0.246734
Hepatocellular carcinoma, Liver site, Liver subgroup, Male	2642	0.436911

From the 2-, 3-, and 4-itemsets above we can obtain the following rules and their corresponding confidence levels:

Candidate 2-itemsets:

Rules	Confidence
Current health authority Fraser → Liver subgroup	1394/1462 = 0.95
Diagnosis Age range 60 through 69 → Liver site	1344/1484 =0.91
Diagnosis Age range 60 through 69 → Liver subgroup	1429/1484 =0.96
Diagnosis health authority Fraser → Liver site	1360/1546 =0.88
Diagnosis health authority Fraser → Liver subgroup	1478/1546 = 0.96
Diagnosis Age range 70 through 79 → Liver site	1456/1671 = 0.87
Diagnosis Age range 70 through 79 → Liver subgroup	1619/1671 = 0.97
Current health authority Vancouver Coastal → Liver site	1598/1755 = 0.91
Current health authority Vancouver Coastal → Liver subgroup	1692/1755 = 0.96
Female → Liver site	1535/1874 = 0.82
Female → Liver subgroup	1795/1874 = 0.96
Diagnosis health authority Vancouver Coastal → Liver site	1719/1883 = 0.91
Diagnosis health authority Vancouver Coastal → Liver subgroup	1817/1883 = 0.96
Male → Hepatocellular carcinoma	2642/4173 = 0.63
Male → Liver site	3820/4173 = 0.92
Male → Liver subgroup	4017/4173 = 0.96

Candidate 3-itemsets:

Rules	Confidence
Current health authority Fraser, Diagnosis health authority Fraser → Liver subgroup	1347/1409 = 0.96
Diagnosis Age range 70 through 79 → Liver subgroup, Liver site	1404/1671 = 0.84
Current health authority Vancouver Coastal → Liver site, Liver subgroup	1535/1755 = 0.88
Current health authority Vancouver Coastal, Diagnosis health authority Vancouver Coastal → Liver site	1553/1706 = 0.91
Current health authority Vancouver Coastal, Diagnosis health authority Vancouver Coastal → Liver subgroup	1645/1706 = 0.96
Female → Liver subgroup, Liver site	1456/1874 = 0.78
Diagnosis health authority Vancouver Coastal → Liver subgroup, Liver site	1653/1883 = 0.88
Male → Liver site, Hepatocellular carcinoma	2642/4173 = 0.63
Male → Liver subgroup, Liver site	3664/4173 = 0.88

Candidate 4-itemsets:

Rules	Confidence
Current health authority Vancouver Coastal, Diagnosis health authority Vancouver Coastal → Liver subgroup, Liver site	1492/1706 = 0.88
Male → Liver subgroup, Hepatocellular carcinoma, Liver site	2642/4173 = 0.63

Below are the association rules that survive with a minimum confidence level of 0.85.

Candidate 2-itemsets:

Rules	Confidence
Current health authority Fraser → Liver subgroup	1394/1462 = 0.95
Diagnosis Age range 60 through 69 → Liver site	1344/1484 =0.91
Diagnosis Age range 60 through 69 → Liver subgroup	1429/1484 =0.96
Diagnosis health authority Fraser → Liver site	1360/1546 =0.88
Diagnosis health authority Fraser → Liver subgroup	1478/1546 = 0.96
Diagnosis Age range 70 through 79 → Liver site	1456/1671 = 0.87
Diagnosis Age range 70 through 79 → Liver subgroup	1619/1671 = 0.97
Current health authority Vancouver Coastal → Liver site	1598/1755 = 0.91
Current health authority Vancouver Coastal → Liver subgroup	1692/1755 = 0.96
Female → Liver subgroup	1795/1874 = 0.96
Diagnosis health authority Vancouver Coastal → Liver site	1719/1883 = 0.91
Diagnosis health authority Vancouver Coastal → Liver subgroup	1817/1883 = 0.96
Male → Liver site	3820/4173 = 0.92
Male → Liver subgroup	4017/4173 = 0.96

Candidate 3-itemsets:

Rules	Confidence
Current health authority Fraser, Diagnosis health authority Fraser → Liver subgroup	1347/1409 = 0.96
Current health authority Vancouver Coastal → Liver site, Liver subgroup	1535/1755 = 0.88
Current health authority Vancouver Coastal, Diagnosis health authority Vancouver Coastal → Liver site	1553/1706 = 0.91
Current health authority Vancouver Coastal, Diagnosis health authority Vancouver Coastal → Liver subgroup	1645/1706 = 0.96
Diagnosis health authority Vancouver Coastal → Liver subgroup, Liver site	1653/1883 = 0.88
Male → Liver subgroup, Liver site	3664/4173 = 0.88

Candidate 4-itemsets:

Rules	Confidence
Current health authority Vancouver Coastal, Diagnosis health authority Vancouver Coastal → Liver subgroup, Liver site	1492/1706 = 0.88

C.1.2. Yukon Prior Data Set:

Minimum support chosen was 3.

Candidate 1-itemsets:

Factors	Support Count	Support
Diagnosis Age range 50 through 59	3	0.176471
Diagnosis Age range 0 through 9	3	0.176471
Malignant neoplasm	3	0.176471
Diagnosis Age range 40 through 49	3	0.176471
Diagnosis Age range 60 through 69	5	0.294118
Female	4	0.235294
Hepatocellular carcinoma	9	0.529412
Male	13	0.764706
Liver site	16	0.941176

Candidate 2-itemsets:

Factors	Support Count	Support
Diagnosis Age range 50 through 59, Hepatocell. carcinoma	3	0.176471
Diagnosis Age range 50 through 59, Male	3	0.176471
Diagnosis Age range 50 through 59, Liver site	3	0.176471
Diagnosis Age range 0 through 9, Female	3	0.176471
Diagnosis Age range 0 through 9, Liver site	3	0.176471
Malignant neoplasm, Male	3	0.176471
Diagnosis Age range 40 through 49, Male	3	0.176471
Diagnosis Age range 40 through 49, Liver site	3	0.176471
Female, Liver site	4	0.235294
Diagnosis Age range 60 through 69, Hepatocell. carcinoma	4	0.235294
Diagnosis Age range 60 through 69, Male	4	0.235294
Age range 60 through 69, Liver site	5	0.294118
Hepatocellular carcinoma, Male	8	0.470588
Hepatocellular carcinoma, Liver site	9	0.529412
Male, Liver site	12	0.705882

Candidate 3-itemsets:

Factors	Support Count	Support
Diagnosis age range 50 - 59, Male, Hepatoc. carcinoma	3	0.176471
Diagnosis age range 50 - 59, Liver site, Hepatoc. carcinoma	3	0.176471
Diagnosis Age range 50 through 59, Liver site, Male	3	0.176471
Diagnosis Age range 0 through 9, Liver site, Female	3	0.176471
Diagnosis Age range 40 through 49, Liver site, Male	3	0.176471
Diagnosis Age range 60 - 69, Male, Hepatoc. carcinoma	4	0.235294
Diagnosis Age range 60 - 69, Liver site, Hepatoc. carcinoma	4	0.235294
Diagnosis Age range 60 through 69, Liver site, Male	4	0.235294
Hepatocellular carcinoma, Liver site, Male	8	0.470588

Candidate 4-itemsets:

Factors	Support Count	Support
Diagnosis Age range 50 through 59, Liver site, Male, Hepatocellular carcinoma	3	0.176471
Diagnosis Age range 60 through 69, Liver site, Male, Hepatocellular carcinoma	4	0.235294

After eliminating all the rules that are not relevant because they do not contain liver cancer, we are left with:

Candidate 2-itemsets

Factors	Support Count	Support
Diagnosis Age range 50 through 59, Hepatocellular carcinoma	3	0.176471
Diagnosis Age range 50 through 59, Liver site	3	0.176471
Diagnosis Age range 0 through 9, Liver site	3	0.176471
Malignant neoplasm, Male	3	0.176471
Diagnosis Age range 40 through 49, Liver site	3	0.176471
Female, Liver site	4	0.235294
Diagnosis Age range 60 through 69, Hepatocellular carcinoma	4	0.235294
Diagnosis Age range 60 through 69, Liver site	5	0.294118
Hepatocellular carcinoma, Male	8	0.470588
Hepatocellular carcinoma, Liver site	9	0.529412
Male, Liver site	12	0.705882

Candidate 3-itemsets:

Factors	Support Count	Support
Diagnosis Age range 50 through 59, Male, Hepatocellular carcinoma	3	0.176471
Diagnosis Age range 50 through 59, Liver site, Hepatocellular carcinoma	3	0.176471
Diagnosis Age range 60 through 69, Male, Hepatocellular carcinoma	4	0.235294
Diagnosis Age range 60 through 69, Liver site, Hepatocellular carcinoma	4	0.235294
Hepatocellular carcinoma, Liver site, Male	8	0.470588

Candidate 4-itemsets:

Factors	Support Count	Support
Diagnosis Age range 50 through 59, Liver site, Male, Hepatocellular carcinoma	3	0.176471
Diagnosis Age range 60 through 69, Liver site, Male, Hepatocellular carcinoma	4	0.235294

From the itemsets above we obtain the following rules and their confidence levels:

Candidate 2-itemsets:

Rules	Confidence
Diagnosis Age range 50 through 59 → Hepatocellular carcinoma	3/3 = 1
Diagnosis Age range 50 through 59 → Liver site	3/3 = 1
Diagnosis Age range 0 through 9 → Liver site	3/3= 1
Malignant neoplasm → Male	3/3 = 1
Diagnosis Age range 40 through 49 → Liver site	3/3 = 1
Female → Liver site	4/4 = 1
Diagnosis Age range 60 through 69 → Hepatocellular carcinoma	4/5 = 0.8
Diagnosis Age range 60 through 69 → Liver site	5/5 = 1
Male → Hepatocellular carcinoma	8/13 = 0.62
Male → Liver site	12/13 = 0.92

Candidate 3-itemsets:

Rules	Confidence
Diagnosis Age range 50 through 59, Male → Hepatocell. carcinoma	3/3 = 1
Diagnosis Age range 50 through 59 → Hepatocell. carcinoma, Liver site	3/3 = 1
Diagnosis Age range 60 through 69, Male → Hepatocell. carcinoma	4/4 = 1
Diagnosis Age range 60 through 69 → Hepatocellular carcinoma, Liver site	4/5 = 0.8
Male → Liver site, Hepatocellular carcinoma	8/13 = 0.62

Candidate 4-itemsets:

Rules	Confidence
Diagnosis Age range 50 through 59, Male → Hepatocellular carcinoma, Liver site	3/3 = 1
Diagnosis Age range 60 through 69, Male → Hepatocellular carcinoma, Liver site	4/4 = 1

Below are the association rules that survive if we set the minimum confidence level at 0.85.

Candidate 2-itemsets:

Rules	Confidence
Diagnosis Age range 50 through 59 → Hepatocellular carcinoma	3/3 = 1
Diagnosis Age range 50 through 59 → Liver site	3/3 = 1
Diagnosis Age range 0 through 9 → Liver site	3/3= 1
Malignant neoplasm → Male	3/3 = 1
Diagnosis Age range 40 through 49 → Liver site	3/3 = 1
Female → Liver site	4/4 = 1
Diagnosis Age range 60 through 69 → Liver site	5/5 = 1
Male → Liver site	12/13 = 0.92

Candidate 3-itemsets:

Rules	Confidence
Diagnosis Age range 50 through 59, Male → Hepatocell. carcinoma	3/3 = 1
Diagnosis Age range 50 through 59 → Hepatocellular carcinoma, Liver site	3/3 = 1
Diagnosis Age range 60 through 69, Male → Hepatocell. carcinoma	4/4 = 1

Candidate 4-itemsets:

Rules	Confidence
Diagnosis Age range 50 through 59, Male → Hepatocellular carcinoma, Liver site	3/3 = 1
Diagnosis Age range 60 through 69, Male → Hepatocellular carcinoma, Liver site	4/4 = 1

C.1.3. BC Survivability Data Set:

C.1.3.1. For the BC survivability data set excluding records where treatment was undefined, the ideal support level was 520.

Candidate 2-itemsets:

Factors	Support Count	Support
Current health authority Vancouver Coastal, Hepatocellular carcinoma	523	0.312239
Current health authority Vancouver Coastal, Deceased	528	0.315224
Current health authority Vancouver Coastal, Liver tumor	608	0.362985
Current health authority Vancouver Coastal, Diagnosis Vancouver Coastal	611	0.364776
Current health authority Vancouver Coastal, Liver site	611	0.364776
Diagnosis health authority Vancouver Coastal, Hepatocell. carcinoma	526	0.31403
Diagnosis health authority Vancouver Coastal, Deceased	531	0.317015
Diagnosis health authority Vancouver Coastal, Liver tumor	613	0.36597
Diagnosis health authority Vancouver Coastal, Liver site	616	0.367761
Male, Hepatocellular carcinoma	989	0.590448
Male, Deceased	1057	0.631045
Male, Liver site	1183	0.706269
Male, Liver tumor	1187	0.708657
Hepatocellular carcinoma, Deceased	1117	0.666866
Hepatocellular carcinoma, Liver tumor	1264	0.754627
Hepatocellular carcinoma, Liver site	1264	0.754627
Liver site, Deceased	1330	0.79403
Liver tumor, Deceased	1370	0.81791
Liver site, Liver tumor	1504	0.89791

Candidate 3-itemsets:

Factors	Support Count	Support
Current health auth. Vancouver Coastal, Hepatocell. carcinoma, Liver tumor	523	0.312239
Current health auth. Vancouver Coastal, Hepatocell. carcinoma, Liver site	523	0.312239
Current health authority Vancouver Coastal, Liver tumor, Diagnosis health authority Vancouver Coastal	589	0.351642
Current health authority Vancouver Coastal, Liver tumor, Liver site	589	0.351642
Current health authority Vancouver Coastal, Diagnosis health authority Vancouver Coastal, Liver site	592	0.353433
Diagnosis health auth. Vancouver Coastal, Hepatocell. carcinoma, Liver tumor	526	0.31403
Diagnosis health auth. Vancouver Coastal, Hepatocell. carcinoma, Liver site	526	0.31403
Diagnosis health authority Vancouver Coastal, Liver tumor, Liver site	593	0.35403
Male, Hepatocellular carcinoma, Deceased	882	0.526567
Male, Hepatocellular carcinoma, Liver site	989	0.590448
Male, Hepatocellular carcinoma, Liver tumor	989	0.590448
Male, Liver site, Deceased	1015	0.60597
Male, Liver tumor, Deceased	1033	0.616716
Male, Liver site, Liver tumor	1140	0.680597
Hepatocellular carcinoma, Liver site, Deceased	1117	0.666866
Hepatocellular carcinoma, Liver tumor, Deceased	1117	0.666866
Hepatocellular carcinoma, Liver tumor, Liver site	1264	0.754627
Liver site, Liver tumor, Deceased	1287	0.768358

Candidate 1-itemsets:

Factors	Support Count	Support
Current health authority Vancouver Coastal	630	0.376119
Current health authority Fraser	520	0.310448
Diagnosis health authority Vancouver Coastal	636	0.379701
Male	1230	0.734328
Hepatocellular carcinoma	1264	0.754627
Deceased	1413	0.843582
Liver site	1582	0.944478
Liver tumor	1597	0.953433

Candidate 4-itemsets:

Factors	Support Count	Support
Current health authority Vancouver Coastal, Hepatocellular carcinoma, Liver site, Liver tumor	523	0.312239
Current health authority Van Coastal, Diagnosis health authority Vancouver Coastal, Liver site, Liver tumor	570	0.340299
Diagnosis health authority Vancouver Coastal, Hepatocellular carcinoma, Liver site, Liver tumor	526	0.31403
Male, Hepatocellular carcinoma, Liver site, Deceased	882	0.526567
Male, Hepatocellular carcinoma, Liver tumor, Deceased	882	0.526567
Male, Hepatocellular carcinoma, Liver tumor, Liver site	989	0.590448
Male, Liver site, Liver tumor, Deceased	991	0.591642
Hepatocellular carcinoma, Liver tumor, Liver site, Deceased	1117	0.666866

Candidate 5-itemsets:

Factors	Supp. Count	Support
Male, Hepatoc. carcinoma, Liver tumor, Liver site, Deceased	882	0.526567

After eliminating all the rules that are not of interest for not containing either alive, or deceased, we are left with:

Candidate 2-itemsets:

Factors	Support Count	Support
Current health authority Vancouver Coastal, Deceased	528	0.315224
Diagnosis health authority Vancouver Coastal, Deceased	531	0.317015
Male, Deceased	1057	0.631045
Hepatocellular carcinoma, Deceased	1117	0.666866
Liver site, Deceased	1330	0.79403
Liver tumor, Deceased	1370	0.81791

Candidate 3-itemsets:

Factors	Support Count	Support
Male, Hepatocellular carcinoma, Deceased	882	0.526567
Male, Liver site, Deceased	1015	0.60597
Male, Liver tumor, Deceased	1033	0.616716
Hepatocellular carcinoma, Liver site, Deceased	1117	0.666866
Hepatocellular carcinoma, Liver tumor, Deceased	1117	0.666866
Liver site, Liver tumor, Deceased	1287	0.768358

Candidate 4-itemsets:

Factors	Support Count	Support
Male, Hepatocellular carcinoma, Liver site, Deceased	882	0.526567
Male, Hepatocellular carcinoma, Liver tumor, Deceased	882	0.526567
Male, Liver site, Liver tumor, Deceased	991	0.591642
Hepatocellular carcinoma, Liver tumor, Liver site, Deceased	1117	0.666866

Candidate 5-itemsets:

Factors	Support Count	Support
Male, Hepatocell. carcinoma, Liver tumor, Liver site, Deceased	882	0.526567

From the 2-, 3-, 4-, and 5-itemsets above we can obtain the following rules and their corresponding confidence levels:

Candidate 2-itemsets:

Rules	Confidence
Current health authority Vancouver Coastal → Deceased	528/630 = 0.84
Diagnosis health authority Vancouver Coastal → Deceased	531/636 = 0.84
Male → Deceased	1057/1230 = 0.86
Hepatocellular carcinoma → Deceased	1117/1264 = 0.88
Liver site → Deceased	1330/1582 = 0.84
Liver tumor → Deceased	1370/1597 = 0.86

Candidate 3-itemsets:

Rules	Confidence
Male, Hepatocellular carcinoma → Deceased	882/989 = 0.89
Male, Liver site → Deceased	1015/1183 = 0.86
Male, Liver tumor → Deceased	1033/1187 = 0.87
Hepatocellular carcinoma, Liver site → Deceased	1117/1264 = 0.88
Hepatocellular carcinoma, Liver tumor → Deceased	1117/1264 = 0.88
Liver site, Liver tumor → Deceased	1287/1504 = 0.86

Candidate 4-itemsets:

Rules	Confidence
Male, Hepatocellular carcinoma, Liver site → Deceased	882/989 = 0.89
Male, Hepatocellular carcinoma, Liver tumor → Deceased	882/989 = 0.89
Male, Liver site, Liver tumor → Deceased	991/1140 = 0.87
Hepatocellular carcinoma, Liver tumor, Liver site → Deceased	1117/1264 = 0.88

Candidate 5-itemsets:

Rules	Confidence
Male, Hepatocell. carcinoma, Liver tumor, Liver site → Deceased	882/989 = 0.89

Below are the association rules obtained by setting the minimum confidence level at 0.85:

Candidate 2-itemsets:

Rules	Confidence
Male → Deceased	1057/1230 = 0.86
Hepatocellular carcinoma → Deceased	1117/1264 = 0.88
Liver tumor → Deceased	1370/1597 = 0.86

Candidate 3-itemsets:

Rules	Confidence
Male, Hepatocellular carcinoma → Deceased	882/989 = 0.89
Male, Liver site → Deceased	1015/1183 = 0.86
Male, Liver tumor → Deceased	1033/1187 = 0.87
Hepatocellular carcinoma, Liver site → Deceased	1117/1264 = 0.88
Hepatocellular carcinoma, Liver tumor → Deceased	1117/1264 = 0.88
Liver site, Liver tumor → Deceased	1287/1504 = 0.86

Candidate 4-itemsets:

Rules	Confidence
Male, Hepatocellular carcinoma, Liver site → Deceased	882/989 = 0.89
Male, Hepatocellular carcinoma, Liver tumor → Deceased	882/989 = 0.89
Male, Liver site, Liver tumor → Deceased	991/1140 = 0.87
Hepatocellular carcinoma, Liver tumor, Liver site → Deceased	1117/1264 = 0.88

Candidate 5-itemsets:

Rules	Confidence
Male, Hepatocell. carcinoma, Liver tumor, Liver site → Deceased	882/989 = 0.89

C.1.3.2. For the BC survivability data set excluding all treatment attributes, the ideal support level was 1600.

Candidate 1-itemsets:

Factors	Support Count	Support
Current health authority Vancouver Coastal	1755	0.290227
Death age range 70 through 79	1613	0.266744
Diagnosis Age range 70 through 79	1671	0.276335
Female	1874	0.309906
Diagnosis health authority Vancouver Coastal	1883	0.311394
Hepatocellular carcinoma	3413	0.564412
Male	4173	0.690094
Liver site	5355	0.885563
Deceased	5508	0.910865
Liver tumor	5812	0.961138

Candidate 2-itemsets:

Factors	Support Count	Support
Death age range 70 through 79, Deceased	1613	0.266744
Death age range 70 through 79, Liver tumor	1619	0.267736
Current health authority Vancouver Coastal, Liver tumor	1692	0.279808
Female, Deceased	1703	0.281627
Current Vancouver Coastal, diagnosis Vancouver Coastal	1706	0.282123
Female, Liver tumor	1795	0.296841
Diagnosis health authority Vancouver Coastal, Deceased	1677	0.277328
Diagnosis health authority Vancouver Coastal, Liver site	1719	0.284273
Diagnosis health authority Vancouver Coastal, Liver tumor	1817	0.30048
Hepatocellular carcinoma, Deceased	3059	0.505871
Male, Hepatocellular carcinoma	2642	0.436911
Male, Deceased	3805	0.629238
Male, Liver site	3820	0.631718
Male, Liver tumor	4017	0.664296
Hepatocellular carcinoma, Liver site	3413	0.564412
Hepatocellular carcinoma, Liver tumor	3413	0.564412
Liver site, Deceased	4834	0.799405
Liver tumor, Deceased	5331	0.881594
Liver site, Liver tumor	5120	0.846701

Candidate 3-itemsets:

Factors	Support Count	Support
Current health authority Vancouver Coastal, Diagnosis health authority Vancouver Coastal, Liver tumor	1645	0.272036
Female, Liver tumor, Deceased	1646	0.272201
Diagnosis health authority Vancouver Coastal, Liver tumor, Deceased	1630	0.269555
Diagnosis health authority Vancouver Coastal, Liver tumor, Liver site	1653	0.273359
Hepatocellular carcinoma, Male, Deceased	2381	0.393749
Hepatocellular carcinoma, Male, Liver site	2642	0.436911
Hepatocellular carcinoma, Male, Liver tumor	2642	0.436911
Hepatocellular carcinoma, Liver site, Deceased	3059	0.505871
Hepatocellular carcinoma, Liver tumor, Deceased	3059	0.505871
Hepatocellular carcinoma, Liver site, Liver tumor	3413	0.564412
Male, Liver site, Deceased	3459	0.572019
Male, Liver tumor, Deceased	3685	0.609393
Male, Liver site, Liver tumor	3664	0.60592
Liver site, Liver tumor, Deceased	4657	0.770134

Candidate 4-itemsets:

Factors	Support Count	Support
Male, Hepatocellular carcinoma, Liver site, Deceased	2381	0.393749
Male, Hepatocellular carcinoma, Liver tumor, Deceased	2381	0.393749
Male, Hepatocellular carcinoma, Liver tumor, Liver site	2642	0.436911
Male, Liver site, Liver tumor, Deceased	3339	0.552175
Hepatocellular carcinoma, Liver tumor, Liver site, Deceased	3059	0.505871

Candidate 5-itemsets:

Factors	Support Count	Support
Hepatocell. carcinoma, Male, Liver tumor, Liver site, Deceased	2381	0.393749

After eliminating all the rules that are not relevant because they do not contain an alive or deceased attribute, we are left with:

Candidate 2-itemsets:

Factors	Support Count	Support
Death age range 70 through 79, Deceased	1613	0.266744
Female, Deceased	1703	0.281627
Diagnosis health authority Vancouver Coastal, Deceased	1677	0.277328
Hepatocellular carcinoma, Deceased	3059	0.505871
Male, Deceased	3805	0.629238
Liver site, Deceased	4834	0.799405
Liver tumor, Deceased	5331	0.881594

Candidate 3-itemsets:

Factors	Support Count	Support
Female, Liver tumor, Deceased	1646	0.272201
Diagnosis health authority Vancouver Coastal, Liver tumor, Deceased	1630	0.269555
Hepatocellular carcinoma, Male, Deceased	2381	0.393749
Hepatocellular carcinoma, Liver site, Deceased	3059	0.505871
Hepatocellular carcinoma, Liver tumor, Deceased	3059	0.505871
Male, Liver site, Deceased	3459	0.572019
Male, Liver tumor, Deceased	3685	0.609393
Liver site, Liver tumor, Deceased	4657	0.770134

Candidate 4-itemsets:

Factors	Support Count	Support
Male, Hepatocellular carcinoma, Liver site, Deceased	2381	0.393749
Male, Hepatocellular carcinoma, Liver tumor, Deceased	2381	0.393749
Male, Liver site, Liver tumor, Deceased	3339	0.552175
Hepatocellular carcinoma, Liver tumor, Liver site, Deceased	3059	0.505871

Candidate 5-itemsets:

Factors	Support Count	
Hepatocellular carcinoma, Male, Liver tumor, Liver site, Deceased	2381	0.393749

From the 2-, 3-, 4-, and 5-itemsets above we can obtain the following rules and their corresponding confidence levels:

Candidate 2-itemsets:

Rules	Confidence
Death age range 70 through 79 → Deceased	1613/1613 = 1
Female → Deceased	1703/1874 = 0.91
Diagnosis health authority Vancouver Coastal → Deceased	1677/1883 = 0.89
Hepatocellular carcinoma → Deceased	3059/3413 = 0.90
Male → Deceased	3805/4173 = 0.91
Liver site → Deceased	4834/5355 = 0.90
Liver tumor → Deceased	5331/5818 = 0.92

Candidate 3-itemsets:

Rules	Confidence
Female, Liver tumor → Deceased	1646/1795 = 0.92
Diagnosis health authority Vancouver Coastal, Liver tumor → Deceased	1630/1817 = 0.90
Hepatocellular carcinoma, Male → Deceased	2381/2642 = 0.90
Hepatocellular carcinoma, Liver site → Deceased	3059/3413 = 0.90
Hepatocellular carcinoma, Liver tumor → Deceased	3059/3413 = 0.90
Male, Liver site → Deceased	3459/3820 = 0.91
Male, Liver tumor → Deceased	3685/4017 = 0.92
Liver site, Liver tumor → Deceased	4657/5120 = 0.91

Candidate 4-itemsets:

Rules	Confidence
Male, Hepatocellular carcinoma, Liver site → Deceased	2381/2642 = 0.90
Male, Hepatocellular carcinoma, Liver tumor → Deceased	2381/2642 = 0.90
Male, Liver site, Liver tumor → Deceased	3339/3664 = 0.91
Hepatocellular carcinoma, Liver tumor, Liver site → Deceased	3059/3413 = 0.90

Candidate 5-itemsets:

Rules	Confidence
Hepatocellular carcinoma, Male, Liver tumor, Liver site → Deceased	2381/2642 = 0.90

C.1.4. Yukon Survivability Data Set:

Candidate 1-itemsets:

Factors	Support Count	Support
Diagnosis age range 60 through 69	5	0.294118
Female	4	0.235294
Death age range 60 through 69	5	0.294118
Alive	5	0.294118
Hepatocellular carcinoma	9	0.529412
Deceased	12	0.705882
Male	13	0.764706
Liver site	16	0.941176

Candidate 2-itemsets:

Factors	Support Count	Support
Diagnosis age range 60 through 69, Hepatocel. carcinoma	4	0.235294
Diagnosis age range 60 through 69, Male	4	0.235294
Female, Liver site	4	0.235294
Diagnosis age range 60 through 69, Liver site	5	0.294118
Death age range 60 through 69, Hepatocellular carcinoma	4	0.235294
Death age range 60 through 69, Male	4	0.235294
Death age range 60 through 69, Deceased	5	0.294118
Death age range 60 through 69, Liver site	5	0.294118
Liver site, alive	5	0.294118
Hepatocellular carcinoma, Deceased	6	0.352941
Hepatocellular carcinoma, Male	8	0.470588
Male, Deceased	10	0.588235
Hepatocellular carcinoma, Liver site	9	0.529412
Liver site, Deceased	11	0.647059
Male, Liver site	12	0.705882

Candidate 3-itemsets:

Factors	Support Count	Support
Diagnosis age range 60 through 69, Male, Hepatocel. carcinoma	4	0.235294
Diagnosis age range 60 through 69, Liver site, Hepatocel. carcinoma	4	0.235294
Diagnosis age range 60 through 69, Liver site, Male	4	0.235294
Death age range 60 through 69, Male, Hepatocellular carcinoma	4	0.235294
Death age range 60 through 69, Hepatocell. carcinoma, Deceased	4	0.235294
Death age range 60 through 69, Male, Deceased	4	0.235294
Death age range 60 through 69, Hepatocel. carcinoma, Liver site	4	0.235294
Death age range 60 through 69, Male, Deceased	4	0.235294
Death age range 60 through 69, Liver site, Deceased	5	0.294118
Hepatocellular carcinoma, Male, Deceased	6	0.352941
Hepatocellular carcinoma, Liver site, Deceased	6	0.352941
Hepatocellular carcinoma, Liver site, Male	8	0.470588

Candidate 4-itemsets:

Factors	Support Count	Support
Diagnosis age range 60 - 69, Liver site, Male, Hepatocel. carcinoma	4	0.235294
Death age range 60 - 69, Male, Hepatocell. carcinoma, Deceased	4	0.235294
Death age range 60 - 69, Liver site, Male, Hepatocellular carcinoma	4	0.235294
Death age range 60 - 69, Liver site, Hepatocellular carcinoma, Deceased	4	0.235294
Death age range 60 - 69, Liver site, Male, Deceased	4	0.235294
Hepatocellular carcinoma, Liver site, Male, Deceased	6	0.352941

Candidate 5-itemsets:

Factors	Support Count	Support
Death age range 60 through 69, Liver site, Male, Hepatocellular carcinoma, Deceased	4	0.235294

After eliminating all the rules that are not relevant because they do not contain an alive or deceased attribute, we are left with:

Candidate 2-itemsets:

Factors	Support Count	Support
Liver site, Alive	5	0.294118
Hepatocellular carcinoma, Deceased	6	0.352941
Male, Deceased	10	0.588235
Liver site, Deceased	11	0.647059

Candidate 3-itemsets:

Factors	Support Count	Support
Death age range 60 through 69, Hepatocellular carcinoma, Deceased	4	0.235294
Death age range 60 through 69, Male, Deceased	4	0.235294
Death age range 60 through 69, Male, Deceased	4	0.235294
Death age range 60 through 69, Liver site, Deceased	5	0.294118
Hepatocellular carcinoma, Male, Deceased	6	0.352941
Hepatocellular carcinoma, Liver site, Deceased	6	0.352941

Candidate 4-itemsets:

Factors	Support Count	Support
Death age range 60 through 69, Male, Hepatocellular carcinoma, Deceased	4	0.235294
Death age range 60 through 69, Liver site, Hepatocellular carcinoma, Deceased	4	0.235294
Death age range 60 through 69, Liver site, Male, Deceased	4	0.235294
Hepatocellular carcinoma, Liver site, Male, Deceased	6	0.352941

Candidate 5-itemsets:

Factors	Support Count	Support
Death age range 60 through 69, Liver site, Male, Hepatocellular carcinoma, Deceased	4	0.235294

From the previous itemsets we can obtain the following rules and their confidence levels:

Candidate 2-itemsets:

Rules	Confidence
Liver site → alive	5/16 = 0.31
Hepatocellular carcinoma → Deceased	6/9 = 0.67
Male → Deceased	10/13 = 0.77
Liver site → Deceased	11/16 = 0.69

Candidate 3-itemsets:

Rules	Confidence
Death age range 60 through 69, Hepatocellular carcinoma → Deceased	4/4 = 1
Death age range 60 through 69, Male → Deceased	4/4 = 1
Death age range 60 through 69, Liver site → Deceased	5/5 = 1
Hepatocellular carcinoma, Male → Deceased	6/8 = 0.75
Hepatocellular carcinoma, Liver site → Deceased	6/9 = 0.67

Candidate 4-itemsets:

Rules	Confidence
Death age range 60 through 69, Male, Hepatocellular carcinoma → Deceased	4/4 = 1
Death age range 60 through 69, Liver site, Hepatocellular carcinoma → Deceased	4/4 = 1
Death age range 60 through 69, Liver site, Male → Deceased	4/4 = 1
Hepatocellular carcinoma, Liver site, Male → Deceased	6/8 = 0.75

Candidate 5-itemsets:

Rules	Confidence
Death age range 60 through 69, Liver site, Male, Hepatocellular carcinoma → Deceased	4/4 = 1

Below are the association rules that survive after setting the minimum confidence level at 0.85:

Candidate 3-itemsets:

Rules	Confidence
Death age range 60 through 69, Hepatocellular carcinoma → Deceased	4/4 = 1
Death age range 60 through 69, Male → Deceased	4/4 = 1
Death age range 60 through 69, Liver site → Deceased	5/5 = 1

Candidate 4-itemsets:

Rules	Confidence
Death age range 60 through 69, Male, Hepatocellular carcinoma → Deceased	4/4 = 1
Death age range 60 through 69, Liver site, Hepatocellular carcinoma → Deceased	4/4 = 1
Death age range 60 through 69, Liver site, Male → Deceased	4/4 = 1

Candidate 5-itemsets:

Rules	Confidence
Death age range 60 through 69, Liver site, Male, Hepatocellular carcinoma → Deceased	4/4 = 1

C.2. Apriori Association Rules

C.2.1. BC Preliminary Information Data Set:

The 20 best rules found with minimum confidence 0.85 and support 1512 (25%) were:

Rules	Confidence
Hepatocellular carcinoma ==> Liver site	3410/3410 = 1
Hepatocellular carcinoma ==> Liver tumour subgroup	3410/3410 = 1
Hepatocellular carcinoma, Liver tumour subgroup ==> Liver site	3410/3410 = 1
Hepatocellular carcinoma, Liver site ==> Liver tumour subgroup	3410/3410 = 1
Hepatocellular carcinoma ==> Liver site, Liver tumour subgroup	3410/3410 = 1
Hepatocellular carcinoma, Male ==> Liver site	2640/2640 = 1
Hepatocellular carcinoma, Male ==> Liver tumour subgroup	2640/2640 = 1
Hepatocell. carcinoma, Male, Liver tumour subgroup ==> Liver site	2640/2640 = 1
Hepatocell. carcinoma, Liver site, Male ==> Liver tumour subgroup	2640/2640 = 1
Hepatocell. carcinoma, Male ==> Liver site, Liver tumour subgroup	2640/2640 = 1
Vancouver Coastal current health authority, Liver tumour subgroup ==> Vancouver Coastal diagnosis health authority	1645/1692 = 0.97
Vancouver Coastal current health authority ==> Vancouver Coastal diagnosis health authority	1706/1755 = 0.97
Vancouver Coastal current health authority, Liver site ==> Vancouver Coastal diagnosis health authority	1553/1598 = 0.97
Diagnosis age range 70 to 79 ==> Liver tumour subgroup	1619/1671 = 0.97
Vancouver Coastal diagnosis health auth.==> Liver tumour subgroup	1817/1883 = 0.96
Vancouver Coastal current health authority, Vancouver Coastal diagnosis health authority ==> Liver tumour subgroup	1645/1706 = 0.96
Vancouver Coastal current health auth. ==> Liver tumour subgroup	1692/1755 = 0.96
Male ==> Liver tumour subgroup	4017/4173 = 0.96
Van. Coastal diagn. health auth., Liver site ==> Liver tumour subgrp.	1653/1719 = 0.96
Vancouver Coastal current health authority, Liver site ==> Liver tumour subgroup	1535/1598 = 0.96

For the Apriori association rules given in this section, the numerator denotes the number of instances for which the antecedent is true, whereas the denominator is the number of instances in which the consequent is also true. The confidence (in parenthesis) is the ratio between the numerator and the denominator (Witten & Frank, 2005).

The rules above were cleaned up to eliminate obvious rules, such as where both precedent and consequent mean that the patient has liver cancer. Indicators of liver cancer were moved to the consequent whenever they appeared in the rule precedent. The rules below are obtained:

Rules	Confidence
Male ==> Liver site, Hepatocellular carcinoma	2640/2640 = 1
Male ==> Liver tumour subgroup, Hepatocellular carcinoma	2640/2640 = 1
Male ==> Liver site, Hepatocellular carcinoma, Liver tumour subgroup	2640/2640 = 1
Male ==> Liver tumour subgroup, Hepatocell. carcinoma, Liver site	2640/2640 = 1
Vancouver Coastal current health authority, Vancouver Coastal diagnosis health authority ==> Liver site	1553/1598 = 0.97
Diagnosis age range 70 to 79 ==> Liver tumour subgroup	1619/1671 = 0.97
Vancouver Coastal diagnosis health authority ==> Liver tumour subgroup	1817/1883 = 0.96
Vancouver Coastal current health authority, Vancouver Coastal diagnosis health authority ==> Liver tumour subgroup	1645/1706 = 0.96
Vancouver Coastal current health authority ==> Liver tumour subgroup	1692/1755 = 0.96
Male ==> Liver tumour subgroup	4017/4173 = 0.96
Vancouver Coastal diagnosis health authority or ==> Liver tumour subgroup, Liver site	1535/1598= 1653/1719 = 0.96

C.2.2. Yukon Preliminary Information Data Set:

The ten best Apriori association rules found with a minimum confidence of 0.85 and a support level of 4 (25%) are:

Rules	Confidence
Hepatocellular carcinoma ==> Liver site	9/9 = 1
Male, Hepatocellular carcinoma ==> Liver site	8/9 = 1
Diagnosis age range 60 to 69 ==> Liver site	5/5 = 1
Female ==> Liver site	4/4 = 1
Diagnosis age range 60 to 69, Hepatocellular carcinoma ==> Male	4/4 = 1
Male, Diagnosis age range 60 to 69 ==> Hepatocellular carcinoma	4/4 = 1
Male, Diagnosis age range 60 to 69 ==> Liver site	4/4 = 1
Diagnosis age range 60 to 69, Hepatocellular carcinoma ==> Liver site	4/4 = 1
Diagnosis age range 60 to 69, Hepatocellular carcinoma, Liver ==> Male	4/4 = 1
Male, Diagnosis age range 60 to 69, Liver site ==> Hepatocellular carcinoma	4/4 = 1

The rules above were cleaned up to eliminate obvious rules, such as where both precedent and consequent mean that the patient has liver cancer. Indicators of liver cancer were also moved to the consequent whenever they appeared in the rule precedent. After those modifications, the Apriori association rules below are obtained:

Rules	Confidence
Hepatocellular carcinoma 9 ==> Liver site 9	1
Male ==> Liver site, Hepatocellular carcinoma	8/8 = 1
Diagnosis age range 60 to 69 ==> Liver site	5/5 = 1
Female ==> Liver site	4/4 = 1
Male, Diagnosis age range 60 to 69 ==> Hepatocellular carcinoma	4/4 = 1
Male, Diagnosis age range 60 to 69 ==> Liver site	4/4 = 1
Diagnosis age range 60 to 69 ==> Liver site, Hepatocellular carcinoma	4/4 = 1
Male, Diagnosis age range 60 to 69 ==> Hepatocellular carcinoma, Liver site	4/4 = 1

C.2.3. BC Survivability Data set:

C.2.3.1. For the BC survivability data set excluding records where treatment was undefined, the best Apriori association rules found with a minimum confidence of 0.85 and a support level of 837 (50%) were:

Rules	Confidence
Hepatocellular carcinoma ==> Liver site	1262/1262 = 1
Hepatocellular carcinoma ==> Liver tumour	1262/1262 = 1
Hepatocellular carcinoma, Liver tumour ==> Liver site	1262/1262 = 1
Hepatocellular carcinoma, Liver site ==> Liver tumour	1262/1262 = 1
Hepatocellular carcinoma ==> Liver site, Liver tumour	1262/1262 = 1
Hepatocellular carcinoma, Deceased ==> Liver site	1115/1115 = 1
Hepatocellular carcinoma, Deceased ==> Liver tumour	1115/1115 = 1
Hepatocel. carcinoma, Liver tumour, Deceased ==> Liver site	1115/1115 = 1
Hepatocel. carcinoma, Liver site, Deceased ==> Liver tumour	1115/1115 = 1
Hepatocellular carcinoma, Deceased ==> Liver site, Liver tumour	1115/1115 = 1
Hepatocellular carcinoma, Male ==> Liver site	988/988 = 1
Hepatocellular carcinoma, Male ==> Liver tumour	988/988 = 1
Hepatocell. carcinoma, Male, Liver tumour subgroup ==> Liver site	988/988 = 1
Hepatocell. carcinoma, Liver site, Male ==> Liver tumour	988/988 = 1
Hepatocellular carcinoma, Male ==> Liver site, Liver tumour	988/988 = 1
Hepatocellular carcinoma, Male, Deceased ==> Liver site	881/881 = 1
Hepatocellular carcinoma, Male, Deceased ==> Liver tumour	881/881 = 1
Hepatocel. carcinoma, Male, Liver tumour, Deceased ==> Liver site	881/881 = 1
Hepatocel. carcinoma, Liver site, Male, Deceased ==> Liver tumour	881/881 = 1
Hepatocel. carcinoma, Male, Deceased ==> Liver site, Liver tumour	881/881 = 1
Male, Deceased ==> Liver tumour	1033/1057 = 0.98
Liver site, Male, Deceased ==> Liver tumour	991/1015 = 0.98
Deceased ==> Liver tumour	1370/1413 = 0.97

Liver site, Deceased ==> Liver tumour	1287/1330 = 0.97
Male ==> Liver tumour	1187/1230 = 0.97
Liver site Male ==> Liver tumour	1140/1183 = 0.96
Male ==> Liver site	1183/1230 = 0.96
Male, Liver tumour ==> Liver site	1140/1187 = 0.96
Male, Deceased ==> Liver site	1015/1057 = 0.96
Male, Liver tumour, Deceased ==> Liver site	1033/991 = 0.96
Liver site ==> Liver tumour	1504/1582 = 0.95
Liver tumour ==> Liver site	1504/1597 = 0.94
Deceased ==> Liver site	1330/1413 = 0.94
Liver tumour, Deceased ==> Liver site	1287/1370 = 0.94
Male, Deceased ==> Liver site, Liver tumour	991/1057 = 0.94
Male ==> Liver site, Liver tumour	1140/1230 = 0.93
Deceased ==> Liver site, Liver tumour	1287/1413 = 0.91
Hepatocellular carcinoma, Male ==> Deceased	881/988 = 0.89
Hepatocellular carcinoma, Liver site, Male ==> Deceased	881/988 = 0.89
Hepatocellular carcinoma, Male ==> Liver site, Deceased	881/988 = 0.89
Hepatocel. carcinoma, Male, Liver tumour ==> Deceased	881/988 = 0.89
Hepatocel. carcinoma, Male ==> Liver tumour, Deceased	881/988 = 0.89
Hepatocel. carcinoma, Liver site, Male, Liver tumour ==> Deceased	881/988 = 0.89
Hepatocel. carcinoma, Male, Liver tumour ==> Liver site, Deceased	881/988 = 0.89
Hepatocel. carcinoma, Liver site, Male ==> Liver tumour, Deceased	881/988 = 0.89
Hepatocel. carcinoma, Male ==> Liver site, Liver tumour, Deceased	881/988 = 0.89
Liver site, Male, Liver tumour, Deceased ==> Hepatocel. carcinoma	881/991 = 0.89
Hepatocellular carcinoma ==> Deceased	1115/1262 = 0.88
Hepatocellular carcinoma, Liver site ==> Deceased	1115/1262 = 0.88
Hepatocellular carcinoma ==> Liver site, Deceased	1115/1262 = 0.88

After eliminating all the rules that are not relevant because they do not contain either alive or deceased in the rule consequent, we are left with:

Rules	Confidence
Hepatocellular carcinoma, Male ==> Deceased	881/988 = 0.89
Hepatocellular carcinoma, Liver site, Male ==> Deceased	881/988 = 0.89
Hepatocellular carcinoma, Male ==> Liver site, Deceased	881/988 = 0.89
Hepatocellular carcinoma, Male, Liver tumour subgroup ==> Deceased	881/988 = 0.89
Hepatocellular carcinoma, Male ==> Liver tumour subgroup, Deceased	881/988 = 0.89
Hepatocellular carcinoma, Liver site, Male, Liver tumour subgroup ==> Deceased	881/988 = 0.89
Hepatocellular carcinoma, Male, Liver tumour subgroup ==> Liver site, Deceased	881/988 = 0.89
Hepatocellular carcinoma, Liver site, Male ==> Liver tumour subgroup, Deceased	881/988 = 0.89
Hepatocellular carcinoma, Male ==> Liver site, Liver tumour subgroup, Deceased	881/988 = 0.89
Hepatocellular carcinoma ==> Deceased	1115/1262 = 0.88
Hepatocellular carcinoma, Liver site ==> Deceased	1115/1262 = 0.88
Hepatocellular carcinoma ==> Liver site, Deceased	1115/1262 = 0.88

C.2.3.2. For the BC survivability data set excluding all treatment attributes, the best association rules found with a minimum confidence of 0.85 and a support level of 2419 (40%) were:

Rules	Confidence
Hepatocellular carcinoma ==> Liver site	3410/3410 = 1
Hepatocellular carcinoma ==> Liver tumour	3410/3410 = 1
Hepatocellular carcinoma, Liver tumour subgroup ==> Liver site	3410/3410 = 1
Hepatocellular carcinoma, Liver site ==> Liver tumour	3410/3410 = 1
Hepatocellular carcinoma ==> Liver site, Liver tumour	3410/3410 = 1
Hepatocellular carcinoma, Deceased ==> Liver site	3056/3056 = 1
Hepatocellular carcinoma, Deceased ==> Liver tumour	3056/3056 = 1
Hepatocellular carcinoma, Liver tumour, Deceased ==> Liver site	3056/3056 = 1
Hepatocellular carcinoma, Liver site, Deceased ==> Liver tumour	3056/3056 = 1
Hepatocellular carcinoma, Deceased ==> Liver site, Liver tumour	3056/3056 = 1
Hepatocellular carcinoma, Male ==> Liver site	2640/2640 = 1
Hepatocellular carcinoma, Male ==> Liver tumour	2640/2640 = 1
Hepatocell. carcinoma, Male, Liver tumour ==> Liver site	2640/2640 = 1
Hepatocellular carcinoma, Liver site, Male ==> Liver tumour	2640/2640 = 1
Hepatocellular carcinoma, Male ==> Liver site, Liver tumour	2640/2640 = 1
Male, Deceased ==> Liver tumour	3685/3805 = 0.97
Deceased ==> Liver tumour	5331/5508 = 0.97
Liver site, Male, Deceased ==> Liver tumour	3339/3459 = 0.97
Liver site, Deceased ==> Liver tumour	4657/4834 = 0.96
Male ==> Liver tumour	4017/4173 = 0.96
Liver site, Male ==> Liver tumour	3664/3820 = 0.96
Liver site ==> Liver tumour	5120/5355 = 0.96
Male, Liver tumour ==> Deceased	3685/4017 = 0.92
Liver tumour ==> Deceased	5331/5812 = 0.92
Male ==> Liver site	3820/4173 = 0.92

Male, Liver tumour ==> Liver site	3664/4017 = 0.91
Male ==> Deceased	3805/4173 = 0.91
Liver site, Male, Liver tumour ==> Deceased	3339/3664 = 0.91
Liver site, Liver tumour ==> Deceased	4657/5120 = 0.91
Male, Deceased ==> Liver site	3459/3805 = 0.91
Male, Liver tumour subgroup, Deceased ==> Liver site	3339/3685 = 0.91
Liver site, Male ==> Deceased	3459/3820 = 0.91
Liver site ==> Deceased	4834/5355 = 0.90
Hepatocellular carcinoma ==> Deceased	3056/3410 = 0.90
Hepatocellular carcinoma, Liver site ==> Deceased	3056/3410 = 0.90
Hepatocellular carcinoma ==> Liver site, Deceased	3056/3410 = 0.90
Hepatocellular carcinoma, Liver tumour ==> Deceased	3056/3410 = 0.90
Hepatocellular carcinoma ==> Liver tumour, Deceased	3056/3410 = 0.90
Hepatocell. carcinoma, Liver site, Liver tumour ==> Deceased	3056/3410 = 0.90
Hepatocellular carcinoma, Liver tumour ==> Liver site, Deceased	3056/3410 = 0.90
Hepatocellular carcinoma, Liver site ==> Liver tumour, Deceased	3056/3410 = 0.90
Hepatocellular carcinoma ==> Liver site, Liver tumour, Deceased	3056/3410 = 0.90
Male ==> Liver tumour, Deceased	3685/4173 = 0.88
Liver tumour ==> Liver site	5120/5812 = 0.88
Male ==> Liver site, Liver tumour	3664/4173 = 0.88
Deceased ==> Liver site	4834/5508 = 0.88
Male, Deceased ==> Liver site, Liver tumour	3805/3339 = 0.88
Liver site, Male ==> Liver tumour, Deceased	3339/3820 = 0.88
Liver tumour, Deceased ==> Liver site	4657/5331 = 0.88
Liver site ==> Liver tumour, Deceased	4657/5355 = 0.88

After eliminating all the rules that are not relevant because they do not contain either alive or deceased in the rule consequent, we are left with:

Rules	Confidence
Male, Liver tumour subgroup ==> Deceased	3685/4017 = 0.92
Liver tumour subgroup ==> Deceased	5331/5812 = 0.92
Male ==> Deceased	3805/4173 = 0.91
Liver site, Male, Liver tumour subgroup ==> Deceased	3339/3664 = 0.91
Liver site, Liver tumour subgroup ==> Deceased	4657/5120 = 0.91
Liver site, Male ==> Deceased	3459/3820 = 0.91
Liver site ==> Deceased	4834/5355 = 0.90
Hepatocellular carcinoma ==> Deceased	3056/3410 = 0.90
Hepatocellular carcinoma, Liver site ==> Deceased	3056/3410 = 0.90
Hepatocellular carcinoma ==> Liver site, Deceased	3056/3410 = 0.90
Hepatocellular carcinoma, Liver tumour subgroup ==> Deceased	3056/3410 = 0.90
Hepatocellular carcinoma ==> Liver tumour subgroup, Deceased	3056/3410 = 0.90
Hepatocellular carcinoma, Liver site, Liver tumour subgroup ==> Deceased	3056/3410 = 0.90
Hepatocellular carcinoma, Liver tumour subgroup ==> Liver site, Deceased	3056/3410 = 0.90
Hepatocellular carcinoma, Liver site ==> Liver tumour subgroup, Deceased	3056/3410 = 0.90
Hepatocellular carcinoma ==> Liver site, Liver tumour subgroup, Deceased	3056/3410 = 0.90
Male ==> Liver tumour subgroup, Deceased	3685/4173 = 0.88
Liver site, Male ==> Liver tumour subgroup, Deceased	3339/3820 = 0.88
Liver site ==> Liver tumour subgroup, Deceased	4657/5355 = 0.88

C.2.4. Yukon Survivability Data Set:

For the Yukon survivability data set, the 25 best Apriori association rules obtained with a minimum confidence of 0.84 and a support level of 5 (30%) were:

Rules	Confidence
Hepatocellular carcinoma ==> Liver site	9/9 = 1
Hepatocellular carcinoma, Male ==> Liver site	8/8 = 1
Hepatocellular carcinoma, Deceased ==> Liver site	6/6 = 1
Hepatocellular carcinoma, Deceased ==> Male	6/6 = 1
Hepatocellular carcinoma, Male, Deceased ==> Liver site	6/6 = 1
Hepatocellular carcinoma, Liver site, Deceased ==> Male	6/6 = 1
Hepatocellular carcinoma, Deceased ==> Liver site, Male	6/6 = 1
Diagnosis age range 60 to 69 ==> Liver site	5/5 = 1
Death age range 60 to 69 ==> Liver site	5/5 = 1
Death age range NA ==> Liver site	5/5 = 1
Alive ==> Liver site	5/5 = 1
Death age range 60 to 69 ==> Deceased	5/5 = 1
Alive ==> Death age range NA	5/5 = 1
Death age range NA ==> Alive	5/5 = 1
Death age range 60 to 69, Deceased ==> Liver site	5/5 = 1
Liver site, Death age range 60 to 69 ==> Deceased	5/5 = 1
Death age range 60 to 69 ==> Liver site, Deceased	5/5 = 1
Death age range NA Alive ==> Liver site	5/5 = 1
Liver site, Alive ==> Death age range NA	5/5 = 1
Liver site, Death age range NA ==> Alive	5/5 = 1
Alive ==> Liver site, Death age range NA	5/5 = 1
Death age range NA ==> Liver site, Alive	5/5 = 1
Male ==> Liver site	12/13 = 0.92
Deceased ==> Liver site	11/12 = 0.92
Male, Deceased ==> Liver site	9/10 = 0.90

After eliminating all the rules that are not of relevant because they do not contain either alive or deceased in the rule consequent, we are left with:

Rules	Confidence
Death age range 60 to 69 ==> Deceased	5/5 = 1
Liver site, Death age range 60 to 69 ==> Deceased	5/5 = 1
Death age range 60 to 69 ==> Liver site, Deceased	5/5 = 1
Liver site, Death age range NA ==> Alive	5/5 = 1
Death age range NA ==> Liver site, Alive	5/5 = 1