



Computationally Enhanced Medical Decision Support for Pancreatic Cancer

An Interdisciplinary Qualifying Project
Submitted to the faculty of
Worcester Polytechnic Institute
In partial fulfillment of the requirements for the
Degree of Bachelor of Science

Submitted By:

Andreea Bodnari
Towa Matsumura

Submitted To:

Project Advisor, Professor Carolina Ruiz

Date: October 29, 2008

Abstract

This project investigated and applied computational techniques to enlarge a pancreatic cancer database and to enhance the medical decision-making process supported by this database. The database was previously developed by the Department of Surgical Oncology of the University of Massachusetts Medical School in conjunction with the Department of Computer Science at WPI. We substantially increased the number of patients included in the database, and conducted data mining experiments. These experiments compared the accuracies of predictions made by medical doctors and by data mining methods for two separate patient outcomes: tumor malignancy and survival time after surgery. The results of our experiments show that data mining techniques can be used to enhance the quality of medical decisions.

Acknowledgement

Our team would like to recognize the following medical doctors for their help and assistance in the completion of this project:

- Giles F. Whalen, MD
- Jennifer F. Tseng, MD, MPH
- Jessica P. Simons, MD
- Melissa M. Murphy, MD, MPH
- Theodore McDade, MD

Our project advisor:

- Professor Carolina Ruiz

*Dedicated to our beloved family members
who have supported us throughout our studies.*

Table of Contents

| | |
|--|----|
| Abstract | 2 |
| Acknowledgement | 3 |
| Dedication | 4 |
| Table of Contents | 5 |
| Table of Tables | 7 |
| Table of Figures | 10 |
| 1 Introduction..... | 11 |
| 2 Background..... | 13 |
| 2.1 Medical Background..... | 13 |
| 2.1.1 Anatomy and Physiology of the Pancreas..... | 13 |
| 2.1.2 Pancreatic Cancer..... | 16 |
| 2.2 Database Background | 26 |
| 2.2.1 Database Schema | 26 |
| 2.2.2 Database Tables | 26 |
| 2.2.3 Database Forms..... | 27 |
| 2.3 Data Mining Background..... | 28 |
| 2.3.1 Target Attribute, Training Data, and Test Data | 28 |
| 2.3.2 Attribute Selection | 28 |
| 2.3.3 Classification Algorithms | 29 |
| 2.3.4 The WEKA System..... | 30 |
| 2.4 Previous Work on the Pancreatic Cancer Database | 30 |
| 2.4.1 Hayward's M.S. Thesis..... | 30 |
| 2.4.2 Floyd's M.S. Thesis | 36 |
| 3 Our Database Work..... | 43 |
| 3.1 Patient Demographics | 44 |
| 3.2 Pre-Operative | 45 |
| 3.3 Peri-Operative | 46 |
| 3.4 Pathology | 47 |
| 3.5 Follow-up..... | 47 |
| 3.6 Final Status of the Database..... | 47 |
| 3.6.1 Current Tables..... | 48 |
| 3.6.2 Current Forms | 50 |

| | | |
|-------|--|----|
| 4 | Our Data Mining Work..... | 53 |
| 4.1 | Patient Outcome Prediction: Doctors vs. Machine Learning Techniques..... | 54 |
| 4.1.1 | Methodology..... | 54 |
| 4.1.2 | Selected Attributes: Malignancy..... | 60 |
| 4.1.3 | Selected Attributes: Survival Time..... | 62 |
| 4.1.4 | Results: Malignancy..... | 65 |
| 4.1.5 | Results: Survival Time..... | 67 |
| 4.1.6 | Discussion..... | 69 |
| 4.2 | Prediction Accuracy Based on Frequency of Selected Attributes..... | 71 |
| 4.2.1 | Methodology..... | 72 |
| 4.2.2 | Repeated Malignancy Attributes..... | 72 |
| 4.2.3 | Repeated Survival Time Attributes..... | 74 |
| 4.2.4 | Results..... | 76 |
| 4.2.5 | Discussion..... | 79 |
| 4.3 | Reducing Test Set Selection Biases..... | 80 |
| 4.3.1 | Methodology..... | 80 |
| 4.3.2 | Results..... | 81 |
| 4.3.3 | Discussion..... | 81 |
| 5 | Conclusions and Future Work..... | 85 |
| 6 | Bibliography..... | 87 |
| | Appendix A: WEKA Parameters..... | 90 |
| | Appendix B: Database Attributes Description..... | 92 |
| | Appendix C: General Medical Terms..... | 97 |

Table of Tables

| | |
|---|----|
| Table 2-1 T Classification..... | 23 |
| Table 2-2 N Classification | 23 |
| Table 2-3 M Classification..... | 23 |
| Table 2-4 R Classification | 24 |
| Table 2-5 Cancer Staging..... | 24 |
| Table 2-6 ECOG Score Explanation..... | 24 |
| Table 3-1 Distribution of Surgical Procedures Among Patients with Pancreatic Cancer | 48 |
| Table 3-2 Database Tables Not in Use..... | 49 |
| Table 3-3 Database Tables Used in the Database Design..... | 49 |
| Table 3-4 Database Tables Used for Storing Patient Data..... | 49 |
| Table 3-5 Forms No Longer Used in the Database..... | 51 |
| Table 3-6 Forms Indirectly Used in the Database and their Associated Complex Forms | 51 |
| Table 3-7 Forms Used for Patient Data Entry..... | 52 |
| Table 4-1 Summary of the different attribute selection and outcome classification methods combination | 56 |
| Table 4-2 WEKA Parameters | 59 |
| Table 4-3 ReliefF and Doctor A's Malignancy Attribute Lists Over Training Set A | 60 |
| Table 4-4 ReliefF and Doctor B's Malignancy Attribute Lists Over Training Set B..... | 61 |
| Table 4-5 ReliefF and Doctor C's Malignancy Attribute Lists Over Training Set C..... | 61 |
| Table 4-6 ReliefF and Doctor C's Malignancy Attribute Lists Over Training Set C..... | 62 |
| Table 4-7 ReliefF and Doctor A's Survival Time Attribute Lists Over Training Set A | 63 |
| Table 4-8 ReliefF and Doctor B's Survival Time Attribute Lists Over Training Set B..... | 63 |
| Table 4-9 ReliefF and Doctor C's Survival Time Attribute Lists Over Training Set C..... | 64 |
| Table 4-10 ReliefF and Doctor D's Survival Time Attribute Lists Over Training Set D | 64 |
| Table 4-11 Human Experts' Prediction Results (Values in % accuracy)..... | 65 |
| Table 4-12 Hybrid Prediction Results: Bayesian Network (Values in % accuracy)..... | 65 |

| | |
|--|----|
| Table 4-13 Hybrid Prediction Results: Logistic Regression (Values in % accuracy)..... | 66 |
| Table 4-14 Machine Learning Prediction Results: Bayesian Network (Values in % accuracy)..... | 66 |
| Table 4-15 Machine Learning Prediction Results: Logistic Regression (Values in % accuracy) | 66 |
| Table 4-16 Human Experts' Prediction Results (Values in % accuracy) | 67 |
| Table 4-17 Hybrid Prediction Results: Bayesian Network (Values in % accuracy)..... | 67 |
| Table 4-18 Hybrid Prediction Results: Logistic Regression (Values in % accuracy)..... | 68 |
| Table 4-19 Machine Learning Prediction Results: Bayesian Network (Values in % accuracy)..... | 68 |
| Table 4-20 Machine Learning Prediction Results: Logistic Regression (Values in % accuracy) | 68 |
| Table 4-21 Accuracy Summary | 69 |
| Table 4-22 Attributes in Common for Malignancy..... | 70 |
| Table 4-23 Attributes in Common for Survival Time..... | 70 |
| Table 4-24 Frequency of Patients per Class Value. Malignancy | 70 |
| Table 4-25 Frequency of Patients per Class Value. Survival Time (Values in number of patients) | 71 |
| Table 4-26 Malignancy Confusion Matrix (Values in number of patients)..... | 71 |
| Table 4-27 Survival Time Confusion Matrix (Values in number of patients) | 71 |
| Table 4-28 Attributes occurring in 4 out of 4 Attribute Lists | 72 |
| Table 4-29 Attributes occurring in 3 out of 4 Attribute Lists | 72 |
| Table 4-30 Attributes occurring in 2 out of 4 Attribute Lists | 73 |
| Table 4-31 Attributes occurring in 1 out of 4 Attribute Lists | 73 |
| Table 4-32 Attributes occurring in 4 out of 4 Attribute Lists | 74 |
| Table 4-33 Attributes occurring in 3 out of 4 Attribute Lists | 74 |
| Table 4-34 Attributes occurring in 2 out of 4 Attribute Lists | 74 |
| Table 4-35 Attributes occurring in 1 out of 4 Attribute Lists | 75 |
| Table 4-36 Repeated Doctor Selected Attributes: Malignancy (Values in % accuracy) | 76 |
| Table 4-37 Repeated Relief-F Selected Attributes: Malignancy (Values in % accuracy) | 77 |
| Table 4-38 Repeated Doctor Selected Attributes: Survival Time (Values in % accuracy) | 78 |

| | |
|--|----|
| Table 4-39 Repeated Relief-F Selected Attributes: Survival Time (Values in % accuracy) | 79 |
| Table 4-40 Malignancy Results (25 fold cross-validation) (Values in % accuracy)..... | 81 |
| Table 4-41 Survival Time Results (6 fold cross-validation) (Values in % accuracy)..... | 81 |
| Table 4-42 Comparison of Results Bayesian Networks Classifier for Malignancy Class (values in % accuracy) | 82 |
| Table 4-43 Comparison of Results Logistic Regression Classifier for Malignancy Class (values in % accuracy) | 83 |
| Table 4-44 Comparison of Results Bayesian Networks Classifier for Survival Time Class (values in % accuracy) | 83 |
| Table 4-45 Comparison of Results Logistic Regression Classifier for Survival Time Class (values in % accuracy) | 84 |

Table of Figures

| | |
|--|----|
| Figure 2-1 Structure of the Pancreas..... | 14 |
| Figure 2-2 Blood Supply for the Pancreas (8) | 15 |
| Figure 2-3 Generic Table in Microsoft Access 2003..... | 26 |
| Figure 2-4 Generic Table Relationship in Microsoft Access 2003..... | 27 |
| Figure 2-5 Generic Form in Microsoft Access 2003 | 28 |
| Figure 2-6 Pancreatic Cancer Presentation Form (1)..... | 31 |
| Figure 2-7 Pancreatic Cancer Medical History Form (1)..... | 31 |
| Figure 2-8 Pancreatic Cancer Diagnosis (Serum Studies) Form (1)..... | 32 |
| Figure 2-9 Pancreatic Cancer Diagnosis (Diagnostic Imaging) Form (1) | 32 |
| Figure 2-10 Pancreatic Cancer Diagnosis (Endoscopy Studies) Form (1)..... | 33 |
| Figure 2-11 Pancreatic Cancer Preliminary Outlook Form (1)..... | 33 |
| Figure 2-12 Pancreatic Cancer Resection Form (1)..... | 34 |
| Figure 2-13 Pancreatic Cancer No Resection Form (1)..... | 35 |
| Figure 2-14 Pancreatic Cancer Pathology Form (1) | 35 |
| Figure 2-15 Pancreatic Cancer Follow-Up Form (1) | 36 |
| Figure 2-16 Pancreatic Cancer Patient Form (2)..... | 37 |
| Figure 2-17 Pancreatic Cancer Pre-Operative Form (2) | 38 |
| Figure 2-18 Pancreatic Cancer Peri-Operative Form (2)..... | 40 |
| Figure 2-19 Pancreatic Cancer Pathology Form (2) | 41 |
| Figure 2-20 Pancreatic Cancer Follow-Up Form (2) | 42 |
| Figure 3-1 Screenshot of MEDITECH | 44 |
| Figure 3-2 Database Tables Relationship | 50 |
| Figure 4-1 Flow Chart..... | 56 |
| Figure 4-2 Experimental Map..... | 58 |

1 Introduction

Medical doctors must go through continuous training to make good clinical decisions. They rely on background knowledge, current research, and professional experience. Since these decisions affect the welfare of patients, the investigation of novel methods that can enhance the quality of medical decision-making is of high importance. The M.S. theses of WPI students John Hayward (1) and Stuart Floyd (2) emphasized the potential of data mining techniques in improving patients' cancer treatment. These theses, done in collaboration between the Department of Computer Science at WPI and the Department of Surgical Oncology of the University of Massachusetts Medical School, developed and populated a database of pancreatic cancer patient data, and conducted data analysis experiments using this database. The experiments demonstrated that novel data mining techniques are comparable, and in some cases superior, in forming predictive models for clinical patient prognosis when compared to standard statistical methods used in the medical community including linear and logistic regression.

The problem statement of this project was to expand this pancreatic cancer database by adding more patients and more data to existing patients in the database, and to conduct experiments over these data to compare the accuracy of predictions made by data mining methods and the accuracy of human expert predictions made by medical doctors.

The UMass pancreatic database stores a larger amount of information (278 attributes) per patient than the main national databases (SEER (3), NIS (4)). Initially, this database included 91 patients with incomplete medical information. We extended it to include 252 patients containing all available information in the database's range of interest. We participated in training covering patient information privacy rules defined in the Health Insurance Portability and Accountability Act (HIPPA) of 1996. We also studied essential topics concerning pancreatic cancer in order to be able to parse and interpret medical information that was to be stored in the database.

In the data analysis part of our project, we investigated two patient outcomes: tumor malignancy and survival time after surgery. We conducted an experiment comparing the accuracy results when doctors and data mining techniques (Relief-F feature selection, Bayesian networks, and standard logistic regression) made their outcome predictions entirely on their own on a randomly selected test set of retrospective patient records, and also conducted a hybrid experiment in which doctors and data mining techniques collaborated in making predictions. Another experiment considered the pattern observed in the way doctors and data mining methods chose patient attributes that were used for making the predictions in the first experiment. In a final experiment, we re-did the first experiment, but used the cross-validation

method rather than the test set method to calculate prediction accuracies in order to reduce the variance in results due to unintended statistical biases in the randomly selected test set. Through these experiments, the usefulness of data mining algorithms in assisting doctors when making medical predictions was assessed.

2 Background

This IQP involves technical concepts that the layman may not understand initially. This chapter is dedicated to give the reader enough background information to comprehend the rest of the report. There are two parts to this chapter. The medical background will cover all the essentials about pancreatic cancer which will help the reader understand the various attributes present in the database and why they are important. The technical background will talk about the past experiments done on the database and will also explain a few database mining concepts which are needed to understand the chapter explaining the experiments done in this IQP.

2.1 Medical Background

The medical background section briefly covers the anatomy and physiology of the pancreas which will help in understanding the section on pancreatic cancer. These are very important concepts to know when looking at the various data fields present in the database. The content of this section should provide the reader with enough medical background in terms of understanding the report.

2.1.1 Anatomy and Physiology of the Pancreas

The pancreas is a glandular organ situated behind the abdomino-pelvic cavity, in the J-shape loop between the stomach and the duodenum. The adult pancreas has a length of 20-25 cm (8-10 in.) and weights about 80 g (2.8 oz) (5).

The pancreas is often described as having five regions. The head is the broad section on the right most part of the pancreas, which abuts the second part of the duodenum (6). The ucinat process is the most inferior end of the head, and it “hooks” behind the superior mesenteric artery and vein (5). The neck is the right upper portion to the left of the head (6). The body is the main region of the pancreas. Finally the tail is the most left end region towards the spleen (6).

Located inside the pancreas, starting at the tail and emptying in the second part of the duodenum, is the pancreatic duct. This duct is formed by the junction of several lobular ducts in the tail and it increases in size as it runs within the pancreas body (6). The pancreatic duct and the common bile duct, the duct coming from the bile, usually converge together and empty into the duodenum at the Ampulla of Vater (7). The general population has just this one pancreatic duct, but some have an additional accessory pancreatic duct (the duct of Santorini) that drains the head and the ucinat process (5). The duct bifurcates from the main pancreatic duct towards the duodenum and empties above the Ampulla of Vater (7). The

main role of these two ducts is to drain the pancreas by gathering the pancreatic juice secreted by the exocrine cells.

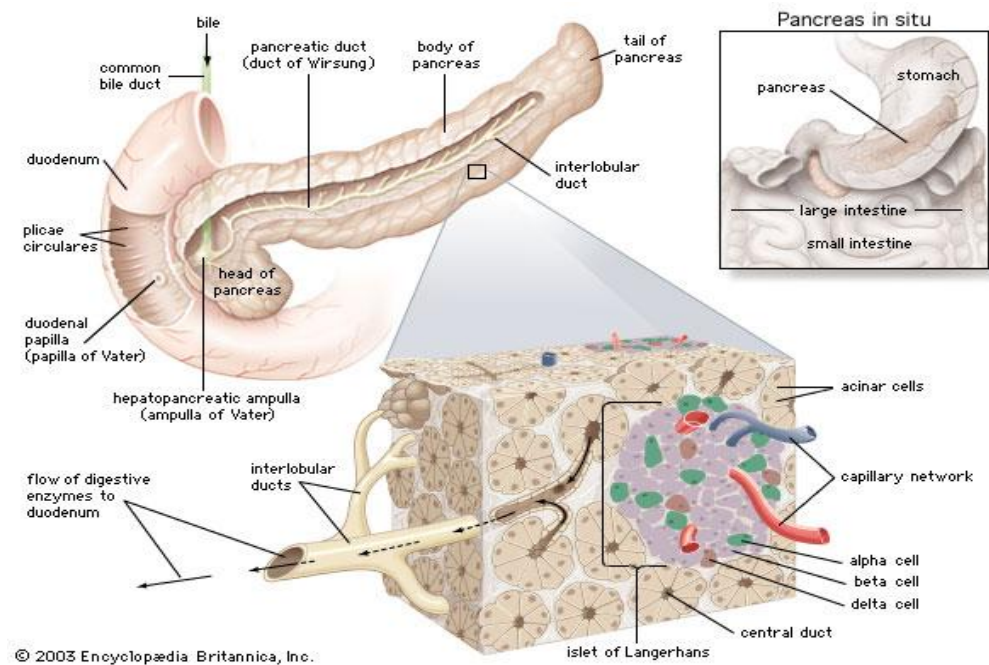


Figure 2-1 Structure of the Pancreas

The main functions of the pancreas are the secretion of hormones into the blood (endocrine) and the secretion of hormones into ducts (exocrine).

The endocrine function is fulfilled by the islets of Langerhans (pancreatic islets; 1% of the pancreatic cell population) which are clusters of cells scattered among the exocrine cells (5). The pancreatic islets secrete two hormones that keep the level of glucose from the blood constant: glucagon and insulin. Glucagon is released when the blood glucose level is too low, triggering the secretion of stored glucose. On the other hand, insulin acts to decrease blood glucose levels (5). Each islet contains four different cell types: alpha cells- produce glucagon, beta cells- produce insulin, delta cells- produce a regulatory hormone identical to somatostatin and F cells- produce pancreatic polypeptide (PP) (5).

The exocrine pancreas (99% of the pancreatic volume) is formed by clusters of gland cells -called pancreatic acini and their attached ducts (5). Each pancreatic acinus consists of more pyramidal acinar cells that enclose a lumen, into which the acinar cells secrete digestive enzymes (7).

The exocrine gland cells and duct cells secrete together the pancreatic juice (about 1000 ml per day), which is a mixture of water, ions- secreted by the duct cells and digestive enzymes (e.g., amylase which

breaks down starch, proteolytic enzymes which breaks down certain proteins, and lipase, which breaks down complex lipids) (5) (7); once secreted, the pancreatic juice is drained into lobular ducts and flows through the main or through the accessory duct into the duodenum (6).

The pancreas is a relatively high vascular organ. Therefore the surrounding vasculature surrounding the pancreas should be of great interest when studying disorders of the pancreas such as pancreatic cancer.

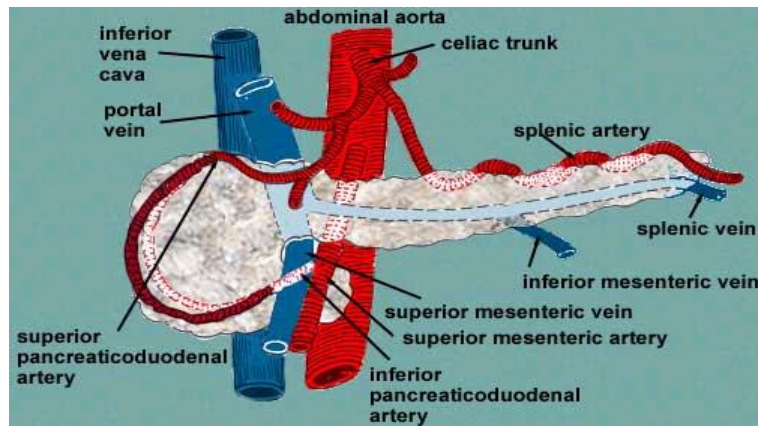


Figure 2-2 Blood Supply for the Pancreas (8)

The celiac artery is the first major artery that branches off the abdominal aorta in below the diaphragm (6). It supplies oxygenated blood to digestive organs, including the stomach, liver, spleen, duodenum and the pancreas (6). The celiac artery is of special importance since three arteries bifurcates from it. One of the arteries that branch off the celiac artery is the superior mesenteric artery (SMA). It supplies oxygenated blood to the pancreas and parts of the intestine starting from the duodenum and ending at the left colic flexure. The similarly named vein, the superior mesenteric vein (SMV), lies next to the SMA also running posterior of the pancreas. It drains deoxygenated blood from the small intestines and is one of the veins that connect to the portal vein (6). The SMV is special in that it contributes the greatest volume of blood compared to any other tributaries of the portal vein (6).

The portal vein is one of the main blood vessels of the hepatic portal venous system (6). It collects blood used by the digestive system organs, and delivers the blood to the liver. The liver also receives blood from the hepatic artery, but this blood is different from the blood contained in the portal vein. The hepatic artery supplies the liver with oxygenated blood, whereas the portal vein delivers blood exiting the digestive organs for detoxification. However, the liver outputs all blood via the hepatic veins to the inferior vena cava, which will then enter the right atrium of the heart (6).

2.1.2 Pancreatic Cancer

As the name implies, pancreatic cancer is the uncontrolled growth of abnormal cells in the pancreas. This section will cover topics ranging from cancer in general to pancreatic cancer–specific materials.

2.1.2.1 Cancer Overview

Cancer is a malignant, out-of-control growth of abnormal body cells. Instead of following a normal cycle (growth, division, and death), cancer cells continue to live and divide thus creating new abnormal cells. The cells abnormality is usually caused by damage in various genes in the DNA^A. When abnormal cells divide, they pass the damaged DNA to the young cells, thus creating an abnormal tissue⁴.

The abnormal tissue can be called tumor (describes an abnormal swelling, lump or mass) or neoplasm (the common term for an abnormally new grown tissue; can be malignant or invasive when cells spread to surrounding tissues or benign when cells stop their growth). Although tumor and neoplasm are almost synonymous, not all neoplasm are tumors (e.g., leukemia is a neoplasm and not a tumor), and the implications are different (9).

The terms pre-malignant, pre-cancer and non-invasive tumor refer to tumors that can potentially become malignant if untreated (e.g., atypia, dysplasia, carcinoma in situ). On the other hand, the term malignant tumor/ neoplasm are synonyms to cancer.

Cancer (implicitly tumors) is named after the tissue/ organ it initially develops in. Malignant tumors are named using the suffixes –carcinoma, -sarcoma or –blastoma (e.g., adenocarcinoma) while benign tumors take –oma as suffix and are referred to as adenomas (e.g., seminoma) (10)

2.1.2.2 Statistics

According to SEER (Surveillance Epidemiology and End Result) of the National Cancer Institute (11), incidence rate of pancreatic cancer in the U.S. is 11.4 per 100,000 every year. As this number suggests, the incidence of this form of cancer is low. However, it is the 4th leading cause of cancer death, having a 10.6 per 100,000 every year mortality rate. The 5 year relative survival rate from 1996-2003 was 5.0%, and based on the data from 2002-2004, 1.31% of the people born today will develop pancreatic cancer during their lifetime, which translates to 1 in every 76 persons. 52% of pancreatic cancer patients are diagnosed when the cancer has already metastasized and it is too late for surgical removal of the tumor.

^A DNA or deoxyribonucleic acid contains coded information vital to the functioning of an organism.

The American Cancer Society estimated that 37,680 persons will be diagnosed and 34,290 persons will die from pancreatic cancer in 2008 (12).

2.1.2.3 Types

Pancreatic cancer is one of the most frequent cases of periampullary neoplasms, which are neoplasms situated around the ampulla (80% of all cases). Other types of periampullary cancer have a smaller rate of occurrence: neoplasms of ampulla of Vater hold 10% of the cases, neoplasms in the duodenum 4% and neoplasms of the common bile duct 3% (13).

Pancreatic cancer is caused by an abnormal growth in either the endocrine or exocrine cells, resulting in endocrine or exocrine cancer. There are also tumors that appear in non pancreatic tissues and once in the malignant stage, spread to the pancreas through the blood or other means.

Non-endocrine tumors account for 98% of the pancreatic tumors and from this group adenocarcinoma is the most common tumor with a frequency of 95%. 90% of the patients diagnosed with adenocarcinoma have mutations in the DNA, more specifically in the Ki-ras oncogene^B on codon^C (14).

One of the characteristic of pancreatic adenocarcinoma is the early extension to contiguous structures and metastasis to lymph nodes and the liver. 80% of these carcinomas are unresectable at the time of diagnosis. The pulmonary, peritoneal and other distant nodal metastases occur later (15).

Another non-endocrine type tumor is the cystic neoplasm. Cystic neoplasm of the pancreas can divide into inflammatory pseudocysts, serous cystic neoplasm, mucinous cystic neoplasm (MCNs), and papillary cystic-solid neoplasm. Inflammatory pseudocysts occur in response to recurrent pancreatitis. They do not become malignant and rarely require pancreatic resection (16). Serous (benign cysts) and mucinous cysts are the true pancreatic cysts.

Serous cystic neoplasm (cystadenoma) is a pancreatic lesion consisting of multiple small cysts that is considered benign as it has a local and indolent growth. Nevertheless, progressive and local growth may have secondary effects (obstruction of the bile duct or duodenum) thus the neoplasm is surgically resected. Even though most serous cysts are benign, there have been reports of malignant serous cystadenocarcinoma.

^B An oncogene is a gene with DNA sequence that causes cancer.

^C A sequence of three adjacent nucleotides in the genetic code.

Mucinous cystic neoplasm is either malignant (cystadenocarcinoma) or premalignant (cystadenoma). In both cases surgical resection is considered to be the best approach (16). The most common approach for mucinous tumors is surgical resection.

Papillary cystic-solid neoplasm is an uncommon lesion that usually occurs in young women. The tumor has a relatively big size but rarely evolves into metastases and is expected to cure after resection.

Also in the list of non-endocrine tumors, intraductal Papillary-Mucinous Neoplasms/ Tumors (IPMN/ IPMT) are characterized by the dilatation of the pancreatic duct and its branches, as well as the secretion of a large amount of mucin by the neoplastic epithelium. These tumors are often confused with mucinous cystic neoplasms. Unlike the latter, IPMN do communicate with the ductal system and are responsible of dilated, mucin-filled ducts (17). IPMN can be in situ (pre-malignant), benign or malignant.

Acinar cell carcinoma is a malignant epithelial neoplasm that can arise in any portion of the pancreas. The tumor is usually big and well-circumcised. Acinar cell carcinoma is rare, encountered in 2% of the malignant cases, more commonly in men than in women.

Some other rare non-endocrine tumors are adenosquamous carcinoma, sarcoma, giant cells tumor, pancreaticoblastoma, papillary epithelial neoplasm, and solid and pseudopapillary tumor (Solid pseudo papillary neoplasm is still under research. It is one of the tumors that have not been categorized in what concerns the direction of differentiation of the neoplastic cells).

Endocrine (Islet cells) tumors occur due to abnormal growth in different endocrine cells specialized in production of hormones. The endocrine (Islet cells) tumors are not as frequent as the non-endocrine tumors. A classification of islet cell tumors include endocrine microadenoma, well-differentiated pancreatic endocrine neoplasms, non-functional pancreatic endocrine neoplasms (PPoma), poorly differentiated endocrine carcinoma (small cell carcinoma and large cell endocrine carcinoma), and mixed endocrine carcinomas.

The well differentiated pancreatic neoplasms include a series of tumors that are categorized related to the cells they initially develop in. Insulinoma is a result of an abnormal growth in the β cells. It has been encountered in all age groups and is considered the most common type of islet cell tumor. Insulinoma spreads unevenly throughout the pancreas. 5% of the cases have malignant potential. One of the clinical findings in patients with insulinoma is hypoglycemia (17). Glucagonoma arises from abnormal growth in the α cells. Even though the incidence of glucagonoma is rare, 70% of the cases have a malignant potential. Some clinical findings for glucagonoma are skin rash, stomatitis, diabetes, and weight loss.

Gastrinoma occurs in the G cells and triggers the Zollinger-Ellison syndrome (a severe peptic ulcer disease). More than 60% of the gastrinomas are cancerous.

VIPoma (Verner-Morrison Syndrome) normally arises within cells in the pancreas, whose original function can no longer be identified due to it being heavily affected, but extra-pancreatic tumor may also occur. Patients with VIPoma present watery diarrhea, hypokalemia and achlorhydria^D and 40% of the cases have malignant potential.

Somatostatinoma arises from the islet cells specialized in producing somatostatin (either in the duodenum or in the pancreas) and occurs simultaneously with diabetes mellitus or cholelithiasis^E (16). The lesion is usually large. It occurs rarely and has 70 % malignancy probability.

Nonfunctional islet cell tumor (PPoma) occur in the cells specialized in the production of pancreatic polypeptide. When diagnosed, these tumors are already malignant with large sizes (15).

The mixed endocrine carcinomas can be mixed ductal-endocrine carcinoma, mixed acinar-endocrine carcinoma or mixed acinar-endocrine-ductal carcinoma (18).

2.1.2.4 Symptoms

Pancreatic cancer is named the “silent” cancer since symptoms do not occur in an early stage but only after the tumor is relatively large and has already spread to nearby organs and lymph nodes (17). The invasion of other organs trigger initial symptoms such as weight loss (as much as 25 pounds/ 11.34 kg and caused by a variety of factors) and epigastric pain that radiates to the back or simply back pain (present in 75-90% of the patients) (10). Other symptoms depend on the part of the pancreas the tumor occurs in. When the tumor is in the head of the pancreas (80% of the cases), thus close to the common bile duct, jaundice^F occurs due to the obstruction of the bile duct. Jaundice is accompanied by general itchiness, pruritus, constipation, and/or diarrhea. Jaundice, epigastric pain and weight loss are encountered with very high frequency (0 ~ 90%). Other symptoms are back pain, nausea and diarrhea, general weakness, itchy skin, light-colored bowel movements, and slow digestion of food. The liver and the gallbladder may swell because of the interaction with the tumor. If the tumor is located in the tail of the pancreas (20% of the cases) it affects the nearby veins. Additionally, if the splenic vein is encased, the

^D Low production of gastric acid in the stomach.

^E The presence of gallstones in the gallbladder.

^F Yellowing of the skin and white of the eyes and darkening of the urine due to an increased level of bilirubin in the blood.

spleen gets swollen. Other symptoms are loss of appetite, back pain that depends on the body position, blood clots in the legs, early satiety, gastrointestinal bleeding and anorexia.

Islet cell cancers cause weakness, dizziness, chills, muscle spasm or diarrhea as a result of an abnormal secretion of hormones. Some diseases (diabetes mellitus without a predisposing cause, pancreatitis) can also occur previous to the diagnosis of pancreatic cancer, though they are not always related.

Pain develops as cancer evolves and spreads to other organs. Usually located in the upper abdomen and lower back, the pain becomes worse as the person eats or lies down. It can be a result of various causes: compression of nearby organs, splenic or celiac invasion, or increase in the pressure of secretion in the ducts.

2.1.2.5 Causes

Pancreatic cancer is sometimes associated with family history, medical conditions, environmental risk factors and others.

Recent studies show that people who have had a case of pancreatic cancer in their first-degree relatives have a two or three times higher chance of developing the same disease than people without family history of pancreatic cancer (probably due to the transmission of mutated genes like K-ras through DNA) (19). Yet, having a mutated K-ras gene or having a pancreatic cancer history in the family is not a determining factor for developing pancreatic cancer. Only 10% of the patients with pancreatic cancer present a hereditary genetic factor (20).

The most important predictor factor for pancreatic cancer is age. About 80% of the cases occur in people aged 60- 80 years. Some cases can develop in people younger than 40 but it is uncommon. Also, gender seems to be a factor that determines the incidence of this type of cancer. The male/female ratio of pancreatic cancer incidences in the United States is about 4/3 (20). In terms of race, the incidence of pancreatic cancer is higher in the black population compared to the white population. Furthermore, there tends to be a lower frequency of pancreatic cancer in Hispanic and Asian population.

People with previous diseases such as pancreatitis, diabetes mellitus, or patients that underwent partial gastrectomy or cholecystectomy are more likely to develop pancreatic cancer (18). Habitual factors may also increase the risks of pancreatic cancer. Some of these factors are smoking (increases the cancer predisposition by two to three folds), nutritional factors (diet with high cholesterol, fats and processed

meat content), dietary carcinogenes^G, alcohol and coffee consumption, occupation, and possibly obesity (gives a 20 increase in the chance of developing cancer).

According to the Pancreatic Cancer Institute, there are some hereditary syndromes associated with pancreatic cancer: familial breast cancer, familial atypical multiple mole melanoma syndrome, Peutz-Jeghers Syndrome (PJS), hereditary pancreatitis, hereditary nonpolyposis colon cancer (Lynch syndrome), multiple endocrine neoplasia type I syndrome (MEN 1), and Von Hippel-Lindau Syndrome (19).

2.1.2.6 Diagnosis Techniques

There are several techniques oncologists use to diagnose and stage pancreatic cancer. Noninvasive radiographic imaging is one of the primary techniques implemented. Of the various forms of imaging techniques, CT, MRI, and PET scans are often used. Computed Tomography (CT) scans are X-rays taken at various angles and then combined to form detailed cross-sectional images of the target. It makes use of contrast agents, either intravenously or orally, to enhance image quality. A special type called the dual-phase helical CT scanning procedure is estimated to be able to diagnose about 98% of all pancreatic cancers and distant metastases (21).

Magnetic Resonance Imaging (MRI) is another form of radiological imaging. Instead of gamma radiation like in the case of CT, MRI uses powerful magnets to create an image, so it can be used on patients who cannot accept radioactive contrast agents. The magnetic field created by the powerful magnet polarizes the hydrogen atoms in the body, and when a pulse of radio wave is initiated, the polarization is scattered. The time it takes for the atoms to realign themselves to the magnetic field differs depending on the tissue, so it is very useful in contrasting soft tissue. Magnetic Resonance Cholangiopancreatography (MRCP) is specially used for noninvasively imaging the pancreas and the biliary ductal system, which are difficult to see with CT or MRI. MRI and MRCP are typically combined (22).

Another form or retrograde cholangiopancreatography is the Endoscopic Retrograde Cholangiopancreatography (ERCP). Unlike MRCP, this procedure is invasive. A thin tube is passed down the throat to the small intestine, and contrast dye is injected into the pancreatic duct. Then an X-ray is taken. Although invasive, there are several advantages in this procedure. A stent that may be placed in order to keep the duct open can be left there, relieving jaundice and related symptoms. Another diagnosis technique using an endoscope is the Endoscopic Ultrasonography (EUS). EUS is ultrasonography taken

^G Pancreatic cancer risks are higher in people that use in excess salt, smoked meat, dehydrated or fried foods, and refined sugars (20).

from the inside, specifically from the stomach or the duodenum. By advancing the sensor this close to the site of interest, a higher frequency ultrasound can be used since the waves do not have to travel through the body to the pancreas, which translates into higher resolution. Laparoscopic Ultrasonography is similar to EUS, but rather than passing down the ultrasound sensor down the throat, a small incision is made in the abdomen. Laparoscopy is performed to view the pancreas and a probe is inserted to perform the ultrasound (20).

Positron Emission Tomography (PET) scan typically uses fluorodeoxy-D-glucose (FDG) to show not only the anatomy but also biological function. Since cells take in glucose to function, a high concentration of FDG indicates higher activity in the area. Since cancer cells often absorb much more glucose than normal cells, a specialized camera will be able to detect the relatively higher concentration of radiation (23) (20).

Whenever an instrument is placed near the pancreas, such as in the case of EUS, ERCP, and laparoscopy, small tissue samples can be taken for biopsy. A biopsy is the only definitive way to diagnose cancer. One way to obtain a tissue sample is called Fine Needle Aspiration. CT or EUS is used to guide a long thin needle to the tumor. The brush biopsy is done together with the ERCP. A small brush is inserted via the endoscope and it collects cell samples by brushing against the site of interest (23).

Serum tests are also conducted to look for signs of cancer. These test are done both preoperatively and postoperatively to see the evolution of the tumor. For pancreatic cancer, cancer antigen 19-9 (CA19-9) and carcinoembryonic antigen (CEA) are the main tumor markers that are produced by tumor cells. However, these marker tests by themselves are not accurate enough to screen for or to make a diagnosis of pancreatic cancer (23).

When the pancreas is affected by a tumor it can influence the liver as well (e.g., through blockage of the ducts). Thus the level of the liver enzymes can sometimes be conclusive towards a malfunction in the pancreas. Some tested enzymes are AST (aspartate aminotransferase, an enzyme secreted into blood when the liver is damaged), ALT (alanine aminotransferase, found in liver, kidney and pancreas), ALK (alkaline phosphatase, high level of this enzyme can show a blockage of the bile ducts), and amylase. The levels of total bilirubin (product that results at the breakdown of hemoglobin) and albumin (a plasma protein) are also examined to make eventual connections to the damages in the liver or gallbladder.

2.1.2.7 Staging

The diagnosis techniques presented can gather enough information for classifying the tumor, whether it is in an initial phase or it has already spread to other sites, and staging cancer. Staging is important since it helps in predicting future treatments and patient's prognosis. The most important things taken into consideration in any staging system are the location of the primary tumor, tumor size and the number of tumors, lymph nodes involved, cell types and presence or absence of metastasis. One commonly accepted tumor classification system is the TNM system, where T stands for tumor, N for lymph and M for metastasis. The TNM system adds digits to the letters T, N and M to further describe the tumor. The R criterion also gives information regarding the existence of residual tumor. The following is a summary of the TNM and R classification system (18).

| Stage | Explanation |
|------------|---|
| TX | The primary tumor cannot be evaluated |
| T0 | No evidence of cancer in the pancreas |
| Tis | Carcinoma in situ (tumor remains in a pre-invasive state and is within the pancreas) |
| T1 | The tumor is in the pancreas only, size ≤ 2 cm |
| T2 | The tumor is in the pancreas only, size > 2 cm |
| T3 | The tumor has spread to surrounding tissue near the pancreas but not to the major blood vessels |
| T4 | The tumor extends beyond the pancreas into major blood vessels near the pancreas |

Table 2-1 T Classification

| Stage | Explanation |
|-----------|--|
| NX | The regional lymph nodes cannot be evaluated |
| N0 | The cancer was not found in the regional lymph nodes |
| N1 | The cancer has spread to regional lymph nodes |

Table 2-2 N Classification

| Stage | Explanation |
|-----------|--|
| MX | Distant spreads of the disease cannot be evaluated |
| M0 | The disease has not spread to distant lymph nodes or to distant organs |
| M1 | The disease has spread to distant lymph nodes or to distant organs |

Table 2-3 M Classification

| Stage | Explanation |
|-----------|----------------------------|
| R0 | No residual tumor |
| R1 | Microscopic residual tumor |
| R2 | Macroscopic residual tumor |

Table 2-4 R Classification

Combining T, N, and M defines the stage of the cancer. The following table summarizes the cancer staging system.

| Stage | Description |
|------------|------------------------|
| O | Tis, N0, M0 |
| IA | T1, N0, M0 |
| IB | T2, N0, M0 |
| IIA | T3, N0, M0 |
| IIB | T1 or T2 or T3; N1; M0 |
| III | T4, any N, M0 |
| IV | Any T, any N, M1 |

Table 2-5 Cancer Staging

The overall patient's performance is included in the ECOG performance status. This scale system assesses the impact the disease has on the patient. Below is an explanation for each of the 6 ECOG stages (24).

| Grade | Explanation |
|----------|--|
| 0 | Fully active, able to carry on all pre-disease performance without restriction. |
| 1 | Restricted in physically strenuous activity but ambulatory and able to carry out work of a light or sedentary nature. |
| 2 | Ambulatory and capable of all self-care but unable to carry out any work activities. Up and about more than 50% of waking hours. |
| 3 | Capable of only limited self-care, confined to bed or chair more than 50% of waking hours. |
| 4 | Completely disabled. Cannot carry on any self-care. Totally confined to bed or chair. |
| 5 | Dead. |

Table 2-6 ECOG Score Explanation

2.1.2.8 Possible Treatment

The only way to cure pancreatic cancer is to remove the tumor, but surgery is only performed when a surgeon believes that the surgery is not presenting too high risk factors. Some major procedures for removing pancreatic tumors are pancreaticoduodenectomy, total pancreatectomy, and distal pancreatectomy. Pancreaticoduodenectomy is also known as the Whipple procedure, and it involves the removal of the pancreatic head, duodenum, gallbladder and the bile duct. There exist two basic Whipple procedure types. The less common type removes the lower part of the stomach. In the more common type called the pylorus-preserving Whipple procedure, the same organs as the first type of Whipple procedure are removed, with the exception that the entire stomach and the first portion of the duodenum are spared. After the removal of these organs, the stomach, the pancreas, the remaining parts of the duodenum and the bile duct are connected to the small intestine. This preserves the normal flow of bile and pancreatic enzymes in to the small intestine with ingested food (20) (12).

Another surgical technique is total pancreatectomy. This is the removal of the entire pancreas and spleen. Because the removal of the pancreas implies the removal of islet cells that produce insulin, the direct result of this procedure is diabetes. This procedure is rarely used. A less extreme version of this procedure is the distal pancreatectomy, which is the standard operation to remove tumors of the body and tail of the pancreas. Surgical oncologists sometimes opt for a central pancreatectomy, which is performed when there is a benign tumor in the head of the pancreas (12). Other procedures used in dealing with pancreatic tumors are enucleation, Berger procedure and Frey's procedure. It should be noted that the above mentioned procedures sometime require venous resections and reconstruction.

If the cancer has metastasized and total surgical removal of the tumor is not feasible, surgery can still be performed to improve the situation of the patient. For example, surgeons can remove or bypass blockages in the pancreatic or bile duct and gastrointestinal tract to remove symptoms such as nausea and jaundice. They can also perform nerve blocks to reduce pain (12).

There exists non-surgical treatment of pancreatic cancer, which includes radiation therapy and chemotherapy. The underlying concept of radiation therapy is to use high-energy X-rays to kill cancer cells and shrink the tumor. The foremost use in the treatment of pancreatic cancer is external beam radiation therapy. In chemotherapy, drugs are used to kill cancer cells. The drug reaches the tumor cells by traveling through the blood stream. A few terms that relate to chemotherapy are neoadjuvant and adjuvant therapy. In neoadjuvant therapy, chemotherapy or radiotherapy are used before surgery, and in adjuvant therapy, they are used after the surgery. These non-surgical treatments have severe side effects

including nausea, vomiting, diarrhea, and fatigue. Some potentially serious side effects such as bleeding, low blood cell counts and infection can occur with chemotherapy (23).

2.2 Database Background

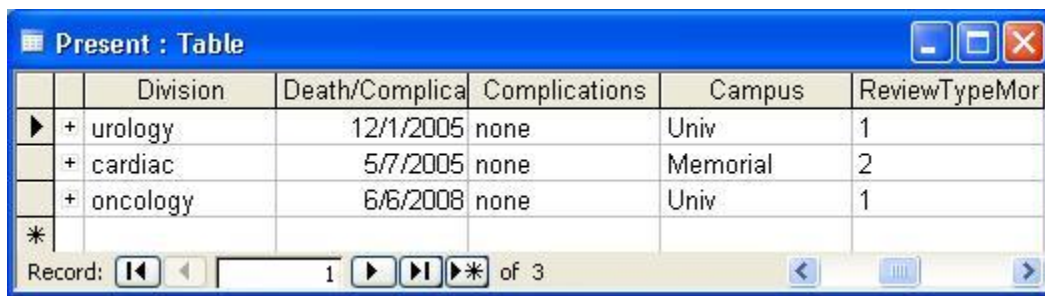
The UMass Medical School pancreatic cancer database has been created in Microsoft Access 2003 database software offered by Microsoft Co. In the following sections we will familiarize the reader with some of the essentials of this database software.

2.2.1 Database Schema

Microsoft Access 2003 is a user-friendly database software largely used in the medical field. Any database built with this software will contain one or more tables (see section 2.2.2) and a various number of forms (see section 2.2.3), even though forms are not mandatory. Tables are the most important part of the database and they serve as foundation of forms. Tables store the entire data of a database, while forms help with entering the data into the tables. Each form can store data from only one table. In case that information from more than one table needs to be stored in one single form, sub-forms (forms included within another form) are being used.

2.2.2 Database Tables

Tables are a data-relationship structure organized in columns identified by a name (called fields), and rows (called records). Tables usually store data on a specific topic (e.g., library collections, students, etc.). The importance of storing data according to a specific topic is the elimination of data-entry errors and the increased efficiency of the database. Each row in a table represents a complex record that is characterized by the correlated value appearing in the table's columns. The columns in a table store the values of the specific attribute mentioned in the column header.



| | Division | Death/Complica | Complications | Campus | ReviewTypeMor |
|-----|----------|----------------|---------------|----------|---------------|
| ▶ + | urology | 12/1/2005 | none | Univ | 1 |
| + | cardiac | 5/7/2005 | none | Memorial | 2 |
| + | oncology | 6/6/2008 | none | Univ | 1 |
| * | | | | | |

Record: 1 of 3

Figure 2-3 Generic Table in Microsoft Access 2003

In order to avoid having errors created by duplicated records in a table, primary keys are created which uniquely characterize or identifies every record. This serves as a universal index for all the tables existing in the database. Furthermore, in a relational database redundant data can be avoided by creating table relationships that enhance the synchronization of records from specific tables (see **Figure 2-4**).

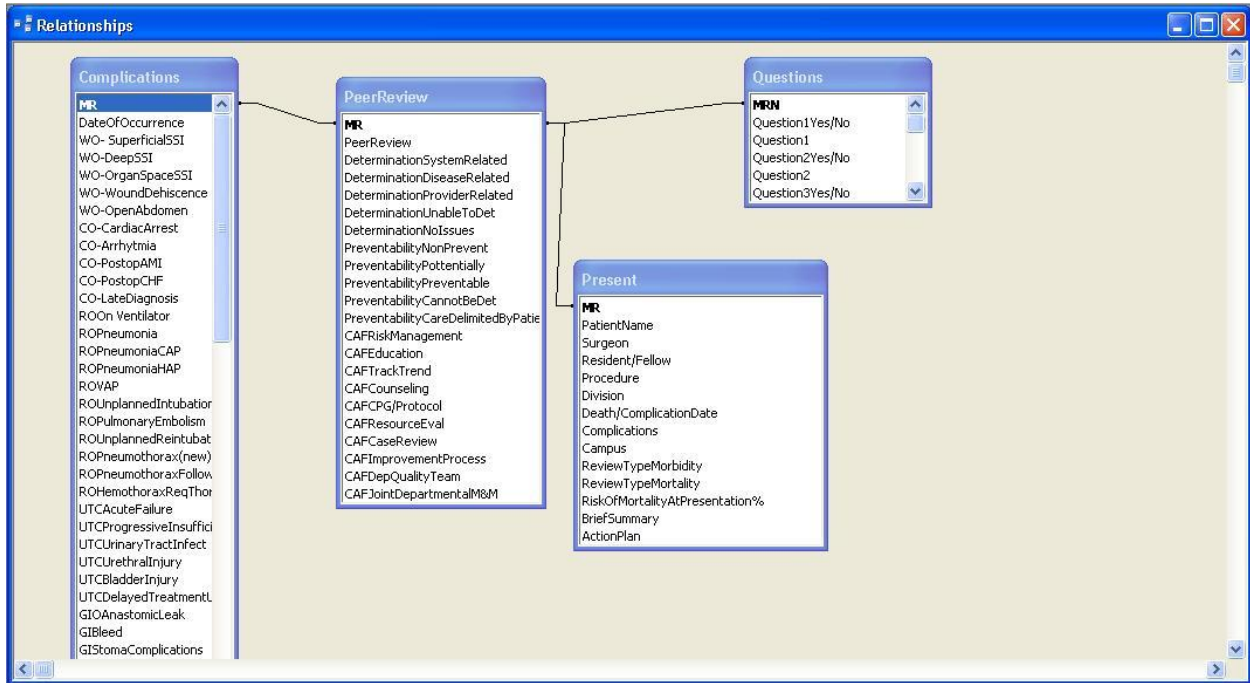


Figure 2-4 Generic Table Relationship in Microsoft Access 2003

2.2.3 Database Forms

Forms are a graphical representation of one or more tables. Records can be more easily modified through a form, which is the reason why most data entry is done through forms. Forms are also preferred for data entry when the database contains many fields that are hard to visualize in a table. Most of the time forms give information on one record at a time. A form can have various field formats for data input: radio buttons, check boxes, drop down list (combo boxes), and text fields.

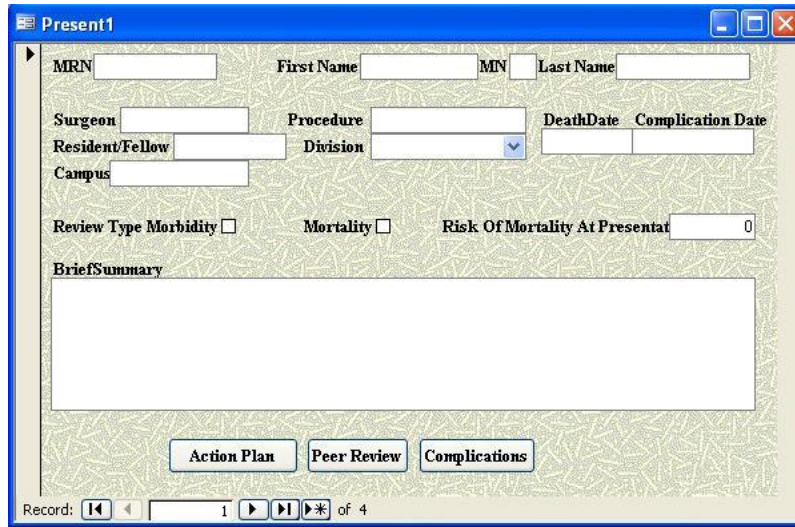


Figure 2-5 Generic Form in Microsoft Access 2003

2.3 Data Mining Background

The data mining background will familiarize the reader with technical concepts regarding data mining topics. Before our data mining work will be described in section 0,53 a few technical topics will be covered in this section to facilitate the understanding of the said section.

2.3.1 Target Attribute, Training Data, and Test Data

In each of the experiments conducted, a prediction is made for a specific patient outcome. This patient outcome is the target class of the experiment. The different values that are possible for this class are called target values. A machine learning algorithm will attempt to classify or predict the value of the target class for each patient, and its accuracy is measured by the number of correctly predicted values out of the total number of predictions made. When employing machine learning algorithms, the algorithm uses the data of a collection of patients called the training set to create a model. Then it makes predictions on a separate collection of patients called the test set.

2.3.2 Attribute Selection

The experiments using machine learning algorithms begin with an attribute selection algorithm. Attribute selection is done to learn which are the most appropriate attributes to use for predicting the target class. The importance of attribute selection can be understood when looking at the following example. Consider that you are assigned on a task to locate WPI, and you are given a world map. The task becomes significantly easier of given a map of Worcester County. Similarly, if all the available information is

filtered first so that only the important ones are left, the classifier algorithm will have a higher chance at creating an efficient model.

Relief-f is the attribute selection algorithm employed in the experiments conducted in this IQP. Relief-F is an extension of Relief, which only handled Boolean concept problems (25). It samples instances (patients) randomly and checks neighboring instances of the same and different classes. The main idea in Relief is to estimate attributes according to how well their values distinguish among instances that are near to each other. Good attributes should differentiate between instances from different classes and should have the same value for instances from the same class (26). It is used in conjunction with Ranker, which ranks attributes and removes the lower ranking ones.

2.3.3 Classification Algorithms

Classification methods use a set of input parameters for characterizing specific objects in relation to a target class. The objects are later ordered or organized according to the specified target class. In addition to the initial classification methods, there were developed classification algorithms that are able to create a model that can correctly predict the class of a targeted object.

A classification algorithm used in this IQP is Bayesian network. Bayesian networks use a network structure to represent probability distributions graphically. It is a network of nodes, where each node is an attribute in the dataset, connected by direct edges. Each node essentially has a table listing the probability of each class value of the attribute associated with the node. This probability is dependent on the class values of the parent node, which is origin node of an edge that is pointing to the node in hand. If there are two or more parents to the node, the probability for each class value is the probability of that particular value occurring, given a specific combination of class values of the parent nodes. The maximum number of parent nodes allowed is a controllable parameter in Bayesian network (27).

Another classification algorithm used in this IQP is the logistic regression. Logistic regression is a technique that builds a linear model based on the logit transform of a target attribute. The logit transform of a certain probability $Pr[1|a_1, \dots, a_k]$ is $\frac{\log(Pr[1|a_1, \dots, a_k])}{1 - Pr[1|a_1, \dots, a_k]}$. This technique is a spin-off of linear regression. In linear regression, a regression is performed for each target class, setting the output to 1 for instances in the training set that belong to the class and 0 for ones that do not. This results in a linear expression for the class. To make a prediction, the value of each linear expression is calculated, and the one with that largest value is chosen. The logit transformation is done to take care of certain drawbacks of

the linear regression techniques, such as assumptions that are made in using this method which are violated when the technique is applied to classification problems (27).

2.3.4 The WEKA System

To apply these algorithms to our data, a Java based software called WEKA was used (27). WEKA is a collection of machine learning algorithms for data mining. It is a highly versatile software, containing tools for data pre-processing, classification, regression, clustering, association rules, visualization and others. It is open source software issued under the GNU General Public License.

2.4 Previous Work on the Pancreatic Cancer Database

The original pancreatic cancer database was created by John Hayward as a part of his Master's thesis (1). The scope of his thesis can be broken down into two distinct goals. The first was to develop a clinical performance database of pancreatic cancer patients. The second was to conduct data mining and machine learning studies on the information contained in the database to develop models for predicting cancer patient medical outcomes. Hayward developed the clinical database in conjunction with the surgeons and oncologists at the UMass Memorial Health Care Center, Worcester.

2.4.1 Hayward's M.S. Thesis

The database was developed using Microsoft Access 2003 with Visual Basic scripting and SQL Server for data storage. Hayward developed databases for six major forms of gastrointestinal cancer (pancreatic, biliary, esophageal, gastric, colorectal, and hepatocellular). Patient information was structured into eight categories in the database:

- Presentation
- Medical History
- Diagnostic Tests
- Preliminary Outlook
- Treatment
- Surgical Resection Details/Reasons for Not Pursuing Resection
- Pathology Reports
- Follow-Up

Since the scope of this project is only on cancer of the pancreas and our project is based on the updated database, only the portion of Hayward's database regarding pancreatic cancer will be covered in this section in a brief manner. The following figures are screen shots of the various forms from the original database.

Pan_1_Present : Form

Presumptive Diagnosis at Onset of Care

[Dropdown]

Presentation

| Date of Evaluation | ECOG Performance Status | Height (in.) | Weight (lbs.) |
|--------------------|---|--------------|---------------|
| [Text] | <input type="radio"/> 0 <input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4 | [Text] | [Text] |

Symptoms

| | | | |
|--|---|---|--|
| <input type="checkbox"/> Weight Loss | <input type="checkbox"/> Biliary Colic | <input type="checkbox"/> Pruritis | <input type="checkbox"/> Indigestion |
| How Much (pounds): [Text] | <input type="checkbox"/> Nausea | <input type="checkbox"/> Abdominal Pain | <input type="checkbox"/> Dysphagia |
| <input type="checkbox"/> Jaundice | <input type="checkbox"/> Vomiting | <input type="checkbox"/> Back Pain | <input type="checkbox"/> Early Satiety |
| <input type="checkbox"/> Cholecystitis | <input type="checkbox"/> Clay Colored Stool | <input type="checkbox"/> Other Specify: | |
| <input type="checkbox"/> Cholangitis | <input type="checkbox"/> Fatigue | [Text] | |

Figure 2-6 Pancreatic Cancer Presentation Form (1)

Pan_2_History : Form

Medical History

| Comorbidities | Cancer History |
|--|---|
| <input type="checkbox"/> Heart Failure <input type="checkbox"/> Ischemic Heart Disease <input type="checkbox"/> Respiratory <input type="checkbox"/> Renal Failure <input type="checkbox"/> Hypertension <input type="checkbox"/> Bleeding Disorder | <input type="checkbox"/> Malnutrition <input type="checkbox"/> Liver Failure/Cirrhosis <input type="checkbox"/> Diabetes <input checked="" type="radio"/> Less than Six Months <input type="radio"/> Greater than Six Months <input type="checkbox"/> Oral Agents <input type="checkbox"/> Diet Control |
| Social History <input type="checkbox"/> Cigarette Use <input type="checkbox"/> Alcohol Use <input type="checkbox"/> Other - Specify: [Text] | <input type="checkbox"/> Irregular Drug Use <input type="checkbox"/> Environmental Exposure |

Cancer History

Patient Prior Dx: [Dropdown]

Chemo Radiation Surgery

Father Dx: [Dropdown]

Mother Dx: [Dropdown]

Other Relation: [Text]

Related Dx: [Dropdown]

Other Relation: [Text]

Related Dx: [Dropdown]

Figure 2-7 Pancreatic Cancer Medical History Form (1)

Pan_3a_Serum : Form

Serum Studies

| | | | | | | | |
|---------|----------------------|-----------------|----------------------|------|----------------------|----------|----------------------|
| CEA: | <input type="text"/> | Albumin | <input type="text"/> | ALK | <input type="text"/> | ALT | <input type="text"/> |
| CA19-9: | <input type="text"/> | Total Bilirubin | <input type="text"/> | AST: | <input type="text"/> | Amylase: | <input type="text"/> |

Figure 2-8 Pancreatic Cancer Diagnosis (Serum Studies) Form (1)

Pan_3b_Diagimg : Form

Diagnostic Imaging Procedures

CT with Pancreatic Protocol/CTA

Date of Procedure:

| | | | | | |
|---|---------------------------------------|--|--|---|--|
| <input type="checkbox"/> Celiac Artery Involvement | <input checked="" type="radio"/> Open | <input checked="" type="radio"/> Abutted | <input checked="" type="radio"/> Encased | <input checked="" type="radio"/> Occluded | <input checked="" type="radio"/> Unknown |
| <input type="checkbox"/> Superior Mesenteric Artery Involvement | <input checked="" type="radio"/> Open | <input checked="" type="radio"/> Abutted | <input checked="" type="radio"/> Encased | <input checked="" type="radio"/> Occluded | <input checked="" type="radio"/> Unknown |
| <input type="checkbox"/> Hepatic Artery Involvement | <input checked="" type="radio"/> Open | <input checked="" type="radio"/> Abutted | <input checked="" type="radio"/> Encased | <input checked="" type="radio"/> Occluded | <input checked="" type="radio"/> Unknown |
| <input type="checkbox"/> Inferior Vena Cava Involvement | <input checked="" type="radio"/> Open | <input checked="" type="radio"/> Abutted | <input checked="" type="radio"/> Encased | <input checked="" type="radio"/> Occluded | <input checked="" type="radio"/> Unknown |
| <input type="checkbox"/> Superior Mesenteric Vein Involvement | <input checked="" type="radio"/> Open | <input checked="" type="radio"/> Abutted | <input checked="" type="radio"/> Encased | <input checked="" type="radio"/> Occluded | <input checked="" type="radio"/> Unknown |
| <input type="checkbox"/> Portal Vein Involvement | <input checked="" type="radio"/> Open | <input checked="" type="radio"/> Abutted | <input checked="" type="radio"/> Encased | <input checked="" type="radio"/> Occluded | <input checked="" type="radio"/> Unknown |

Nodes: Celiac Nodal Disease Other Nodal Disease No Nodal Assessment or Mention

Tumor Size (cm): by

Chest X-Ray (CXR)

Percutaneous Transhepatic Cholangiography (PTC)

Date of Procedure:

Stenting Type: Internal External

Figure 2-9 Pancreatic Cancer Diagnosis (Diagnostic Imaging) Form (1)

Pan_3c_Endoscopy : Form

Endoscopy Procedures

Endoscopic Ultrasound (EUS)

Date of Procedure:

| | | | | | |
|---|---------------------------------------|--|--|---|--|
| <input type="checkbox"/> Celiac Artery Involvement | <input checked="" type="radio"/> Open | <input checked="" type="radio"/> Abutted | <input checked="" type="radio"/> Encased | <input checked="" type="radio"/> Occluded | <input checked="" type="radio"/> Unknown |
| <input type="checkbox"/> Superior Mesenteric Artery Involvement | <input checked="" type="radio"/> Open | <input checked="" type="radio"/> Abutted | <input checked="" type="radio"/> Encased | <input checked="" type="radio"/> Occluded | <input checked="" type="radio"/> Unknown |
| <input type="checkbox"/> Hepatic Artery Involvement | <input checked="" type="radio"/> Open | <input checked="" type="radio"/> Abutted | <input checked="" type="radio"/> Encased | <input checked="" type="radio"/> Occluded | <input checked="" type="radio"/> Unknown |
| <input type="checkbox"/> Inferior Vena Cava Involvement | <input checked="" type="radio"/> Open | <input checked="" type="radio"/> Abutted | <input checked="" type="radio"/> Encased | <input checked="" type="radio"/> Occluded | <input checked="" type="radio"/> Unknown |
| <input type="checkbox"/> Superior Mesenteric Vein Involvement | <input checked="" type="radio"/> Open | <input checked="" type="radio"/> Abutted | <input checked="" type="radio"/> Encased | <input checked="" type="radio"/> Occluded | <input checked="" type="radio"/> Unknown |
| <input type="checkbox"/> Portal Vein Involvement | <input checked="" type="radio"/> Open | <input checked="" type="radio"/> Abutted | <input checked="" type="radio"/> Encased | <input checked="" type="radio"/> Occluded | <input checked="" type="radio"/> Unknown |

Nodes: Celiac Nodal Disease Other Nodal Disease No Nodal Assessment or Mention

Tumor Size (cm): by

EUS Staging: T N **FNA Cytology**

Endoscopic Retrograde Cholangiopancreatogram (ERCP)

Date of Procedure:

Stenting Type: Plastic Metal

Figure 2-10 Pancreatic Cancer Diagnosis (Endoscopy Studies) Form (1)

Pan_4_Prelim : Form

Pre-Surgical Outlook

Potentially Resectable

Locally Advanced/Unresectable

Metastatic or Equivocal Findings

Figure 2-11 Pancreatic Cancer Preliminary Outlook Form (1)

Pan_6a_Res : Form

If Resection is Performed

Surgery

Date of Admission: Date of Surgery:

Procedure Type: OR Time (hr):

Venous Resection
 Venous Reconstruction
 Arterial Resection
 Arterial Reconstruction

Other Organs Resected: Estimated Blood Loss (cc):

Transfusion
If Yes, Units:
Methods:
 FFP
 Cell Saver

Resection Attempt: Successful
 Unsuccessful - Reason:

Post-Op

Days in ICU:

Post-Op Care Path: Congruent Divergent

NG/Gastrostomy Drainage > 7days
 Abdominal Collection

Pulmonary Complications
 Wound Infection
 Leak

Liver Insufficiency (Total Bilirubin > 5)
If Yes, Total Bilirubin:

Date of Discharge: Discharge Status:

Figure 2-12 Pancreatic Cancer Resection Form (1)

Pan_6b_NoRes : Form

If Resection is Not Performed

Date of Decision:

Reasons (select all that apply):

Clinical Decision

- Patient Couldn't Handle Proposed Treatment
- Patient Refused Treatment
- Proposed Magnitude of Treatment and Risks Not Worth Likely Benefit

Vascular Involvement

- Celiac Artery Involvement
- Superior Mesenteric Artery Involvement
- Hepatic Artery Involvement
- Inferior Vena Cava Involvement
- Superior Mesenteric Vein Involvement
- Portal Vein Involvement

Additional Disease

- Cirrhosis
- Evidence of Metastasis

Figure 2-13 Pancreatic Cancer No Resection Form (1)

Pan_7_Path : Form

Final Tumor Histology

(from best of imaging, FNA, pathology, etc..)

Pathology (if available)

Tumor Size (cm): by

TMN Staging: T N M R:

Figure 2-14 Pancreatic Cancer Pathology Form (1)

Pancreatic Tumor Follow-up Information

MR: Follow-up Window:

Date of Visit: Weight (pounds): QOL score:

ECGO performance status: 0 1 2 3 4

Lab Value:

CEA: Albumin: Alkaline Phosphatase:
 CA19-9: Total Bilirubin:

Redeveloped Symptoms:

Weight loss how much (pounds) Biliary colic Pruritis Back pain
 Nausea Abdominal pain Indigestion
 Vomiting Other Specify:
 Cholecystitis Clay colored stool Fatigue
 Cholangitis

Status:

Died Death Date:
 N.E.D.
 A.W.D., Method of Detection: Lab Radiologic Evidence Clinical Evidence

Figure 2-15 Pancreatic Cancer Follow-Up Form (1)

2.4.2 Floyd's M.S. Thesis

After Hayward, Stuart Floyd also did a Master's thesis (2) using the UMass Medical School Pancreatic Cancer database. Using Hayward's database as a foundation, Stuart focused on pancreatic cancer and applied data mining techniques for pancreatic cancer survival time prognosis. All material discussed from this point on considers Floyd's updated version of the database. He made changes to the database with inputs from doctors at UMass Memorial Hospital. The new pancreatic cancer database displays information in a different organizational format. Once the patient is selected either by name, medical record number, or associated primary key number unique to the database, there are four forms that display patient information:

- Pre-Operative
- Peri-Operative
- Surgical Pathology
- Follow-Up

In order to ease the access to patient information, the database consists of an initial page that opens automatically. This form contains basic information about the patient such as medical record number (MRN), first name, last name and middle initial, demographic data (race, gender), date of birth, data of last data entry, and the date of death (if the patient has expired). From this main form, the user can choose to view any of the four sections: Pre-Operative, Peri-Operative, Surgical Pathology and Follow-Up Information.

The screenshot shows a software window titled "Patient" with a sub-header "Patient Information". Below the sub-header are two buttons: "Find Record" and "Print Record". The main content area is titled "Pancreatic Tumor" and includes a "Date of Last Data Entry:" field and a "History Later" checkbox. The form contains several input fields: "MR#:", "First Name", "MI", "Last Name", "Date of Birth", "Date of Death", "Race:" (with a dropdown arrow), and "Gender" (with radio buttons for "Male" and "Female"). On the right side of the form, there are four stacked buttons: "Pre-operative Information", "Peri-operative Information", "Surgical Pathology", and "Follow Up Information". At the bottom of the window, there is a record navigation bar that says "Record: 1 of 1" with navigation icons.

Figure 2-16 Pancreatic Cancer Patient Form (2)

The organization into these four sections follows a logical and systematic flow mirroring the sequence of events that takes place in the treatment of a patient.

Patient Enrollment

Pancreatic Tumor Race:

MR#: First Name MI Last Name

Date of Birth Gender Male Female

Date of Initial Evaluation:

Presumptive Diagnosis at Onset of Care:

Other:

Presenting Signs/Symptoms

Abdominal Pain Cholangitis Fatigue Pruritis
 Back Pain Cholecystitis Indigestion Vomiting
 Biliary Colic Dysphagia Jaundice Weight Loss Pounds:
 Clay Colored Stool Early Satiety Nausea
 Other Specify:

Weight(lbs): Height(inches):

ECOG Performance Status Not Specified
 0 1 2 3 4

Medical History - Comorbidities(check all that apply)

Heart Failure Respiratory Bleeding Disorder Diabetes
 Renal Failure Hypertension Ischemic Heart Disease Less than Six Months
 Malnutrition EtOH Abuse Liver Failure/Cirrhosis Greater than Six Months
 Pancreatitis Obesity Other Major Comorbidity Oral Agents
 Current Tobacco Pack Years: Diet Control

Cancer History: Patient History of Any Cancer Details:
 Family History of Pancreatic Cancer Relationship to Patient:

PreOp Serum Studies: Test Date:

CEA: Albumin ALK ALT
CA19-9: Total Bilirubin AST: Amylase:

CT with Pancreatic Protocol/CTA: # of Tumors:

Date of Procedure: Max Tumor Size (cm): by

Celiac Artery Involvement Clear Abutted Encased Occluded Unknown
 Superior Mesenteric Artery Involvement Clear Abutted Encased Occluded Unknown
 Hepatic Artery Involvement Clear Abutted Encased Occluded Unknown
 Inferior Vena Cava Involvement Clear Abutted Encased Occluded Unknown
 Superior Mesenteric Vein Involvement Clear Abutted Encased Occluded Unknown
 Portal Vein Involvement Clear Abutted Encased Occluded Unknown

Endoscopic U/S Date of Procedure: # of Tumors:

Celiac Artery Involvement Clear Abutted Encased Occluded Unknown
 Superior Mesenteric Artery Involvement Clear Abutted Encased Occluded Unknown
 Hepatic Artery Involvement Clear Abutted Encased Occluded Unknown
 Inferior Vena Cava Involvement Clear Abutted Encased Occluded Unknown
 Superior Mesenteric Vein Involvement Clear Abutted Encased Occluded Unknown
 Portal Vein Involvement Clear Abutted Encased Occluded Unknown

EUS Staging: T N FNA Cytology

Max Tumor Size (cm): by

Nodes > 1cm: Peripancreatic Celiac Other

ERCP Date:

ERCP Stenting: None Metal Plastic Material Unknown

FNA Cytology: Malignant Not Definitive for Malignancy Not Done

Anticipated Operation:

Pancreaticoduodenectomy Distal Pancreatectomy Central Pancreatectomy Enucleation Other

Figure 2-17 Pancreatic Cancer Pre-Operative Form (2)

Information of the patient before any definitive diagnosis is contained under the pre-operative section. The date of initial evaluation and the presumptive diagnosis at the onset of care are included since it may take a long time between the initial evaluation and the actual diagnosis of pancreatic cancer. Also, there exist situations where patients may be diagnosed with a slightly different form of pancreatic cancer due to insufficient or misleading symptoms and test results.

The section continues to check for symptoms related to pancreatic cancer, the ECOG performance status and physical data (weight and height). Medical history is an important section of the pre-operative section, since some percentages of the patients show relational patterns between their newly developed pancreatic cancer and comorbidities and/or family history (pancreatic cancer in a family member or other form of cancer in the patient).

Diagnostic tests done prior to surgery are recorded in the pre-operative section. These include pre-operative serum studies (see section 3.2), CT with pancreatic protocol or CTA, EUS, ERCP and FNA cytology results. All the above mentioned diagnosis techniques provide information about the patient's initial status, thus it is important that the database has included the data of the test for each of them. Both the CT and EUS look for the number of visible tumors, register the size of the biggest tumor and its involvement with major blood vessels around the pancreas. Depending on the position and development of the tumor, the blood vessel can be clear, encased, abutted or opened. Aside from the fields common to a CT scan, the EUS provides a staging of the tumor (in the TNM system, out of which just TN is used in the database), and information about abnormal lymph nodes. M classification was not included since it does not give more information than T and N. FNA results give important information about the type of cells from the abnormal tumor, thus enabling for a classification (malignant, malignant with adenocarcinoma, insufficient for diagnosis, inadequate, other) so space has been allocated in the section to record these information. FNA is sometimes done for the abnormal lymph nodes for concluding whether they are malignant or not. Since ERCP is usually, but not always, accompanied but a stent placement, information concerning stents in relation to ERCP is also recorded. Finally, there is a field for recording anticipated surgery, which is based on doctors' decision using all the diagnostic test results.

Patient

UMass Surgical Outcomes and Analysis Research (SOAR) Pancreatic Cancer Database

Peri-Operative

MR#: First Name MI Last Name

I. Pre-Op Information: Neoadjuvant Chemotherapy - Agent:

Neoadjuvant Radiation

Endoscopic U/S Date of Procedure: # of Tumors:

Celiac Artery Involvement Clear Abutted Encased Occluded Unknown

Superior Mesenteric Artery Involvement Clear Abutted Encased Occluded Unknown

Hepatic Artery Involvement Clear Abutted Encased Occluded Unknown

Inferior Vena Cava Involvement Clear Abutted Encased Occluded Unknown

Superior Mesenteric Vein Involvement Clear Abutted Encased Occluded Unknown

Portal Vein Involvement Clear Abutted Encased Occluded Unknown

EUS Staging: T N FNA Cytology

Max. Tumor Size (cm): by

Nodes > 1cm: Peripancreatic Celiac Other

ERCP Date:

ERCP Stenting: None Metal Plastic Material Unknown

FNA Cytology: Malignant Not Definitive for Malignancy Not Done

II. Surgery (primary resection): Date of Admission: Date of Surgery:

Whipple (standard) Berger Venous Resection: None SMV Portal Vein

Whipple (pylorus preserving) Frey's SMV-Portal Vein Confluence Unknown/Other

Total Pancreatectomy Enucleation Venous Reconstruction: None Interposition graft

Distal Pancreatectomy Other: Greater Saphenous Vein Patch Saphenous vein

Central Pancreatectomy End-to-end SMV to Portal anastomosis IJ

Other Organs Resected: Unknown/Other Synthetic conduit

Pancreatic Stent Biliary Stent Splenic Vein Preservation

Estimated Blood Loss (cc): Transfusions: PRBC, Units: FFP Cell Saver

Feeding Tubes: g tube j tube g-j tube none Drains: 0 1 2

III. Post-operative:

Extubated in OR Extubated in ICU: POD 0 POD 1 POD 2 >POD 2

POD Tube Feeds Started: POD TPN Started:

POD Clears Started: POD Reg Diet:

POD tube feeds stopped: POD TPN Stopped:

POD Drain Removed: POD NG tube removed:

Complications:

Delayed Gastric Emptying ileus Prolonged PO Intolerance: TPN Started ICU Re-Admission

Feeding Tube Dislodged Feeding Tube Clogged Tube Feeds Not Tolerated Post Op Bleeding

Wound Infection ARF: Abdominal Collection Leak

Any Post-Op Transfusion DVT Cardiac Complications/MI Pulmonary Complications

Any Reoperation PE Liver Insufficiency (T. bili > 5) Central Line Sepsis

C. Diff Colitis UTI Other Complications:

Date of Discharge: Discharged To:

Discharged With: Drains Feeding Tube

TPN New Insulin Requirement

Figure 2-18 Pancreatic Cancer Peri-Operative Form (2)

The next section is based on the surgery and the post-operative course. Some diagnostic test result information is displayed again at the beginning of this section. They are followed by fields such as date of admission, date of operation, the type of surgery, venous resection and venous reconstruction. There are various types of surgeries that can be performed (Whipple standard or pylorus preserving, total, distal or central pancreatectomy, Berger or Frey’s procedure, enucleation or other less known surgeries. See section 3.3 for details.). During the surgery some organs (spleen, gallbladder) or veins can be resected (the portal vein, the SMV, other types). The database includes the option to record some of the main reconstruction methods like the interposition graft (saphenous vein, IJ, synthetic conduct), greater saphenous vein patch, end-to-end SMV portal anastomosis, other reconstruction methods or in some cases no reconstruction after a vein resection. During the surgery the patients might have a placement of a stent (biliary or pancreatic), feeding tubes (gastric tube, jejunostomy tube or gastric- jejunostomy tube) or drains. The database also has space to record information regarding blood loss during the surgery and transfusion option used (packets of red blood cells- PRBC, fresh frozen plasma- FFP or cell saver.^H).

The peri-operative section concludes with information on post-operative care of the patients, such as where the extubation took place, type of care started at the intensive care unit including the use of TPN (Total Parenteral Nutrition), and any complications experienced by the patient after surgery. Discharge information is also recorded.

The screenshot shows a software window titled "Pan_7_Path : Form". At the top, there are input fields for "MR#", "First Name", "MI", and "Last Name". Below these is a section titled "Final Tumor Histology" with a dropdown menu and the text "(from best of imaging, FNA, pathology, etc..)". Underneath is another section titled "Pathology (if available)" which includes fields for "Tumor Size (cm):" followed by two input boxes and the word "by", and "TNM Staging:" followed by dropdown menus for "T", "N", "M", and "R".

Figure 2-19 Pancreatic Cancer Pathology Form (2)

If concrete final diagnosis can be given, using results from biopsy, they are recorded in the third section. The pathology section includes information on final tumor histology, tumor size and TNM staging.

^H The cell saver procedure consists of the washing and filtering of the patient’s blood; the procedure is done in order for the patient to receive his blood back.

Patient

UMass Surgical Outcomes and Analysis Research (SOAR) Pancreatic Cancer Database

Peri-Operative MR#: First Name MI Last Name

Follow-up

Date of Visit: Weight (pounds):

Adjuvant Treatment:

Adjuvant Chemotherapy: Currently Planned Stopped Completed Agent:

Adjuvant Radiation: Currently Planned Stopped Completed

ECOG performance status: 0 1 2 3 4 Not Specified

Signs / Symptoms(check all that apply):

| | | |
|--|---|--|
| <input type="checkbox"/> Abdominal pain | <input type="checkbox"/> Clay colored stool | <input type="checkbox"/> Jaundice |
| <input type="checkbox"/> Back pain | <input type="checkbox"/> Dysphagia | <input type="checkbox"/> Nausea |
| <input type="checkbox"/> Biliary colic | <input type="checkbox"/> Early Satiety | <input type="checkbox"/> Pruritis |
| <input type="checkbox"/> Cholangitis | <input type="checkbox"/> Fatigue | <input type="checkbox"/> Vomiting |
| <input type="checkbox"/> Cholecystitis | <input type="checkbox"/> Indigestion | <input type="checkbox"/> Weight loss (lbs): <input type="text"/> |
| <input type="checkbox"/> Other: <input type="text"/> | | |

Serum Studies (include date if different from current):

| | |
|--|---|
| CEA: <input type="text"/> Date: <input type="text"/> | CA19-9: <input type="text"/> Date: <input type="text"/> |
| Albumin <input type="text"/> Date: <input type="text"/> | Total Bilirubin <input type="text"/> Date: <input type="text"/> |
| Alk. Phos. <input type="text"/> Date: <input type="text"/> | AST: <input type="text"/> Date: <input type="text"/> |
| ALT: <input type="text"/> Date: <input type="text"/> | Amylase: <input type="text"/> Date: <input type="text"/> |

Drains / TPN / Feeding Tubes

If drains present:

Number of Drains:

24 hour drainage(cc.):

Remove Drains Leave Drains

If currently receiving TPN:

Continue TPN Discontinue TPN

If currently receiving tube feeds:

Continue Tube Feeds Discontinue Tube Feeds Remove Feeding Tube

Status:

Died

No Evidence of Disease

Alive with Disease as evidenced by: Lab Radiologic Evidence Clinical Evidence

Record: of 1

Figure 2-20 Pancreatic Cancer Follow-Up Form (2)

The follow-up form, the final section, concerns the course of the patient after surgery. If the patient is going through adjuvant radiation or chemotherapy, the therapy status is given. ECOG performance status and signs/symptoms are recorded to see if the patient is recovering after the surgery. Results of post-operative serum studies are also given for comparison to pre-operative results. In the case that the patient was discharged with drains, TPN, or feeding tubes, the current status regarding them is mentioned on this form. Finally, the current status of the patient concludes the section. The patient status can be dead, no evidence of disease, or alive with disease.

3 Our Database Work

A pattern or technique to search through MEDITECH for desired information was naturally developed as data fields for first several patients were filled in. This was highly important for the purpose of efficiency since there were many patients to be input into the database. There were a total of 91 patients before additional patients were added as a part of this project. The remaining patients that were included into the database have been provided by one of the nurse practitioners from the Surgical Oncology Division. When we started working on the database, the patients' Medical Record Number (MRN) as well as the patients' names had already been input into the Access Database. Please refer to section 3.6 for the status of the database at the time of this writing. Without a systematic procedure for looking through the large variety of reports and datasheets available for each patient in MEDITECH, the amount of time taken to completely fill in the database would have been longer. In this section we will discuss the systematic procedure mentioned above, procedure that has been organized into four sections (pre-operative, peri-operative, pathology, and follow-up) as to match the organization of the actual database. However, it is helpful to first know the organization of the MEDITECH database in order to understand the rest of this section.

MEDITECH (28) is an interface that enables access to UMass Hospital patient records, and it is comprised of many sections which vary in length depending on the patient and their medical situation. For example, patients with minor sport injuries may have few sections in the database, compared to a patient with pancreatic cancer, who may easily have more than 10 sections. The sections of interest for our purpose of inputting pancreatic cancer patient data are:

- Admissions Demographic Data
- Laboratory Data
- Anatomical Pathology Reports
- Diagnostic Imaging
- Diagnostic Imaging (After 10/10/06)
- Endoscopy Reports
- Transcription Reports

Under Transcription Reports, there exist several different types of reports. There are transfer summaries, pre-admit summaries, letters, operative reports, discharge reports, consultations, clinical notes, outpatient consultation reports, and radiology/oncology therapy completion report.

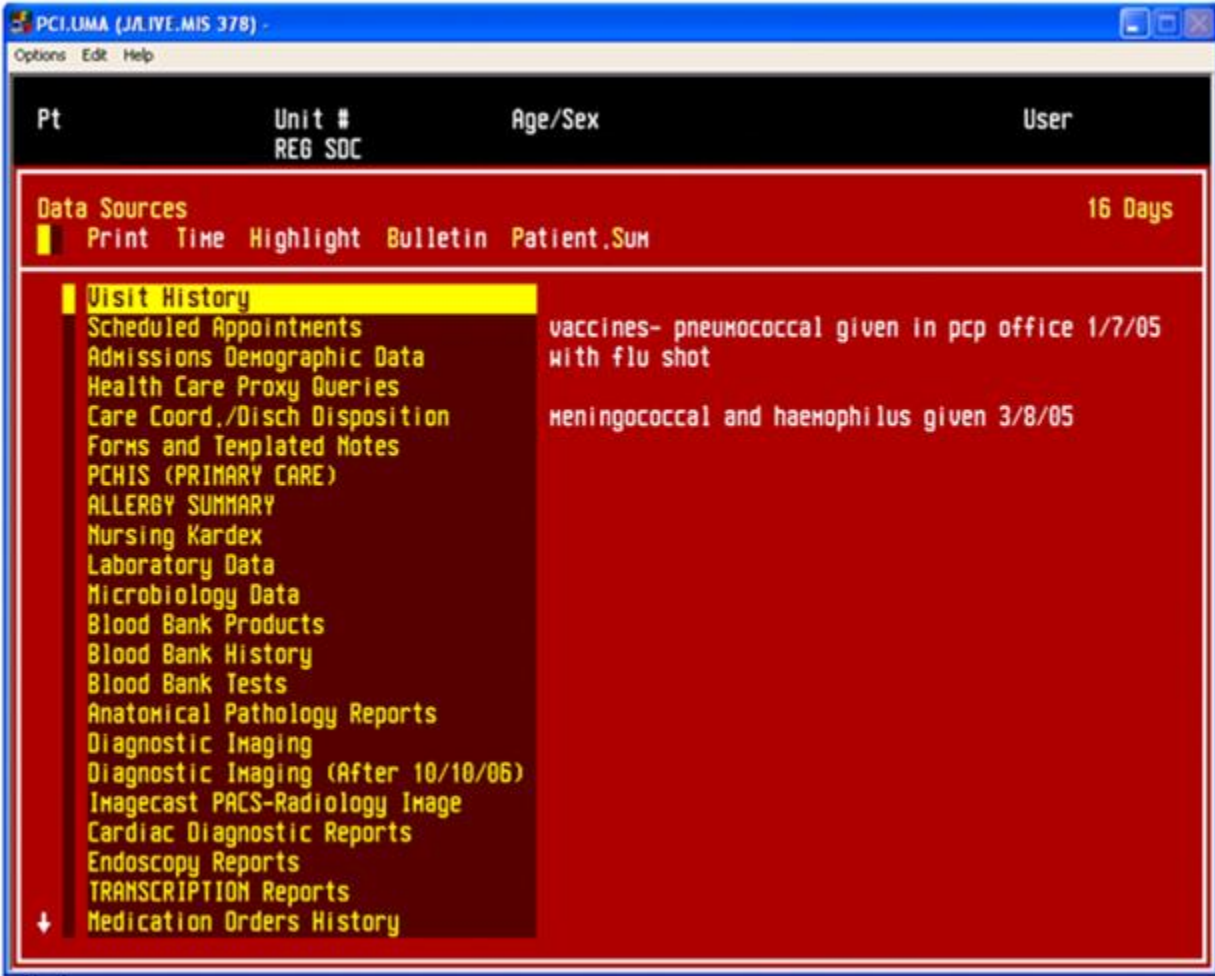


Figure 3-1 Screenshot of MEDITECH

3.1 Patient Demographics

The patient demographics information is stored in the Pancreatic Cancer Database under the form “Patient”. For each patient, we looked at the MRN stored in the Access database and then we did a query in MEDITECH (28) for that specific MRN. Once the query was retrieved we could access the patients’ medical records for demographic information. Under the Admissions Demographic Data we could find all the fields we were interested in: the medical record number, the name of the patient (the first two data fields we had already been stored in the Access database), date of birth (DOB), date of death, race and gender. The death dates that were not present in MEDITECH were taken from Social Security Death

Index¹. In order to keep track of the data updates done on the database, we also filled in a field with the last date of data entry for the current patient. The last date of data entry represents the last time the patient information was updated into the Pancreatic Cancer Database, not in MEDITECH. In case the patient was known to still be under medical observation, we would fill in the field “History later”. This field informs the person who will work on the database next that the patient’s medical case was not solved at the time of the last data entry.

3.2 Pre-Operative

The first series of information that is crucial in starting to load a patient are dates. To find the date of initial medical evaluation related to pancreatic cancer, we looked through the entire list of transcription reports and read the very first report that is related to pancreatic cancer symptoms. In order to confirm this, several reports may have to be read. Starting from the oldest report, we went through report content and found the report with the first mentioning of pancreatic cancer. This report usually begins with initial symptoms. Then we backtrack into older reports to see if the same symptoms were previously reported. The date of this report is used as the initial date of evaluation.

From the same report, we can usually find the presumptive diagnosis at the onset of care. If the report was not done by an oncologist then the presumptive diagnosis is not mentioned. Thus, in order to retrieve the presumptive diagnosis we would have to look into the first report done by an oncologist. Patient weight and height are occasionally but not always noted in the report, depending on who wrote the report. Patients that were seen only for consultations and not admitted into the hospital would not have the weight and height available. Therefore there is a slight variability on the availability of information depending on the patient. If any ECOG performance status was recorded, it is usually mentioned in the same report or a report that is very close in date to the initial report. Medical and cancer history (comorbidities) can also be obtained from the first couple of reports after the date of initial evaluation.

Results of all serum studies can be found in the subsection Chemistry which is under the section Laboratory Tests. These values are straightforward to obtain. However, if the patient has undergone surgery, the serum study results recorded in the pre-operative report must be filed on a date prior to the surgery. Therefore we checked the date of surgery first and then looked for serum study results filed between the initial date of evaluation and the operation date. The serum test date was also cross-referenced to the date of initial evaluation, since the purpose of these tests is to describe the patient’s

¹The Social Security Death Index is and up to date database that stores the death dates for people registered with the Social Security System.

overall initial condition at the onset of care. In conclusion, the serum tests date should be closest to the initial evaluation date and not exceeding the date of surgery.

If a patient has undergone CT scan, the results are recorded in the subsection Body CT under Diagnostic Imaging. Similar to the serum studies, special attention must be paid to the date. The abovementioned also apply for EUS and ERCP results, except for the fact that they are found under Endoscopy Reports. The results available from CT and endoscopy are the size of the pancreatic tumor, the number of pancreatic tumors, any vein involvement with the tumor. The ERCP information consists of the type of the stent applied and the date of the procedure. To check if FNA was performed, we searched through the Anatomical Pathology Reports section for any FNA results with a date close to the EUS/ ERCP date, and read the report to confirm the biopsy was from the specific procedure.

By reading through the reports prior to the date of operation, we checked to see if the doctors mentioned anticipated course of action for the patient after various diagnostic tests. This information is usually in transcription reports that are dated closer to the date of operation.

3.3 Peri-Operative

Information belonging to the peri-operative section is found in operative and discharge reports. The type of surgery is mentioned towards the very beginning of an operative report. If any venous resection or reconstruction was done, it will also be noted in the report. In the case of a venous reconstruction, there is another operative report with the same date that separately covers the reconstruction procedure. The reason behind this is that a doctor of different specialty is called into the operation room to perform the procedure, and this doctor is required to write a separate report. Other information such as the use of stents, feeding tubes, drains, blood transfusions and estimated blood loss are embedded in the operative report, sometimes summarized at the very end. Availability of this information varies depending on the doctor who wrote the report.

We had to look at both the operative and the discharge report to confirm information regarding extubation after surgery. Discharge reports also contain information about the care given immediately after surgery (tube feeds and TPN for example). Some patients experience post operative complications, and this is noted in the body of the discharge report. If the patient was discharged with special needs like drain, feeding tubes, TPN or insulin regiment, it is mentioned towards the end of the report as well as the date of discharge and the location the patient is discharged to.

3.4 Pathology

The final tumor histology is found under the Anatomical Pathology Report section. It is important to mention that the histology is available only when a type of pancreas resection was performed. In the subsection Pancreatic Surgery under the Anatomical Pathology Report section we found important data like the size of the pancreatic tumor, TNM staging and the final histology.

3.5 Follow-up

Based on the surgeons' decision, the database provides only for a single follow-up instance for each patient. Also, patients that did not undergo surgery have no information for follow-up, as the follow-up section was strictly designed as follow-up after surgery. The selection of the follow-up report from MEDITECH depends on the recurrence of cancer. In case the patient did not present with cancer recurrence after surgery, the follow-up information is taken from the latest follow-up report from MEDITECH that is related to pancreatic cancer. In the case that a patient's cancer was recurrent, the first follow-up report that indicated this recurrence was used to fill out the follow-up form. These follow-up reports usually mention the date of visit, the patient's health status (ECOG, weight, symptoms) as well as the status of possible drains, feeding tubes or TPN. Serum study results for this section of the database are found under the Chemistry section. In case that no serum tests were performed on the date of the selected follow-up visit, we would choose the serum tests with the closest date to the date of the follow-up we were interested in. The date of each individual serum test was stored in the form, thus keeping accuracy of the data selection. The selected follow-up report provided information about any possible cancer recurrence for the patient under observation, thus we were able to fill out the Access form with the status of the patient (alive with disease, death or no evidence of disease).

3.6 Final Status of the Database

The database schema was not modified from that of Floyd's (2). However the database was continuously updated with more patients, even after the data analysis was being conducted. Therefore the size of the database is larger from what was available at the time the experiments began being conducted. Currently, the database is populated with 261 patients, as compared to 252 patients that were available at the time our analysis started. Of the 261 patients, 150 have gone through surgery. Of the 150 that had surgery, 130 of the procedures were resections. The surgery and resection procedures are explained in section 2.1.2.8. **Table 3-1** shows the distribution of surgeries performed. Of the 261 patients 114 have the date of death

confirmed^J by either MEDITECH or the Social Security Death Index. Out of the patients with confirmed survival time, 62 have undergone surgery and 52 had no surgery. Included in the database cohort, there are 120 male and 141 female patients. The database includes several forms that are no longer being used, but there are totally 278 attributes that are being currently filled in and updated for each patient.

| Distribution | Resection Procedure |
|--------------|--|
| 53 | Standard Whipple |
| 28 | Pylorus preserving Whipple |
| 12 | Total Pancreatectomy |
| 23 | Distal Pancreatectomy |
| 3 | Central Pancreatectomy |
| 2 | Berger procedure |
| 5 | Frey procedure |
| 1 | Enucleation |
| 1 | Enucleation and Berger procedure |
| 1 | Total Pancreatectomy and standard Whipple |
| 1 | Distal Pancreatectomy and pylorus preserving Whipple |

Table 3-1 Distribution of Surgical Procedures Among Patients with Pancreatic Cancer

3.6.1 Current Tables

The UMass Pancreatic Cancer database contains 17 tables^K, originally created during Hayward's (1) and Floyd's projects (2). There are 11 functional tables that store the patient information inputted through the forms later mentioned in section 3.6.2. Of the 17 tables, 3 of them are used for storing information needed in the design of the just mentioned functional tables and another 3 out of 17 tables are no longer being used. It should be noted that the tables contain more attributes than the ones that are being collected through the forms. Thus the attributes not present on the forms will not get updated, since the person entering the data is not aware of their existence.

^J Because the MEDITECH database was not updated in what concerns the patients' death dates, we have decided on referring to a second source for death dates. The Social Security Death Index has been generally accepted by the medical field as a reliable death dates source and that motivated us in including it as a data source for our experiments.

^K Refer to section 2.2.2 for further details on Access tables.

| Tables Not in Use |
|-------------------|
| “Cancer Type” |
| “Clinician” |
| “Pan_6b_NoRes” |

Table 3-2 Database Tables Not in Use

| Tables Used in Database Design |
|--------------------------------|
| “Lookup” |
| “Pan_Lookup” |
| “Race” |

Table 3-3 Database Tables Used in the Database Design

| Tables Used for Data Storage | Related Forms | Description of Table Fields included on Related Forms ^L |
|------------------------------|---|--|
| “Pan_1_Present” | “Patient” | Symptoms and initial medical evaluation data |
| “Pan_2_History” | “Short_1_PreOperativeData” | Comorbidities, social history and family history |
| “Pan_3a_Serum” | “Short_1_PreOperativeData” | Pre-operative serum tests results |
| “Pan_3b_DiagImg” | “Short_1_PreOperativeData” | Pre-operative CT data |
| “Pan_3c_Endoscopy” | “Short_1_PreOperativeData” “Short_2_PerioperativeData” | Pre-operative EUS, ERCP and FNA data |
| “Pan_4_Prelim” | “Short_1_PreOperativeData” | Anticipated operation |
| “Pan_5_Treatment” | “Short_2_PerioperativeData” | Therapeutic chemotherapy or radiation |
| “Pan_6a_Res” | “Short_2_PerioperativeData” | Resection information |
| “Pan_7_Path” | “Pan_7_Path” | Pathological information |
| “Pan_8_FU” | “Short_3_FollowUpData” | Follow-up information |
| “Patient” | “Patient” | Patient Demographics |

Table 3-4 Database Tables Used for Storing Patient Data

In order to avoid data redundancy (e.g., repeated patients), table relationships were created into the database. This was done by placing common fields in the tables that are related. The dependencies of our database are presented below. Notice that **Figure 2-1** shows only 14 out of the total of 17 tables. The main reason for this is that the table relationships have not been updated after some of the tables were taken out of use.

^L Refer to section 3.6.2 for a description of the Database forms.

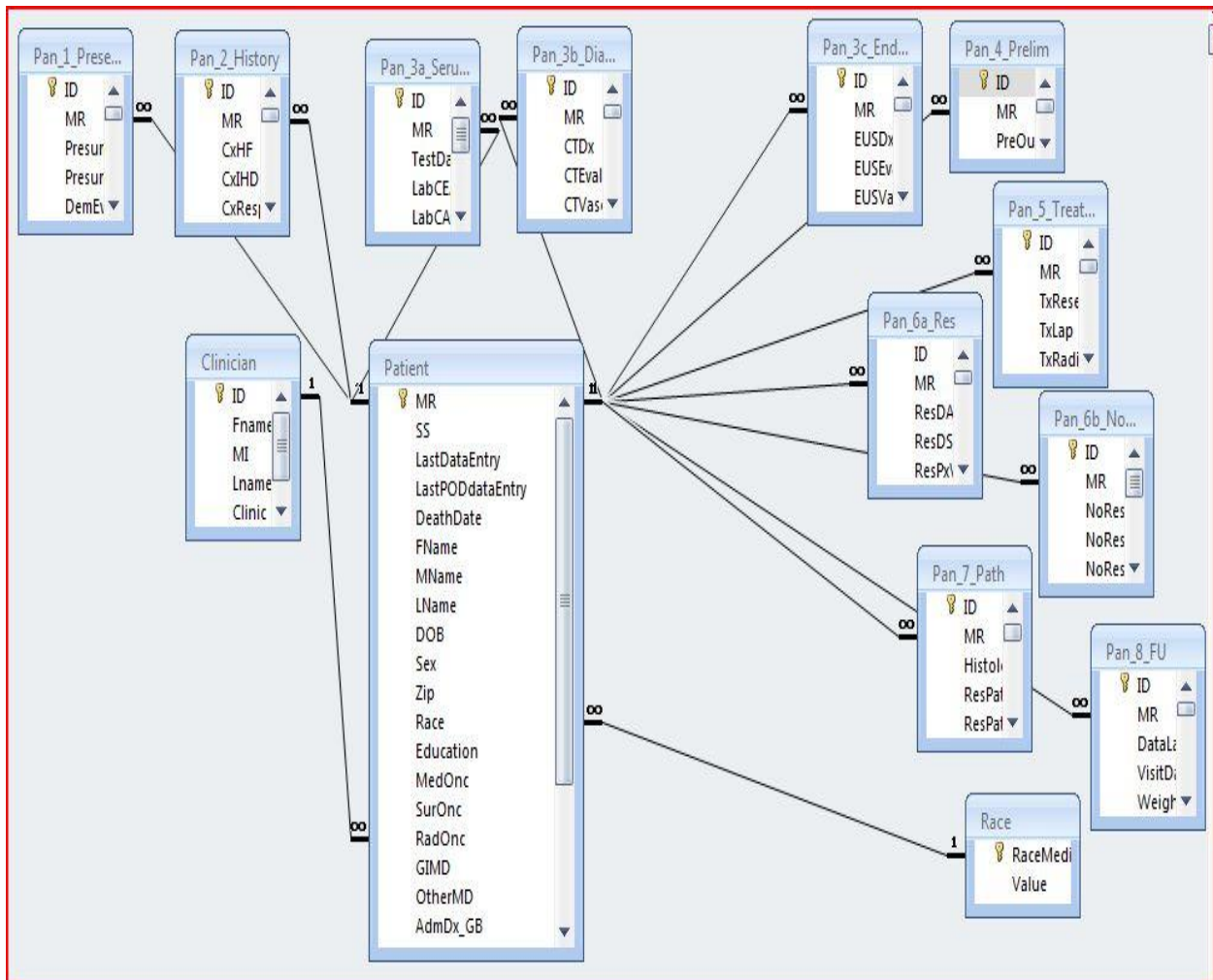


Figure 3-2 Database Tables Relationship

3.6.2 Current Forms

The current database consists of 24 forms^M that were initially designed during Hayward’s (1) and Floyd’s (2) projects. Six of the 24 forms are more complex forms physically used for data entry. 10 out of the 24 forms function only as sub-forms^N as they represent just constituent parts of the more complex forms mentioned above and they are never used as individual entities. We will name these 10 forms “indirectly used forms”. 8 out of the 24 forms are no longer used in the database.

^M Refer to section 2.2.3 for further details on Access forms.

^N A subform is simply a form contained within another form.

| Forms Not In Use |
|---------------------------|
| “Main” |
| “Clinicians” |
| “Pan_6b_NoRes” |
| “Pan_Main” |
| “Pan_Op” |
| “Patient_First_Resection” |
| ”Patient_Tumor_Histology” |
| “Print Blank History” |

Table 3-5 Forms No Longer Used in the Database

| Indirectly Used Forms | Associated Complex Form |
|-----------------------|--|
| “Pan_1_Present” | “Short_1_PreOperativeData” |
| “Pan_2_History” | “Short_1_PreOperativeData” |
| “Pan_3a_Serum” | “Short_1_PreOperativeData” |
| “Pan_3b_DiagImg” | “Short_1_PreOperativeData” |
| “Pan_3c_Endoscopy” | “Short_1_PreOperativeData”; “Short_2_PeriOperativeData” |
| “Pan_4_Prelim” | “Short_1_PreOperativeData” |
| “Pan_5_Treatment” | “Short_2_PeriOperativeData” |
| “Pan_6a_Res” | “Short_2_PeriOperativeData” |
| “Pan_6a_Res_PO” | “Short_2_PeriOperativeData” |
| “Pan_8_Follow_Up” | “Short_3_FollowUpData” |

Table 3-6 Forms Indirectly Used in the Database and their Associated Complex Forms

| Data Entry Forms | Functions | Related Tables |
|------------------------------------|---|--|
| "Patient" | <ul style="list-style-type: none"> data entry for basic patient information link to the other database forms | "Patient" |
| "Patient_FindRecord" | <ul style="list-style-type: none"> allows the search for a specific patient using the medical record number, first or last name the patient is searched throughout the "Patient" form, which, is then opened with the searched patient data if patient is not found, the "Patient" form is opened with the first stored record | None |
| "Short_1_PreOperativeData" | <ul style="list-style-type: none"> data entry for pre-operative information | "Pan_1_Present" "Pan_2_History" "Pan_3a_Serum" "Pan_3b_DiagImg" "Pan_3c_Endoscopy" "Pan_4_Prelim" |
| "Short_2_PeriOperativeData" | <ul style="list-style-type: none"> data entry for peri-operative information | "Pan_3c_Endoscopy" "Pan_5_Treatment" "Pan_6a_Res" |
| "Pan_7_Path" | <ul style="list-style-type: none"> data entry for pathology information | "Pan_7_Path" |
| "Short_3_FollowUpData" | <ul style="list-style-type: none"> data entry for follow-up information | "Pan_8_FU" |

Table 3-7 Forms Used for Patient Data Entry

4 Our Data Mining Work

After the UMass Pancreatic database was populated with all the available patient information from the institutional centralized database, analysis was conducted using the stored data. Both John Hayward and Stuart Floyd conducted experiments for their Master's theses (1) (2) using the partially completed database (further information available in section 2.2). In a nutshell, they concluded that machine learning algorithms can match or even surpass the accuracies of linear and logistic regression techniques that are the statistical techniques traditionally accepted in the medical domain. Building off from this result, the following new experiments were formulated. Part of the experiments was conducted with the help of four medical surgeons from the Department of Surgery of the University of Massachusetts Medical School. Two of the doctors specialize in surgical oncology (pancreatic cancer) meanwhile the other two surgeons were pancreatic cancer research fellows. Thus, there is a variation in the experience level among the medical doctors involved in our research.

Two different patient outcomes were tested. The first one tested was tumor malignancy. The possible values are "TRUE"- the patient presents malignant cancer and "FALSE"- the patient does not present malignancy. The second outcome tested for survival time, defined by the number of months the patient lived after surgery. The possible values determined by the medical doctors were "0 ~ 2 months", "3 ~ 5 months", "6 ~ 8 months", and "9 months or more".

At the time the experiments conducted, there were 252 patients in the database. The data of all 252 patients in the database were used in the malignancy experiments. Since there were only 62 recorded patients in the database that had undergone surgery and have a confirmed date of death⁰, the training dataset for all the survival time experiments consist of only 62 instances. The reason for trimming the dataset was the definition of survival time as the time interval between the date of surgery and date of death. The data used for the survival time class consists of information available up to the surgery and including the surgery information.

The aim of these experiments is to compare the performance of doctors' predictions against data mining algorithms'. In making these comparisons we have looked at the accuracy of the target classes' prediction, which is calculated as an average of correctly classified instances for each of the two classification target.

⁰ Confirmed implies that the death date was either stored in MEDITECH or in the Social Security Death Index.

4.1 Patient Outcome Prediction: Doctors vs. Machine Learning Techniques

The purpose of this experiment is to determine whether doctors can better predict a patient's outcome compared to machine learning techniques. Based on their experience each doctor had to choose a restricted set of patient attributes that would help one in making a clinical prediction. Then, a prediction was made by each doctor by using only the previously selected patient attributes. Overall, we tried to determine the doctors' performance in classifying patient cases, and whether using an automated prediction process alone or along with doctors' experience would improve the patient's case classification.

4.1.1 Methodology

Three sets of experiments were conducted: human expert prediction, hybrid prediction and machine learning prediction. The following procedure holds true for both the malignancy and survival time experiments. As mentioned before, 252 patients were included in the total-dataset for the malignancy experiments, and 62 patients were included in the total-dataset for the survival time experiments. Refer to **Figure 4-1** for a flow chart of the entire series of experiments. **Figure 4-2** displays the entire decision making mapping of the experiment. Also refer to **Table 4-2**.

The human experts' prediction experiment was done with the help of doctors from the Department of Surgery of the University of Massachusetts Medical School. Each doctor was asked to select from a master list of attributes, 20 attributes that they would like to have access to if asked to predict patient outcome, specifically malignancy and survival time. For the malignancy experiments, only pre-operative attributes (76 attributes) were included in the master list. For the survival time experiments, all attributes except for the follow-up data were included in the master list. The survival time attributes thus summed up to 142 attributes. The experiments using the malignancy class is explained later on. The experiments for the survival time class have the same flow except for the number of patients included. The patients used for the malignancy test-train sets are all the patients from the database (252 patients), meanwhile for the survival time test-train sets we used only the 62 patients that underwent surgery. After the doctors made their attribute selection, four attribute lists were created for the malignancy class, attribute list A, B, C, and D. We then created a test set made up of 10 patients and a training set made up of the remainder of the patients. The training set will be used in the future experiments. We repeated these steps four times to form four test and training set pairs A, B, C, and D. The four test sets are disjoint. One test set was assigned to each of the four doctors. For each doctor, the assigned test set was reduced to contain only the attributes from his/her corresponding attribute list. For example, follow the first branch on the left from

Figure 4-1. Doctor A generates attribute list A and also is assigned test set A. These two form an attribute list and test set pair. From this test set, attributes not included in attribute list A are eliminated, resulting in a test set of 10 patients with 20 attributes that the doctor requested. Then doctor A made predictions of the patient outcome, given test set A. This procedure was applied to all attribute lists and test set pairs. The doctors' predictions were compared against the actual values stored in the database to see how accurate they were.

The hybrid experiment was conducted on the four previously created training sets. Each training set is reduced accordingly by the corresponding attribute list. That is, training set A was reduced to contain only the attributes selected by doctor A. Bayesian networks and logistic regression were used to create a model using the training set and this model was tested on the corresponding test set created during the human expert's prediction experiment. This procedure is shown by the horizontal lines diverging from the main vertical flow line in Figure 4-1. For example, an arrow connects test set A with Bayesian network and logistic regression. This flow path represents the hybrid experiment.

The machine learning prediction experiment followed a similar approach. Our automatic attribute selection algorithm, Relief-F, chose 20 attributes using each of the four training sets, and created four attribute lists R-A, R-B, R-C, and R-D. Each R-X attribute list was used to reduce the corresponding training and test set pair. Finally Bayesian networks and logistic regression techniques were each applied on each of the four training sets to create models which were then tested on their paired test set.

It should be noted that no two doctors saw the same patient when asked to predict patient outcome. Therefore, each set of experiments (human expert prediction, hybrid prediction, and machine learning prediction) were repeated on four different sets of 10 patients. All experiments involving machine learning algorithms were conducted using WEKA 3.5.7. The parameter settings for the different machine learning algorithms are displayed in **Table 4-2** (see **Appendix A: WEKA Parameters** for explanation of these parameters).

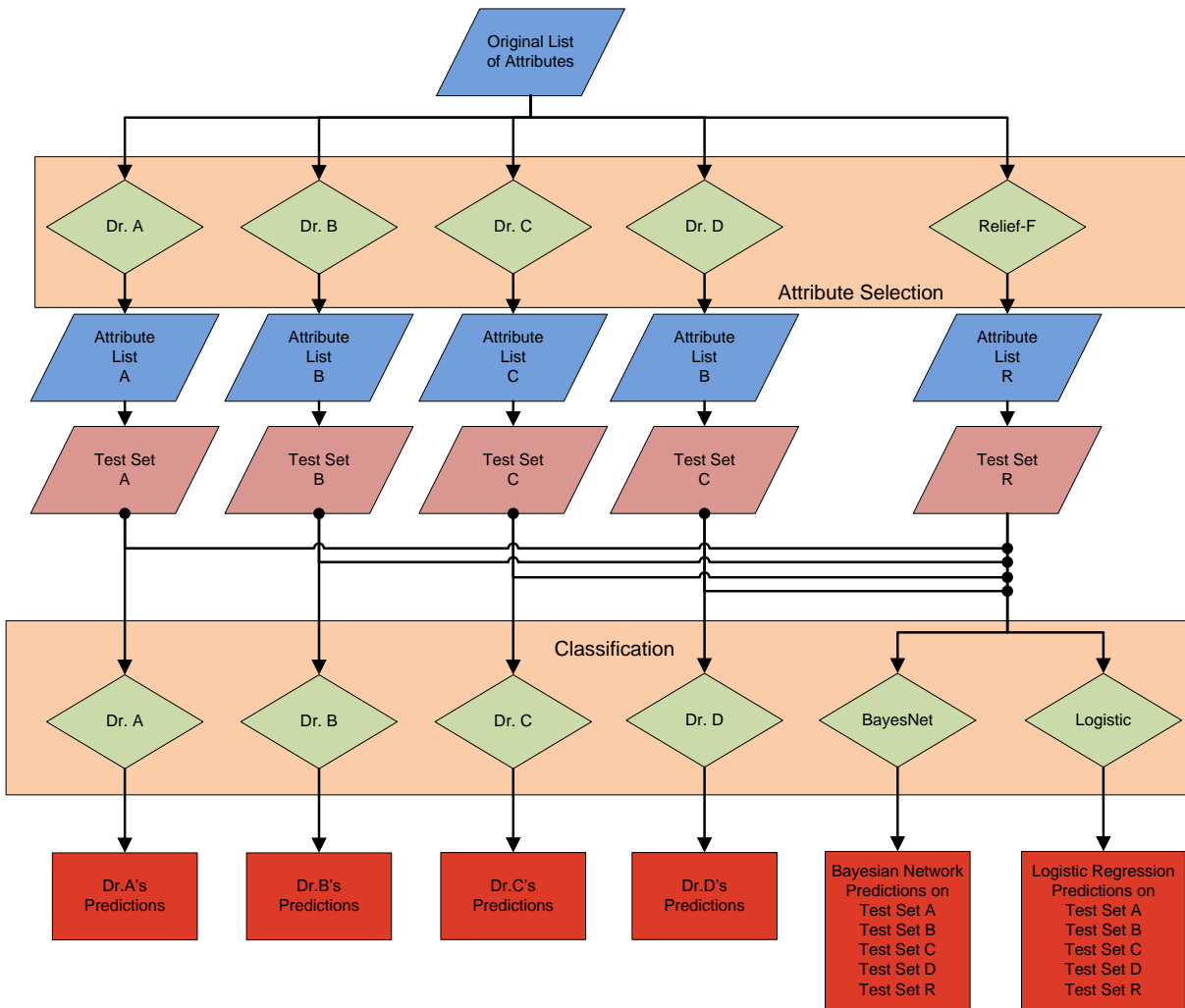


Figure 4-1 Flow Chart

| Prediction Type | Target Class | Attribute Selection | Classification Method |
|-----------------|---------------|---------------------|-----------------------|
| Human | Malignancy | Doctors A~D | Doctors A~D |
| Hybrid | Malignancy | Doctors A~D | Bayesian Network |
| Hybrid | Malignancy | Doctors A~D | Logistic Regression |
| Machine | Malignancy | Relief-F | Bayesian Network |
| Machine | Malignancy | Relief-F | Logistic Regression |
| Human | Survival Time | Doctors A~D | Doctors A~D |
| Hybrid | Survival Time | Doctors A~D | Bayesian Network |
| Hybrid | Survival Time | Doctors A~D | Logistic Regression |
| Machine | Survival Time | Relief-F | Bayesian Network |
| Machine | Survival Time | Relief-F | Logistic Regression |

Table 4-1 Summary of the different attribute selection and outcome classification methods combination

Figure 4-2 represents another perspective in understanding the procedure of this experiment. The total data set with n instances (where n equals 252 for the malignancy experiment and 62 for the survival time experiment), is split into two sets: a training set and a test set. The test set contains 10 patients while the training set contains $(n-10)$ patients. This separation of the total dataset is made four times, creating four pairs of training and test set. These are labeled training and test set A, B, C, and D. To simplify the rest of the explanation, we will assume that we are working with the set pair X. Relief-F selects 20 attributes after looking through training set X, and the training data set is reduced (R-Reduced Training Set X) so that it only includes the 20 previously chosen attributes by Relief-F. The corresponding test set X is also accordingly reduced (R-Reduced Test Set X) so that it only has the same attributes with the training. Using this test- training pair, Bayesian networks and logistic regression machine learning techniques each makes its predictions.

For the hybrid experiment, the original training and test set pair X is reduced, not by the attribute list created by Relief-F, but by the attribute list created by Doctor X. Therefore, the resulting set pair is X-Reduced Training Set X, and X-Reduced Test Set X. The outcome predictions are made in the same way as in the machine learning technique experiment. The human expert predictions are made by giving Doctor X the X-Reduced Test Set X and allowing the doctor to predict the patient outcome using his/her medical experience.

This whole experiment is repeated for training and test set pairs A, B, C, and D, and also for the two target classes: malignancy and survival time. Another thing to note to avoid confusion is that the attribute lists A, B, C, and D are only associated with their corresponding set pair. This means that the list generated by doctor A is uniquely used to reduce the training and test set pair A for the hybrid and human expert experiments. This is not the case for machine learning experiments because the set pairs are reduced according to the list created by Relief-F. Since there exist set pairs A~D, the machine learning experiment had to be conducted on each of the set pairs.

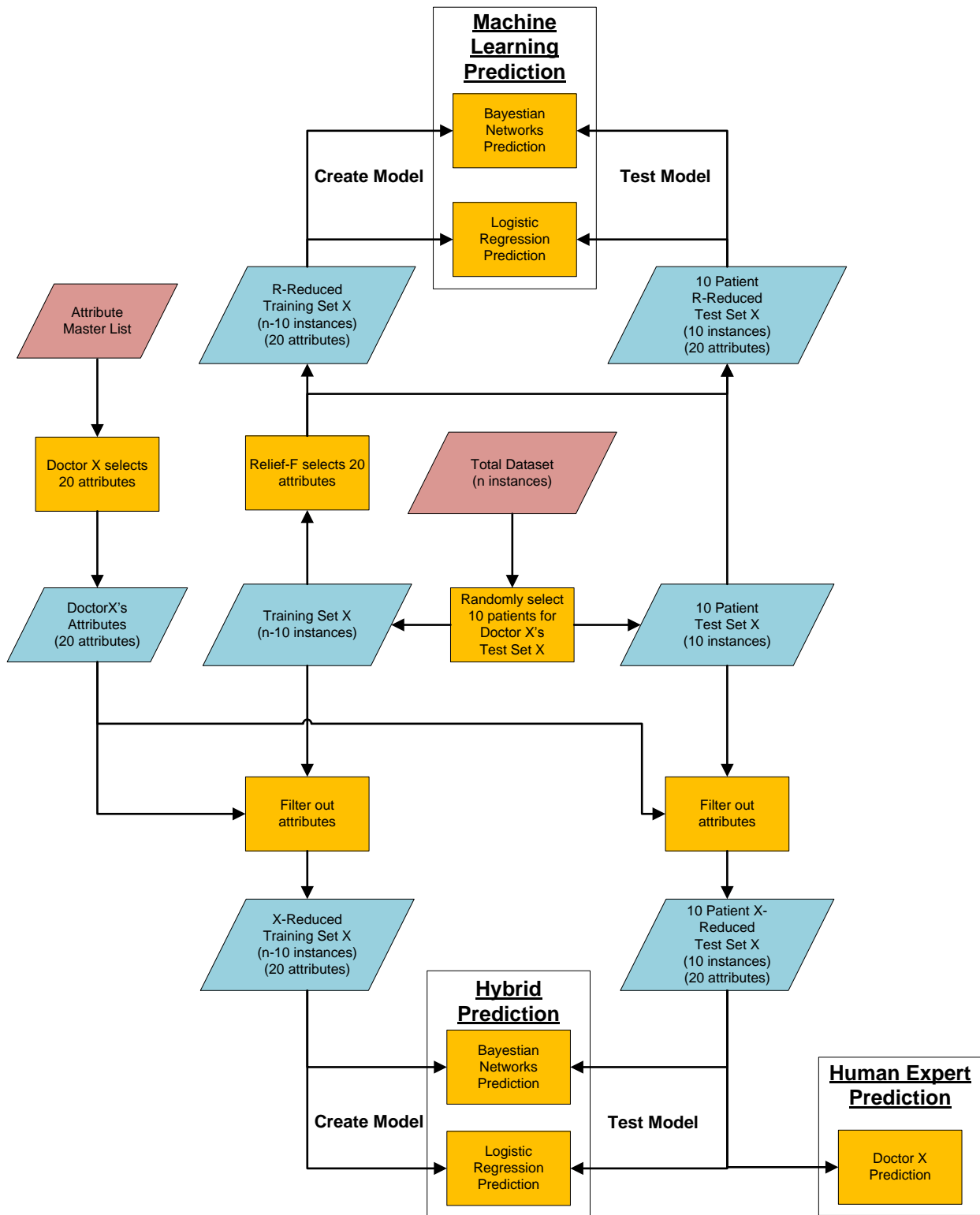


Figure 4-2 Experimental Map

| | |
|--|------------------------|
| <u>ReliefAttributeEval</u> | |
| numNeighbours | 10 |
| sampleSize | -1 |
| seed | 1 |
| sigma | 2 |
| weightByDistance | FALSE |
| <u>Ranker(*Required in order to conduct Relief-F)</u> | |
| generateRankinf | TRUE |
| numToSelect | 20 |
| startSet | (blank) |
| threshold | 1.7976931348623157E308 |
| <u>BayesNet</u> | |
| debug | FALSE |
| estimator | |
| alpha | 0.5 |
| <u>searchAlgorithm</u> | |
| initAsNaiveBayes | FALSE |
| markovBlanketClassifier | FALSE |
| maxNrParents | 1~10 |
| randomOrder | FALSE |
| scoreType | BAYES |
| useAdTree | FREE |
| <u>Logistic</u> | |
| debug | FALSE |
| maxIts | -1 |
| ridge | 1.0E-08 |

Table 4-2 WEKA Parameters

4.1.2 Selected Attributes: Malignancy

This section displays the attributes selected by the doctors and by the Relief-F attribute selection algorithm over the training sets for the malignancy class. Data on a total of 252 patients were included in the malignancy experiments. For each doctor a set of ten patients was selected and the corresponding remaining 242 patients were given as input for Relief-F. The highlighted attributes are the ones that are common between what the doctor and Relief-F chose.

| Attributes Selected by Relief-F Over Training Set A | Attributes Selected by Doctor A |
|---|---|
| Sex Presumptive Diagnosis ECOG Initial Symptoms: Weight Loss Initial Symptoms: Weight Loss in lbs Initial Symptoms: Jaundice Initial Symptoms: Abdominal Pain Comorbidity: Pancreatitis Other Major Comorbidity Pre-Operative Laboratory: CEA Pre-Operative Laboratory: CA19-9 Diagnostic Procedure CT: SMV Involvement Diagnostic Procedure CT: Tumor Size X Diagnostic Procedure CT: Tumor Size Y Diagnostic Procedure EUS: Tumor Size X Diagnostic Procedure EUS: Tumor Size Y Diagnostic Procedure EUS: Number of Tumors Therapeutic ERCP: Stent Type Anticipated Operation Diagnostic Procedure: FNA Cytology Number of Attributes in Common = 8 | Initial Symptoms: Weight Loss Initial Symptoms: Jaundice Initial Symptoms: Nausea Initial Symptoms: Vomiting Initial Symptoms: Clay Colored Stool Initial Symptoms: Abdominal Pain Initial Symptoms: Early Satiety Comorbidity: Ethanol(Alcohol) Abuse Family History of Pancreatic Cancer: Relationship to Patient Social History: Cigarettes (significant use) Social History: Cigarette Pack Years Pre-Operative Laboratory: CA19-9 Pre-Operative Laboratory: Bilirubin Diagnostic Procedure CT: Tumor Size Y Diagnostic Procedure EUS: Tumor Size X Diagnostic Procedure EUS: Hepatic Vein Involvement Diagnostic Procedure EUS: Portal Vein Involvement Diagnostic Procedure EUS: No Node Diagnostic Procedure EUS: Tumor Size X Diagnostic Procedure EUS: Tumor Size Y |

Table 4-3 ReliefF and Doctor A's Malignancy Attribute Lists Over Training Set A

| Attributes Selected by Relief-F Over Training Set B | Attributes Selected by Doctor B |
|--|--|
| Sex Presumptive Diagnosis Initial Symptoms: Jaundice Initial Symptoms: Weight Loss Initial Symptoms: Weight Loss in lbs Social History: Cigarettes (significant use) Initial Symptoms: Vomiting Initial Symptoms: Abdominal Pain Anticipated Operation Pre-Operative Laboratory: Albumin Pre-Operative Laboratory: CA19-9 Comorbidity: Diabetes w/ Oral Agents Comorbidity: Pancreatitis Therapeutic ERCP: Stent Type Diagnostic Procedure EUS: Tumor Size X Diagnostic Procedure EUS: Tumor Size Y Diagnostic Procedure EUS: Number of Tumors Diagnostic Procedure: FNA Cytology Diagnostic Procedure CT: Tumor Size X Diagnostic Procedure CT: Tumor Size Y | Initial Symptoms: Weight Loss Initial Symptoms: Jaundice Initial Symptoms: Back Pain Initial Symptoms Early Satiety Family History of Pancreatic Cancer: Relationship to Patient Social History: Cigarette Pack Years Pre-Operative Laboratory: CA19-9 Diagnostic Procedure CT: SMA Involvement Diagnostic Procedure CT: SMV Involvement Diagnostic Procedure CT: Portal Vein Involvement Diagnostic Procedure EUS: Celiac Artery Involvement Diagnostic Procedure EUS: SMA Involvement Diagnostic Procedure EUS: Hepatic Vein Involvement Diagnostic Procedure EUS: SMV Involvement Diagnostic Procedure EUS: Portal Vein Involvement Diagnostic Procedure EUS: Celiac Node Disease Diagnostic Procedure EUS: Peripancreatic Node Disease Diagnostic Procedure EUS: Tumor Size X Diagnostic Procedure EUS: Tumor Size Y Age at Diagnosis |
| Number of Attributes in Common = 5 | |

Table 4-4 ReliefF and Doctor B's Malignancy Attribute Lists Over Training Set B

| Attributes Selected by Relief-F Over Training Set C | Attributes Selected by Doctor C |
|---|--|
| Sex Weight Initial Symptoms: Weight Loss Initial Symptoms: Back Pain Comorbidities: Diabetes w/ Oral Agents Comorbidities: Onset of Diabetes Comorbidity: Malnutrition Comorbidity: Ethanol(Alcohol) Abuse Social History: Cigarettes (significant use) Pre-Operative Laboratory CEA Pre-Operative Laboratory Albumin Pre-Operative Laboratory ALT Pre-Operative Laboratory AST Diagnostic Procedure CT: Tumor Size X Diagnostic Procedure CT: Tumor Size Y Diagnostic Procedure CT: Number of Tumors Diagnostic Procedure EUS: Tumor Size X Diagnostic Procedure EUS: Tumor Size Y Anticipated Operation Age at Diagnosis | Presumptive Diagnosis Initial Symptoms: Weight Loss in lbs Initial Symptoms: Back Pain Initial Symptoms: Dysphagia Initial Symptoms: Malnutrition Pre-Operative Laboratory CA19-9 Diagnostic Procedure CT: Celiac Artery Involvement Diagnostic Procedure CT: SMA Involvement Diagnostic Procedure CT: Portal Vein Involvement Diagnostic Procedure CT: Tumor Size X Diagnostic Procedure CT: Tumor Size Y Diagnostic Procedure CT: Number of Tumors Diagnostic Procedure EUS: Celiac Artery Involvement Diagnostic Procedure EUS: SMA Involvement Diagnostic Procedure EUS: SMV Involvement Diagnostic Procedure EUS: Portal Vein Involvement Diagnostic Procedure EUS: Tumor Size X Diagnostic Procedure EUS: Tumor Size Y Anticipated Operation Age at Diagnosis |
| Number of Attributes in Common = 9 | |

Table 4-5 ReliefF and Doctor C's Malignancy Attribute Lists Over Training Set C

| Attributes Selected by Relief-F Over Training Set D | Attributes Selected by Doctor D |
|--|--|
| Sex Presumptive Diagnosis ECOG Initial Symptoms: Weight Loss Initial Symptoms: Weight Loss in lbs Initial Symptoms: Jaundice Initial Symptoms: Abdominal Pain Comorbidity: Pancreatitis Pre-Operative Laboratory: CEA Pre-Operative Laboratory: CA19-9 Diagnostic Procedure CT: SMV Involvement Diagnostic Procedure CT: Tumor Size X Diagnostic Procedure CT: Tumor Size Y Diagnostic Procedure EUS: Portal Vein Involvement Diagnostic Procedure EUS: Tumor Size X Diagnostic Procedure EUS: Tumor Size Y Diagnostic Procedure EUS: Number of Tumors Therapeutic ERCP: Stent Type Anticipated Operation Diagnostic Procedure: FNA Cytology | ECOG Initial Symptoms: Weight Loss in lbs Initial Symptoms: Jaundice Initial Symptoms: Fatigue Initial Symptoms: Abdominal Pain Initial Symptoms: Back Pain Comorbidity: Onset of Diabetes Pre-Operative Laboratory: CEA Pre-Operative Laboratory: CA19-9 Pre-Operative Laboratory: Albumin Pre-Operative Laboratory Bilirubin Diagnostic Procedure CT: Celiac Artery Involvement Diagnostic Procedure CT: SMA Involvement Diagnostic Procedure CT:SMV Involvement Diagnostic Procedure CT: Portal Vein Involvement Diagnostic Procedure EUS: Celiac Node Disease Diagnostic Procedure EUS: No Node Therapeutic ERCP: Stent Type Anticipated Operation Age at Diagnosis |
| Number of Attributes in Common = 8 attributes | |

Table 4-6 ReliefF and Doctor C's Malignancy Attribute Lists Over Training Set C

4.1.3 Selected Attributes: Survival Time

This section displays the attributes selected by the doctors and by the Relief-F attribute selection algorithm over the training sets for the survival time class. Data on a total of 62 patients were included in the survival time experiments. For each doctor a set of ten patients was selected and the remaining 52 patients were given as input for Relief-F. The highlighted attributes are the ones that are common between what the doctor and Relief-F chose .

| Attributes Selected by Relief-F Over Training Set A | Attributes Selected by Doctor A |
|--|---|
| Initial Symptoms: Back Pain Initial Symptoms: Abdominal Pain Initial Symptoms: Jaundice Other Major Comorbidity Post-Operative: Days in ICU Resection: Number of Drains Post-Operative: Discharge Status Post-Operative: Days Until Regular Diet TNM Staging: N Surgery: Blood Loss in cc Post-Operative: Bleeding Post-Operative: Days Until Drain Removed Surgery: Extubated in OP Post-Operative: Prolonged PO Intolerance Histology TNM Staging: T Post-Operative: Feeding Tube Complications Post-Operative: Pulmonary Complications Pathology: Tumor Size X Post-Operative: Any Reoperation | Initial Symptoms: Weight Loss Initial Symptoms: Jaundice Initial Symptoms: Early Satiety Pre-Operative CA19-9 Diagnostic Procedure CT: Celiac Artery Involvement Diagnostic Procedure CT: Hepatic Vein Involvement Diagnostic Procedure CT: Portal Vein Involvement Diagnostic Procedure CT: Tumor Size X Diagnostic Procedure CT: Tumor Size Y Diagnostic Procedure EUS: No Node Diagnostic Procedure EUS: Tumor Size X Diagnostic Procedure: FNA Histology Neoadjuvant Treatment: Radiation Neoadjuvant Treatment: Chemotherapy Surgery Type Neoadjuvant Treatment: Chemotherapy Specify Agent Surgery: Blood Loss in cc Surgery: Transfusion PRBC Units Post-Operative: ICU Re-Admission Post-Operative: Leak |
| Number of Attributes in Common = 2 | |

Table 4-7 ReliefF and Doctor A's Survival Time Attribute Lists Over Training Set A

| Attributes Selected by Relief-F Over Training Set B | Attributes Selected by Doctor B |
|---|--|
| Presumptive Diagnosis Initial Symptoms: Jaundice Initial Symptoms: Vomiting Initial Symptoms: Abdominal Pain Initial Symptoms: Weight Loss Post-Operative: Days in ICU Post-Operative: Discharge Status TNM Staging: N Post-Operative: Blood Loss in cc Post-Operative: Number of Drains Histology Post-Operative: Bleeding Post-Operative: Days Until Drain Removed Post-Operative: Extubated in OR Post-Operative: Length of Stay Post-Operative: Days Until Clears Started Post-Operative: Feeding Tube Complications Post-Operative: Days Until Tube Feeds Started Surgery: Transfusion PRBC Units Post-Operative: Days Until Regular Diet | ECOG Comorbidity: Ischemic Heart Disease Comorbidity: Respiratory Comorbidity: Renal Failure Comorbidity: Liver Failure/Cirrhosis Comorbidity: Malnutrition Pre-Operative Laboratory CA19-9 Social History: Cigarettes (significant use) Surgery Type Age at Diagnosis Surgery: Venous Resection Post-Operative: Days in ICU Post-Operative: ICU Re-Admission Post-Operative: Cardiac Complications/MI Post-Operative: Leak Post-Operative: Pulmonary Complications Post-Operative: Liver Insufficiency Histology TNM Staging: N TNM Staging: M |
| Number of Attributes in Common = 2 | |

Table 4-8 ReliefF and Doctor B's Survival Time Attribute Lists Over Training Set B

| Attributes Selected by Relief-F Over Training Set C | Attributes Selected by Doctor C |
|---|---|
| Initial Symptoms: Back Pain Initial Symptoms: Abdominal Pain Initial Symptoms: Weight Loss in lbs Initial Symptoms: Vomiting Other Major Comorbidity Neoadjuvant Treatment: Chemotherapy Post-Operative: Bleeding Post-Operative: Discharge Status Post-Operative: Days in ICU Post-Operative: Any Transfusion Surgery: Blood Loss in cc Surgery: Number of Drains Post-Operative Extubated in OP Post-Operative: Days Until Drain Removed Post-Operative: Any Reoperation TNM Staging: R Post-Operative: Length of Stay Pathology: Tumor Size X Histology Post-Operative: Days Until Clears Started | Sex ECOG Initial Symptoms: Weight Loss in lbs Initial Symptoms :Abdominal Pain Comorbidity: Ischemic Heart Disease Comorbidity :Malnutrition Pre-Operative Laboratory: CA19-9 Pre-Operative Laboratory: Albumin Pre-Operative Laboratory: Bilirubin Diagnostic Procedure CT: Portal Vein Involvement Diagnostic Procedure CT: Tumor Size X Diagnostic Procedure EUS: Cytology Surgery Type Age at Diagnosis Surgery: Blood Loss in cc Post-Operative: Days in ICU Post-Operative: Length of Stay Histology TNM Staging: N TNM Staging: R |
| Number of Attributes in Common = 6 | |

Table 4-9 ReliefF and Doctor C's Survival Time Attribute Lists Over Training Set C

| Attributes Selected by Relief-F Over Training Set D | Attributes Selected by Doctor D |
|---|--|
| Initial Symptoms: Back Pain Initial Symptoms: Weight Loss Initial Symptoms: Weight Loss in lbs Initial Symptoms: Abdominal Pain Initial Symptoms: Vomiting Other Major Comorbidity Neoadjuvant Therapy: Radiation Neoadjuvant Therapy: Chemotherapy Post-Operative: Discharge Status Surgery: Number of Drains Post-Operative: Days in ICU Post-Operative: Any Transfusion Post-Operative: Bleeding Surgery: Blood Loss in cc TNM Staging: N Post-Operative: Extubated in OP Post-Operative: Days Until Drain Removed Post-Operative: Any Reoperation Post-Operative: Pulmonary Complications Post-Operative: Prolonged PO Intolerance | Presumptive Diagnosis ECOG Initial Symptoms: Weight Loss in lbs Initial Symptoms: Abdominal Pain Initial Symptoms: Back Pain Comorbidity: Liver Failure/Cirrhosis Pre-Operative Laboratory: CEA Pre-Operative Laboratory: CA19-9 Pre-Operative Laboratory: Albumin Diagnostic Procedure CT: SMA Involvement Diagnostic Procedure: FNA Histology Anticipated Operation Neoadjuvant Therapy: Radiation Surgery Type Surgery: Blood Loss in cc Post-Operative: Any Reoperation Post-Operative: Leak Post-Operative: Liver Insufficiency Histology TNM Staging: M |
| Number of Attributes in Common = 5 | |

Table 4-10 ReliefF and Doctor D's Survival Time Attribute Lists Over Training Set D

4.1.4 Results: Malignancy

The classification accuracy results for the various experiments done with malignancy as the target attributes are summarized in this section. When doctors made their classification using the data from the 20 attributes they chose at the beginning of the experiment, they averaged 75% accuracy. When the computer was used to make classification using the same information, the best average using Bayesian networks was 73%, with maximum of 1 parent, and 75% using logistic regression. For the machine experiment, the accuracy was 85% using Bayesian network with 2 maximum parents and 85% using logistic regression.

4.1.4.1 Human Expert Prediction Results

The table below shows the results for when doctors selected the attribute and made the predictions.

| | | |
|-----------------|------------|----|
| Doctor A | Test Set A | 80 |
| Doctor B | Test Set B | 70 |
| Doctor C | Test Set C | 70 |
| Doctor D | Test Set D | 80 |
| | Average | 75 |

Table 4-11 Human Experts' Prediction Results (Values in % accuracy)

4.1.4.2 Hybrid Prediction Results

The following is the result when machine learning algorithms made the predictions using the attribute list the doctors created. The number of parents represents the maximum number of parents allowed per node while constructing the Bayesian network.

| Bayesian Network | Number of Parents | | | | | | | | | |
|-------------------|-------------------|----|------|------|----|------|----|----|----|----|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| Test Set A | 50 | 50 | 60 | 60 | 70 | 60 | 70 | 70 | 70 | 70 |
| Test Set B | 80 | 70 | 70 | 70 | 70 | 70 | 70 | 70 | 70 | 70 |
| Test Set C | 60 | 60 | 60 | 60 | 60 | 60 | 60 | 60 | 60 | 60 |
| Test Set D | 100 | 80 | 80 | 80 | 80 | 80 | 80 | 80 | 80 | 80 |
| Average | 72.5 | 65 | 67.5 | 67.5 | 70 | 67.5 | 70 | 70 | 70 | 70 |

Table 4-12 Hybrid Prediction Results: Bayesian Network (Values in % accuracy)

| Logistic Regression | |
|---------------------|----|
| Test Set A | 80 |
| Test Set B | 70 |
| Test Set C | 80 |
| Test Set D | 70 |
| Average | 75 |

Table 4-13 Hybrid Prediction Results: Logistic Regression (Values in % accuracy)

The following is the result when the entire process of attribute selection and classification is automated by machine learning algorithms.

4.1.4.3 Machine Learning Prediction Results

| Bayesian Network | Number of Parents | | | | | | | | | |
|------------------|-------------------|----|------|------|------|------|------|------|------|------|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| Test Set A | 80 | 80 | 80 | 80 | 80 | 80 | 80 | 80 | 80 | 80 |
| Test Set B | 90 | 90 | 90 | 90 | 90 | 90 | 90 | 90 | 90 | 90 |
| Test Set C | 70 | 80 | 80 | 80 | 80 | 80 | 80 | 80 | 80 | 80 |
| Test Set D | 80 | 90 | 80 | 80 | 80 | 80 | 80 | 80 | 80 | 80 |
| Average | 80 | 85 | 82.5 | 82.5 | 82.5 | 82.5 | 82.5 | 82.5 | 82.5 | 82.5 |

Table 4-14 Machine Learning Prediction Results: Bayesian Network (Values in % accuracy)

| Logistic Regression | |
|---------------------|-----|
| Test Set A | 70 |
| Test Set B | 100 |
| Test Set C | 80 |
| Test Set D | 90 |
| Average | 85 |

Table 4-15 Machine Learning Prediction Results: Logistic Regression (Values in % accuracy)

4.1.5 Results: Survival Time

The classification accuracy results for the various experiments done with survival time as the target classes are summarized in this section. When doctors made their classification using the data from the 20 attributes they chose at the beginning of the experiment, they averaged 35% accuracy. When the computer was used to make classification using the same information, the best average using Bayesian networks was 55%, with no dependencies on the number of maximum parents, and 35% using logistic regression. When the machine selected the attributes using Relief-F and made the predictions, the accuracy was 78% using Bayesian network with maximum parents greater than 2, and 73% using logistic regression.

4.1.5.1 Human Expert Prediction Results

The table below shows the results of when doctors selected the attribute and made classifications.

| | | |
|-----------------|------------|----|
| Doctor A | Test Set A | 20 |
| Doctor B | Test Set B | 10 |
| Doctor C | Test Set C | 60 |
| Doctor D | Test Set D | 50 |
| | Average | 35 |

Table 4-16 Human Experts' Prediction Results (Values in % accuracy)

4.1.5.2 Hybrid Prediction Results

The following is the result when machine learning algorithms made classifications using the attribute lists the doctors created.

| Bayesian Network | Number of Parents | | | | | | | | | |
|-------------------|-------------------|----|----|----|----|----|----|----|----|----|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| Test Set A | 60 | 60 | 60 | 60 | 60 | 60 | 60 | 60 | 60 | 60 |
| Test Set B | 60 | 60 | 60 | 60 | 60 | 60 | 60 | 60 | 60 | 60 |
| Test Set C | 40 | 40 | 40 | 40 | 40 | 40 | 40 | 40 | 40 | 40 |
| Test Set D | 60 | 60 | 60 | 60 | 60 | 60 | 60 | 60 | 60 | 60 |
| Average | 55 | 55 | 55 | 55 | 55 | 55 | 55 | 55 | 55 | 55 |

Table 4-17 Hybrid Prediction Results: Bayesian Network (Values in % accuracy)

| Logistic Regression | |
|---------------------|------|
| Test Set A | 40 |
| Test Set B | 20 |
| Test Set C | 40 |
| Test Set D | 50 |
| Average | 37.5 |

Table 4-18 Hybrid Prediction Results: Logistic Regression (Values in % accuracy)

4.1.5.3 Machine Learning Prediction Results

The following is the result when the entire process of attribute selection and classification is automated by machine learning algorithms.

| Bayesian Network | Number of Parents | | | | | | | | | |
|------------------|-------------------|----|------|------|------|------|------|------|------|------|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| Test Set A | 80 | 80 | 90 | 90 | 90 | 90 | 90 | 90 | 90 | 90 |
| Test Set B | 60 | 70 | 70 | 70 | 70 | 70 | 70 | 70 | 70 | 70 |
| Test Set C | 60 | 60 | 60 | 60 | 60 | 60 | 60 | 60 | 60 | 60 |
| Test Set D | 90 | 90 | 90 | 90 | 90 | 90 | 90 | 90 | 90 | 90 |
| Average | 72.5 | 75 | 77.5 | 77.5 | 77.5 | 77.5 | 77.5 | 77.5 | 77.5 | 77.5 |

Table 4-19 Machine Learning Prediction Results: Bayesian Network (Values in % accuracy)

| Logistic Regression | |
|---------------------|------|
| Test Set A | 80 |
| Test Set B | 70 |
| Test Set C | 60 |
| Test Set D | 80 |
| Average | 72.5 |

Table 4-20 Machine Learning Prediction Results: Logistic Regression (Values in % accuracy)

4.1.6 Discussion

The experiment using our newly updated database brought out some interesting points. The machine learning algorithms provided accuracies that met or surpassed the doctors' performance. The best accuracy was obtained when Bayesian networks were used to make the predictions using attributes selected by Relief-F for both malignancy and survival time experiments.

| Prediction Type | Target Attribute | Attribute Selection | Classification Method | Accuracy |
|-----------------|------------------|---------------------|-----------------------|----------|
| Human | Malignancy | Doctors A~D | Doctors A~D | 75% |
| Hybrid | Malignancy | Doctors A~D | Bayesian Network | 69% |
| Hybrid | Malignancy | Doctors A~D | Logistic Regression | 75% |
| Machine | Malignancy | Relief-F | Bayesian Network | 82.5% |
| Machine | Malignancy | Relief-F | Logistic Regression | 85% |
| Human | Survival Time | Doctors A~D | Doctors A~D | 35% |
| Hybrid | Survival Time | Doctors A~D | Bayesian Network | 55% |
| Hybrid | Survival Time | Doctors A~D | Logistic Regression | 37.5% |
| Machine | Survival Time | Relief-F | Bayesian Network | 76.75% |
| Machine | Survival Time | Relief-F | Logistic Regression | 72.5% |

Table 4-21 Accuracy Summary

It is also interesting to note the varying degrees of correlation between the attribute lists for each target class. When selecting the 20 attributes from the malignancy experiments, there were 7 attributes that the Relief-F repeatedly chose and there were 2 attributes that all the doctors chose (see **Table 4-22**). For the survival time experiments, Relief-F chose 8 attributes repeatedly, and 2 attributes were chosen by all doctors (see **Table 4-23**). 35% ~ 40% overlap of Relief-F selected attributes suggests that this machine learning method for attribute selection produces relatively consistent results. Combined with the fact that the Bayesian networks algorithm predicted patient outcomes with high accuracy especially after Relief-F attribute selection, machine learning methods may be very useful tools in assisting doctors in predicting patient outcomes. Although still controversial, machine learning algorithms have advanced to the point where their predictions may be a good basis for determining clinical decision making.

| Relief-F | Doctors |
|---|---|
| Initial Symptoms: Weight Loss | Initial Symptoms: Weight Loss in lbs |
| Diagnostic Procedure CT: Tumor Size X | Pre-Operative Laboratory: CA19-9 |
| Diagnostic Procedure CT: Tumor Size Y | |
| Diagnostic Procedure EUS: Tumor Size X | |
| Diagnostic Procedure EUS: Tumor Size Y | |
| Anticipated Operation | |
| Sex | |
| Number of Attributes in Common = 7 | Number of Attributes in Common = 2 |

Table 4-22 Attributes in Common for Malignancy

| Relief-F | Doctors |
|---|---|
| Initial Symptoms: Abdominal Pain | Pre-Operative Laboratory: CA19-9 |
| Post-Operative: Days in ICU | Surgery |
| Post-Operative: Discharge Status | |
| Surgery: Number of Drains | |
| Post-Operative: Days Until Drain Removed | |
| Post-Operative: Bleeding | |
| Surgery: Blood Loss in cc | |
| Post-Operative: Extubated in Operating Room | |
| Number of Attributes in Common = 8 | Number of Attributes in Common = 2 |

Table 4-23 Attributes in Common for Survival Time

A breakdown of the actual value and predicted value for each doctor is shown in **Table 4-24** and **Table 4-25**. A better experiment may have been to distribute the target class values equally when choosing the 10 patients for the doctors to look at, so that the results do not become biased to a specific class value. The experiments in section 4.3 try to reduce this bias in the hybrid and machine learning experiments.

| Class | Doctor A | Doctor B | Doctor C | Doctor D |
|--------------|----------|----------|----------|----------|
| FALSE | 2 | 4 | 4 | 2 |
| TRUE | 8 | 6 | 6 | 8 |

Table 4-24 Frequency of Patients per Class Value. Malignancy

| Class | Doctor A | Doctor B | Doctor C | Doctor D |
|------------|----------|----------|----------|----------|
| < 2 months | 3 | 2 | 0 | 0 |
| 3~5 months | 0 | 0 | 5 | 5 |
| 6~8 months | 1 | 2 | 1 | 1 |
| >9 months | 6 | 6 | 4 | 4 |

Table 4-25 Frequency of Patients per Class Value. Survival Time (Values in number of patients)

| Malignancy Doctor's Prediction | | | | |
|--------------------------------|--------------------|---------------|----------------------|----------------|
| Prediction | Correct Prediction | | Incorrect Prediction | |
| | "Yes" | "No" | "Yes" | "No" |
| | True Positive | True Negative | False Positive | False Negative |
| Doctor A | 6 | 2 | 2 | 0 |
| Doctor B | 4 | 3 | 2 | 1 |
| Doctor C | 3 | 4 | 3 | 0 |
| Doctor D | 6 | 2 | 2 | 0 |
| Total | 19 | 11 | 9 | 1 |

Table 4-26 Malignancy Confusion Matrix (Values in number of patients)

| Survival Time Doctor's Prediction | | | | | | | | |
|-----------------------------------|---------------|------------|------------|------------|-----------------|------------|------------|------------|
| Prediction | Correct Value | | | | Incorrect Value | | | |
| | < 3 months | 3-5 months | 6-8 months | > 9 months | < 3 months | 3-5 months | 6-8 months | > 9 months |
| Doctor A | 0 | 0 | 0 | 2 | 2 | 1 | 1 | 4 |
| Doctor B | 0 | 0 | 1 | 5 | 2 | 0 | 1 | 1 |
| Doctor C | 0 | 0 | 0 | 1 | 0 | 5 | 1 | 3 |
| Doctor D | 0 | 0 | 0 | 5 | 0 | 0 | 2 | 3 |
| Total | 0 | 0 | 1 | 13 | 4 | 6 | 5 | 11 |

Table 4-27 Survival Time Confusion Matrix (Values in number of patients)

4.2 Prediction Accuracy Based on Frequency of Selected Attributes

After reviewing the results of experiment 1 and looking at the data, several more questions surfaced. As it can be seen from the list of attributes chosen by the doctors and by Relief-F for the different test sets (sections 4.1.2 and 4.1.3), there are several attributes that all doctors chose and several attributes that Relief-F repeatedly chose for all test sets. Experiment 2 expounds on this realization. This experiment attempts to determine how significant the attributes are in predicting patient outcome, depending on their frequency of appearance in the doctors' list and the Relief-F's list.

4.2.1 Methodology

First attributes were organized in terms of the frequency they appeared in the attribute lists. The attributes that the doctors' chose and Relief-F chose were processed separately. Then using each of these newly formed list of attributes, Bayesian networks and logistic regression classifiers were each used to create a decision making model. The experiment with Bayesian networks was done with 3 maximum number of parents, since according to the results from experiment 1, there is not too much accuracy to be gained by increasing maximum parent numbers in general. The accuracies resulting from each list were compared for malignancy and for survival time target classes.

4.2.2 Repeated Malignancy Attributes

The following lists are for the malignancy target class.

| Relief-F | Doctors |
|---|----------------------------------|
| Sex Initial Symptoms: Weight Loss Diagnostic Procedure CT: Tumor Size X Diagnostic Procedure CT: Tumor Size Y Diagnostic Procedure EUS: Tumor Size X Diagnostic Procedure EUS: Tumor Size Y Anticipated Operation | Pre-Operative Laboratory: CA19-9 |

Table 4-28 Attributes occurring in 4 out of 4 Attribute Lists

| Relief-F | Doctors |
|--|---|
| Presumptive Diagnosis Initial Symptoms: Weight Loss in lbs Initial Symptoms: Jaundice Initial Symptoms: Abdominal Pain Initial Symptoms: Pancreatitis Pre-Operative Laboratory: CEA Pre-Operative Laboratory: CA19-9 Diagnostic Procedure EUS: Number of Tumors Therapeutic ERCP: Stent Type Diagnostic Procedure: FNA Cytology | Initial Symptoms: Jaundice Initial Symptoms: Back Pain Diagnostic Procedure EUS: Tumor Size X Diagnostic Procedure EUS: Tumor Size Y Diagnostic Procedure CT: SMA Involvement Diagnostic Procedure CT: Portal Vein Involvement Age at Diagnosis |

Table 4-29 Attributes occurring in 3 out of 4 Attribute Lists

| Relief-F | Doctors |
|---|--|
| ECOG Social History: Cigarettes (significant use) Pre-Operative Laboratory: Albumin Comorbidity: Diabetes w/ Oral Agents Diagnostic Procedure CT: SMV Involvement | Initial Symptoms: Weight Loss Initial Symptoms: Weight Loss in lbs Initial Symptoms: Abdominal Pain Family History of Pancreatic Cancer: Relationship to Patient Social History: Cigarette Packs Year Pre-Operative Laboratory: Bilirubin Diagnostic Procedure CT: Tumor Size X Diagnostic Procedure CT: Tumor Size Y Diagnostic Procedure CT: SMV Involvement Diagnostic Procedure CT: Celiac Artery Involvement Diagnostic Procedure EUS: Hepatic Vein Involvement Diagnostic Procedure EUS: No Node Diagnostic Procedure EUS: Celiac Artery Involvement Diagnostic Procedure EUS: SMA Involvement Diagnostic Procedure EUS: SMV Involvement Diagnostic Procedure EUS: Celiac Node Disease Anticipated Operation |

Table 4-30 Attributes occurring in 2 out of 4 Attribute Lists

| Relief-F | Doctors |
|--|---|
| Weight Initial Symptoms: Vomiting Initial Symptoms: Back Pain Other Symptoms Comorbidity: Onset of Diabetes Comorbidity: Malnutrition Comorbidity: Ethanol(Alcohol) Abuse Pre-Operative Laboratory: ALT Pre-Operative Laboratory: AST Diagnostic Procedure CT: Number of Tumors Diagnostic Procedure EUS: Portal Vein Involvement Age at Diagnosis | Presumptive Diagnosis ECOG Initial Symptoms: Nausea Initial Symptoms: Vomiting Initial Symptoms: Clay Colored Stool Initial Symptoms: Fatigue Initial Symptoms: Dysphagia Social History: Cigarettes (significant use) Comorbidity: Ethanol(Alcohol) Abuse Comorbidity: Malnutrition Comorbidity: Onset of Diabetes Pre-Operative Laboratory: CEA Pre-Operative Laboratory: Albumin Diagnostic Procedure EUS: Peripancreatic Node Disease Diagnostic Procedure CT: Number of Tumors Therapeutic ERCP: Stent Type |

Table 4-31 Attributes occurring in 1 out of 4 Attribute Lists

4.2.3 Repeated Survival Time Attributes

The following lists are for the survival time target class.

| Relief-F | Doctors |
|--|--|
| Initial Symptoms: Abdominal Pain Post-Operative: Days in ICU Surgery: Number of Drains Post-Operative: Discharge Status Surgery: Blood Loss in cc Post-Operative: Bleeding Post-Operative: Days Until Drain Removed Post-Operative: Extubated in OR | Pre-Operative Laboratory: CA19-9 Surgery Type |

Table 4-32 Attributes occurring in 4 out of 4 Attribute Lists

| Relief-F | Doctors |
|--|--|
| Initial Symptoms: Back Pain Other Major Comorbidity TNM Staging: N Histology Post-Operative: Any Reoperation Initial Symptoms: Vomiting | ECOG Surgery: Blood Loss in cc Histology Post-Operative: Leak |

Table 4-33 Attributes occurring in 3 out of 4 Attribute Lists

| Relief-F | Doctors |
|--|--|
| Initial Symptoms: Jaundice Post-Operative: Days Until Regular Diet Post-Operative: Prolonged Intolerance Post-Operative: Feeding Tube Complications Post-Operative: Pulmonary Complications Pathology: Tumor Size X Initial Symptoms: Weight Loss Post-Operative: Length of Stay Post-Operative: Days Until Clears Started Initial Symptoms: Weight Loss in lbs Neoadjuvant Therapy: Chemotherapy Post-Operative: Any Transfusion | ECOG Surgery: Blood Loss in cc Histology Post-Operative: Leak |

Table 4-34 Attributes occurring in 2 out of 4 Attribute Lists

| Relief-F | Doctors |
|--|--|
| TNM Staging: T Presumptive Diagnosis Post-Operative: Days Until Tube Feed Started Surgery: Transfusion PRBC Units TNM Staging: R Neoadjuvant Therapy: Radiation | Presumptive Diagnosis Sex Initial Symptoms: Weight Loss Initial Symptoms: Jaundice Initial Symptoms: Early Satiety Social History: Cigarettes (significant use) Initial Symptoms: Back Pain Comorbidity: Respiratory Comorbidity: Renal Failure Pre-Operative Laboratory: CEA Pre-Operative Laboratory: Bilirubin Anticipated Operation Diagnostic Procedure CT: SMA Involvement Diagnostic Procedure CT: Celiac Artery Involvement Diagnostic Procedure CT: Hepatic Vein Involvement Diagnostic Procedure CT: Tumor Size Y Diagnostic Procedure EUS: No Node Diagnostic Procedure EUS: Cytology Diagnostic Procedure EUS: Tumor Size X Neoadjuvant Therapy: Chemotherapy Neoadjuvant Therapy: Chemotherapy Agent Surgery: Venous Resection Surgery: Transfusion PRBC Units Post-Operative: Cardiac Complication/MI Post-Operative: Pulmonary Complications Post-Operative: Length of Stay TNM Staging: R Post-Operative: Any Reoperation |

Table 4-35 Attributes occurring in 1 out of 4 Attribute Lists

4.2.4 Results

The following tables summarize the results of experiment 2.

| Doctor Selected Attributes | | | |
|----------------------------|------------------|----------|----------|
| Attribute Frequency | Test Set/ Doctor | Accuracy | |
| | | BayesNet | Logistic |
| Frequency 1 | Test Set A | 90 | 80 |
| | Test Set B | 90 | 90 |
| | Test Set C | 60 | 70 |
| | Test Set D | 80 | 90 |
| | Average | 80 | 82.5 |
| Frequency 2 | Test Set A | 70 | 70 |
| | Test Set B | 80 | 80 |
| | Test Set C | 70 | 70 |
| | Test Set D | 60 | 70 |
| | Average | 70 | 72.5 |
| Frequency 3 | Test Set A | 70 | 60 |
| | Test Set B | 70 | 60 |
| | Test Set C | 70 | 70 |
| | Test Set D | 80 | 70 |
| | Average | 72.5 | 65 |
| Frequency 4 | Test Set A | 90 | 80 |
| | Test Set B | 70 | 50 |
| | Test Set C | 60 | 60 |
| | Test Set D | 70 | 80 |
| | Average | 72.5 | 67.5 |

Table 4-36 Repeated Doctor Selected Attributes: Malignancy (Values in % accuracy)

| Relief-F Selected Attributes | | | |
|------------------------------|------------------|----------|----------|
| Attribute Frequency | Test Set/ Doctor | Accuracy | |
| | | BayesNet | Logistic |
| <u>Frequency 1</u> | Test Set A | 70 | 80 |
| | Test Set B | 60 | 80 |
| | Test Set C | 70 | 80 |
| | Test Set D | 80 | 80 |
| | Average | 70 | 80 |
| <u>Frequency 2</u> | Test Set A | 50 | 80 |
| | Test Set B | 60 | 90 |
| | Test Set C | 70 | 80 |
| | Test Set D | 70 | 80 |
| | Average | 62.5 | 82.5 |
| <u>Frequency 3</u> | Test Set A | 90 | 90 |
| | Test Set B | 100 | 90 |
| | Test Set C | 90 | 90 |
| | Test Set D | 90 | 90 |
| | Average | 92.5 | 90 |
| <u>Frequency 4</u> | Test Set A | 80 | 80 |
| | Test Set B | 80 | 80 |
| | Test Set C | 100 | 90 |
| | Test Set D | 100 | 80 |
| | Average | 90 | 82.5 |

Table 4-37 Repeated Relief-F Selected Attributes: Malignancy (Values in % accuracy)

| Doctor Selected Attributes | | | |
|----------------------------|------------------|----------|----------|
| Attribute Frequency | Test Set/ Doctor | Accuracy | |
| | | BayesNet | Logistic |
| Frequency 1 | Test Set A | 60 | 40 |
| | Test Set B | 60 | 30 |
| | Test Set C | 70 | 50 |
| | Test Set D | 90 | 50 |
| | Average | 70 | 42.5 |
| Frequency 2 | Test Set A | 60 | 30 |
| | Test Set B | 60 | 50 |
| | Test Set C | 70 | 30 |
| | Test Set D | 90 | 50 |
| | Average | 70 | 40 |
| Frequency 3 | Test Set A | 60 | 60 |
| | Test Set B | 60 | 50 |
| | Test Set C | 70 | 60 |
| | Test Set D | 90 | 90 |
| | Average | 70 | 65 |
| Frequency 4 | Test Set A | 60 | 60 |
| | Test Set B | 60 | 50 |
| | Test Set C | 70 | 70 |
| | Test Set D | 90 | 90 |
| | Average | 70 | 67.5 |

Table 4-38 Repeated Doctor Selected Attributes: Survival Time (Values in % accuracy)

| Relief-F Selected Attributes | | | |
|------------------------------|------------------|----------|----------|
| Attribute Frequency | Test Set/ Doctor | Accuracy | |
| | | BayesNet | Logistic |
| Frequency 1 | Test Set A | 60 | 60 |
| | Test Set B | 60 | 50 |
| | Test Set C | 70 | 50 |
| | Test Set D | 90 | 60 |
| | Average | 70 | 55 |
| Frequency 2 | Test Set A | 40 | 50 |
| | Test Set B | 60 | 50 |
| | Test Set C | 60 | 50 |
| | Test Set D | 90 | 50 |
| | Average | 62.5 | 50 |
| Frequency 3 | Test Set A | 60 | 50 |
| | Test Set B | 60 | 60 |
| | Test Set C | 70 | 30 |
| | Test Set D | 80 | 70 |
| | Average | 67.5 | 52.5 |
| Frequency 4 | Test Set A | 60 | 30 |
| | Test Set B | 70 | 80 |
| | Test Set C | 60 | 50 |
| | Test Set D | 100 | 90 |
| | Average | 72.5 | 62.5 |

Table 4-39 Repeated Relief-F Selected Attributes: Survival Time (Values in % accuracy)

4.2.5 Discussion

We would expect to see higher accuracy when more attributes are included in the training data set. However, looking at the Malignancy experiment result for attributes commonly chosen by the four doctors, an interesting observation can be made. There was only one attribute that all the doctors chose (Lab CA19-9), but when the machine learning algorithm made a model using solely that attribute, the accuracy was comparable to the accuracy the model gave when using the attributes that appeared with a frequency of 2 and 1. The best accuracy when using the Relief-F lists is achieved when the predictions were made with the attributes that appeared with a frequency of 3 (see **Table 4-36**). Interestingly, this is the list that included the Lab CA19-9 attribute. This may be an indication that this particular attribute, Lab CA19-9 is an important attribute when classifying patient outcome for malignancy.

The results of the survival time experiment were not as widely distributed as the malignancy experiment. One thing that can be noted is that even though there were only two attributes that were chosen by all doctors, these attributes were enough to predict the survival time with a better accuracy compared to the other frequency lists (see **Table 4-37**).

4.3 Reducing Test Set Selection Biases

Many interesting observations were made throughout the two already mentioned experiments. However, the selection of test sets A~D has introduced bias into the previous experiments. This experiment will try to reduce this bias by randomly selecting test sets using stratified sampling. Stratified sampling is a sampling technique that preserves the distribution of the target attribute. That is each random sample produced has approximately the same distribution of the target attribute (in our case either malignancy or survival time) as that of the full dataset. By using the same prediction algorithms but changing the attribute selector, the effectiveness of the attribute selector can be assessed. It should be noted that only the hybrid and the machine learning experiments are repeated with these unbiased test sets. The method used to calculate accuracy in this series of experiments is cross-validation. This method allows for a different random test set of roughly 10 patients to be selected several times (25 times for malignancy as the malignancy cohort contains 252 patients that can be divided into approximately 25 disjoint groups of 10 patients, and 6 times for survival time as the survival time cohort contains 62 patients that can be divided into approximately 6 disjoint groups of 10 patients) rather than testing on a single set of 10 patients. This in effect diminishes the variance in the results due to the choice of the test set that appeared in our very first experiment. Another advantage of cross-validation is that this method allows every patient to be part of the test set once.

4.3.1 Methodology

This experiment was conducted using n-fold cross-validation. Cross-validation is a method where the instances are separated into number of groups defined by the fold number, and the model is created using as training set all but one of the folds. The excluded single fold is used as a test set. Then the folds are cycled through so that each fold is used as a test set. For this experiment, the number of folds was made to roughly equal the total number of instances divided by 10. This was done so that each fold contained roughly 10 patients, just like the patient outcome prediction experiment described in section 4.1.

For each target attribute and for each of the n iterations of n-fold cross-validation (n=25 for malignancy, and n=6 for survival time), Relief-F was applied to the (n-1) folds forming the training dataset. A model

was constructed from the resulting Relief-F reduced training set, using either logistic regression, or Bayesian networks with maximum of 3 parents (see the introduction of section 4.2.1). Then this model was tested on the excluded fold to determine its accuracy. The average of these n accuracies over the n repetitions was output as the accuracy of the cross-validation experiment.

A similar experiment was done with the attributes that each doctor chose. The total-dataset was filtered so that only the attributes each doctor chose were included. Then the same machine learning algorithms were applied with the same validation scheme. Since there were four different lists of attributes created by the four doctors, four sets of results were obtained from this part of the experiment.

4.3.2 Results

The following tables summarize the results of experiment 3.

| Prediction Type | Attributes | Bayesian Networks | Logistic Regression |
|-----------------|-------------------------|-------------------|---------------------|
| Hybrid | Attribute List A | 74.9004 | 73.3068 |
| Hybrid | Attribute List B | 76.0956 | 73.7052 |
| Hybrid | Attribute List C | 82.4701 | 76.8924 |
| Hybrid | Attribute List D | 73.3068 | 74.1036 |
| Hybrid | Average A~D | 76.6932 | 74.5020 |
| Machine | Attribute List Relief-F | 82.0717 | 78.8845 |

Table 4-40 Malignancy Results (25 fold cross-validation) (Values in % accuracy)

| Prediction Type | Attributes | Bayesian Networks | Logistic Regression |
|-----------------|-------------------------|-------------------|---------------------|
| Hybrid | Attribute List A | 66.1290 | 48.3871 |
| Hybrid | Attribute List B | 59.6774 | 30.6452 |
| Hybrid | Attribute List C | 66.1290 | 50.0000 |
| Hybrid | Attribute List D | 58.0645 | 41.9355 |
| Hybrid | Average A~D | 62.5000 | 42.7420 |
| Machine | Attribute List Relief-F | 62.9032 | 43.5484 |

Table 4-41 Survival Time Results (6 fold cross-validation) (Values in % accuracy)

4.3.3 Discussion

Comparing against the average cross-validation accuracy over the doctors' selected attribute lists, it is clear that the cross-validation accuracy of the Relief-F attribute selection was superior. That said, it should be noted that the attributes chosen by Doctor C provided a cross-validation accuracy comparable

to that of Relief-F for malignancy. Furthermore, the attributes chosen by Doctor A and by Doctor C provided higher cross-validation accuracies than that of Relief-F for survival time. Interestingly, the average cross-validation accuracy of the doctors' survival time attributes was almost the same as that of Relief-F.

Trends worth considering come from comparing the results obtained in this experiment with the results obtained in the patient outcome prediction experiments in section 4.1. Notice that for Bayesian Networks we will compare only the results obtained with three parents since the current experiment has been run with only three parents.

| Prediction Type | Attribute Selection | Bayesian Networks (3 parents) Taken from Table 4-12 and Table 4-14 | | Bayesian Networks (3 parents) 25 cross-validation accuracy Taken from Table 4-40 |
|-----------------|-------------------------|--|----------|--|
| | | Test Set | Accuracy | |
| Hybrid | Attribute List A | Test Set A | 60 | 74.9004 |
| Hybrid | Attribute List B | Test Set B | 70 | 76.0956 |
| Hybrid | Attribute List C | Test Set C | 60 | 82.4701 |
| Hybrid | Attribute List D | Test Set D | 80 | 73.3068 |
| Hybrid | Attribute List A~D | Test Set A~D | 67.5 | 76.6932 |
| Machine | Attribute List Relief-F | Test Set A | 80 | N/A |
| Machine | Attribute List Relief-F | Test Set B | 90 | N/A |
| Machine | Attribute List Relief-F | Test Set C | 80 | N/A |
| Machine | Attribute List Relief-F | Test Set D | 80 | N/A |
| Machine | Attribute List Relief-F | Test Set A~D | 82.5 | 82.0717 |

Table 4-42 Comparison of Results Bayesian Networks Classifier for Malignancy Class (values in % accuracy)

| Prediction Type | Attribute Selection | Logistic Regression Taken from Table 4-13 and Table 4-15 | | Logistic Regression 25 cross-validation accuracy Taken from Table 4-40 |
|-----------------|-------------------------|--|----------|--|
| | | Test Set | Accuracy | |
| Hybrid | Attribute List A | Test Set A | 80 | 73.3068 |
| Hybrid | Attribute List B | Test Set B | 70 | 73.7052 |
| Hybrid | Attribute List C | Test Set C | 80 | 76.8924 |
| Hybrid | Attribute List D | Test Set D | 70 | 74.1036 |
| Hybrid | Attribute List A~D | Test Set A~D | 75 | 74.502 |
| Machine | Attribute List Relief-F | Test Set A | 70 | N/A |
| Machine | Attribute List Relief-F | Test Set B | 100 | N/A |
| Machine | Attribute List Relief-F | Test Set C | 80 | N/A |
| Machine | Attribute List Relief-F | Test Set D | 90 | N/A |
| Machine | Attribute List Relief-F | Test Set A~D | 85 | 78.8845 |

Table 4-43 Comparison of Results Logistic Regression Classifier for Malignancy Class (values in % accuracy)

| Prediction Type | Attribute Selection | Bayesian Networks (3 parents) Taken from Table 4-17 and Table 4-19 | | Bayesian Networks (3 parents) 6 cross-validation accuracy Taken from Table 4-41 |
|-----------------|-------------------------|--|----------|---|
| | | Test Set | Accuracy | |
| Hybrid | Attribute List A | Test Set A | 60 | 66.129 |
| Hybrid | Attribute List B | Test Set B | 60 | 59.6774 |
| Hybrid | Attribute List C | Test Set C | 40 | 66.129 |
| Hybrid | Attribute List D | Test Set D | 60 | 58.0645 |
| Hybrid | Attribute List A~D | Test Set A~D | 55 | 62.5 |
| Machine | Attribute List Relief-F | Test Set A | 90 | N/A |
| Machine | Attribute List Relief-F | Test Set B | 70 | N/A |
| Machine | Attribute List Relief-F | Test Set C | 60 | N/A |
| Machine | Attribute List Relief-F | Test Set D | 90 | N/A |
| Machine | Attribute List Relief-F | Test Set A~D | 77.5 | 62.9032 |

Table 4-44 Comparison of Results Bayesian Networks Classifier for Survival Time Class (values in % accuracy)

| Prediction Type | Attribute Selection | Logistic Regression Taken from Table 4-18 and Table 4-20 | | Logistic Regression 25 cross-validation accuracy Taken from Table 4-41 |
|-----------------|-------------------------|--|----------|--|
| | | Test Set | Accuracy | |
| Hybrid | Attribute List A | Test Set A | 40 | 48.3871 |
| Hybrid | Attribute List B | Test Set B | 20 | 30.6452 |
| Hybrid | Attribute List C | Test Set C | 40 | 50 |
| Hybrid | Attribute List D | Test Set D | 50 | 41.9355 |
| Hybrid | Attribute List A~D | Test Set A~D | 37.5 | 42.742 |
| Machine | Attribute List Relief-F | Test Set A | 80 | N/A |
| Machine | Attribute List Relief-F | Test Set B | 70 | N/A |
| Machine | Attribute List Relief-F | Test Set C | 60 | N/A |
| Machine | Attribute List Relief-F | Test Set D | 80 | N/A |
| Machine | Attribute List Relief-F | Test Set A~D | 72.5 | 43.5484 |

Table 4-45 Comparison of Results Logistic Regression Classifier for Survival Time Class (values in % accuracy)

In general, cross-validation accuracies are more reliable as they are obtained over multiple test sets, each one having approximately the same target attribute distribution of the full dataset.

5 Conclusions and Future Work

It should be noted that this IQP has been greatly influenced and benefited from the feedback and suggestions offered by the medical personnel from the UMass Medical School Division of Surgical Oncology.

Many suggestions from the doctors have had an important role in making the final approach towards predicting the survival time. The initial aim for the survival time class was to predict a patient's survival time using only the pre-operative information. This prediction type would have then helped the doctors decide whether operating on a patient would be worthwhile considering the post-operative survival time and quality of life. However, one of the doctors specifically mentioned that a medical doctor would not base his/her decision on a presumptive survival time, as it would be an unreliable factor. Additionally, studies have shown that patients that undergo surgery have a better survival time. For example patients that have "small" pancreatic cancer and undergo resection are more likely to get cured/ have a better survival time and their post-operative mortality has decreased to 5% (29) (30) (31). Yet, from the sample of patients that are diagnosed with pancreatic cancer only 10 -20% are eligible for resection.

A doctor pointed out that predicting survival time as a numeric value is more of a lucky guess- issue, a matter also addresses by another doctor. There is not sufficient information from pre-operative and operative reports in order to actually make a good numeric prediction. His medical experience motivated us to break up the survival time predicted class into time four intervals (less than 3 months, 3- 6 months, 6- 9 months, more than 9 months). Another issue that was discussed was the actual type of cancer and its impact on the survival time prediction. The doctor mentioned that each type of cancer has its own features and results in a specific tendency for survival time (e.g., neuroendocrine tumor patients will have a relatively improved survival time (32).). Thus the prediction accuracy for both malignancy and survival time class may improve if the doctors were given the type of pancreatic cancer the patient has. This suggestion and other ideas that came to mind were not pursued as they would have been very time consuming for the medical doctors involved. Limiting our time demands on their extremely busy schedule was a constant concern of this project.

One of the doctors has also suggested for our future research to consider allowing the doctors to make a selection of only five attributes for both the malignancy and survival time and then base the final patient classification using only these five attributes. This approach might be better because the doctor would be constrained to select only what is tremendously relevant towards the target class. Such a strict selection might prove that doctors have the ability to correctly determine the markers of a disease, and at the same

time get a high prediction accuracy using these markers. Another thing that comes to mind is the varying distributions of target class values for the patients that each doctor saw in the patient outcome prediction experiment (see section 4.1). As it can be seen from **Table 4-24** and **Table 4-25** class values have varying distributions. This could have been a biasing factor for the doctors when they made their classification, affecting their accuracy either positively or negatively.

From the three experiments conducted throughout this IQP we have concluded that machine learning techniques perform as well or better than medical doctors in making patient outcomes predictions. Thus, the patient outcome prediction experiment (see section 4.1) proved that doctors are surpassed by machine learning techniques (Bayesian Networks and logistic regression) in both outcomes predictions (malignancy and survival time; see Table **4-21**). When looking at the prediction accuracy based on frequency of selected attributes (see section 4.2) the best prediction accuracy for the malignancy class was obtained with the set of attributes encountered in three of the lists of attributes coming from Relief-F as attribute selector and in one list of attributes coming from the doctors as attribute selectors. For the survival time class, the best accuracy when using both Relief-F and the doctors as attribute selectors was obtained using the attributes encountered in four of the selected attribute lists. The last experiment (see section 4.3) showed that the malignancy hybrid experiment accuracies conducted using the cross-validation method are equivalent or better compared to the individual test set validation method. On the other hand, the accuracies for survival time were better in the machine experiments ran with the individual test set validation method (see Table **4-42**, Table **4-43**, Table **4-44**, and Table **4-45**). In conclusion, all experiments conducted in this IQP showed that machine assisted predictions had similar or higher accuracies than those predictions made by doctors alone.

6 Bibliography

1. **Hayward, John.** *Mining Oncology Data: Knowledge Discovery in Clinical Performance of Cancer Patients*, M.S. Thesis. Worcester : Department of Computer Science, Worcester Polytechnic Institute, 2006.
2. **Floyd, Stuart.** *Data Mining Techniques for Prognosis in Pancreatic Cancer.* M.S. Thesis. Worcester : Department of Computer Science, Worcester Polytechnic Institute, 2007.
3. <http://seer.cancer.gov/>. <http://seer.cancer.gov/>. [Online] [Cited: October 27, 2008.]
4. <http://www.hcup-us.ahrq.gov/db/nation/nis/nisdbdocumentation.jsp>. [Online] [Cited: October 27, 2008.]
5. **Martini, Freidrich.** *Fundamentals of Anatomy and Physiology.* New Jersey : Prentice hall, 1998.
6. **Standring, Susan.** *Gray's Anatomy: Anatomical Basis of Clinical Practice.* Edinburgh : Elsevier Churchill Livingstone, 2005.
7. **Fox, Stuart Ira.** *Human Physiology.* s.l. : McGraw-Hill Higher Education, 2008.
8. **Norman, Wesley.** Pancreas. [Online] 1999. [Cited: April 10, 2008.] <http://home.comcast.net/~wnor/pancreas.htm>.
9. *Tumor versus neoplasm: Isn't it time to use the correct term-neoplasm?* **Sarr, Michael G. and Warshaw, Andrew L.** 3, March 2005, Surgery, Vol. 137, p. 297.
10. **Steen, Grant J.** *The Conspiracy of Cells- the basic science of cancer.* New York : Plenum Press, 1993.
11. **National Cancer Institute.** Cancer of the Pancreas, SEER Stat Fact Sheets. *Cancer of the Pancreas.* [Online] [Cited: February 12, 2008.] <http://seer.cancer.gov/statfacts/html/pancreas.html>.
12. **American Cancer Society.** Detailed Guide: Pancreatic Cancer- Surgery. [Online] [Cited: April 15, 2008.] http://www.cancer.org/docroot/cri/content/cri_2_4_4x_surgery_34.asp?.sitearea=cri.
13. **Schwartz, Seymour I.** *Principles of Surgery.* New York : McGraw-Hill Book Company, 1989.
14. *Values of mutations of K-ras oncogene at codon 12 in detection of pancreatic cancer: 15- year experience.* **Mu, De-Quing, Peng, You-Shu and Xu, Qiao-Jian.** 2004, World Journal of Gastroenterology, pp. 471-475.
15. **Way, Lawrence W.** *Current Surgical Diagnosis & Treatment.* Norwalk, CN : Appleton & Lange, 1994.
16. **Cameron, John L.** *Current Surgical Therapy.* St. Louis : Mosbt, Inc., 1998.
17. **Prucha, Edward J.** *Cancer Sourcebook.* Detroit : Omnigraphics, 2000.

18. **Borfitz, D. and Getz, K.** *Informed Consent. A Guide to the Risks and Benefits of Volunteering for Clinical Trials.* Boston : Thomson, 2003.
19. **The Lustgarten Foundation.** *Understanding Pancreatic Cancer.* USA : The Lustgarten Foundation for Pancreatic Cancer Research, 2007.
20. **Hoff, Von; D., Daniel; Evans, B. Douglas; Hruban, H. Ralph;.** *Pancreatic Cancer.* USA : Jones and Bartlett, 2005.
21. **Freelove, R. and Walling, AD.** Pancreatic Cancer: Diagnosis and Management. *American Family Physician.* Feb 1, 2006, pp. 485-92.
22. **Miller, J. C.** Magnetic Resonance Cholangiopancreatography. [Online] [Cited: April 13, 2008.] http://www.massgeneralimaging.org/newsletter/june_2004.
23. **American Cancer Society.** Pancreatic Cancer. *American Cancer Society.* [Online] [Cited: April 14, 2008.] <http://documents.cancer.org/116.00>.
24. *Toxicity and Response Criteria of the Eastern Cooperative Oncology Group.* **Oken, M.M., Creech, R.H., Tormey, D.C., Horton, J., Davis, T.E., McFadden, E.T., Carbone, P.P.** 1982, *American Journal of Clinical Oncology*, pp. 649-655.
25. *Application of Relief-F Feature Filtering Alogorithm to Selecting Informative Genes for Cancer Classification Using Micrarray Data.* **Wang, Yuhang.** 2004, *Proceedings of the 2004 IEEE Computational Systems.*
26. **Kononenko, Igor.** Estimating Attributes: Analysis and Extensions of RELIEF. *Machine Learning: ECML-94.* Berlin : Springer Berlin / Heidelberg, 1994, Vol. Volume 784/1994, pp. 171 - 182.
27. **Witten, Ian H. and Frank, Eibe.** *Data Mining: Practical Machine Learning Tools and Techniques.* San Fransisco : Elsevier Inc., 2005.
28. MEDITECH. [Online] <http://www.meditech.com/>.
29. *The Impact of Curative Intent Surgery on the Survival of Pancreatic Cancer Patients: A US Polpulation-Based Study.* **Sahib, Yasser.** 2007, *The American Journal of Gastroenterology.*
30. *Improved Survival in Small Pancreatic Cancer.* **Pantalone, Desiree.** 2001, *Digestive Surgery*, pp. 41-46.
31. **Nunes, Quentin M.** Pancreatic cancer. *Surgery (Oxford).* s.l. : Elsevier Ltd, 2007, pp. 87-94.
32. *Pancreatic neuroendocrine tumors (PNETs): incidence, prognosis and recent trend toward improved survival.* **Halfdanarson, T. R.** 2008, *Annals of Oncology* .
33. **Ruggeri, F.** Bayesian Networks. *Encyclopedia of Statistics in Quality and Reliability.* s.l. : John Wiley & Sons, p. 2007.

34. http://www.cancer.org/docroot/CRI/content/CRI_2_6x_the_history_of_cancer_72.asp.
www.cancer.org. [Online] [Cited: October 21, 2008.]

Appendix A: WEKA Parameters^P

Relief Attribute Evaluation: Evaluates the worth of an attribute by repeatedly sampling an instance and considering the value of the given attribute for the nearest instance of the same and different class.

numNeighbours -- Number of nearest neighbors for attribute estimation.

sampleSize -- Number of instances to sample. Default (-1) indicates that all instances will be used for attribute estimation.

seed -- Random seed for sampling instances.

sigma -- Set influence of nearest neighbors. Used in an exp function to control how quickly weights decrease for more distant instances. Use in conjunction with *weightByDistance*. Sensible values = 1/5 to 1/10 the number of nearest neighbors.

weightByDistance -- Weight nearest neighbors by their distance.

Ranker: Ranks attributes by their individual evaluations

generateRanking -- A constant option. Ranker is only capable of generating attribute rankings.

numToSelect -- Specify the number of attributes to retain. The default value (-1) indicates that all attributes are to be retained. Use either this option or a threshold to reduce the attribute set.

startSet -- Specify a set of attributes to ignore. When generating the ranking, Ranker will not evaluate the attributes in this list. This is specified as a comma separated list of attribute indexes starting at 1. It can include ranges. Eg. 1,2,5-9,17.

threshold -- Set threshold by which attributes can be discarded. Default value results in no attributes being discarded. Use either this option or *numToSelect* to reduce the attribute set.

BayesNets: Bayes Network learning using various search algorithms and quality measures.

debug -- If set to true, classifier may output additional info to the console.

estimator -- Select Estimator algorithm for finding the conditional probability tables of the Bayes Network.

alpha -- Alpha is used for estimating the probability tables and can be interpreted as the initial count on each value. *searchAlgorithm* -- Select method used for searching network structures.

initAsNaiveBayes -- When set to true (default), the initial network used for structure learning is a Naive Bayes Network, that is, a network with an arrow from the classifier node to each other node. When set to false, an empty network is used as initial network structure

^P Description of parameters taken from the WEKA 3.0 Help Manual

markovBlanketClassifier -- When set to true (default is false), after a network structure is learned a Markov Blanket correction is applied to the network structure. This ensures that all nodes in the network are part of the Markov blanket of the classifier node.

maxNrOfParents -- Set the maximum number of parents a node in the Bayes net can have. When initialized as Naive Bayes, setting this parameter to 1 results in a Naive Bayes classifier. When set to 2, a Tree Augmented Bayes Network (TAN) is learned, and when set >2, a Bayes Net Augmented Bayes Network (BAN) is learned. By setting it to a value much larger than the number of nodes in the network (the default of 100000 pretty much guarantees this), no restriction on the number of parents is enforced

randomOrder -- When set to true, the order of the nodes in the network is random. Default random order is false and the order of the nodes in the dataset is used. In any case, when the network was initialized as Naive Bayes Network, the class variable is first in the ordering though.

scoreType -- The score type determines the measure used to judge the quality of a network structure. It can be one of Bayes, BDeu, Minimum Description Length (MDL), Akaike Information Criterion (AIC), and Entropy.

useADTree -- When ADTree (the data structure for increasing speed on counts, not to be confused with the classifier under the same name) is used learning time goes down typically. However, because ADTrees are memory intensive, memory problems may occur. Switching this option off makes the structure learning algorithms slower, and run with less memory. By default, ADTrees are used.

Logistic- Class for building and using a multinomial logistic regression model with a ridge estimator

debug -- Output debug information to the console.

maxIts -- Maximum number of iterations to perform.

ridge -- Set the Ridge value in the log-likelihood.

Appendix B: Database Attributes Description

Initial Symptom: symptom presented by patient at initial medical consultation. These symptoms might span on a time length of up to 6 months prior to the initial medical visit.

Jaundice: abnormal yellow pigmentation of the skin, white of the eyes, and mucous membranes caused by an increased level of bilirubin in the blood.

Abdominal Pain: pain present in the abdominal region of a patient.

Clay Colored Stool: unusual change of the stool color towards a clay color due to possible problems in the biliary system.

Dysphagia: symptom of difficulty in swallowing.

Early Satiety: abnormal early feeling of fullness while eating or after a meal.

Nausea: sensation of discomfort in the stomach, usually present with an urge to vomit.

Vomiting: tendency to expulse the content of one's stomach.

Weight Loss: weight loss occurred throughout past 6 months up to initial medical consultation; value stored as true or false.

Weight Loss in lbs: weight loss occurred prior to initial medical consultation; value stored as the number of pounds lost.

Comorbidities: presence or the effect of additional disorders present along with a primary disease or disorder.

Pancreatitis: inflammation of the pancreas.

Diabetes w/ Oral Agents: presence of diabetes that is treated through oral agents.

Ethanol (Alcohol): abuse in the intake of alcohol.

Ischemic Heart Disease: heart disease caused by reduced blood supply to the heart.

Liver Failure/Cirrhosis: malfunction in the function of the liver/ progressive liver disease resulting in liver failure.

Malnutrition: improper or insufficient diet resulting in a disease.

Onset of Diabetes: onset of diabetes related to the time of diagnosis: onset in less than six months or onset in more than six months.

Other Major Comorbidities: presence of a comorbidities not included as possible comorbidities in the database forms fields.

Renal Failure: inadequate function of the kidney.

Respiratory: disease related to the respiratory system.

Pre-Operative Laboratory: laboratory tests done before surgery (pre-operatively) on patient's serum.

Albumin: blood plasma protein produced in the liver.

ALT: (alanine transaminase) is an enzyme used for tracking the liver health.

AST: (aspartate aminotransferase) is an enzyme released into the blood when certain organs are injured, particularly the liver and heart.

Bilirubin: yellow breakdown product of a form of hemoglobin, a main component of the blood cells.

CA19-9: (carbohydrate antigen 19-9) a blood marker for pancreatic cancer. We pay special attention to this class since it was selected by all the doctors as a key attribute in making predictions for both malignancy and survival time. It is measured in units/mL where "units" are clinically standardized form of measurement. This class was stored as continuous numerical values. The range of values present in the database was 0 ~ 402242 units/mL. 104 instances were in the 0 ~ 100 units/mL range, 62 in the 101 ~ 1000 units/mL range, and 40 values above 1001 units/mL (These ranges were selected just for convenience in this description but were not used to discretize the actual data set). 56 patients out of the total of 262 patients in the database do not have values for this attribute.

CEA: (carcinoembryonic antigen) a protein encountered at low levels in the blood; it is generally used as a tumor marker for pancreatic cancer.

Diagnostic Procedure: method used for determining or analyzing the nature of a disease or disorder.

EUS: No Node: no presence of lymph node disease noted during the EUS.

CT: Celiac Artery Involvement: involvement of the celiac artery into the tumor as shown by the CT; the involvement can be unknown, open, encased, occluded, abuts.

CT: Hepatic Vein Involvement: involvement of the hepatic vein into the tumor as shown by the CT; the involvement can be unknown, open, encased, occluded, abuts.

CT: Portal Vein Involvement: involvement of the portal vein into the tumor as shown by the CT; the involvement can be unknown, open, encased, occluded, abuts.

CT: SMA Involvement: involvement of the SMA into the tumor as shown by the CT; the involvement can be unknown, open, encased, occluded, abuts.

CT: SMV Involvement: involvement of the SMV into the tumor as shown by the CT; the involvement can be unknown, open, encased, occluded, abuts.

CT: Tumor Size X: the size of the pancreatic tumor as noted during a CT scan measured on the X axis.

CT: Tumor Size Y: the size of the pancreatic tumor as noted during a CT scan measured on the Y axis.

EUS: Celiac Artery Involvement: involvement of the celiac artery into the tumor as shown by the EUS; the involvement can be unknown, open, encased, occluded, abuts.

EUS: Celiac Node Disease: presence of celiac lymph node disease as noted during EUS.

EUS: Cytology: cytology of the tumor cells retrieved during the EUS.

EUS: Hepatic Vein Involvement: involvement of the hepatic vein into the tumor as shown by the EUS; the involvement can be unknown, open, encased, occluded, abuts.

EUS: Number of Tumors: number of tumors in the pancreas noted during an EUS scan.

EUS: Peripancreatic Node Disease: presence of peripancreatic lymph node disease as noted during EUS.

EUS: Portal Vein Involvement: involvement of the portal vein into the tumor as shown by the EUS; the involvement can be unknown, open, encased, occluded, abuts.

EUS: SMA Involvement: involvement of the SMA into the tumor as shown by the EUS; the involvement can be unknown, open, encased, occluded, abuts.

EUS: SMV Involvement: involvement of the SMV into the tumor as shown by the EUS the involvement can be unknown, open, encased, occluded, abuts.

EUS: Tumor Size X: the size of the pancreatic tumor as noted during an EUS scan measured on the X axis.

EUS: Tumor Size Y: the size of the pancreatic tumor as noted during an EUS scan measured on the Y axis.

FNA Cytology: procedure used in extracting and investigating cells of a specific tumor.

Therapeutic ERCP: Stent Type: placement of a stent into the duct that drains the liver and the pancreas during an ERCP procedure.

Other Attributes

Family History of Pancreatic Cancer: Relationship to Patient: relationship to the patient under observation of any family member with pancreatic cancer.

Age at Diagnosis: age of the patient at diagnosis with pancreatic cancer.

Anticipated Operation: operation suggested to the patient by the medical oncologist after one or more clinical consultations.

Chemotherapy: drug treatment applied to patients.

Chemotherapy Specify Agent: specification of agent used during for chemotherapy.

Neoadjuvant: treatment applied to people with cancer prior to surgery with the intent of reducing the tumor size.

Pathology: Tumor Size X: size of the pancreatic tumor revealed after pathological analysis.

Presumptive Diagnosis: diagnosis given at initial medical visit.

Radiation: various type of radiation treatment applied to patients.

Social History: Cigarette Pack Years: number of cigarettes packs used by the patient per year. If the patient quit smoking the number of packs is still recorded.

Social History: Cigarettes (significant use): patient's use of cigarettes in the past or started in the past and continued to present.

Surgery Type: type of pancreatic surgery applied as treatment to pancreatic cancer patient.

Weight: patient's weight at the time of first medical visit.

Post-Operative: events related to the patient's overall health occurring after the patient's surgery.

Any Reoperation: reoperation done after the main pancreatic procedure.

Any Transfusion: complications occurring post-operatively that required transfusion (blood or serum).

Bleeding: post-operative wound bleeding.

Cardiac Complications/MI: cardiac complications or myocardial infarct.

Days in ICU: number of days spent in the Intensive Care Unit (ICU).

Days until Clears Started: number of days after surgery when the patient was given clear liquids.

Days until Drain Removed: number of days after the surgery when the patient's drains were removed.

Days until Regular Diet: number of days after the surgery when the patient was transferred to a regular diet.

Days until Tube Feed Started: number of days after surgery when the patient was started on feeding tubes.

Discharge Status: place of discharge for the patients (home, Subacute Rehabilitation Center, Acute Rehabilitation Center).

Extubated in OR: extubation of patient in the Operative Room (OR).

Feeding Tube Complications: complications in patient's health status due to intolerance of Feeding Tubes.

Histology: anatomical study of the pancreatic resected tissue.

ICU Re-Admission: re-admission to the Intensive Care Unit.

Leak: leak from the site of the resection.

Length of Stay: number of days spent in hospital after surgery until discharge.

Liver Insufficiency: dysfunction of the liver occurring after the pancreatic cancer surgery.

Prolonged PO Intolerance: prolonged intolerance to the treatment applied after surgery.

Pulmonary Complications: complications of the pulmonary system.

Resection: events occurring during the pancreatic cancer related surgery.

Blood Loss in cc: blood loss during the surgery, measured in cubic centimeters (cc).

Number of Drains: number of drains placed inside the patient's abdominal cavity during the surgery.

Transfusion PRBC Units: transfusion of Packets of Red Blood Cells (PRBC) units during surgery.

Venous Resection: resection of any vein during the pancreatic resection due to vein involvement with the tumor.

Appendix C: General Medical Terms

Ampulla of Vater: an enlargement of the ducts from the pancreatic and common bile duct at the point where they enter the small intestine.

Anorexia: eating disorder characterized by voluntary starvation, vomiting, etc.

Atrium: chamber of the heart where de-oxygenated blood is received.

Beger Procedure: surgery where the pancreatic head is resected while preserving the duodenum.

Benign: medical term to describe mild or non-progressive disease.

Celiac Artery: first major branch of the abdominal aorta.

CT (Computed Tomography): medical imaging technique used for generating a three-dimensional image of a specific inner region of the body.

Cyst: closed sac having its own membrane and remain separate from nearby tissue.

Cytology: study of the cellular disease and cellular changes leading to disease diagnosis.

Duodenum: first part of the small intestine, connecting the stomach to the jejunum.

ECOG: Eastern Cooperative Oncology Group Performance Status; scales and criteria used by doctors in assessing how a patient's disease is evolving.

Endocrine cell: cell that produces hormones into the bloodstream.

Enucleation: surgery where the target tumor is shelled out from the pancreas without removing any pancreatic tissue.

Epigastrium: region of the upper central abdomen in between the costal margins and the subcostal plane.

ERCP (Endoscopic retrograde cholangiopancreatography): study of the duct that drains the liver and the pancreas.

ERCP(Endoscopic Retrograde Cholangiopancreatography): technique using endoscopy and fluoroscopy in diagnosing and treating problems related to liver, gallbladder, bile ducts, and pancreas.

EUS(Endoscopic Ultrasound): technique combining endoscopy and ultrasound to obtain images and information about the digestive tract and surrounding tissues.

Exocrine Cell: cell that produces enzymes into ducts

Familial atypical multiple mole melanoma syndrome: an inherited condition where one or more first or second degree relatives have malignant melanoma, and had many moles.

FDG: Flourodeoxyglucose; glucose analog used in medical imaging (PET).

Feeding Tubes: medical device used in providing nutrition to patients that cannot obtain nutrients independently.

FNA Biopsy: Fine Needle Aspiration biopsy is a procedure for removing cells from a targeted tissue for performing analysis on them.

Frey's Procedure: surgery where the diseased part of the pancreatic head is cored out.

Glandular Organ: organ that synthesizes a substance for release (eg. Liver).

Glucose: simple sugar, used by living cells as source of energy and also is a metabolic intermediate.

Hepatic Vein: vein that drain blood from the liver into the superior vena cava.

Hereditary nonpolyposis colon cancer (Lynch syndrome): inherited cancer particularly of the colon and rectum, and increases the risk of other cancers such as stomach, small intestine, liver, etc.

Hormone: chemicals released by cells that affect other parts of the body.

Hypokalemia : condition in which potassium concentration of the blood is too low.

Islet cells: endocrine cells of the pancreas.

Left colic flexure: sharp bend of the large intestine at the left upper quadrant of humans; also called the splenic flexure.

Lumen: the space inside of lining of a tubular structure (eg. an artery).

Malignant: medical term to describe severe or worsening disease (eg. a malignant tumor is basically cancer).

MRCP (Magnetic Resonance Cholangiopancreatography): a MRI alternative for ERCP.

MRI (Magnetic Resonance Imaging): medical diagnosis technique used for visualization of the structure and function of the body.

Mucinous tumor: tumor of the mucous glands.

Peptic: referring to any part of the body that usually has an acidic lumen.

Periampullary: located in the Ampulla of Vater.

Peritoneum: the serous membrane that covers most of the intra-abdominal organs.

PET(Positron Emission Tomography): technique for producing a three dimensional image of the functional processes of the body.

Peutz- Jeghers Syndrome (PJS): an inherited disease of the gastrointestinal tract, characterized by hamartoma development in the small intestine.

Polypeptide: a peptide, or short polymers, consisting of multiple amino acids.

Portal Vein: large vein that carries blood from the stomach and intestines to the liver.

Pruritus: itch, or a sensation that causes a person to want to scratch.

Serous tumor: tumor of the serous glands.

SMA (Superior Mesenteric Artery): blood vessel in charge with supplying blood to the intestine from the lower part of the duodenum to the left colic flexure and the pancreas.

SMV (Superior Mesenteric Vein): blood vessel in charge of the blood drainage from the small intestine.

Somatostatin: hormone that regulates the endocrine system, inhibiting numerous secondary hormones.

Stomatitis: an inflammation of the mucous lining of any structure in the mouth.

TNM Staging: staging for the Tumor, Nodes and Metastasis and margins.

Von Hippel-Lindau Syndrome: an inherited condition characterized by abnormal growth of tumors in parts of the body which are rich in blood supply.