Mining Oncology Data: Knowledge Discovery in Clinical Performance of Cancer Patients

By John Hayward (Computer Science), Prof. Carolina Ruiz (CS), and Dr. Giles Whales (UMass Medical School)

Cancer research: "*the intense scientific effort to understand the development of cancer and identify potential therapies*" [Cancer Genome Project]
Database and data mining techniques have become vital tools in studying clinical oncology data

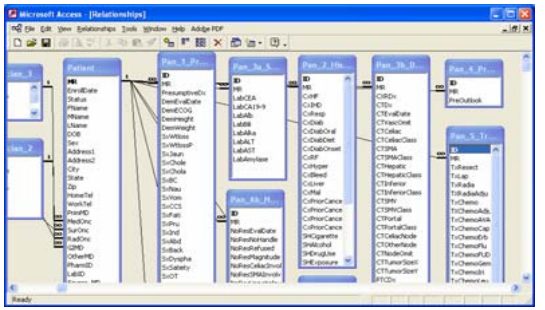
Our joint WPI-UMass Medical School Research on Gastrointestinal Cancer and Breast Cancer:

(1) **Design and program cancer database modules** and (2) **Discover patterns in clinical data using data mining and machine learning techniques**

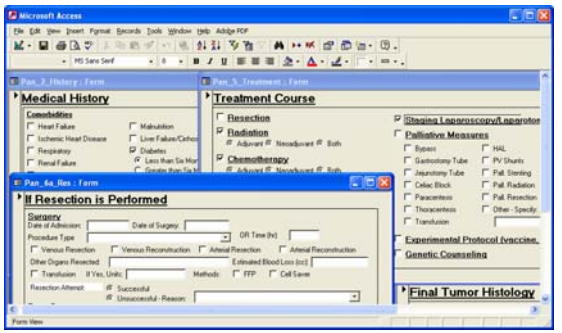
Collect: Patient Demographics, Health Factors, Medical History, Diagnostic Studies, Treatment Courses, Surgery, Patient follow-up

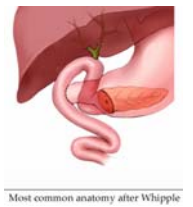
Study health patterns in along population divisions: Tumor Resection/No Resection, Chemo/Radiotherapy, Tumor Progression, Redeveloped Symptoms


Organs removed during a Whipple


Database Captures Heterogeneous Structure of Patient Narrative

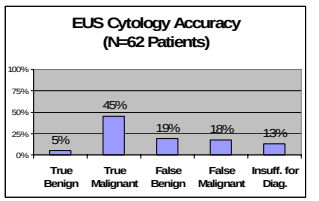
DB User Interface



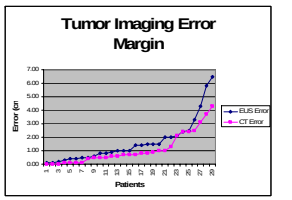

Most common anatomy after Whipple

Initial Data Analysis Results:

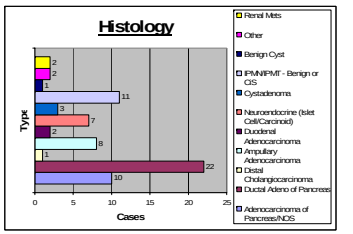
Contributions of Endoscopic Ultrasound and CT Scan to Tumor Staging and Preoperative Assessment of Pancreatic Tumor Resectability



EUS: Endoscopic Ultrasound

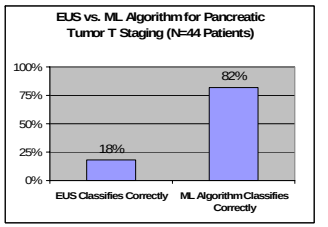


EUS: Endoscopic Ultrasound. CT: CT Scan

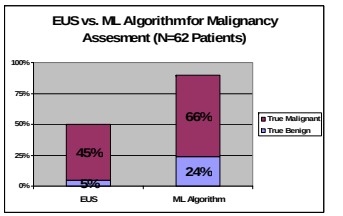


Further Data Analysis Using Machine Learning (ML) & Data Mining Techniques

ML algorithms have demonstrated higher accuracy rates over major oncology diagnostic tools like Endoscopic Ultrasound (EUS)

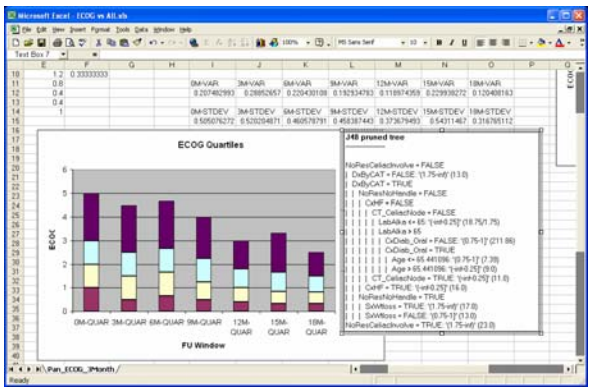


Comparative Accuracy of T-Staging Pancreatic Tumors



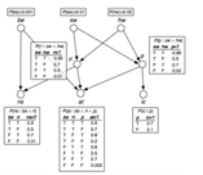
Comparative Accuracy of Assessing Tumor Malignancy

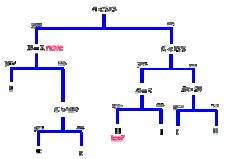
Data Mining & ML Analysis of Quality of Life Performance using Decision Trees



Our Major ML & Data Mining Objectives

- Patient Longevity
- Quality of Life/Performance Status
- Surgical Prospects
- Imaging and Diagnostic Test Accuracy
- Patient Demographics (breast cancer)


Naïve Bayes and General Bayesian Models


Regression Trees, Model Trees, Decision Trees

RStatus (tumor growth on margins of surgically excised tissue)
Prediction – Using Metalearning Classifiers

Data Set	Algorithm	Percent Correct
Conclu	J48	84.35%
All	NaiveBayes	86.25%
Conclu.CFS	Bagging.NaiveBayes	92.00%
All.CFS	Bagging.NaiveBayes	92.00%
*	ZeroR (Benchmark)	78.00%
*	OneR (Benchmark)	70.25%

Data Mining & ML Analysis of Age vs. Longevity using Regression Trees