THE UNIVERSITY OF HULL

The use of knowledge discovery databases in the identification of patients with colorectal cancer.

being a Thesis submitted for the Degree of Doctor of Medicine

in the University of Hull

by

Jonathan Bowes Cowley, MB ChB, MRCS

July 2012

Abstract:

Colorectal cancer is one of the most common forms of malignancy with 35,000 new patients diagnosed annually within the UK. Survival figures show that outcomes are less favourable within the UK when compared with the USA and Europe with 1 in 4 patients having incurable disease at presentation as of data from 2000.

Epidemiologists have demonstrated that the incidence of colorectal cancer is highest on the industrialised western world with numerous contributory factors. These range from a genetic component to concurrent medical conditions and personal lifestyle. In addition, data also demonstrates that environmental changes play a significant role with immigrants rapidly reaching the incidence rates of the host country.

Detection of colorectal cancer remains an important and evolving aspect of healthcare with the aim of improving outcomes by earlier diagnosis. This process was initially revolutionised within the UK in 2002 with the ACPGBI 2 week wait guidelines to facilitate referrals form primary care and has subsequently seen other schemes such as bowel cancer screening introduced to augment earlier detection rates. Whereas the national screening programme is dependent on FOBT the standard referral practice is dependent upon a number of trigger symptoms that qualify for an urgent referral to a specialist for further investigations. This process only identifies 25-30% of those with colorectal cancer and remains a labour intensive process with only 10% of those seen in the 2 week wait clinics having colorectal cancer.

This thesis hypothesises whether using a patient symptom questionnaire in conjunction with knowledge discovery techniques such as data mining and artificial

neural networks could identify patients at risk of colorectal cancer and therefore warrant urgent further assessment. Artificial neural networks and data mining methods are used widely in industry to detect consumer patterns by an inbuilt ability to learn from previous examples within a dataset and model often complex, non-linear patterns. Within medicine these methods have been utilised in a host of diagnostic techniques from myocardial infarcts to its use in the Papnet cervical smear programme for cervical cancer detection.

A linkert based questionnaire of those attending the 2 week wait fast track colorectal clinic was used to produce a 'symptoms' database. This was then correlated with individual patient diagnoses upon completion of their clinical assessment. A total of 777 patients were included in the study and their diagnosis categorised into a dichotomous variable to create a selection of datasets for analysis. These data sets were then taken by the author and used to create a total of four primary databases based on all questions, 2 week wait trigger symptoms, Best knowledge questions and symptoms identified in Univariate analysis as significant. Each of these databases were entered into an artificial neural network programme, altering the number of hidden units and layers to obtain a selection of outcome models that could be further tested based on a selection of set dichotomous outcomes. Outcome models were compared for sensitivity, specificity and risk. Further experiments were carried out with data mining techniques and the WEKA package to identify the most accurate model. Both would then be compared with the accuracy of a colorectal specialist and GP

Analysis of the data identified that 24% of those referred on the 2 week wait referral pathway failed to meet referral criteria as set out by the ACPGBI. The incidence of those with colorectal cancer was 9.5% (74) which is in keeping with other studies

and the main symptoms were rectal bleeding, change in bowel habit and abdominal pain. The optimal knowledge discovery database model was a back propagation ANN using all variables for outcomes cancer/not cancer with sensitivity of 0.9, specificity of 0.97 and LR 35.8. Artificial neural networks remained the more accurate modelling method for all the dichotomous outcomes.

The comparison of GP's and colorectal specialists at predicting outcome demonstrated that the colorectal specialists were the more accurate predictors of cancer/not cancer with sensitivity 0.27 and specificity 0.97, (95% CI 0.6-0.97, PPV 0.75, NPV 0.83) and LR 10.6. When compared to the KDD models for predicting the same outcome, once again the ANN models were more accurate with the optimal model having sensitivity 0.63, specificity 0.98 (95% CI 0.58-1, PPV 0.71, NPV 0.96) and LR 28.7.

The results demonstrate that diagnosis colorectal cancer remains a challenging process, both for clinicians and also for computation models. KDD models have been shown to be consistently more accurate in the prediction of those with colorectal cancer than clinicians alone when used solely in conjunction with a questionnaire. It would be ill conceived to suggest that KDD models could be used as a replacement to clinician- patient interaction but they may aid in the acceleration of some patients for further investigations or 'straight to test' if used on those referred as routine patients.

Table of Contents

List of tables

List of figures:

Acknowledgement

I would like to thank John Hartley for his time, patience, guidance and advice over the past few years, without which I would not have got to this stage. I would also like to acknowledge the help and support of all those at the Academic Surgical Unit at Castle Hill and specifically the valuable, worldly advice from James Gunn, John Monson and Graeme Duthie in in the early stages.

Thanks also to Mandy Bulmer, Judy East and Nicky Stocks for their help and support in collating patient questionnaires.

A special thanks to my wife Laura who has continually provided support, understanding and encouragement throughout, tolerating hours spent in front of the computer.

Finally I would like to thank my parents, who have been supportive and positive in everything I have done. I am especially grateful for their patience and understanding over the past few years.

Author's Declaration:

The work presented in this thesis was performed entirely by the author except as acknowledged. This thesis has not previously been submitted for a degree or diploma at this or any other institution.

Jonathan Bowes Cowley

July 21012

List of abbreviations

ANN	Artificial Neural Network
DM	Data Mining
CEA	Carcinoembryonic Antigen
CT	Computer Tomography
FAP	Familial Adenomatous Polyposis
MRI	Magnetic Resonance Imaging
US	Ultrasound
KDD	Knowledge Discovery in databases
FOB	Faecal Occult Blood
FOBT	Faecal Occult Blood test
EMR	Endomucosal resection
EMD	Endomucosal dissection
DCBE	Double contrast barium enema
TNM	Tumour/Node/Metastases
TME	Total mesorectal excision
TEMS	Transanal endoscopic microsurgery
5-FU	5-Flurouricil
ACPGBI	Association of coloproctology of Great Britain and Ireland
OPD	Out patient department
2WW	Two week wait
SOFM	Self organising feature map
MLP	Multilayered perceptron

Introduction

1.1 Epidemiology

Colorectal cancer remains one of the most common forms of malignancy, with over 1 million individuals being affected worldwide [1] Specifically within the United Kingdom colorectal is the second most common form of malignancy with approximately 35,000 new patients being diagnosed annually. [2-5]

It has been shown that survival from the disease within the UK is less favourable compared with the USA and other European Countries [2, 3, 6, 7]. Whilst causative factors may range from fewer doctors per capita to healthcare expenditure, the time of presentation has been demonstrated to play a significant role. A study in 2000 demonstrated that 1 in 4 patients presenting with colorectal cancer had incurable disease at diagnosis[2, 4, 8, 9]

The incidence of colon cancer is at its highest in the industrialised western world with epidemiological studies focusing on the identification of factors that influence the risk of an individual acquiring colorectal cancer. Whilst there is a genetic aspect to developing colorectal cancer and an increased prevalence amongst certain medical conditions a number of dietary and lifestyle factors have also been identified and proven to modify risk.

International data demonstrates that colorectal cancer is highly sensitive to environmental changes with immigrants rapidly reaching the incident rates of the host country [3-5] [6]

1.2 Colorectal Pathology

1.2.1 Benign Colonic Pathology

Polyp is a term used to clinically describe any elevated tumour and covers a variety of histologically different tumours. They occur either individually, in small numbers or can be found ‘carpeting’ the colon in conditions such as Familial Adenomatous Polyposis (FAP). Whilst the term polyp can encompass a clinical description it is the histological conformation of the polyp that is important as they can be subdivided into inflammatory, metaplastic, harmartomatous and neoplastic variants.

Specifically focusing on the neoplastic variant once again gives rise to further sub classification of adenomatous polyps which vary from tubular adenomas to the villous adenoma. Both of these variants differ in their symptomatology and also the potential risk for colorectal cancer. The tubular adenoma is generally identified incidentally through investigations for colonic bleeding and has a risk of malignancy that increases as the size of the polyp itself increases, a 1cm diameter tubular adenoma carrying a 10% risk of colorectal cancer. Villous adenomas tend to present with slightly different symptoms, usually those of diarrhoea, mucous and potentially hypokalaemia. Tumours of this variety carry a 15% chance of carcinoma if they are greater than 2cm in size.

The specific incidence of these polyps in the general population is difficult to estimate but autopsy studies have been performed to try and assess their prevalence [7] [8] [9]. Willians et al is the only UK study and examined 365 cases in which the colon was examined for hyperplastic / metaplastic polyps and neoplastic adenomas.

It found a general prevalence of 36.87% in men and 28.74% in women, values that may be higher in society today. Similar rates have been found in studies from Norway and the USA.

The rate at which a neoplastic polyp undergoes malignant transformation was examined retrospectively by Stryker [10] et al in the Mayo Clinic over a period of 6 years. They examined 226 cases where polyps $\geq$ 10mm were observed and obtained a mean follow up period of 108 months and demonstrated a 37% increase in size and at 5 years a 2.5 % transformation to invasive cancer. Further follow up at 10 and 20 years illustrated rates of 8% and 24% respectively for cancerous change

Those with a history of polyps and who have undergone excision are at higher risk of further polyps when compared with an individual who has never been diagnosed with polyps [11-13] [14]

1.2.2 Inflammatory Bowel Disease

The term inflammatory bowel disease encompasses two different entities, specifically ulcerative colitis and Crohn's disease. The first association between IBD and colorectal cancer was documented by Crohn and Rosenberg [15] in 1925, an association which nowadays is widely accepted. Ten to fifteen percent of all deaths in those with IBD is due to colorectal cancer [16] with the age at diagnosis of CRC being 15-20 years earlier when compared to the general population [17] Ulcerative colitis is universally accepted as increasing risk for the development of colorectal cancer with and is demonstrated in the meta-analysis by Eden et al [17] . The risk of developing colorectal cancer in those with UC increases with time and rates of 1.6% at 10 years, 8% at 20 years and 18% at 30 years having been quoted. Disease distribution of UC has also been shown to influence risk of development of CRC when compared with the general population, proctitis 1.7 times the risk, left sided colitis 2.8 times and pancolitis 14.8 times [18, 19]. More recent data have shown Crohn's disease patients to be at increased risk also [20]

1.2.3 Diverticular Disease

Diverticular disease is a benign condition that typically is acquired and affects the distal colon. Whilst it is not confined to these areas and can, in rare cases be found congenitally in cases of meckel's diverticulae, approximately 95% affects the sigmoid colon. The diverticulae are a herniation of the mucosa through the muscularis propria and while the specific aetiology of this condition is unknown the theories are that increase intraluminal pressure and weakness within the colonic wall can lead to herniation or that defective collagen consistency or defective muscular structure may lead to weakness. Primarily it is a disease of western society and it is hypothesised that diet is a prime contributing factor with its incidence increasing markedly with an ageing population.

Clinically its presentation can vary widely, presenting with generalised abdominal pain, alteration in bowel habit, bleeding PR, diverticulitis and complications of diverticular disease. These symptoms and the population that diverticulosis is commonly found in can make distinguishing it from someone with a colonic cancer difficult.

1.2.4 Haemorrhoids

Haemorrhoids in the general population are very common and can in many cases be the cause of unnecessary individual anxiety. These vascular cushions become symptomatic when inflamed, enlarged, prolapsed or thrombosed and it is at these times that individuals generally seek medical advice. While common the specific aetiology is poorly understood, many authors concur that low fibre diets and straining at defecation increases pressure resulting in engorgement of the haemorrhoidal cushion, primarily through reduced venous return. The typical ‘bright red’ appearance of haemorrhoidal bleeds and the arterial pH support the theory that haemorrhoidal bleeding is actually arterial in origin. Anatomically the dentate line is the division between internal and external haemorrhoids and histological differences being evident in the epithelial covering, internal having columnar and external having squamous. The relation to the dentate line is also important in the innervation, and thus the potential discomfort caused. Symptomatically the difference between internal and external haemorrhoids that can be appreciated on clinical evaluation, with external haemorrhoids predominantly causing trouble with anal hygiene and redundant skin tags.

1.3 Risk Factors for Colorectal Cancer

1.3.1 Non Modifiable

1.3.1.1 Age related

Ninety percent of colorectal cancers are classed as sporadic in their occurrence, making the risk of developing the disease at a young age very low, increasing in later years. It is generally accepted that the development of colorectal cancer is from a pre-existing adenoma within the colon wall [21] [22]. The incidence of adenoma formation also increases with age, one in three people having at least one adenoma at the age of 60 years. Studies have examined the natural progression of these lesions, demonstrating the progression to adenocarcinoma to be slow, taking up to 10years in some instances [10] with small, flat adenomas progressing somewhat faster. Other inherent factors in the progression of these lesions are size, number, histological type and also the presence of epithelial dysplasia.

1.3.1.2 Genetic

The remaining 10% of cancers can generally be attributed to two main hereditary conditions, Familial Adenomatous polyposis (FAP) and Hereditary non-polyposis colorectal cancer (HNPCC). FAP is caused by a mutation of the Adenomatous polyposis coli (APC) gene and leads to the development of multiple polyps within the bowel between 10 and 30 years of age, histologically identical to sporadic occurrences it is the sheer volume of polyps within the colon almost guarantees developing colorectal cancer by the age of 40 years. HNPCC is a dominantly inherited condition resulting in an alteration in a mismatch repair gene, diagnosed using Amsterdam Criteria with affected individuals at risk of developing colorectal cancer predominantly in the proximal colon and in the absence of multiple polyps [23, 24]. The most common germline defects in HNPCC are mutations in the nMLH1 and hMSH2 genes, essential in the nucleotide mismatch repair system and have also been associated with the development of extra colonic tumours. In addition to these genetic conditions the personal or family history of colorectal cancer or adenomatous polyps increases the risk of developing colorectal cancer and is modified by the age and number of family member affected, specifically first degree relatives. [23] .

1.3.1.3 Inflammatory bowel diseases

Those patients with ulcerative colitis carry an increased risk of developing colorectal cancer, up to ten times higher than those in the general population [18] with Crohn's disease being implicated in recent evidence as a risk factor also. [25] Diseases of the endocrine system are also linked to an increased risk of colorectal cancer, specifically those with Diabetes mellitus who have 1.3-1.5 times increased risk [26] and also those with acromegaly who have a 2.5x increased risk [27] Both conditions are thought to increase risk via excessive levels of insulin like growth factor (IGF) stimulating the proliferation of colonic mucosa.

1.3.2 Modifiable

1.3.2.1 Diet

1.3.2.1.1 Vegetables and Fruit

A number of studies have examined the role or fruit and vegetable consumption in relation to colorectal cancer but findings have been limited. [28] [29] [30]. A follow up study in Sweden showed that low fruit and vegetable consumption in women had an associated relative risk of 1.65[31] however this conflicted with a larger study in both men and women that did not show any relationship [32] . Raw, green and cruciferous vegetables have been shown, when consumed , to lower the risk of colon cancer [28] [30], and a meta-analysis [33, 34] demonstrated a relative risk of 0.48

1.3.2.1.2 Vitamins

The ACS cancer prevention study II did not show multivitamins to reduce risk of colorectal cancer when used as a baseline marker however, reported use of vitamins 10 years earlier did show a relative risk of 0.71 [35]. A further study in 2002 had shown a lower risk in men who took vitamin E [36] supplements and higher selenium levels in serum have been associated with a lower risk of colonic polyps [37]

1.3.2.1.3 Meat

The association between meat and colorectal cancer has been variable. The Cancer Prevention study II showed no difference in the risk of colorectal cancer death in men or women when comparing the uppermost and lower most quintiles [35] . More recent data from three western society studies suggest that fresh and processed meat are each associated with an elevated risk [38] [39] [40] . More recent studies have suggested causal agents within meat as an explanation, such as Haeme, Nitrosation and O6 carboxymethyl guanine [41] [42] [43] .

1.3.2.1.4 Calcium and Vitamin D

Most studies addressing the role of calcium and vitamin D in colorectal cancer have shown a reduced risk or no association. [44] [45] [46] [47] Interestingly, a high serum vitamin D level had a reduced risk of adenoma only when in association with calcium supplements [48]

1.3.2.2 Exercise

There is a high, consistent association with a reduced risk of colon cancer in those undertaking physical activity [49] [50] [51] . This is attributed to physical activity stimulating peristalses thus reducing the time that faecal matter is in contact with the epithelium. Conversely rectal cancer does not seem to be modified by exercise.

1.3.2.3 Obesity

Obesity in association with reduced physical activity increases the risk of developing colorectal cancer by 2 [49] [51] [50] [52] . Data from the Framingham study showed that waist size rather than BMI was a better predictor for lifetime risk of colorectal cancer [53]

1.3.2.4 Alcohol

Both colon and rectal cancer have been shown to have a dose response relationship to alcohol [54] [30]. This is thought to be due to the inhibition of DNA repair [55] ,formation of DNA adducts through Acetaldehyde or the associated deficiency of nutrients [56] [57]

1.3.2.5 Smoking

The association between smoking and colon cancer is thought to be through microsatellite instability colon cancer [58, 59] and tumours with the loss of MLHI expression [60]

1.4 Presentation of Colorectal Cancer

It is necessary to recognise the cancers within the colonic tract present with different symptoms depending upon their level.

1.4.1 Rectal Bleeding

Significant challenges are faced when trying to identify patients with symptoms indicative of colorectal cancer and who thus require urgent investigation. Studies have shown consultation rates up of four -- sixteen per thousand patients a year in primary care presenting with bleeding per rectum, [61] [62] [63] [64] abdominal pain [65] and alterations in bowel habit [66] . Within the community, these symptoms are very high when compared to the actual incidence of colorectal cancer.

Approximately 19% of patients within general practice reported rectal bleeding in the previous year [64] and it is estimated that 97% of these will not have colorectal cancer [67]. The prevalence of altered bowel habit and abdominal pain within the community are even higher [68] thus less specific at predicting colorectal cancer.

Studies undertaken in the late 1990's aimed to determine the predictive value of rectal bleeding in the community for colorectal cancer [63]. This concluded that painless rectal bleeding, alteration in bowel habit and dark red bleeding; factors previously attributed to a higher risk of colorectal cancer are present in many people within their studied community.

Patients with no anal symptoms but who suffer from rectal bleeding are 3 to 4 times more likely to have cancer as opposed to those who have anal symptoms alone [69] and this finding is independent of any alteration in bowel habit. In patients who have

symptoms of rectal bleeding, bright red rectal bleeding is less predictive as opposed to blood mixed with stool. Whilst this combination of symptoms has been shown to be of more diagnostic value when compared to other attributes, it is of little diagnostic aid [70]. Studies in primary care both in Australia and England have shown a 10% prevalence of cancer within the general community [70] [71] The studies made further suggestions that all those over 40 with rectal bleeding should be referred for further specialist consultation. Symptoms of rectal bleeding and finding a palpable rectal mass are generally indicative of rectal cancers [72]

1.4.2 Alteration in Bowel Habit

Along with rectal bleeding, patients commonly notice changes in their bowel habit. Whilst there can be numerous causes for an alteration in an individual's habit of defecation studies have shown a fivefold increase in the risk of cancer when combined with rectal bleeding than if either symptom occurred on its own. [70] [69] . Increased frequency of defecation along with a change in bowel habit to loose motions has demonstrated a cancer prevalence of one in seven with those tending toward constipation having a prevalence of 1 in 36 [73, 74] One particular study found that all patients with colorectal cancer presented with alteration in the bowel habit and rectal bleeding [73] giving a positive predictive value of 9.2% as opposed to 0% in those with rectal bleeding and no alteration in bowel habit. This study also showed a higher predictive value of colorectal cancer in those with rectal bleeding with no perianal symptoms when compared with those with perianal symptoms. No predictive value was found in the dark or bright red rectal bleeding. . More than 90% of those with rectal and sigmoid cancers have alteration in bowel habit resulting in loose stool or an increased frequency of defecation.

1.4.3 Abdominal Pain

The presence of abdominal pain remains an imprecise diagnostic marker. When associated with rectal bleeding and alteration in bowel habit two studies have demonstrated a reduction in the probability of cancer [69, 70] with only one study showing it to be of benefit in the diagnosis of serious disease.

1.4.4 Iron Deficiency Anaemia

The presence of iron deficiency anaemia with a haemoglobin below 10 g can be found in a large proportion of patients with colorectal cancer type of presentation [74] [75-78] [79] and 50% of these individuals will have no symptoms or clinical signs.

1.5 Diagnosis of Colorectal Cancer

The diagnosis of colorectal cancer is histological; however this tissue diagnosis usually requires a colonoscopy which is not without risk. Given this a number of other diagnostic tools are used to facilitate the identification of those likely to have positive findings at colonoscopy. These range from simple FOB tests to invasive procedures.

1.5.1 FOB tests

These tests are simple and non-invasive, requiring a series of stool samples from the individual following adherence to specific pre-test instructions. They are used as part of the UK screening programme as well as being more widely available. Most FOBT testing is undertaken with a guaiac based test such as the Haemoccult 2 which have a sensitivity of 40-60% and specificity of 90-98% dependent on dietary adherence of the individual before taking the test and rehydration of the sample prior to laboratory analysis.[80].

1.5.2 Flexible Sigmoidoscopy

Flexible sigmoidoscopy utilises fibre optic technology and is commonly used for evaluation of the distal colon. Whilst not the 'gold standard' it is relatively easier to undertake, generally without full bowel preparation and in some areas by non-medical personnel [81-86]thus making its availability greater. As an assessment tool it holds a valuable place, detecting 7 adenocarcinomas and 60 high risk adenomas per 1000 examinations [87]. Given the distribution of colonic malignancies flexible sigmoidoscopy can effectively be used to identify 80% of colonic cancers and both detect and remove 70% of adenomas [88]. Whilst not as extensive as a full colonoscopy, flexible sigmoidoscopy carries with it as an endoscopic procedure, risks of morbidity and mortality, even though they are very small. [89-91]

1.5.3 Colonoscopy

Colonoscopy uses the same technology as flexible sigmoidoscopy, allowing the endoscopist to visualise the whole colon and in some instances intubate the terminal ileum. As with flexible sigmoidoscopy it has the benefit of tissue sampling at time of test, thus allowing histological diagnosis as well as providing the option of therapeutic treatment in the form of polypectomy, EMR or EMD. [92-99] The sensitivity of colonoscopy for adenomas ranges from 90% for large to 75% for smaller lesions [100] and its sensitivity for detecting colorectal cancer is greater than 90%. Whilst a more accurate investigation it does however have some negative aspects such as a higher rate of morbidity and mortality as compared with flexible sigmoidoscopy and an increased cost. The cost implication is generally attributed to the length of the procedure, necessity for sedation and thus monitoring and the expertise required to perform the test but there is also the need to provide the patient with full bowel preparation prior to undertaking the procedure, a factor that needs to be carefully evaluated in some individuals.

1.5.4 Radiological Imaging

The use of radiological procedures in the evaluation of the colon remains popular with modern techniques augmenting older practices. Double contrast barium enemas are of use in fully evaluating the colon in those unable to tolerate endoscopic techniques. While full bowel preparation may be required prior to the procedure being undertaken there is greater tolerance of the insufflation and contrast and there is little need for sedation. There are drawbacks to this method however as direct visualisation of the colonic mucosa is not obtained, as such the test has a lower sensitivity and specificity than colonoscopy detecting only 48% of polyps >10mm [101] with some studies identifying 'miss rates' of cancer up to 22.4% [102, 103]. An alternative to DCBE and colonoscopy in individuals not deemed fit is that of Virtual colonoscopy, a technique that utilises modern CT images in conjunction with intravenous contrast and CO2 insufflation per rectum to image the colon. Using complex software the images are able to be formatted allowing the intraluminal mucosa to be reconstructed in 3D. Studies have shown it to be accurate in detecting polyps >10mm in size although there is variation in the percentage accuracies based on seniority of reporting radiologist/technician and complexity of the scan. [104] [105-111]

1.6 Staging of Colorectal cancer

Staging of colorectal cancer is currently done via the TNM classification system developed by the American Joint committee on Cancer, assessing tumour depth, node status and metastatic disease [112]. This is commonly used in conjunction with the Dukes classification system, classifying the disease into A, B (B1, B2) [113], C1, C2 and D [114].

The TNM classification allows the disease to be staged (ranging from 0- 4) with various subdivisions based on TNM status. All of these classification systems are used to allow clinical planning of treatment and to aid in the overall prognosis of the disease.

1.7 Treatment of Colorectal Cancer

The preoperative staging of colorectal cancer is important as this affects the treatment pathway, more significantly at present with rectal cancer however with the use of preoperative chemoradiotherapy or short course radiotherapy.

1.7.1 Preoperative treatment

Specifically in rectal cancer there has been an increased use of pre-operative oncological treatments to optimise the patients before any surgical intervention is undertaken. This is in the form of short course radiotherapy or combined radiotherapy and chemotherapy which, has been shown in numerous studies to improve patient outcome and survival but is associated with slightly higher post-operative morbidity [115-120] [121]. The benefits of pre-operative chemotherapy in patients with colon cancer have not been fully evaluated at this time however there are on-going studies assessing the benefits of this in the patient cohort.

1.7.2 Surgical Treatment

Surgery is the mainstay treatment option for cancers of the colon and rectum with many approaches to the segmental resection of the colon. Surgical techniques vary between Open and Laparoscopic approaches, with studies demonstrating no oncological difference between the two [122] [123] [124].

TME dissection of rectal tumour has been shown to have improved oncological outcomes[125] [126] and is the widely accepted approach for the removal of rectal tumours. TEMS procedures have been used in the treatment of small rectal tumours [127] [128] [129] although this has been in limited cases and the long term outcomes have not been assessed by a large study at this time.

The specific operation that is undertaken is dependent on numerous factors such as stage of disease, patient co-morbidities and location of tumour. The operations can either be curative in intent or palliative, resecting the necessary amount of colon or rectum to ensure good vascularity in the remnants for anastomosis. Above the peritoneal reflection commonly performed procedures are right hemicolectomy, extended right hemicolectomy, left hemicolectomy and sigmoid colectomy. Below the peritoneal reflection for tumours of the upper, mid and at times lower rectum an anterior resection is performed, ensuring the distal remnant is of sufficient length to allow a healthy anastomosis. Should this not be the case then abdomino perineal excision of the rectum can be performed, this non-sphincter saving procedure leaves the individual with a permanent end colostomy. If palliation is considered then it may be appropriate to defunction the patient and leave the tumour in situ thus relieving any obstruction that may be occurring but reducing the operative morbidity and mortality.

1.7.3 Adjuvant treatment

Post-operative treatment is determined by the histological stage of the specimen in association with the radiological staging of the disease. Stage I disease has a 95% 5 year survival [130] however the presence of lymph node involvement (Stage III disease) reduced 5 year survival to between 30-60% with surgery alone. This survival rate can be improved by 10-15% with the addition of chemotherapy, for which there are many combinations however the main stay remains 5-FU based treatments [131] .

The role of chemotherapy in Stage II disease is becoming more popular, especially if there are adverse prognostic factors within the specimen such as vascular invasion. Trials have shown an improved survival rate [131] with the use of chemotherapy in this cohort however the risks and benefits in this treatment group need discussing on an individual patient basis.

1.8 Referral Pathways

1.8.1 Introduction

Referral pathways for those suspected of having colorectal cancer range from direct primary care referrals including both 2 week wait and routine OPD referrals, interspeciality referrals due to incidental findings during the investigation of other complaints, acute referrals generated from emergency admissions and referrals from screening programmes.

1.8.2 Presentation from primary care

1.8.2.1 2 Week wait referrals

In 2002 guidelines were published by the Association of Coloproctology of Great Britain and Ireland at the request of the Department of Health. The aim of this guidance was to assist those in primary care to refer the most appropriate individuals under the ‘2 week wait’ process assisting in allowing everyone with suspected cancer to be seen by a specialist within two weeks. By defining the criteria it was important to ensure that only those at high risk of colorectal cancer would be identified and therefore referred on the urgent two-week basis. Key facts highlighted in this process were that whilst patients with lower gastrointestinal symptoms are recommended to be referred for prompt investigation in hospital there is no evidence that a delay of two or three months after the onset of symptoms is likely to adversely affect the outcome [132] [133]. Adverse outcomes of investigating all of those with vague symptoms have also been explored on a physical and psychological level [134] [135].

Whilst patient symptoms are of importance other attributes have also been shown to aid in the diagnosis. 85% of colorectal cancers are in the age group of those over 60 with only 1.5% being in those less than 40 years of age. This variation in prevalence of the different age groups alters the management of these individuals, with those over 60 possibly being investigated with more subtle symptoms than someone under 40.

The development of guidelines remains important due to the high prevalence of rectal bleeding within the community [136] [64] [137] previous studies have shown an increased risk of cancer and rectal bleeding occurs in association with alteration in bowel habit of giving a predictive value of 12% for colorectal cancer [73]

The risk of cancer in patients suffering from rectal bleeding varies in accordance with their population. The prevalence within the community is one in 700, in primary care this increases to 1 in 30 and for those in hospital surgical clinics one in 16

It was suggested that 85 to 90% of all patients with symptoms present in table 1 presenting via the two-week wait referral system would be positive for colorectal cancer. The ACPGBI at the time also emphasised the importance of identifying those at low risk of rectal cancer who experienced symptoms as defined in table 2. It was felt that these individuals could be observed and referred as routine patients to specialist services.

Table 1.1: Department of Health higher risk criteria

Criteria	Age threshold
Rectal bleeding with a change in bowel habit to loose stools and/or increased frequency of defecation persistent for 6 weeks	All Ages
Change in bowel habit as above without rectal bleeding and persistent for 6 weeks	Over 60 Years
Recta bleeding persistently without anal symptoms	Over 60 years
A Definite palpable right sided abdominal mass	All Ages
A definite palpable rectal mass	All Ages
Unexplained iron deficiency anaemia Below 11g/dl in men Below 10g/dl in women	 All Ages Post menopausal women

Table 1.2: Low risk criteria (ASGBI guidelines)

Criteria	Age Threshold
Rectal Bleeding WITH anal symptoms	All Ages
Rectal bleeding with an obvious external cause for bleeding on simple examination of the perineum. E.g. an anal fissure, thrombosed or external pile and rectal prolapse	All Ages
Transient changes in bowel habit, particularly to harder stools and/or decreased frequency of defecation	All Ages
Abdominal pain as a single symptom WITHOUT other high risk/age/symptoms/sign profiles, an abdominal mass, an iron deficiency anaemia or intestinal obstruction	All Ages

If patients presented to the GP with any of these symptoms they could then be referred to a hospital specialist and seen within a set two-week time period. This particular referral pathway became known as the ‘2 week wait’ and, as with most guidelines has been subject to revision since its introduction. The most recent alteration occurred in 2005 with the introduction of the 31/62 pathway [138]

The aim of the guidelines and pre determined referral criteria was to 'identify up to 90% of patients with colorectal cancer[139]. This figure of 90% however, over the years that the system has been in place has not been emulated in clinical practice. Recent studies have demonstrated that only 10% of patients referred under the two-week wait criteria have colorectal cancer, with a review article examining the subject finding an average of 10.3% when comparing six different studies [142][141]this accounts for only approximately 30% of those with the disease [140] approximately one quarter of those with colorectal cancer continue to present acutely with the disease, with the remainder presenting via alternative routes[141]. Reasons for the variation in rates of presentation are multiple; some have advocated that pressure within a primary care setting, with an average of seven minutes per consultation makes accurate referral of only high risk individuals unachievable. Other factors that must be taken into account are patients themselves, some failing to seek medical advice for their symptoms until they present to acute services and others who find it too embarrassing [142-144].

1.8.2.2 Routine OPD referrals

There are several variations in the approach to the 2WW process, the more traditional being dedicated clinic time, in which patients are reviewed by a consultant, one of their team or a nurse specialist. Following such a consultation and, based on patient history and clinical signs further investigations may be undertaken. At this point, unless a definitive sign is found at examination, a rectal lesion for example that can be biopsied in an outpatient setting, a further delay will occur prior to definitive histological diagnosis. This, as already alluded to, may not clinically bear any significance to outcome of disease, but will undoubtedly have some psychological implications for the individual [145].

1.8.2.3 Straight to test

The above scenario has evolved quite significantly over recent years with a 'push' towards a 'straight to test' situation. These new routes of access have taken many different guises, but all have an underlying theme of diagnostic test at first hospital visit. Policies adopted range from the use of dedicated 2 week wait clinics, where medical staff not only take a thorough history and examine the patient but also undertake an endoscopic examination at this first instance (generally a flexible sigmoidoscopy). Whilst not the 'gold standard' the benefits of a flexible sigmoidoscopy will be examined further later in this chapter. This method is not far from the traditional referral route and, allows clinical evaluation by a hospital specialist as well as a potentially diagnostic examination to occur simultaneously. Whilst remaining labour intensive and somewhat costly in terms of resources (the need for dedicated sessions in endoscopy and the trained staff) it is beneficial in reducing anxiety and definitive diagnosis of a range of conditions, not only colorectal cancer.

The use of non-medical screening for 2WW referrals has been evaluated by Hemingway et al [140]. This particular method utilised a pre-determined protocol, based on the ACPGBI guidelines and agreed by both specialists within the hospital and the local primary care Services. (Table 1.3)

Table 1.3: Non medical screening of 2WW patients

Presenting Symptoms	**Age**	**Diagnostic Intervention**
Rectal bleeding with change in bowel habit for at least 6 weeks	All Ages	Same day fibre optic sigmoidoscopy and barium enema Colonoscopy, CT Colonography
Rectal Bleeding without anal symptoms	Recommended over 60 Discretionary over 45	Fibreoptic Sigmoidoscopy
Change in bowel habit; increased frequency and/or looser stools for at least 6 weeks	Over 60	Barium Enema CT Colonography Colonoscopy
Palpable abdominal mass	All Ages	USS / CT Scan
Palpable intraluminal rectal mass	All Ages	Sigmoidoscopy and biopsy
Unexplained Iron Deficiency Anaemia HB <11g/dl in men Hb <10g/dl in post menopausal women	All Ages	Barium Enema, CT colonography, Colonoscopy

The outcome of the above interventions achieved 95% of all diagnoses, not just colorectal cancer within a 31 day period from initial referral and reduced the number of patient clinical consultations. Interestingly, and as has been shown in previous examples the detection rate of colorectal cancer in this population was 12%.

A novel but somewhat more expensive route of referral was assessed by Maruthachalam et al, using direct access colonoscopy from primary care as part of the 2ww assessment process [146]. They utilised the DOH high risk criteria in conjunction with a specialised proforma to allow GP's to refer patients directly for colonoscopy or for an urgent out patient appointment. The study demonstrated a reduction in time of diagnosis to 14 days from point of referral and a high level (98%) of patient satisfaction but cancer detection rates were comparable to more traditional 2ww referral routes. Whilst both a reduction in time to diagnosis and high level of patient satisfaction may reduce patient anxiety there is no evidence to support a reduction from 31 to 14 days has any benefit to patient survival. This is compounded by the associated risks of mortality and morbidity that go with colonoscopy and the overall expense of this referral process.

Whilst the above alternative methods for patient access to rapid access services are aimed at improving performance, all appear to have a similar rate of detection. The use of flexible sigmoidoscopy, a procedure that is regarded as being quicker and generally less technically demanding than colonoscopy is, for these reasons easier to access. Whilst it may only (on average) reach the splenic flexure, statistically this should detect over 70% of colon cancers [147] and can be used to detect and remove 70% of adenomas [148]. It, like colonoscopy however is associated with risks of

mortality and morbidity, even if they are small. The ‘gold standard’ test for colonic evaluation is without doubt colonoscopy. It is however also the most expensive method and carries with it risks of mortality and morbidity considerably higher than flexible sigmoidoscopy.

Taking the above information into account and examining the evidence from the numerous trials that have evaluated the 2WW process since its creation in 2000 there appears to be a constant theme throughout. 10% appears to be the recurrent level of colorectal cancer detection with this particular route of referral, a far cry from the original 90% detection rate [139].A great amount of research and one can only presume resources have gone into the improvement of this practice but these figures remain around the same mark. The utilisation of alternative techniques in increasing the speed of diagnosis appears to have altered little other than achieving pre-set government targets. Taking this as point in case alternative methods of detection have been evaluated and in certain cases put into practice as illustrated below.

1.8.3 National Bowel Cancer Screening programme

The philosophy of early patient identification is one that remains important for any malignancy, and whilst the 2 week wait may not have achieved its 90% target other methods of early cancer identification have been explored. The most recent of these is the Bowel Screening Programme, something that will be explored in further detail later in this chapter. Whatever method used in this process is however dependent on two main factors, One being patient participation, something without which stops any process before it has started, the second being the identification of factors pertaining to risk for colorectal cancer.

Mandel et al. [80, 149] demonstrated a high level of screening compliance with a reduction in mortality at 18 years, improved survival and detection of cancer at an earlier stage. Studies undertaken in both Denmark [150] and the UK [151, 152] also demonstrated a survival advantage with community screening using FOBT. This investigation forms the backbone of the recently introduced UK screening programme on a biennial basis in those individuals between the ages of 60 and 69

1.8.4 Non two week wait colorectal cancer detection

An increased urgency in investigation of those referred via the 2WW system has led to an increase in demand for primary diagnostic tests such as flexible sigmoidoscopy or colonoscopy. Whilst these investigations may duly be warranted by the individuals concerned, due to the method of referral all must be undertaken within a pre defined timescale, leading to a diagnosis at 31 days from referral. With only 10% of these cases having colorectal cancer, in the region of 25-30% of all colorectal cancers over a year, 70% of colorectal cancers will present via alternative routes, and thus have to access these scarce resources in an alternative manner. Bowel cancer screening is one method of reducing deaths from colorectal cancer within the UK and is currently being rolled out. The aim is to detect bowel cancer at an early stage [153] and is supported by a 16 percent reduction in death by colorectal cancer. Currently the process of bowel cancer screening is underway with all those aged between 60 and 69 being invited to participate on a two yearly basis via letter and information booklet. This will be followed by a Faecal Occult Blood test which has been quoted in some literature to have a 60% compliance rate [150] and as high as 81% in others. For polyps ≥ 1 cm, sensitivity estimates range from 13 to 31% [154] [155]. Unlike sensitivity, the specificity of Hemoccult II is relatively good, ranging from 98–99% in large screening studies [155-158]. If a patient has a positive FOB test then they will proceed to further diagnostic tests. This process, whilst of benefit albeit to a small age range has some negative aspects, specifically the psychological anguish faced by those with a positive FOB and normal colonoscopy, leaving the individual and in some cases GP's wondering how best to further investigate this

finding. These criticisms apart there is the undoubted positive aspect of earlier cancer detection as some of those who agree to screening will harbour a colorectal cancer and be asymptomatic from it, thus this modality of cancer detection is likely to grow in popularity over time.

1.8.5 Success of the 2ww system

1.8.5.1 How the two-week wait criteria have worked

Since its introduction in 2000 the two-week wait pathway has been intensively monitored with waiting times and diagnostic delay being comprehensively audited [159] . Not ignoring the benefits of early cancer detection, this process and pressure is only on clinics but also on diagnostic services within the hospital setting. With the detection rate of 10% the economic benefit of the system must be questioned. In addition to this further pressure was added in 2005 with the diagnostic and treatment targets 31/62 days respectively [138]. Chohan et al demonstrated that 92% of those with colorectal cancer presented with symptoms that have filled the high risk criteria [160]. The appropriateness of those referred under the two week rule has at times been questioned [161, 162]. Discrepancies have also been illustrated between referral letters and symptoms elicited within clinics [163].

Detailed analysis of referral criteria were undertaken by Flashman et al in which they reviewed all patients diagnosed with cancer in a 1 year period, a total of 249 individuals. 41% of their cases were assessed in the two week wait clinics which on analysis proved to be a statistically higher diagnostic yield as opposed to routine clinics. They further showed that 39% of all those referred under 2 week wait criteria failed to fulfil at least one of the high risk criteria [164]. Although this study was carried out shortly after the introduction of the two week wait rule it highlights the fact that more than 50% of cancers in that institution still presented via alternative

routes, something that has been demonstrated in studies carried out since this time. Promisingly, however, as far as the high risk guidelines perform they found that 85% of all cancers presenting to outpatients had at least one of the high risk criteria. The point outlined above probably represents the main failing of the 2 week wait system, not the fact that the formulated guidelines are inappropriate, but rather that the implementation and use of these guidelines has been flawed in some respects. Perhaps improved adherence to the criteria within a primary care setting can address this problem [67, 165].

1.8.5.2 Effect on Survival

Whilst it is generally accepted that the two-week wait criteria have assisted in the identification of those with colorectal cancer the relationship of those to overall survival has not been ascertained. One study has shown more advanced disease to be more likely in those referred under the two-week wait system [163]. A review evaluating the impact of intensive follow-up on long-term survival reported a mean of 24 months following surgery for a relapse to occur [166]. Walsh et al did not demonstrate any statistically significant medium-term survival benefit in those patients presenting via the acute pathway[167]. In a similar study Bevis et al, who studied referral source in relation to stage of disease found that whilst a significant delay was recorded in time to see a specialist and initiation of treatment, no association with advanced disease or reduction in curative surgery was found [168].

Whether the two week wait system has grossly affected outcome in colorectal cancer may be unclear but it has provided primary care with a dedicated rapid access point to specialist hospital services. This however does come at a cost and, whilst open to abuse from certain quarters is a route of referral that is constantly under review, with additional methods of detection, such as bowel cancer screening programmes coming into operation.

1.8.6 Alternative detection methods

Alternatives to both screening and the 2WW system have been under evaluation since the late 1990's, one particular area of study has been the use of patient questionnaires. Logically this would seem a sensible avenue to explore, given the detection rate of the 2WW system and a paper by Selvachandran evaluated the accuracy of this method [169]. They assess all hospital referrals from primary care with distal colonic symptoms and provided these patients with a questionnaire to complete prior to hospital attendance, grading each according to a weighted scoring system (known as the Selva score). The patient questionnaire was extensive and probed not only a history of patient symptoms but also family history relevant to colorectal cancer as already explained earlier in this chapter. Whilst some of the questions were similar to the ACPGBI guidelines on the whole the questionnaire was more comprehensive in relation to a true patient history. They demonstrated that the patient questionnaire, in correlation with a weighted numerical scoring system allowed accurate assessment of all referrals from primary care, prioritising those with symptoms indicative of colorectal cancer. Whilst this pathway has not been widely accepted into clinical practice it highlights that a system able to comprehensively assess all colorectal referrals and prioritise those at risk of colorectal cancer is achievable. A more recent study undertaken in Leicestershire further assessed the feasibility of the afore mentioned scoring system and compared it to the 2WW system in practice. The showed that the scoring system, when used with a cut off value of 70 had a similar sensitivity but greater specificity in detecting colorectal cancer when compared with the 2WW system [170].

Evidence has clearly shown that the 2WW pathway has been unable to reach its intended target of 90% detection of those with colorectal cancer. Whether this is due to system abuse, patient history at initial presentation or an inherent wish for all patients to be seen as soon as possible is unclear. What has been shown is that with all the monetary investment and time that is ploughed into the 2WW system, the detection rate has remained a constant and, furthermore little has changed in terms of patient outcome. It would be correct to assume that the concept is very genuine but one has to question whether this concept has been somewhat muddied by political interference and the dreaded word 'targets'. The notion of litigation within the medical sector is also something that has increased exponentially over the past decade and it would be ignorant not to assume that this has not played a role in both patient referral and investigation requests

What has been demonstrated by studies is that the use of patient targeted questionnaires can increase the sensitivity and maintain a comparable specificity to the current system. Obtaining a sensitivity and specificity of 100 percent is, in all practical terms an impossible feat to obtain but it should be possible to increase current practice levels utilising the afore mentioned targeted questionnaires. Whilst the concept of a weighted, questionnaire based patient scoring system may appear unattractive to clinicians, the available evidence at this time illustrates that it is an area worthy of further investigation. No system will achieve perfection, but with advances in technology and novel techniques in data exploration and analysis it should be possible to develop a system that achieves an improved sensitivity and specificity compared to current practice

1.9 Data Mining

1.9.1 Introduction

Over the past two decades there has been a rapid increase in the amount of medical data available for research. This can be attributed not only to advances in new molecular genetics techniques such as protein identification genomic sequencing but also due to the increased use of computerised technology within hospital setting. Digitisation of medical information such as blood results, radiological investigations and patient information have resulted in vast quantities of data specific to patient care being available for research. Whilst this data exists its rate of accumulation is far greater than the rate of interpretation for research purposes.

In order to utilise this information as effectively as possible new techniques within medicine have been developed such as data mining, text mining and knowledge management. Whilst these processes are used effectively in government and business settings [171-174] the uptake from a medical point of view has been somewhat slower.

Data mining is primarily a knowledge discovery process, analysing given set of data in order to identify potentially novel and useful patterns [175]. Techniques utilised range from Bayesian models to artificial neural networks and are used to illustrate patterns within the data that are unknown and unrecognised to the users [176, 177] Whilst data mining is an important component in the analysis of the status previously unrecognised patterns it can be used in conjunction with text mining, with an aim to extract information from textual data documents [178, 179] and also as part of generalised knowledge management [179]

Since the advent of the first computer there has been an array of systems built for engineering, business decision making and medical diagnoses [180]. The primary drawback to the vast majority of these systems is the manual acquisition of knowledge which, is an exceedingly labour-intensive and time-consuming process. These systems also draw heavily on human experts for data analysis. To try to address this somewhat lengthy process machine learning has been developed to acquire this knowledge automatically. This process has been defined as "any process by which a system improves its performance." [181] something which in medical fields would be classed as data analysis and primarily done using Bayesian statistics.

Data is omnipresent and whilst the amount of data that is collected can to the human eye appear overwhelming within a substantial quantity of it lies valuable information. The extraction of this information, given the colossal amount of variables is something that requires the assistance of an automated computational process. The technique of data mining is primarily about problem solving by analysing data already present within a database. The process searches the characteristics within the data set allowing distinguishing characteristics to be extracted. It is a process of discovering data patterns leading to a meaningful outcome measure. The ultimate goal of data mining is prediction, usually consisting of three distinct stages, and initial exploration, model building and deployment.

Exploration of data combines data preparation and some preliminary feature selection. This process may involve the cleaning of data, transformation of data and selection of subsets. Depending upon the nature of the problem further analyses may

be required using a variety of statistical and graphical methods allowing identification of the most relevant variables and to provide information on the complexity and nature of the models. Following appropriate data exploration various data mining models can be considered, the most appropriate model being chosen based on outcome prediction. A variety of techniques exist to achieve this step many based on competitive evaluation of models, a process essentially of comparing the same data set on numerous different models and comparing the performance. These techniques include bagging, boosting, stacking and meta-learning. Following successful modelling, deployment of the appropriately identified data mining techniques is undertaken. This applied to a virgin data set and is utilised to generate predictors of expected outcome. The most important difference between data mining and exploratory data analysis and data mining focuses more on application than on the basic underlying phenomena. Data mining is therefore less concerned with the identification of relations between variables, focusing more on the production of a solution that can be utilised in the generation of accurate predictors. A black box approach is therefore generally accepted utilising not only traditional techniques but also techniques such as artificial neural networks

1.9.2 Neural Networks

McCulloch and Pitt in the 1940s [182] were the first to exploit computational mathematical models, consisting of a single neurone utilised to construct a network able to analyse basic Boolean logical functions. While a fundamentally important step these networks in conjunction with very early designs computers were too inflexible to be used as cognitive models. Most current neural networks have learning rules arising from statistical correlation analysis and gradient descent search procedures. In addition, work by Hebb [183] using learning rules that incrementally modify the connection weights based on the ON/OFF allocation to the two connected nodes is still used with some modifications.

It was not until the 1950s when Rosenblatt, a psychologist [184] added to the development of 'artificial neural networks' viewing the brain as an associate of learning stimuli and trying to simulate this electronically. To achieve this he postulated a new class of networks based on the 'perceptron' neural model and utilising association learning rules based on descent gradients, in its simplest form three layers of designated cells. Learning is undertaken as source material in the first layer of cells connected 'randomly' to the central otherwise known as 'association'. The output response is not only influenced by positive neuronal association but also from inhibitory association as the result of a lack of input. This process enabled Rosenblatt to demonstrate that such runs have the ability to not only generalise but were also capable of learning, using pre-entered data. This method of learning further subdivided into two categories, forced learning and competitive learning. Forced learning utilises a specific pattern of inputs to activate a particular response, allowing the neurone to grow in strength with recurrent cycles of exposure, ultimately to the point where the response neurone is activated appropriately.

Competitive learning conversely utilises the continued activation of association units whilst various responses gain in strength. This allows for increased sensitivity to particular input types.

The further development of Neural Networks with ADALINE (adaptive linear neurone) followed by multiple network ADALINE's known as MADALINE (multiple adalines) by Hoff and Widrow took the concept further [185, 186]. These methods differed from Rosenblatts work by using a simple neural element in addition to developing the least mean square supervised learning procedure.

The theory of associated memory in the 1970s was once again an important step in the development of neural networks. This theory is based on the stored pairing of patterns, the presentation of one pattern evoking the associated pattern therefore allowing the content to be regarded as addressable. Further work with linear associated models allowed within the network's output is preventing infinite growth as the model strives to identify a solution.

Further modelling of networks was undertaken by Fukushima [187, 188] based on biological visual systems. These feed forward networks learn through both supervised and unsupervised methods, utilising connected layers such that vague features can be recognised and thus cumulatively combined into an identifiable output object. Such methods are used routinely for the recognition of handwriting.

Werbos, [189] with the development of the back propagation algorithm made one of the most important developments in neural network research. This algorithm has the ability to adjust the weights in a multi layer feed forward network. The technique has since been described further and is vital in the use of artificial neural networks to solve nonlinear problems [190-192].

The structure of a NN was distinguished by Rumelhart and McClelland [193] as:

- a set of processing units
- a state of activation for each unit
- connections between units, defined by a weight that effects the output signal
- a propagation rule
- an activation function
- an external input
- a learning rule
- a working environment

Within the network, units can be further defined as input units (i.e. receive input data from external source), hidden units (input and output signals are 'hidden' within the network) and output units. Units are connected such that the total input is 'weighted' via a mathematical rule before a 'threshold' is reached and the unit fires in either a linear, semi-linear, sigmoid or hyperbolic tangential function.

1.9.2.1 Network Topology

Connection patterns within networks fall into two categories, the Feed-Forward network where data is processed over multiple layers in a forward direction only with no feedback connections and the Recurrent network where feedback connections are used. The Perceptron and Adaline networks constitute feed forward networks whereas recurrent networks have been presented by Anderson and Hopfield [194, 195] .

1.9.2.2 Network training

Training of networks can be either by a priori knowledge or providing teaching sets of data and allowing the network to evolve and thus alter the weights according to the learning rule. This can be undertaken in a supervised or unsupervised manner depending on whether the input and output data is provided or simply input data inserted. Training and adjustment of weights is then undertaken, commonly using a variant of Hebbian learning or occasionally using the delta or Widrow-Hoff rule.

1.9.2.3 Network types

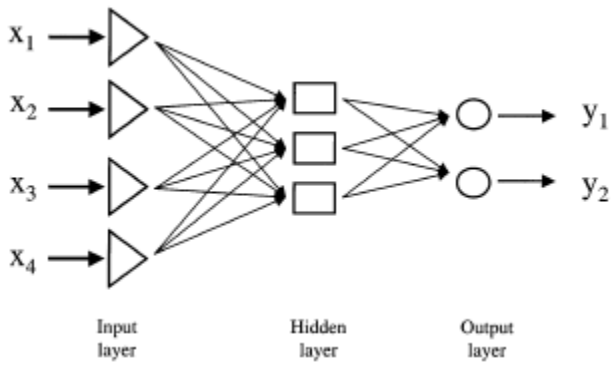
1.9.2.3.1 Multilayer feed forward networks

These networks have a number of layers, with the hidden layers taking input from the previous layer and sending it directly to the following layer (see diagram). No connections exist within the layers and the activation is related to the outcome of the function attributed to the unit based on the weighted inputs.

$$y_{k(t+1)} = F_{k(s_k\ (t))= F_k}\left(\sum_j \omega_{jk}\ (t)\ y_j\ (t) + \theta_k(t)\right)$$

.

Back-propagation learning rules can be applied to allow the network to adjust the weights within the hidden layer and this improves the functionality of the network. Whilst this rule can be applied to networks of any number of layers it has been shown that provided the activation functions are non linear, only one layer of hidden units is required [196, 197].



Feedforward Network

1.9.2.3.2 Recurrent networks

Recurrent networks differ from feed forward networks as they allow connections between the hidden units. These can be based on reaching a stable point (attractor based) or ones where a learning rule is used after each propagation is performed. Examples of this type of network are the Jordan Network [198] where output unit values are fed back into the input layer as ‘state units’ and the Hopfield Network [195] which consist of a set of interconnected neurones which update their activation values asynchronously and independently of the other neurones with binary activation values.

1.9.2.3.3 Radial Basis Function Networks

These networks use a radial base function to measure the distance between unit points and the centre resulting in a Gaussian function [199-201]. The hidden layer of units models the bell shaped response surface and as the functions are non-linear, more than one hidden layer is unnecessary. Whilst these networks have the advantage of modelling nonlinear functions with only one hidden layer they require the number of radial units to be decided initially with the centres and deviations being set.

1.9.2.3.4 Self Organising Feature Map (SOFM)

These networks are designed for unsupervised learning [200, 202, 203] allowing them to recognise clusters of data and relate these clusters to each other. They only have two layers, an input layer and output layer, also known as a topological map.

1.9.2.4 Effect of Hidden Units

Variation in the number of hidden units is undertaken to identify the best fitting network for the input and output data. A large number of hidden units however can impede the ability of the network, causing 'overtraining'. This is an effect whereby the network trains itself to fit the 'noise' of the training data rather than approximating it, resulting in a high error rate when the network is used with the test set.

1.9.2.5 Effect of Hidden layers

There is no 'hard and fast' rule regarding the number of hidden layers required in a neural network. Linear models and even mildly non-linear models have been shown to have better generalisation with no hidden layers [204]. Auer has also advocated the use of a single layer of weight in association with a parallel delta rule on grounds that it is a more realistic alternative in the modelling of biological circuits [205]. Conversely Sontag suggest two hidden layers in multi-layered perceptron's with Heaviside/step/threshold functions and one hidden layer in MLP's with a variety of non-linear activation functions are a more appropriate modelling method. [197].

1.9.3 Decision Trees

The origins of decision trees can be traced back to the late 1950s and the work of Hunt, Quinlan and Breiman [206-208]. A decision tree has three main components: nodes, arcs and leaves, nodes containing a feature attribute, arcs being labelled with a feature value and the Leaf labelled with a class or category. Most decision trees use a top-down algorithm i.e. from the branch to the leaf. In addition a technique used as pruning is used to simplify decision tree by removing useless information. The structure of the decision tree allows it to be easily converted to a classification rule

Amongst symbolic learning and rule induction techniques learning by example is shown to be the most promising approach for data mining. The concept behind this technique is the application of an algorithm that tries to best describe numerous classes within a training example. The ID3 decision tree algorithm [209] and the more recent variation C4 .5 [210] are the most widely utilised symbolic learning techniques. These methods use a decision tree and attempt to classify all objects correctly; finding the attribute the most appropriately splits data into different classes of information uncertainty. Once all attributes have been used the algorithm displays the results as a decision tree. Whilst these techniques may not be as powerful as neural networks or support vector machines with their accuracy they are more efficient and produce outcomes that are easier to interpret.

1.9.3.1 Classification of machine learning techniques

Classification of the five main paradigms of machine learning occurred in 2004 [176]. These five categories were probabilistic and statistical models, symbolic learning and rule induction, neural networks, evolution based models and analytical learning and fuzzy logic.

Statistical Models

Probabilistic models and techniques have the strongest foundation of all methods for data analysis. The statistical analysis using popular techniques such as regression or multidimensional scaling commonly used with the medical research and in papers that have utilised data mining techniques previously have been used as benchmarks for comparison.

Bayesian classification

With its roots in pattern recognition research, the Bayesian model [211] is likely to be the most popular probabilistic model utilised in medical research. Used to classify objects into predefined groups utilising specific features it defines the likelihood of each class, each feature and each feature giving each class-based on training data. Using these predetermined probabilities when a new instance is encountered the model attempts to classify it accordingly [212]. Variants of this model exist, specifically one called naive Bayes. In this variant all features are deemed mutually

independent in each class. Whilst Bayesian models have been widely used in medical data mining, research involving the models has been developed in more recent years. One such model is called a support vector machines [213] which uses statistical learning to identify and model the best separates classes within the data. This particular technique has been shown to perform well in document classification [214] and is also being used in medical research to classify disease states or identify specific diagnoses utilising patient information.

Evolution based algorithms are analogous to Darwinian survival of the fittest and analogies of other natural processes. Genetic algorithms [215] are based on genetic principle, with population data undergoing a set of operations known as crossover and mutation. Crossover is a process aimed specifically at exploitation while mutation is aimed at exploration of the data. As with the Darwinian theory of evolution, there is a continuous filtering process selecting better solutions followed by repetition of the above sequence in order to produce a further generation, with selection of the best solution undertaken once again. Such algorithms are of great use in medical research being one of the most robust techniques for feature selection due to their global search capabilities.

Analytical learning utilises logical rules on which it performs reasoning in order to search for proofs. These proofs can then be arranged into more complex rules in order to solve similar problems. Whilst these traditional learning systems rely upon computing rules generally there is no distinction between the values and classes in the real world. This has been tackled by proposing fuzzy Logic Systems, allowing true **or** false values operate over numbers from 0 to 1[216].

1.9.3.2 Methods of evaluation

It is necessary to thoroughly evaluate any data mining system before it is put into practice. In cases of limited data availability estimating system accuracy is difficult to undertake [217] . Several methods are used for the evaluation process including bootstrap something, cross validation, holdout sampling and leave one out [218, 219]. Each of the above methods has both strengths and weaknesses, many studies have compared in terms of their accuracy.

The bootstrap method takes an independent and random sample from the original data set. These samples were then used to train the system allowing fresh data (in the remaining samples) to be used to test the system [219].

Cross validation randomly divided the data into X subsets all of roughly equal size, generally this is 10 subsets and a process called ' tenfold cross validation ' is undertaken. In this process training and testing is undertaking with 10 iterations, nine subsets of data used for training and the 10th remaining subsets used for testing. This process is performed in rotation with the accuracy of the system being the average accuracy over 10 cycles.

The holdout method splits the dataset into two subsets, the training set and the testing set. Generally speaking two thirds of the data is put forward as the training set with the remaining third being used for testing. Once trained the system will use the testing set to predict the outcome and accuracy determined by comparison with the real output value. Leave one out is a variant of cross validation whereby the original dataset is split into multiple subsets of equal size. . Training is undertaken for N

iterations and as before n -1 instances are used for training purposes, the remaining used for testing. As before the accuracy is the average over the N cycles.

Studies have compared the accuracy of the above methods however the ability to implement them should be taken into account. The easiest method to implement is that of holdout sampling however the training set and testing steps are not mutually independent. With up to one third of the data being removed from training the efficiency of this technique has been questioned [217] . The most unbiased method has been shown to the leave one out [220, 221] but its estimations have high variances and this is more pronounced with a small dataset. Independent comparison of all methods showed tenfold cross validation to be the most appropriate model selection.

Rule induction

Rule induction systems are based on a process of if -- then rules, if X then Y

For an example to be correctly predicted by the above rule, its attributes must fulfil the ‘if’ conditions.

Decision tree induction

Decision trees are constructed by choosing the most informative attribute of each step. Construction stops are when all data examples in a specific node are of the same class. This mode is known as a leaf and is labelled by the value of the class variable. Ideally each leaf has one class name label however some leaves may be empty if no training examples have attribute values leading to it all can be labelled by more than one class name. The most important feature for handling noisy data is a mechanism known as tree pruning. This is aimed at producing decision trees that do not over fit potentially erroneous data. Unreliable parts of the tree are eliminated therefore increasing classification accuracy of the tree on unseen cases. Techniques pruning are based on expected/predicted classification accuracy or expected classification error[210].

Instance-based learning

Algorithms of this type require specific instances in order to perform classification tasks. These differ from the rule induction method which uses generalisations based upon if -- then rules. Instance-based learning on occasion is referred to as a lazy learning algorithm as they save part or the entire training set postponing inductive generalisation until the time of classification. They are based on the assumption that similar instances have similar classifications. As an algorithm they are derived from nearest neighbour classifier algorithms. In these algorithms all attributes are treated as a dimensional within a space with examples and specific points within this space.

1.9.4 Uses of Data Mining

Recent years have seen an increase in the use of artificial neural networks and support vector machines within the medical field. The use of data mining techniques within health care is augmented by their predictive power. Algorithms have the ability to learn from prior examples within a clinical dataset; they then have the ability to model often complex nonlinear relationships between variables. Such patterns may very well be unclear when other analytical techniques are undertaken. The most extensively used data mining technique in the medical field is that of classification, used to analyse various signals and their relationship with diseases or symptoms. Neural networks have been utilised to classify outcome in post-operative colorectal cancer patients [222] and also to classify lung sounds in two distinct categories to assist diagnoses [223] data mining has also been used to extract diagnostic rules for breast cancer data [224] and also to identify new medical knowledge [225].

This process utilises patient data and corresponding diagnoses, allowing data mining techniques to diagnose outcome in new cases. This is undertaken using a predefined set of examples with known classifications. Each example is described by a fixed collection of features (known as attributes) each attribute can be discreet or continuous data. To correctly classify new cases different data mining processes can take different approaches. Sets of symbolic rules to generalise training cases can be constructed, and further analysed for accuracy when used to predict outcome in each separate data cohort.

1.9.4.1 Data Mining in Medicine

The last decade has seen exponential increase in the use of data mining techniques in the field of medicine [226]. Initial studies into the use of artificial neural networks and medical purposes centred on the diagnosis of myocardial infarction[227]). Subsequent prospective studies have illustrated the ability of new networks to be able to outperform alternative computer packages statistical techniques and clinicians achieving sensitivities and specificity in excess of 95% [228]. Artificial neural networks have subsequently been both in the assessment of protein function [229, 230] used as supportive systems in the diagnosis of gastrointestinal bleeding, GORD [231], pancreatitis [232], obesity[233], pulmonary emboli[234], tuberculosis and cancer outcome. Within medical fields neural networks have further been utilised in imaging recognition[235], Cardiology [236, 237], Gastroenterology [238] Histopathology and cytology [239, 240]. The ability of neural networks at data recognition has also led to the development of their use in analysis of waveforms such as electrocardiograms, electromyograms and electro encephalograms. A further area within medicine that artificial neural networks have proven their worth is that of outcome prediction with the associated strengths and weaknesses of such practice[241]. The use of such techniques has been examined in many medical fields such as cardiac surgery [242-244], colorectal surgery [245, 246], anaesthesia [247] ,breast cancer [248-251] and oncology [252] .

The use of logistical regression has since the early 1990's been utilised in outcome prediction in a host of surgical specialities. Copeland's initial work [253] , based on the weighting of input factors followed by logistical regression analysis has been modified, initially be Prytherch to formulate the P possum model [254] and subsequently by a number of surgical specialities [255, 256]. Whilst the validation of these scoring systems has been addressed in a number of studies it consistently over compensates for mortality and morbidity in numerous circumstances.

Neural networks have been used in intensive care setting with cardiology patients to try and predict mortality in outcome. With the continued increase in delivery costs the state of rationing within the health system specifically with expensive services the ability of the physician to identify patients who would benefit most from treatment courses is important. This not only optimises outcome for individuals but equally reduces costs across the board and along with it wastage.

The first use of artificial neural networks in the area of chest pain was in 1989 [257]. It analysed 174 patients with anterior chest pain using a multi-layered network and categorised them into one of three diagnostic groups, high-risk low risk and non-cardiac. Another application utilised was based on a retrospective analysis of 356 patients admitted to the cardiac ICU. 120 of these had myocardial infarctions and the network was trained utilising back propagation on half of the patients with and without myocardial infarctionss prior to being tested on the remainder of the patients

who had not been exposed. Sensitivity for this procedure was 92% and the specificity 96% [228] [258]. Prior to this study the most accurate computer aided method of diagnosing myocardial infarction was 88% sensitivity and 74% specificity [259] A further study was undertaken independently analysing two types of network, one maximum likelihood and the other a least squares method.. Sensitivity specificity and accuracy were in the region of 86 to 80%. Further prospective analysis of 320 patients presenting with acute chest pain compared the diagnostic accuracy of the physician with that of the neural network. This demonstrated the physician's accuracy of sensitivity 78% specificity 85% and the network had a sensitivity of 97% and specificity of 96% respectively [227].

Possibly the most well known commercially available medical use for neural networks it the Papnet cervical smear programme. [260] This network has the ability to constantly assess cells taken from the cervix and assess for signs of precancerous or cancerous change. This has the benefit of allowing greater numbers of smears to be assessed and, in conjunction with clinical lab staff assessment allows more accurate assessment than human assessment alone. As with most cancers this is important as the early detection of cervical cancer allows prompt treatment and results in an almost 100% chance of cure.

Artificial Neural Networks have also been analysed in the prediction of cancer survival in both breast and colorectal cancer [245, 261]. In the case of colorectal cancer an improvement in predicting mortality was achieved with a neural network when compared to clinical estimation (90% vs. 79%). Similar work in breast cancer

using backpropogation proved a more accurate predictor of survival compared to clinical evaluation alone.

Further studies have evaluated the potential of Artificial Neural Networks at making a correct diagnosis. Fraser et al used radial base function networks to diagnose myocardial infarction, achieving sensitivities of 85.7% and specificity of 86.1%, results suggesting that this technique can accurately be used in clinical diagnosis [262]. Another study from Sweden trained and assessed a Neural Net in assessing an MI by examining the ECG and compared the results with an experienced cardiologist. Results demonstrated that in all but the most obvious MI the neural network was better at identifying abnormalities than the cardiologist [263].

Chu et al have utilised ANN techniques in the creation of a system to aid in managing those who present with GI bleeding to hospital. Pre-determining a set of input criteria they proceeded to assess the ability of a range of data mining and ANN algorithms in obtaining the correct diagnosis. Their results showed that whilst most of the computer models were effective, the RandomForest, a form of decision tree model proved to be the most accurate. [264] .

The above examples demonstrate that these ‘black box’ methods of data analysis are able to draw valid and accurate conclusions to clinical scenarios. What they also show is that the outcome is dependent on the data used to train the model and the data model itself. This was demonstrated most effectively by Chu, proving that whilst all of the algorithms were effective in prediction there was one that was more effective. Accepting the above points then means that it is necessary to carefully

collect data in the first instance, ensuring that you collate all variables that are deemed relevant to the outcome and then assess them in multiple data mining algorithms, a process facilitated with programmes such as WEKA [265].

1.10 Summary of Introduction

As explained in chapter one, the current technique for identifying those with colorectal cancer, the 2WW system, whilst providing a rapid access service only detects between 25 and 30 % of colorectal cancers, only 10 % of those seen via this pathway. Many alternatives in modality of assessment once referred have been explored such as telephonic triage, straight to test or the ability of the GP to send straight to colonoscopy. Once again, these improvements show similar detection rates, even though they do reduce patient waiting time. The basis of the 2WW system is a set of high risk criteria, identified by a panel of experts to be the most accurate way of a primary care physician identifying someone at risk of colorectal cancer. Evidence shows that these criteria, whilst indicative of colorectal cancer are also found in a wide range of other benign conditions, which are more prevalent in the community. There has been evidence to show that the use of a weighted scoring system can assist in increasing the diagnosis of those with colorectal cancer but this as yet has not been widely accepted or put into practice.

Data mining techniques are of benefit to medical practice and studies have shown that they can be accurate in both image assessment, prediction of survival and clinical diagnosis. Utilisation of this technique has been shown to be of clinical use in acute GI bleeding, facilitating diagnosis and destination of referral. Current high risk patient selection uses information obtained via the GP. By its nature this is second hand upon reaching the hospital and, as shown in studies already mentioned,

variations in GP interpretation of patient symptoms exist. Using the above data mining techniques, in conjunction with a pre-formulated questionnaire to explore the potential of creating an algorithm that is more accurate than the current process in identification of those with colorectal cancer or polyps would be worthy of assessment. The benefits of a successful algorithm are not only more accurate patient identification but on a wider spectrum would reduce the number of unnecessary 'urgent' appointments thus freeing up clinical staff to undertake alternative duties**.**

1.11 Hypothesis

The use of data knowledge discovery databases within industry and other medical areas indicates that it has functional uses in the detection of patterns within datasets that are not visible via standard statistical techniques. This may prove to be beneficial in the detection of patients with colorectal cancer therefore the aims of this thesis will be to:

1. Establish a prospective database of patient symptoms with basic demographic data and diagnostic outcome
2. Assess the data for referral patterns, symptoms associated with adenocarcinoma and symptoms associated with polyps as well as distribution of diagnoses within the data.
3. Construct a logistical regression model for the data, assessing the accuracy of fit
4. Using ANN, experiment with the datasets and different outcome classifications to determine the optimum model for outcome classification, altering the hidden layers and units.
5. Using DM techniques, model the data further to assess whether alternative methods provide a more optimal modelling technique for the data.
6. Compare the above techniques with that of two primary care physicians and 2 post CCT Colorectal surgeons.

Methods

2 Methods

2.2 Prospective Data Collection

The author (JC) gained approval from the local Ethics Committee and approval within the trust. All patients who were attending the '2ww' outpatient clinics at Castle Hill Hospital over a 12 month period were identified and invited to participate in the completion of an internal symptom questionnaire (Appendix A). The questionnaire is a linkert based questionnaire that covers most of the common symptoms that are seen in patients with colorectal carcinoma as per 1.1.4 with the addition of further information. Data collection was facilitated with the assistance of the colorectal nurse specialists (JE and MB) who aided clinic attendees in the completion of the questionnaires where necessary.

Only patients attending the '2ww' clinics were included in this data collection as they were deemed to have fulfilled the current referral criteria for this route. The collected data was entered into an Access database (Microsoft, Seattle USA) and Excel spread sheet (Microsoft, Seattle USA).

Those attending clinic were then investigated as per local practice via Colonosocpy, BE, CT or MRI to determine the cause of their symptoms before attending a surgical outpatient clinic for review by one of the Colorectal consultants (JEH, JG, JRTM, RB, KC) or one of their team. All investigation results were reported or undertaken by appropriately qualified medical personnel working for Hull and East Yorkshire NHS trust. The diagnosis made at this clinic was taken as the final diagnosis and was

retrieved and correlated with the completed questionnaire. Where available the Haemoglobin of the patient at time of referral was also recorded

Once correlation of the questionnaire to the diagnosis was completed the data was anonymised by the author with each dataset given a unique reference number.

2.3 Referral pattern analysis

In October 2006 the 2ww referral pathway for suspected lower GI cancers was changed to facilitate earlier diagnosis or discharge. This resulted in dedicated 2ww referral clinics led by the colorectal specialist nurses (JE, MB and MH). These clinics comprise of initial clinical assessment and examination followed by flexible sigmoidoscopy at this initial visit. Following initial assessment and flexible sigmoidoscopy further investigations were initiated as clinically indicated prior to the patietns being reviewed by a consultant with all results.

Data collected from this study was analysed, looking at outcomes such as fulfilment of referral criteria, rates of anaemia, factors associated with polyps and factors associated with adenocarcinoma.

2.4 Data Cleaning

All data from the 777 patients was included in the study. As stated in 2.1 data was anonymised with only basic demographic data being included. Data was coded in a binary fashion where it was dichotomous. The final diagnosis was coded into the following data sets:

	1	0
Cancer/NoCacer	Cancer	Not Caner
CancerPolyp/ Not	Cancer or Polyp	Not Cancer or Polyp
Urgent/NonUrgent	Urgent	Non urgent
Normal/Abnormal	Normal	Abnormal

Table 2.1: Table demonstrating data sets and binary outcomes

The categorising of outcome diagnosis was undertaken to increase the breadth of models to be assessed given the variation in diagnoses commonly seen in those referred via the 2ww pathway. The classification 'Urgent' included patients with Cancer, Polyps, and IBD. The classification Normal / Abnormal related to those with any pathology other than haemorrhoids, this group included those with conditions such as cancer, polyps, inflammatory bowel conditions and diverticular disease..

Further sets of data were created for model analysis by varying the number of input variables. Input variables were adjusted based on:

1. All variables
2. ‘2ww’ based variables
3. Variables based on univariate analysis (V2T)
4. Variables selected on clinical knowledge

Each set of outcomes was then modelled with each set of input variables.

2.5 Neural Network Design

All data sets were taken by the author (JC) and entered into Artificial Neural Network software (Alyula Neurointelligence. USA). A multi-layered feed forward network was selected for experimenting with back-propagation for error reduction and learning. Outcome target measures were set and data was processed within the software package ensuring that its format was compatible for further analysis. Experiments were then undertaken on each dataset to evaluate the optimal architecture of the ANN. This was performed by altering the number of hidden units and hidden layers. A maximum of 2 hidden layers were used in modelling with the number of hidden units varying according to the number of input variables. A logistic (Sigmoid) activation function was used in modelling with the 5 best models being assessed further. Data was divided into a test, validation and training set for analysis with outcomes being assessed for sensitivity, specificity and Risk.

2.6 Data Mining Methods

Using WEKA explorer and experimenter (WEKA), data sets were converted by the author to the necessary .arff format. This software uses JAVA code and is available on a general user licence. It is a powerful piece of software allowing analysis of data models based on a number of different model classifiers, ranging from simple linear regression to complex decision trees.

All datasets were then run through the software using the following classifiers in the experimenter:

Table 2.2: WEKA Classifiers

Key	
1	.ZeroR
2	OneR
3	BayesNet
4	NaiveBayes
5	NaiveBayesUpdateable
6	Logistic
7	MultilayerPerceptron
8	RBFNetwork
9	SimpleLogistic
10	SMO
11	SPegasos
12	VotedPerceptron
13	ADTree
14	IB1
15	IBk
16	KStar
17	LWL
18	AdaBoostM1
19	AttributeSelectedClassifier
20	Bagging
21	ClassificationViaClustering
22	ClassificationViaRegression
23	CVParameterSelection
24	Dagging
25	Decorate
26	END
27	FilteredClassifier

28	Grading
29	LogitBoost
30	MultiBoostAB
31	MultiClassClassifier
32	MultiScheme
33	ConjunctiveRule
34	DecisionTable
35	DTNB
36	JRip
37	NNge
38	PART
39	Ridor
40	BFTree
41	DecisionStump
42	FT
43	J48graft
44	LADTree
45	LMT
46	NBTree
47	RandomForest
48	RandomTree
49	REPTree

2.7 Analysis of Methods

2.7.1 Comparison of Methods

Outcome data from this analysis was compared using the t test for significant difference in outcome prediction, with the most accurate models being further assessed in the explorer GUI

2.7.2 Comparison with specialists

Data from 100 respondents in an anonymous form was provided to two independent practicing primary care physicians (GP1 and GP2) and two post CCT Colorectal surgeons (C1 and C2). All variables recorded were provided from the study questionnaires. The assessing GP's and Colorectal Surgeons assessed the questionnaires and were invited to identify those likely to have colorectal cancer based on questionnaire data. The responses were collected by the author (JC) and compared to the optimal data mining model.

2.8 Statistical Analysis

All statistical analysis was undertaken using either SPSS17 (SPSS inc, Chicago, Il) or the WEKA software by the author (JC).

Univariate analysis was be performed on basic demographics using descriptive statistics and frequencies and 2x2 contingency tables for Chi squared analysis. Analysis of the variables with an outcome of Adenocarcinoma was undertaken using logistical regression analysis by the author using Hosmer and Lemenshow's χ^2, Nagelkerke's χ^2 and the Wald statistic for variable association with output.

Graphical display of demographic data was performed illustrating distribution of Hb levels and the distribution of actual outcomes within the study population. Further graphical analysis of models was be performed with ROC curves.

The sensitivity, specificity, PPV, NPV and LR of each model was calculated and tabulated where appropriate.

Comparison of models was undertaken with t tests and chi squared analysis. All tests were undertaken with a $p < 0.05$.

Results

3 Results

3.2 Patients

Data was collated over the 12 month study period from July 2007 to July 2008. A total of 1212 patients were referred via the 2ww pathway, 777 completed the questionnaire successfully. A further 100 samples were collected for testing of models independently. Analysis was undertaken assessing referral patterns, Symptoms commonly found in those with adenocarcinoma, symptoms commonly found in those with polyps and the role of anaemia in the identification of those with colorectal cancer.

3.2.1 Univariate analysis

Mean age	67 years	Range 20-96
Sex Distribution	57% Male	

Frequencies are demonstrated in Table 3.1

Table 3.1: Symptom frequency

	Yes	No
PR Bleeding	382 (49%)	395 (51%)
Dark Red	731 (64%)	46 (36%)
Bright Red	440 (57%)	337 (43%)
On Motion	690 (89%)	87 (11%)
On Toilet Paper	458 (59%)	319 (41%)
Mixed with stool	650 (84%)	127 (16%)
More than once in 6 weeks	486 (63%)	291 (37%)
Mucous PR	616 (80%)	161 (20%)
Pus PR	769 (99%)	8 (1%)
Alteration in Bowel Habit	240 (31%)	537 (69%)
Change in 12 months	731 (94%)	46 (6%)
Constipated	601 (77%)	176 (23%)
Loose Stool	451 (58%)	326 (42%)
Diarrhoea	559 (72%)	218 (28%)
Straining at defecation	624 (80%)	153 (20%)
Complete Evacuation	275 (35%)	502 (65%)
Urgency	483 (62%)	294 (38%)
Pain at defecation	681 (88%)	96 (12%)
Incontinence	701 (90%)	76 (10%)
Abdominal Pain	415 (53%)	362 (47%)
Lethargy	544 (70%)	233 (30%)
SOB at activity	666 (86%)	111 (14%)
SOB on stairs	684 (88%)	93 (12%)

Change in Weight	548 (71%)	229 (29%)
Loss of Weight	595 (77%)	182 (23%)
Loose Clothing	680 (88%)	97 (12%)
Increased Weight	739 (95%)	38 (5%)
Increased Appetite	751 (97%)	26 (3%)
Decreased Appetite	638 (82%)	139 (18%)
Aspirin	700 (90%)	77 (10%)
Painkillers	682 (88%)	95 (12%)
Polyp	752 (97%)	25 (3%)
Ca Colon	773 (99.5%)	4 (0.5%)
Ca Elsewhere	754 (97%)	23 (3%)
Family Polyp	759 (98%)	18 (2%)
Family Ca Colon	704 (91%)	73 (9%)
Family Ca Elsewhere	660 (85%)	117 (15%)
Relative polyp	775 (99.7%)	2 (0.3%)
Relative Ca Colon	729 (94%)	48 (6%)
Relative Ca Elsewhere	730 (94%	47 (6%)
Crohns / UC	766 (99%)	11 (1%)
Family Hx IBD	752 (97%)	25 (3%)
Smoker	611 (79%)	166 (21%)
Ex Smoker	480 (62%)	297 (38%)

There was a significant association between lower GI adenocarcinoma and the following variables following χ^2 :

- Blood mixed with stool $\chi^2 = 13.1$ $p<0.01$
- Mucus PR $\chi^2= 5.2$ $p<0.05$
- Alteration in bowel habit $\chi^2= 15.59$ $p<0.01$
- Loose stools $\chi^2= 12.87$ $p<0.01$
- Abdominal pain $\chi^2= 10.8$ $p<0.01$
- Decreased weight $\chi^2= 5.1$ $p<0.05$
- Ex smoker $\chi^2= 4.79$ $p<0.05$

Table 3.2: Significant variables at Univariate analysis – used as V2T group for model analysis

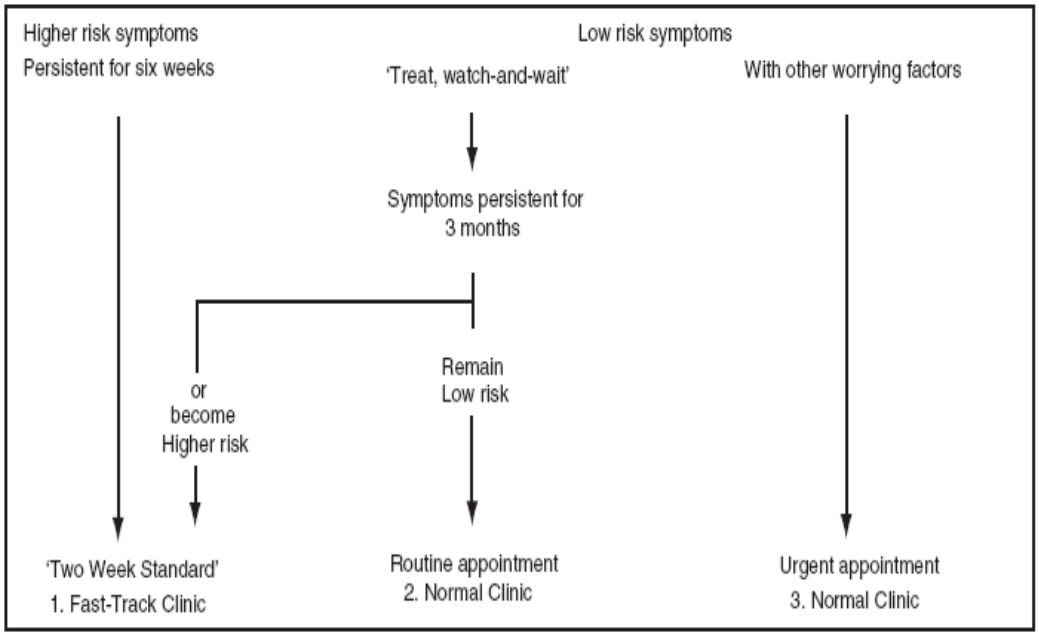
These variables were isolated into a single data set and assessed against all models.

3.2.2 Referral Patterns

Using the Referral pathway (fig 3.1) as the optimal model the route of referral of each patient was examined to assess local compliance with referral guidance.

From the 777 referred as 2ww patients 174 (24%) failed to meet the high risk referral criteria as published [139].

Figure 1: Diagram of referral pathways for those with lower GI symptoms



Further analysis was performed on those who did fulfil the high risk criteria to assess the number who met each measurable criterion and determine the number of polyps and cancers identified within each of those groups (table 3.3)

Table 3.3: Table demonstrating frequency of Cancer / Polyps found based on 2WW referral statements.

	Total number		number
Rectal bleeding with a change in bowel habit to looser stools and/or increased frequency of defecation persisting for more than 6 weeks	163	Adenocarcinoma	16
		Polyp	18
Change in bowel habit as above without rectal bleeding persisting for more than 6 weeks (over 60)	123	Adenocarcinoma	10
		Polyp	12
Rectal bleeding without anal symptoms (over 60)	286	Adenocarcinoma	36
		Polyp	27
Hb <11g/dl in men*	54	Adenocarcinoma	1
		Polyp	8
Hb <10g/dl women* (post menopause)	44	Adenocarcinoma	5
		Polyp	4
* Not all patients had recorded Hb			

3.2.3 Symptoms associated with Adenocarcinoma

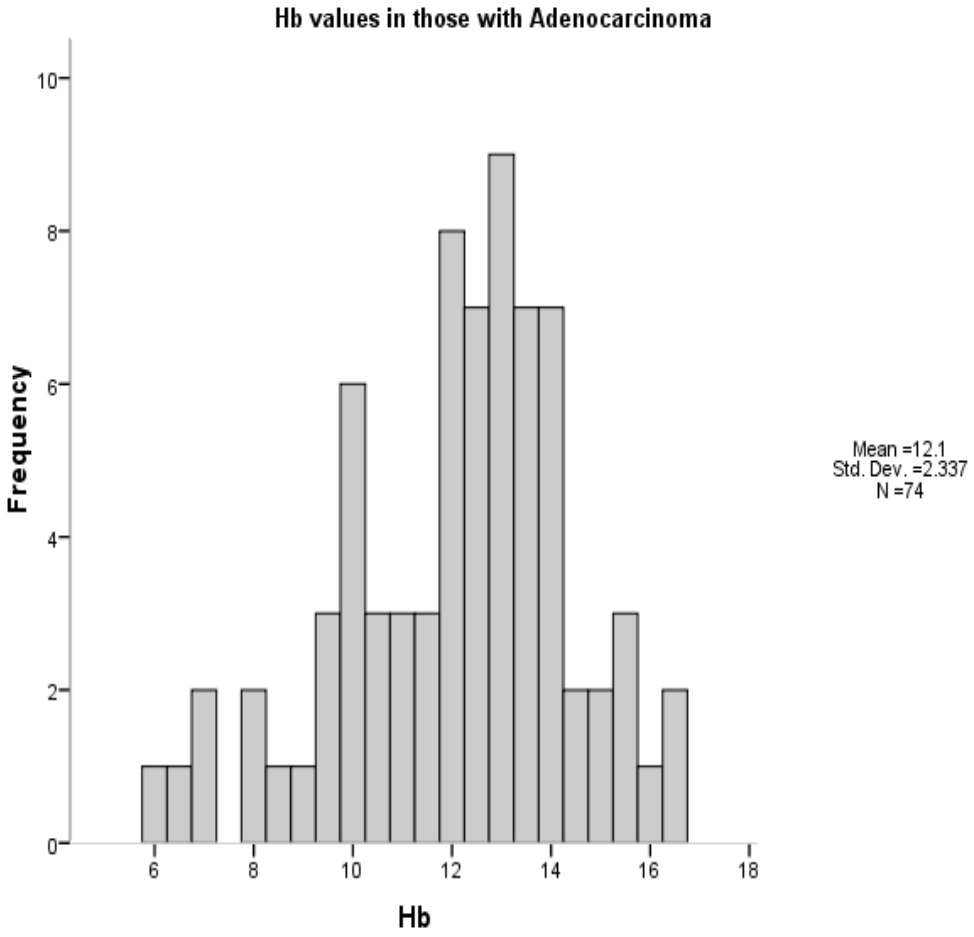
The dataset of those patients with a diagnosis of Colorectal Adenocarcinoma was analysed to assess the frequency of symptoms associated with this diagnosis. A total of 74 Adenocarcinomas were diagnosed in the study, accounting for 9.5% of the group. This level is in keeping with the findings in other studies of the percentage of patients referred via the 2ww pathway who have an underlying adenocarcinoma.

Table 3.4: Frequency of symptoms in those with diagnosis of Adenocarcinoma

Demographics (n=74)

Mean age	70 years (range 39-89)
Males	46 (62%)
Rectal Bleeding	50 (68%)
Bright Red	35 (70%)
Dark Red	15 (30%)
Change in Bowel Habit	50 (68%)
Constipation	17 (34%)
Diarrhoea	16 (32%)
Loose stool	17 (34%)
Abdominal Pain	35 (47%)
Weight Loss	28 (38%)
Anaemic per guidelines	6(8%)
NEITHER rectal bleeding NOR CIBH	11 (15%)

Figure 2: Hb levels in those diagnosed with adenocarcinoma



3.2.4 Symptoms associated with polyps

Table 3.5: Table demonstrating frequency of symptoms in those found to have colonic polyps

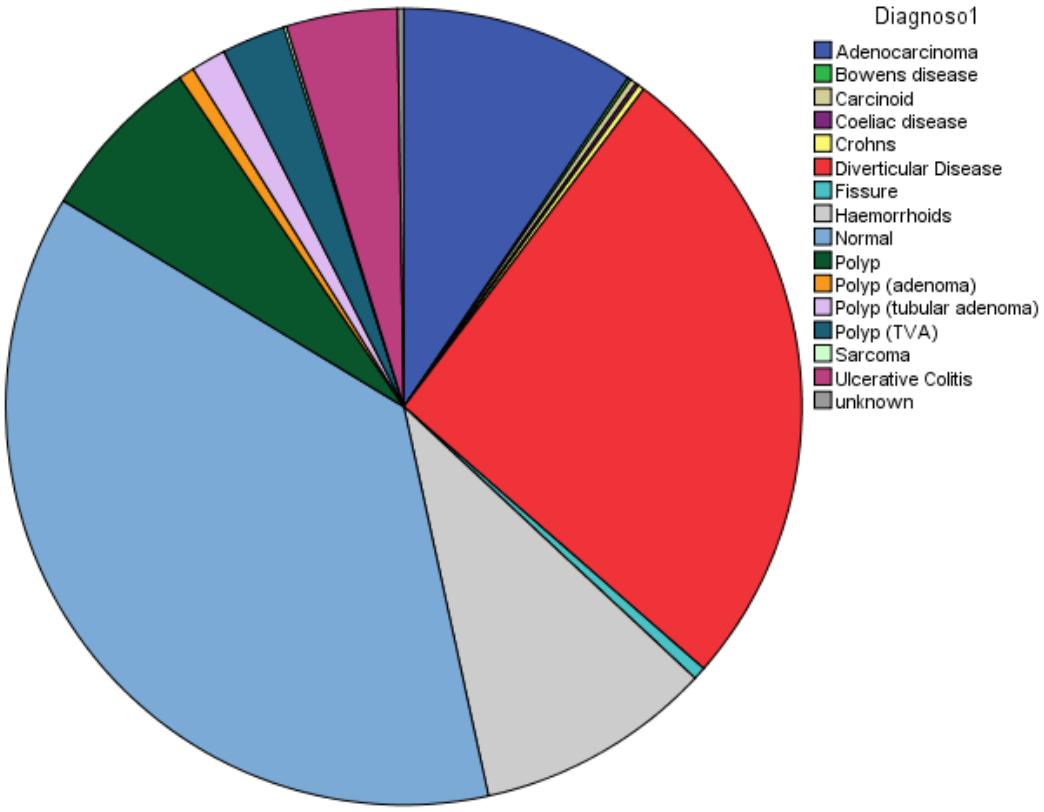
Demographics (n=89)

Mean age	67 years (range 23-92)
Males	44 (50%)
Rectal Bleeding	51 (57%)
Bright Red	38 (43%)
Dark Red	5 (6%)
Change in Bowel Habit	56 (63%)
Constipation	20 (36%)
Diarrhoea	18 (32%)
Loose stool	18 (32%)
Abdominal Pain	42 (47%)
Weight Loss	27 (19%)
Anaemic per guidelines	6 (7%)
NEITHER rectal bleeding NOR CIBH	11 (13%)

3.2.5 Disease Distribution within cohort

Figure 3: Chart demonstrating frequency of diagnoses in those referred to 2ww clinic

Graph showing distribution of diagnoses



Logistical Regression analysis

Logistical regression is a statistical method modelling data with a dichotomous categorical outcome variable (i.e. binary) and input variables that are either continuous or categorical. It enables the transformation of the data using a logarithmic function thus creating the effect of a linear relationship that is necessary for regression modelling [266]. As illustrated in fig 3.4 the logistic regression equation is similar to a linear regression equation, but in logarithmic terms.

Figure 4: logistic regression equation

$$P(Y) = \frac{1}{1 + e^{-(b_o + b_1 X_1 + b_2 X_2 + \cdots \ldots b_n X_n + \varepsilon_i)}}$$

Data was analysed using logistical regression and tested for accuracy of fit using Hosmer and Lemenshow's Chi squared R^2_L and Nagelkerke's R^2_N. The contribution in the model of each independent input variable was assessed with the Wald statistic. Modelling was undertaken using SPSS 17.0 using the Forced entry model [267] to reduce the influence from any random data variation therefore provide more replicable results

Table 3.6: Table demonstrating accuracy of logistical regression model

Chi Squared	-2 Log likelihood	Nagelkerke R Square	Hosmer and Lemeshow Chi S
116.54 45 df Sig <0.01	372.179[a]	.298	1.29. 8df. Sig .829

Table 3.7: Table demonstrating weighting of each variable in logistical regression analysis

Variables in the Equation

	B	S.E.	Wald	df	Sig.	Exp(B)
Male(1)	-.796	.299	7.080	1	.008	.451
PRBleeding(1)	-.489	.747	.428	1	.513	.613
DarkRed(1)	-1.395	.635	4.827	1	.028	.248
BrightRed(1)	-.143	.639	.050	1	.823	.867
OnMotion(1)	.488	.505	.936	1	.333	1.629
ONToiletPaper(1)	-.074	.453	.026	1	.871	.929
Mixedwithstool(1)	-1.278	.450	8.067	1	.005	.279
Morethanoncein6weeks(1)	.488	.440	1.230	1	.267	1.630
MucousPR(1)	-.594	.354	2.809	1	.094	.552
PusPR(1)	-3.685	1.385	7.077	1	.008	.025
AlterationinBowelHabit(1)	.026	.478	.003	1	.957	1.026
Changein12months(1)	.485	.739	.431	1	.512	1.625
Constipated(1)	.319	.534	.357	1	.550	1.376
Loosestool(1)	.093	.430	.047	1	.829	1.097
Diarrhoea(1)	.601	.387	2.408	1	.121	1.823
Strainingatdefecation(1)	-.330	.498	.440	1	.507	.719
Completeevacuation(1)	.405	.317	1.634	1	.201	1.499
Urgency(1)	.253	.370	.466	1	.495	1.288
Painondefectation(1)	2.838	.883	10.335	1	.001	17.074
incontinence(1)	-.133	.546	.060	1	.807	.875
abdominalpain(1)	-.199	.312	.405	1	.525	.820
lethargy(1)	-.020	.383	.003	1	.959	.980
SOBonactivities(1)	-.221	.541	.167	1	.683	.802
SOBonstairs(1)	-.580	.579	1.001	1	.317	.560
ChangeinWt(1)	1.314	1.110	1.403	1	.236	3.723

LossofWt(1)	-2.563	1.109	5.337	1	.021	.077
Looseclothing(1)	-.035	.503	.005	1	.944	.965
IncWt(1)	-1.568	1.193	1.728	1	.189	.208
AppetiteInc(1)	.629	1.190	.280	1	.597	1.876
AppetiteDec(1)	.156	.422	.137	1	.711	1.169
Aspirin(1)	-.062	.538	.013	1	.909	.940
Painkiller(1)	.296	.520	.324	1	.569	1.344
Polyp(1)	19.486	6616.913	.000	1	.998	2.901E8
CAColon(1)	19.283	18982.887	.000	1	.999	2.369E8
Cancerelsewhere(1)	.481	1.151	.174	1	.676	1.617
FamilyPolyp(1)	-.480	1.175	.167	1	.683	.619
FamilyCaColon(1)	.906	.808	1.256	1	.262	2.474
FamilyCaElsewhre(1)	-.234	.458	.262	1	.609	.791
RelativePolyp(1)	17.309	25782.017	.000	1	.999	3.291E7
RelativeCaColon(1)	19.100	4854.938	.000	1	.997	1.973E8
RelativeCaElsewhere(1)	1.297	.878	2.179	1	.140	3.657
CrohnsUC(1)	18.317	10418.510	.000	1	.999	9.017E7
FamilyHxIBD(1)	.637	1.208	.278	1	.598	1.891
Smoker(1)	-.059	.372	.025	1	.874	.942
ExSmoker(1)	-.096	.322	.088	1	.766	.909
Constant	-93.212	34654.905	.000	1	.998	.000

3.3 Artificial Neural Networks

For a detailed review of Artificial neural networks please refer to sections 1.3.2, a brief summary follows. Artificial Neural Networks (ANN) are computer based models that are able to model data by computing weights between variables and use internal algorithms to learn from errors within the analysis thus improving efficiency.

For these experiments a Multi-layered Feed forward network was selected with back propagation for error reduction. Alyuda Neurointelligence 2.2 software (Alyuda, California, USA) was used to assist the author in the design and testing of the different ANN models. This is an industrial software package, used by both the research community and multinational corporations such as Boeing and NASA. It is very versatile and, unlike other software or trial packages is able to analyse data sets with more than 200 cases.

The number of hidden layers and units was varied per dataset. A maximum of two hidden layers was used in model selection to try and avoid any 'over fitting' of the data. An Exhaustive search pattern was employed when assessing models allowing exploration of all topologies within the defined number of layers and units. The number of hidden units within the model was varied depending on the number of input variables and a logistic activation function was used for data analysis with a cross entropy network error function. The accept level was >0.5 and reject level <0.5. Output data was binary and related to the defined values within the data set being assessed.

3.3.1 Comparison of networks

Attributes were selected by the author as stated in table 3.2. Three of the four groups were based on best knowledge and the remaining attribute selection was based on Univariate analysis outcomes per section 3.1.

Data was analysed using ANN software. Following reprocessing of the data multiple experiments were undertaken to assess the optimal design of the network. This involved the variation of hidden layers and nodes to find the optimal network design. A comprehensive record of outcomes from this process can be seen in appendix 1. The best performing network design was then trained, validated and tested on the data using 500 iterations with comparisons done between two different algorithms; Quick Propagation and Online back propagation. The same process was undertaken for all data sets for each outcome variation. Results of this analysis are documented below.

3.3.1.1 All Variables

Using the 46 attributes related to the dataset assessment of different network designs was undertaken. Variation in the number of hidden units and layers was as below:

Hidden layers: 1-2

Hidden units:

Layer 1: 7 - 115

Layer 2: 4 - 76

Network designs, covering all possible combinations of units and layers per each outcome measure were assessed and verified with the top 10 networks compared for sensitivity, specificity, PPV, NPV and LR as per tables 3.8-3.11

3.3.1.1.1

Table 3.8: Top 10 neural networks and accuracy at modelling prediction for all variables against outcome Urgent / Not Urgent

Order	Design	Method	Sensitivity	Specificity	PPV	NPV	LR
1	46-98-1	BP	.63	.86	.11	.98	4.8
		OBP	.57	.86	.01	.98	4.3
2	46-94-1	BP	.88	.96	.79	.98	26.2
		OBP	.44	.88	.23	.95	3.8
3	46-93-1	BP	.93	.96	.77	.99	24.6
		OBP	.74	.88	.18	.98	6.1
4	46-9-1	BP	.76	.92	.52	.97	10.1
		OBP	.8	.88	.18	.99	6.6
5	46-10-1	BP	.75	.87	.08	.99	5.6
		OBP	.66	.87	.12	.98	5.2

3.3.1.1.2

Table 3.9: Top 10 neural networks and accuracy at modelling prediction for all variables against outcome Normal / Abnormal

Order	Design	Method	Sensitivity	Specificity	PPV	NPV	LR
1	46-61-4-1	BP	.65	.68	.75	.57	2.1
		OBP	.56	.75	.93	.22	2.3
2	46-72-4-1	BP	.63	.66	.76	.51	1.9
		OBP	.81	.72	.70	.82	2.9
3	46-99-4-1	BP	.64	.66	.73	.56	1.9
		OBP	.89	.78	.79	.90	4.4
4	46-70-5-1	BP	.64	.68	.77	.53	2.1
		OBP	.86	.68	.61	.89	2.7
5	46-51-5-1	BP	.63	.69	.79	.50	2.1
		OBP	.89	.69	.63	.92	2.9

3.3.1.1.3

Table 3.10: Top 10 neural networks and accuracy at modelling prediction for all variables against outcome Cancer / Not Cancer

Order	Design	Method	Sensitivity	Specificity	PPV	NPV	LR
1	46-70-1	BP	.88	.94	.43	.99	16.5
		OBP	.90	.96	.62	.99	24.3
2	46-77-1	BP	.89	.94	.48	.99	16.9
		OBP	.85	.95	.55	.99	18.8
3	46-89-1	BP	.90	.97	.6	.99	35.9
		OBP	.90	.96	.66	.99	26.2
4	46-23-2-1	BP	.5	.91	.14	.98	5.9
		OBP	.88	.95	.59	.99	21.3
5	46-45-2-1	BP	.66	.94	.41	.97	11.1
		OBP	.92	.93	.31	.99	13.5

3.3.1.1.4

Table 3.11: Top 10 neural networks and accuracy at modelling prediction for all variables against outcome Cancer or Polyp / Not Cancer or polyp

Order	Design	Method	Sensitivity	Specificity	PPV	NPV	LR
1	46-61-4-1	BP	.66	.82	.22	.96	3.7
		OBP	.68	.92	.69	.91	8.2
2	46-51-5-1	BP	.71	.92	.71	.92	9.04
		OBP	.58	.83	.34	.93	3.66
3	46-99-4-1	BP	.58	.83	.34	.93	3.66
		OBP	X	.77	0	1	x
4	46-61-5-1	BP	.62	.84	.34	.94	3.89
		OBP	.77	.92	.71	.94	10.2
5	46-7-1	BP	.63	.80	.13	.98	3.2
		OBP	.82	.89	.59	.96	7.9

X= Incalculable

3.3.1.2 2ww selected variables

Using the 18 attributes related to the dataset assessment of different network designs was undertaken. Variation in the number of hidden units and layers was as below:

Hidden layers: 1-2

Hidden units:

Layer 1: 3-45

Layer 2: 2-36

1548 network designs, covering all possible combinations of units and layers per each outcome measure were assessed and verified with the top 10 networks as per tables3.12-3.15

3.3.1.2.1

Table 3.12: Top 10 neural networks and accuracy at modelling prediction for 2ww selected variables against outcome Urgent / Not Urgent

Order	Design	Method	Sensitivity	Specificity	PPV	NPV	LR
1	18-5-5-1	BP	.42	.85	.02	.99	3.0
		OBP	X	.79	0	1	X
2	18-43-1	BP	X	.85	0	1	X
		OBP	.79	.87	.17	.99	6.5
3	18-31-1	BP	.66	.86	.04	.99	4.8
		OBP	.78	.90	.40	.98	8.5
4	18-30-1	BP	.82	.90	.37	.98	8.6
		OBP	.85	.87	.15	.99	6.8
5	18-19-1	BP	.53	.86	.07	.98	3.9
		OBP	.87	.88	.24	.99	7.7

X= Incalculable

3.3.1.2.2

Table 3.13: Top 10 neural networks and accuracy at modelling prediction for 2ww variables against outcome Normal / Abnormal

Order	Design	Method	Sensitivity	Specificity	PPV	NPV	LR
1	18-6-15-1	BP	.75	.72	.74	.72	2.6
		OBP	.58	.71	.87	.32	2.1
2	18-6-20-1	BP	.63	.69	.80	.40	2.1
		OBP	.65	.66	.74	.56	1.9
3	18-23-23-1	BP	.65	.68	.75	.56	2.0
		OBP	.59	.78	.92	.61	2.8
4	18-32-25-1	BP	.64	.67	.57	.55	1.9
		OBP	.66	.72	.80	.55	2.3
5	18-31-29-1	BP	.64	.66	.74	.55	1.9
		OBP	.57	.81	.95	.21	3.0

3.3.1.2.3

Table 3.14: Top 10 neural networks and accuracy at modelling prediction for 2ww variables against outcome Cancer / Not Cancer

Order	Design	Method	Sensitivity	Specificity	PPV	NPV	LR
1	18-4-2-1	BP	.35	.92	.23	.95	4.5
		OBP	.33	.90	.02	.99	3.5
2	18-30-2-1	BP	.54	.92	.24	.97	7.2
		OBP	.50	.90	.04	.99	5.4
3	18-10-4-1	BP	.70	.95	.51	.97	14.13
		OBP	0	.90	.0	.99	0
4	18-17-6-1	BP	.68	.93	.29	.98	9.8
		OBP	1	.90	.02	.1	10.7
5	18-9-9-1	BP	.55	.93	.40	.96	8.6
		OBP	.66	.90	.05	.99	7.3

3.3.1.2.4

Table 3.15: Top 10 neural networks and accuracy at modelling prediction for 2ww variables against outcome Cancer or polyp / Not Cancer or polyp

Order	Design	Method	Sensitivity	Specificity	PPV	NPV	LR
1	18-27-5-1	BP	.73	.82	0.25	0.98	4.3
		OBP	.67	.89	0.58	0.92	6.1
2	18-32-15-1	BP	.71	.83	0.28	0.97	4.3
		OBP	X	.78	0	1	x
3	18-12-20-1	BP	.67	.83	0.27	0.96	3.9
		OBP	X	.79	0	1	x
4	18-43-5-1	BP	.67	.83	0.29	0.96	4.0
		OBP	.61	.91	0.67	0.88	6.7
5	18-23-6-1	BP	.69	.85	0.4	0.95	4.8
		OBP	.68	.87	0.5	0.94	5.4

X= Incalculable

3.3.1.3 Selected variables through best knowledge

Using the 22 attributes related to the dataset assessment of different network designs was undertaken. Variation in the number of hidden units and layers was as below:

Hidden layers: 1-2

Hidden units:

Layer 1: 3-55

Layer 2: 2-36

1908 network designs, covering all possible combinations of units and layers per each outcome measure were assessed and verified with the top 10 networks as per tables 3.16 – 3.19

3.3.1.3.1

Table 3.16: Top 10 neural networks and accuracy at modelling prediction for 2ww selected variables against outcome Urgent / Not Urgent

Order	Design	Method	Sensitivity	Specificity	PPV	NPV	LR
1	22-6-31-1	BP	.66	.88	.21	.98	5.6
		OBP	.82	.89	.28	.98	7.6
2	22-42-31-1	BP	.86	.98	.77	.98	23.5
		OBP	.90	.93	.61	.98	16.8
3	22-13-1	BP	.75	.78	.16	.99	6.1
		OBP	.90	.95	.69	.98	18.4
4	22-48-2-1	BP	.84	.88	.18	.99	7.1
		OBP	.89	.90	.36	.99	9.3
5	22-46-4-1	BP	.88	.91	.42	.99	10.0
		OBP	.90	.91	.43	.99	10.4

3.3.1.3.2

Table 3.17: Top 10 neural networks and accuracy at modelling prediction for 2ww variables against outcome Normal / Abnormal

Order	Design	Method	Sensitivity	Specificity	PPV	NPV	LR
1	22-39-9-1	BP	.87	.81	.82	.86	4.6
		OBP	.85	.71	.68	.87	3.1
2	22-28-29-1	BP	.77	.78	.80	.74	3.5
		OBP	.83	.77	.78	.83	3.7
3	22-53-10-1	BP	.86	.84	.85	.85	5.4
		OBP	.71	.74	.79	.66	2.8
4	22-3-15-1	BP	.70	.68	.71	.66	2.1
		OBP	.62	.67	.78	.48	1.9
5	22-37-19-1	BP	.70	.71	.76	.64	2.5
		OBP	.88	.76	.74	.89	3.6

3.3.1.3.3

Table 3.18: Top 10 neural networks and accuracy at modelling prediction for 2ww variables against outcome Cancer / Not Cancer

Order	Design	Method	Sensitivity	Specificity	PPV	NPV	LR
1	22-42-2-1	BP	.61	.91	.10	.99	7.1
		OBP	1	.91	.12	1	11.8
2	22-53-2-1	BP	.88	.90	.72	.99	31.6
		OBP	.81	.91	.12	.99	9.6
3	22-32-4-1	BP	.64	.91	.12	.99	7.5
		OBP	.92	.96	.64	.99	25.7
4	22-36-4-1	BP	.87	.97	.74	.98	32.8
		OBP	.87	.93	.37	.99	14.1
5	22-44-4-1	BP	.54	.91	.12	.98	6.1
		OBP	1	.92	.28	1	14.2

3.3.1.3.4

Table 3.19: Top 10 neural networks and accuracy at modelling prediction for 2ww variables against outcome Cancer or polyp / Not Cancer or polyp

Order	Design	Method	Sensitivity	Specificity	PPV	NPV	LR
1	22-4-1	BP	.68	.84	.33	.95	4.3
		OBP	.29	.84	.58	.61	1.8
2	22-28-6-1	BP	X	.78	0	1	X
		OBP	X	.78	0	1	X
3	22-6-8-1	BP	X	.78	0	1	X
		OBP	X	.78	0	1	X
4	22-8-6-1	BP	X	.78	0	1	X
		OBP	X	.78	0	1	X
5	22-45-12-1	BP	X	.78	0	1	X
		OBP	X	.78	0	1	X

X= Incalculable

3.3.1.4 Univariate selected variables

Using the 7 attributes, selected due to Univariate analysis the assessment of different network designs was undertaken. Variation in the number of hidden units and layers was as below:

Hidden layers: 1-2

Hidden units:

Layer 1: 1 - 18

Layer 2: 1 - 12

234 network designs, covering all possible combinations of units and layers per each outcome measure were assessed and verified with the top 5 networks as per tables 3.20-3.23

3.3.1.4.1

Table 3.20: Top 10 neural networks and accuracy at modelling prediction for Univariate selected variables against outcome Urgent / Non Urgent

Order	Design	Method	Sensitivity	Specificity	PPV	NPV	LR
1	7-14-4-1	BP	.65	.88	.23	.97	5.6
		OBP	.72	.87	.11	.99	5.9
2	7-4-6-1	BP	.58	.87	.18	.97	4.6
		OBP	1	.85	.01	1	7.1
3	7-15-8-1	BP	.70	.91	.18	.99	8.1
		OBP	73	.86	.10	.99	5.5
4	7-14-11-1	BP	.60	.87	.16	.98	4.8
		OBP	.77	.86	.60	.99	5.7
5	7-12-12-1	BP	.68	.86	.10	.99	5.2
		OBP	.73	.86	.10	.99	5.5

3.3.1.4.2

Table 3.21: Top 10 neural networks and accuracy at modelling prediction for Univariate selected variables against outcome Normal / Abnormal

Order	Design	Method	Sensitivity	Specificity	PPV	NPV	LR
1	7-6-9-1	BP	.62	.62	.70	.53	1.6
		OBP	.57	.70	.88	.29	1.9
2	7-10-10-1	BP	.65	.60	.61	.63	1.6
		OBP	.58	.70	.88	.60	1.9
3	7-8-6-1	BP	.62	.60	.55	.57	1.6
		OBP	.58	.71	.88	.30	2.0
4	7-15-11-1	BP	.60	.59	.58	.50	1.4
		OBP	.58	.70	.87	.32	1.9
5	7-11-12-1	BP	.60	.59	.68	.50	1.4
		OBP	.58	.71	.89	.29	2.0

3.3.1.4.3

Table 3.22: Top 10 neural networks and accuracy at modelling prediction for Univariate selected variables against outcome Cancer / Not Cancer

Order	Design	Method	Sensitivity	Specificity	PPV	NPV	LR
1	7-1-1	BP	X	.90	0	1	X
		OBP	X	.90	0	1	X
2	7-2-1	BP	X	.90	0	1	X
		OBP	X	.90	0	1	X
3	7-3-1	BP	X	.90	0	1	X
		OBP	X	.90	0	1	X
4	7-4-1	BP	.33	.90	.01	.99	3.5
		OBP	X	.90	0	0	X
5	7-5-1	BP	X	.90	0	0	X
		OBP	X	.90	0	0	X

X= Incalculable

3.3.1.4.4

Table 3.23: Top 10 neural networks and accuracy at modelling prediction for Univariate selected variables against outcome Cancer or polyp / Not Cancer or polyp

Order	Design	Method	Sensitivity	Specificity	PPV	NPV	LR
1	7-1-10-1	BP	.52	.81	.23	.94	2.8
		OBP	X	.78	0	0	x
2	7-9-1	BP	.65	.81	.18	.97	3.5
		OBP	.31	.86	.62	.63	2.3
3	7-10-1	BP	.74	.81	.15	.98	3.9
		OBP	.24	.79	.22	.80	1.1
4	7-11-1	BP	.73	.81	.19	.98	4.0
		OBP	X	.78	0	1	X
5	7-13-1	BP	.71	.81	.19	.97	3.8
		OBP	X	.78	0	1	X

X= Incalculable

3.4 Data Mining

Data mining was undertaken using the WEKA platform by the author (JC). Data was cleaned as per the method. Each experimental dataset was assessed using the experimenter function using the classifiers as listed in table 3.24. Each variation in dependent variables was analysed with each listed classifier to identify the optimal model for this data. Classifier outcomes were compared using t tests for significant differences in correctly predicting outcome against the baseline classifier (ZeroR). Detailed results of this analysis can be seen in tables 3.25-3.28. The top 5 classifiers were then assessed further, examining the sensitivity and specificity of the model at predicting the dependent variable when tested. Model comparisons are illustrated below in tables 3.29-3.32. The numbers in the variables column relates to the name of the classifier used (listed in table 3.24).

Table 3.24: Table illustrating WEKA classifiers

Key	
1	ZeroR
2	OneR
3	BayesNet
4	NaiveBayes
5	NaiveBayesUpdateable
6	Logistic
7	MultilayerPerceptron
8	RBFNetwork
9	SimpleLogistic
10	SMO
11	SPegasos
12	VotedPerceptron
13	ADTree
14	IB1
15	IBk
16	KStar
17	LWL
18	AdaBoostM1
19	AttributeSelectedClassifier
20	Bagging
21	ClassificationViaClustering
22	ClassificationViaRegression
23	CVParameterSelection
24	Dagging
25	Decorate
26	END
27	FilteredClassifier

28	Grading
29	LogitBoost
30	MultiBoostAB
31	MultiClassClassifier
32	MultiScheme
33	ConjunctiveRule
34	DecisionTable
35	DTNB
36	JRip
37	NNge
38	PART
39	Ridor
40	BFTree
41	DecisionStump
42	FT
43	J48graft
44	LADTree
45	LMT
46	NBTree
47	RandomForest
48	RandomTree
49	REPTree

Table 3.25: Table comparing accuracy of WEKA classifiers as predictors (1)

Dataset	V2T				2ww			
Outcome	CA/ Polyp /N	Ca / No Ca	Norm / Abnor	Urgent / Non Urgent	CA/ Polyp/ N	Ca / No Ca	Norm / Abnor	Urgent / Non Urgent
1	78.6	90.5	52.3	85.7	78.6	90.5	52.2	85.7
2	78.6	90.5	57.2	85.7	73.7	86.5	61.3	81.1
3	78.6	90.5	52.1	85.7	76.4	88.9	56.4	84.4
4	78.4	90.5	52.1	85.7	77.8	89.1	58.4	85
5	78.4	90.5	59.3	86.1	77.8	89.1	58.4	85
6	78.3	90.5	58.3	85.9	74.7	87.1	58.1	82.7
7	77.4	89	58.4	85.9	75	87.3	55.8	82.6
8	78.2	90.3	58.6	85.4	77.6	89.3	56.8	84.6
9	78.2	90.5	58.2	85.7	78.4	90.2	61.2	85.6
10	78.6	90.5	56.2	85.7	76.9	89.4	58.9	85.1
11	78.6	90.5	57.1	85.7	76.9	89.4	58.9	85.1
12	78.6	90.4	58.2	85.6	78.6	90.5	52.3	85.7
13	78.7	90.3	59.3	85.8	79.2	90.3	61	85.7
14	69.7	83.4	54.5	77.5	69.7	87.2	51.3	79.7
15	77.1	88.5	58.2	84.8	69.3	86.8	51.7	79.5
16	78.5	90.4	58.1	85.3	73.5	88.9	55.6	82.6
17	78.6	90.5	58.8	85.7	78.6	90.5	61.4	85.7
18	78.7	90.5	58.6	85.9	78.6	90.5	62.5	85.5
19	78	90.4	57.9	85.7	78.6	90.5	60.2	85.7
20	78.3	90.4	57.9	85.7	76.7	90.3	54.6	85.4
21	58	62.7	52.5	60.8	55.5	53.4	50.3	54.5
22	78.6	90.5	58.2	85.5	75.2	88.3	57.3	83.4
23	78.6	90.5	52.3	85.7	78.6	90.5	52.3	85.7
24	78.7	90.5	58.8	85.7	78.8	90.3	59.8	85.4

	= Statistically worse at the 0.05 level compared to ZeroR		= Statistically better at the 0.05 level compared to ZeroR

Table 3.26: Table comparing accuracy of WEKA classifiers as predictors (2)

Dataset	V2T				2ww			
Outcome	CA/ Polyp /N	Ca / No Ca	Norm / Abnor	Urgent / Non Urgent	CA/ Polyp /N	Ca / No Ca	Norm / Abnor	Urgent / Non Urgent
26	77.4	90.5	56.7	85	78.2	90.5	60.4	85.7
27	78.6	90.5	52.1	85.7	78.6	90.5	61.5	85.7
28	78.6	90.5	52.3	85.7	78.6	90.5	52.3	85.7
29	78.4	90.5	58.6	86	77.9	90.2	62.1	84.9
30	78.6	90.5	56.3	85.7	78.6	90.5	61.2	85.7
31	78.3	90.5	58.3	85.9	74.7	87.1	58.1	82.7
32	78.6	90.5	52.3	85.7	78.6	90.5	52.3	85.7
33	78.6	90.5	52.8	85.7	78.6	90.5	60.4	85.7
34	78.6	90.5	56.9	85.6	78.1	90.5	61.2	85.7
35	78.9	90.5	58.4	85.6	78	90.4	63	84.8
36	77.9	90.5	58.4	85.3	78.7	89.6	61.4	84.5
37	68.3	82.7	52.1	77	80.1	84.1	52.2	78.9
38	77.3	89.8	59.1	85.7	73.1	88.9	55.8	83
39	78.2	90.3	57.9	85.5	79.2	90.1	58.8	85.2
40	77.7	90.4	57.8	85.2	78.6	90.5	54.8	85.5
41	78.6	90.5	52.9	85.7	78.6	90.5	61.5	85.7
42	77.7	90.5	59.9	86.3	73	87.1	55.6	80.3
43	77.4	90.5	56.7	85	78.3	90.5	60.4	85.7
44	78.4	90	59.5	85.8	78.3	88.8	60.8	84.6
45	78.2	90.5	57.6	85.4	78.1	90.1	61.1	85.6
46	78.6	90.5	56.6	85.7	77.7	89.8	57.3	85.1
47	77.3	88.8	58.3	85.1	75.4	89.6	53.9	84.2
48	76.6	88	58.1	84.4	69.8	86.5	52.1	78.5
49	78	90.4	57.9	85.5	78.4	90.4	51.8	85.7

	= Statistically worse at the 0.05 level compared to ZeroR		= Statistically better at the 0.05 level compared to ZeroR

Table 3.27: Table comparing accuracy of WEKA classifiers as predictors (3)

Dataset	Best Knowledge				All			
Outcome	CA/ Polyp /N	Ca / No Ca	Norm / Abnor	Urgent / Non Urgent	CA/ Polyp /N	Ca / No Ca	Norm / Abnor	Urgent / Non Urgent
1	78.6	90.5	52.3	85.7	78.6	90.5	52.3	85.7
2	73.7	86.5	61.3	81.1	73.8	86.5	61.3	81.1
3	76.4	88.5	56.4	84.5	76.4	88.5	56.4	84.5
4	76.9	90.8	58.9	85.1	75.1	90.7	57.4	85.5
5	76.9	90.8	58.9	85.1	75.4	90.7	57.4	85.5
6	74.9	87.5	57.7	82.7	73.3	85.4	57.4	81.8
7	74.7	87.5	56.1	81.8	74.1	87.7	57.6	81.8
8	77.5	90.3	58.4	85.4	77.9	90.5	52.3	85.6
9	78.4	90.4	61.4	86.5	78.3	90.6	61.4	86.5
10	77.3	90.2	59.7	86.1	76.3	89.5	59.4	84.8
11	76.9	88.3	58.1	83.8	74.4	87.6	57.8	81.9
12	78.6	90.5	52.2	85.7	78.6	90.5	52.3	85.7
13	79	90.1	62.6	86.1	78.5	90.2	62.2	86.3
14	72.1	88.3	56.1	80.7	72.1	86.7	56.6	81.3
15	71.9	88.1	56.2	80.5	71.9	86.7	56.1	81.3
16	76.9	89.1	59.2	82.6	74.5	88.5	56.3	83.2
17	78.6	90.5	61.2	85.7	78.6	90.5	61.1	85.7
18	79.2	90.5	62.4	85.6	79.2	90.5	62.3	85.5
19	78.6	90.5	60.2	85.7	78.6	90.5	60.1	85.7
20	76.6	90.3	54	85.3	76.8	90.3	53.1	85.2
21	54	51.5	49.1	52.9	56.2	56.4	49.4	56.8
22	75.8	87.8	58.1	83.8	75.9	88.1	57.1	83.5
23	78.6	90.5	52.3	85.7	78.6	90.5	52.3	85.7
24	79.1	90.6	60.6	86.3	77.8	90.8	59.7	86.1

	= Statistically worse at the 0.05 level compared to ZeroR		= Statistically better at the 0.05 level compared to ZeroR

Table 3.28: Table comparing accuracy of WEKA classifiers as predictors (4)

Data set	Best Knowledge					All			
Outcome	CA/ Polyp /N	Ca / No Ca	Norm / Abnor	Urgent / Non Urgent		CA/ Polyp /N	Ca / No Ca	Norm / Abnor	Urgent / Non Urgent
26	78.2	90.5	60.4	85.6		78.4	90.7	59.8	85.5
27	78.6	90.5	61.5	85.7		78.6	90.5	61.5	85.7
28	78.6	90.5	52.3	85.7		78.6	90.5	52.2	85.7
29	78.4	90.8	64.1	86.3		78.1	90.8	61.6	86.3
30	78.6	90.5	61.2	85.7		78.6	90.5	61.1	85.7
31	74.9	87.5	57.7	82.7		73.3	85.4	57.3	81.8
32	45.3	60.5	52.3	85.7		78.6	90.5	52.3	85.7
33	78.6	90.5	60.4	85.7		78.6	90.5	60.2	85.7
34	78.2	90.4	61.1	85.8		78.1	90.4	60.8	85.8
35	78.1	90.1	63.5	83.3		78	90	62.9	82.5
36	78.8	90.6	63.2	84.5		78.3	90.6	62.1	84.6
37	70.7	84.3	51.6	79.1		71.1	84.5	51.9	79.2
38	70.9	89.3	54.2	82.4		70.5	88.5	54.2	81.8
39	78.6	90.4	55.4	85.4		78.6	90.7	59.5	85.4
40	78.6	90.5	61.5	85.7		78.6	90.4	55	85.6
41	78.6	90.5	61.5	85.7		78.6	90.5	61.5	85.7
42	73.1	87.1	55.8	81.4		70.1	87.6	56.7	80.8
43	78.3	90.7	60.3	85.6		78.1	90.9	59.8	85.6
44	78.4	88.9	62.1	84.6		77.5	88.9	61.7	84.3
45	78.2	90.6	61.6	86.4		78.3	90.7	61.5	86.4
46	77.4	90.7	56.9	85.1		76.5	90.7	57.1	85.2
47	76.5	90.1	54.9	84.8		77.1	90.4	55.2	85.3
48	71.2	86.9	54.4	80.1		70.7	85.3	53.5	78.2
49	78.5	90.4	51.5	85.7		78.4	90.4	51.3	85.7

	= Statistically worse at the 0.05 level compared to ZeroR		= Statistically better at the 0.05 level compared to ZeroR

3.4.1 Model comparison

3.4.1.1

Table 3.29: Top 5 WEKA models demonstrating predictive accuracy for data outcome Cancer / No Cancer.

	Variables	% CC	Sensitivity	Specificity	PPV	NPV	LR
1	All 42	88.4	.34	.92	.23	.95	4.3
2	All 25	90.86	.62	.91	.11	.99	7.1
3	All 24	90.34	.45	.90	.06	.99	5.0
4	Jc 4	89.6	.18	.91	.03	.98	1.9
5	Jc 5	89.7	.42	.92	.22	.97	5.4

3.4.1.2

Table 3.30: Top 5 WEKA models demonstrating predictive accuracy for data outcome Cancer or polyp / No Cancer or polyp.

Outcome B	Variables	% CC	Sensitivity	Specificity	PPV	NPV	LR
1	2ww 37	72.4	.33	.81	.29	.74	1.8
2	2ww 13	79.8	.60	.81	.17	.96	3.1
3	Jc 18	79.2	.65	.79	.07	.99	3.1
4	All 18	78.9	.63	.79	.03	.99	2.9
5	Jc 24	79.1	.58	.80	.09	.98	2.8

3.4.1.3

Table 3.31: Top 5 WEKA models demonstrating predictive accuracy for data outcome Urgent / Not Urgent

Outcome C	Variables	% CC	Sensitivity	Specificity	PPV	NPV	LR
1	Jc 9	86.4	.62	.91	.12	.99	6.7
2	All 45	86.5	.67	.87	.10	.99	5.1
3	All 29	86.4	.61	.87	.13	.99	4.7
4	All 13	85.0	.42	.87	.14	.97	3.3
5	V2T 42	86.0	.55	.87	.15	.98	4.4

3.4.1.4

Table 3.32: Top 5 WEKA models demonstrating predictive accuracy for data outcome Normal / Abnormal

Outcome D	Variables	% CC	Sensitivity	Specificity	PPV	NPV	LR
1	Jc 29	63.1	.63	.63	.371	.54	1.7
2	Jc 35	62.4	.62	.62	.72	.52	1.7
3	Jc 36	62.6	.63	.62	.69	.55	1.6
4	2ww 18	61.2	.60	.63	.76	.45	1.6
5	Jc 13	61.9	.62	.61	.69	.51	1.6

3.4.2 Assessment of 'best fit'

The top 5 overall models in predicting outcome for each data set are illustrated below.

3.4.2.1

Table 3.33: Best performing KDD models for Cancer / Not Cancer

	Model	Topology	Sensitivity	Specificity	LR
1	ANN with all variables	46-89-1 BP	.90	.97	35.8
2	ANN with best knowledge selected variables	22-36-4-1 BP	.87	.97	32.8
3	ANN with best knowledge selected variables	22-53-2-1 BP	.88	.97	31.6
4	ANN with all variables	46-89-1 OBP	.90	.96	26.2
5	ANN with best knowledge selected variables	22-32-4-1 OBP	.92	.96	25.7

3.4.2.2

Table 3.34: Best performing KDD models for Cancer or polyp / Not Cancer or polyp

	Variables	Model	Sensitivity	Specificity	LR
1	ANN with all variables	46-61-5-1 BP	.77	.92	10.2
2	ANN with all variables	46-51-5-1 BP	.71	.92	9.04
3	ANN with all variables	46-61-4-1 OBP	.68	.91	8.2
4	ANN with all variables	46-7-1 OBP	.82	.89	7.9
5	ANN with 2ww selected variables	18-43-5-1 OBP	.61	.90	6.7

3.4.2.3

Table 3.35: Best performing KDD models for Urgent / Non Urgent

	Variables	Model	Sensitivity	Specificity	LR
1	ANN with all variables	46-94-1 BP	.88	.96	26.2
2	ANN with all variables	16-93-1 BP	.93	.96	24.6
3	ANN with best knowledge selected variables	22-42-31-1 BP	.86	.96	23.5
4	ANN with best knowledge selected variables	22-13-1 OBP	.90	.95	18.4
5	ANN with best knowledge selected variables	22-42-31-1 OBP	.90	.93	14.8

3.4.2.4

Table 3.36: Best performing KDD models for outcome Normal / Abnormal

	Variables	Model	Sensitivity	Specificity	LR
1	ANN with best knowledge selected variables	22-53-10-1 BP	.86	.84	5.4
2	ANN with best knowledge selected variables	22-39-9-1 BP	.86	.81	4.6
3	ANN with all variables	46-99-4-1 OBP	.89	.79	4.4
4	ANN with best knowledge selected variables	22-28-29-1 OBP	.83	.77	3.7
5	ANN with best knowledge selected variables	22-37-19-1 OBP	.88	.76	3.6

3.5 Specialist comparison

100 independent datasets were provided to two GP partners (GP1 + 2) and two post CCT hospital specialists in Colorectal Surgery (Sp1 + 2) for assessment. The information provided was per patient response to questionnaires thus the same level of detail as the KDD models received. The requested outcome for this analysis was simply Adenocarcinoma or Not Adenocarcinoma. No further information was provided to the assessors. The age and sex distribution within this cohort was similar to the main group, mean age was 66 years (range 23-90) and number of males within cohort was 48 (48%).

Figure 5: ROC curve comparing accuracy of GP's with Colorectal Specialists at predicting outcome

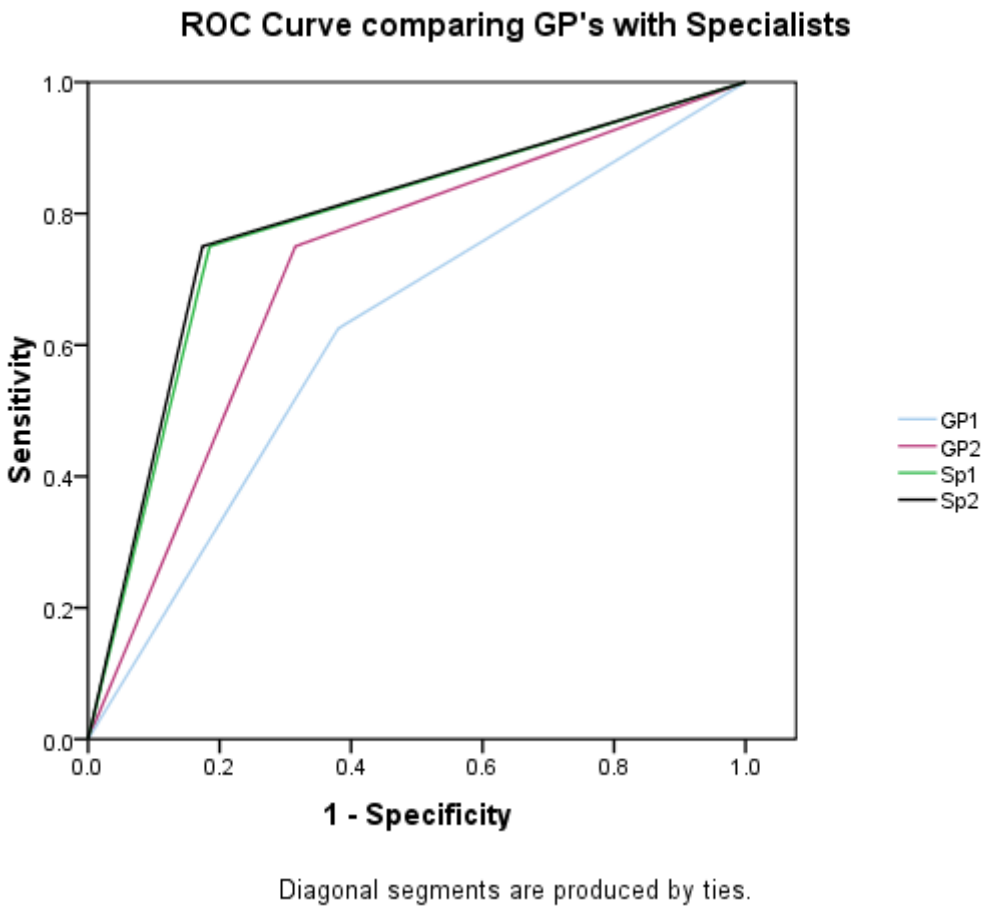


Table 3.37: Table demonstrating accuracy of clinicians in identifying those with lower GI cancer from questionnaire data

Specialist	% Correct	Sensitivity	Specificity	95% CI	PPV	NPV	LR
Gp1	62	0.12	0.95	.42-.83	.63	.62	2.5
Gp2	68	0.15	0.95	.53-.90	.63	.68	3.2
Sp1	81	0.26	0.97	.60-.96	.75	.82	10.0
Sp2	82	0.27	0.97	.60-.97	.75	.83	10.6

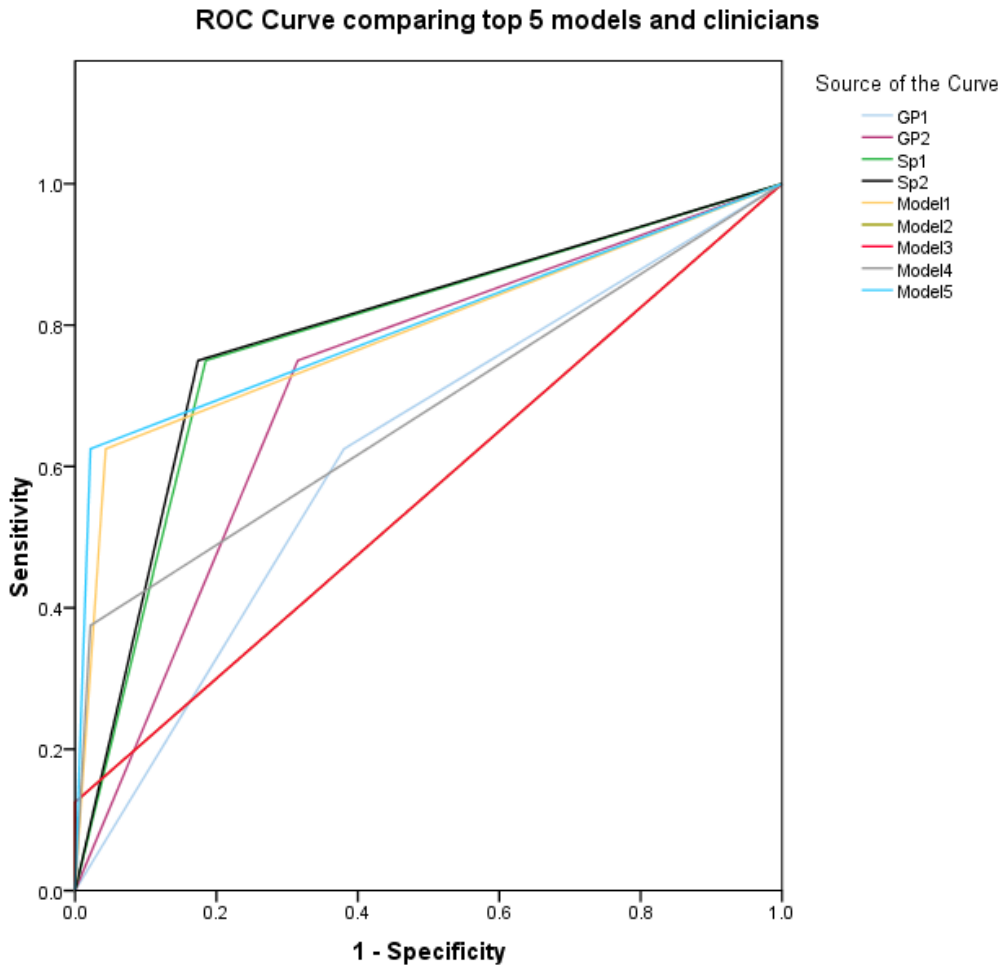
3.5.1 Comparison of Clinicians with Models

Assessment of the models against the specialists was performed using the same data that had been provided to the Clinical assessors. As the clinical assessment had been to solely predict the presence of a lower GI cancer the only dataset assessed was Set A (outcome Cancer / Not Cancer).

Table 3.38: Comparison of all clinicians and the top 5 KDD models

Specialist	% Correct	Sensitivity	Specificity	95% CI	PPV	NPV	LR
Gp1	62	0.12	0.95	.42-.83	.63	.62	2.5
Gp2	68	0.15	0.95	.53-.90	.63	.68	3.2
Sp1	81	0.26	0.97	.60-.96	.75	.82	10.0
Sp2	82	0.27	0.97	.60-.97	.75	.83	10.6
Model 1	93	0.55	.97	.58-.99	.62	.95	16.8
Model 2	93	0.13	1	.34-.79	1	.92	X
Model 3	93	0.13	1	.34-.79	1	.92	X
Model 4	94	.43	.98	.44-.91	.6	.95	19.9
Model 5	95	.63	.98	.58-1	.71	.96	28.7

Figure 6: ROC curve comparing top 5 KDD models and clinicians at accuracy of prediction



Discussion

4 Discussion

4.1 Assessment of Referral patterns

This thesis has confirmed that the proportion of patients referred via the 2ww pathway with colorectal cancer is approximately 10%. Univariate analysis using Chi Squared showed blood mixed with stool, mucus pr, alteration in bowel habit, loose stools, abdominal pain, decreased weight and ex-smoker to be significantly associated with colorectal adenocarcinoma ($p<0.05$) which are all symptoms that would clinically be associated with an increased suspicion of colorectal malignancy. Previous studies have shown abdominal pain, change in bowel habit and occult blood in the stool to be the most common presenting symptoms in colorectal cancer [268]. Rectal bleeding has been found to be present in up to 25% of cases of colon cancer [269] [270] with variation in quantity and colour. 5 demonstrated an incidence of 17.5% of colorectal adenocarcinoma in a series of patients 50 years old and younger.

In total 164 patients (21%) of those in this cohort failed to meet any of the 2ww criteria as defined for urgent colorectal assessment. This is a significant number of individuals being assessed on an urgent basis with the associated utilisation of resources, not only clinical but radiological and managerial resources to ensure that the 31/62 breech date timeline is complied with. Similar percentages have been found in other studies evaluating the success of the 2 week wait process, Leung et al found 15% of referrals over a 12 month period in the West midlands to fail to meet referral criteria [271] and Smith et al found 49.6% of referrals to their colorectal

practice failed to conform to the guidelines[272]. Similar service evaluations in Cambridge [160] found 27% of 2WW referrals failed to meet the criteria, however of those with colorectal adenocarcinoma only 8% failed to meet the criteria.

The use of anaemia as an accurate surrogate for lower GI cancer appears, certainly in this study to be a poor prognostic indicator. Only 8% of those with a colorectal malignancy were found to be anaemic per the current guidelines. Whilst it may be an important factor to consider it does not appear to be an accurate prognostic indicator in identifying those with a lower GI malignancy. Gastrointestinal blood loss remains the most common cause of iron deficiency anaemia in men and postmenopausal women with 5-10% attributed to colonic carcinoma [79] [273] [274] [275] [276] with numerous other pathologies identified on endoscopic evaluation.

4.2 Reflection of KDD methods

The use of data mining software such as WEKA provides the user with a powerful tool in the search for patterns within the data set. This can be explored with an array of different classifiers which incorporate methods of attribute selection. Whilst it may seem logical that a data mining classifier such as a decision tree would be useful in determining a pathway for the identification of those with a lower GI adenocarcinoma they did not perform well within this cohort. It is possible that the data was too 'noisy' for the machine learning tools to define an accurate model for prediction or that there were insufficient actual cases of lower GI cancer within the cohort for these classifiers to make an accurate prediction. Previous studies comparing decision trees with ANN and logistical regression for diagnosing gastro-oesophageal reflux also found the performance of decision trees to be inferior to ANN although they failed to hypothesise as to reasons for this[231].

Whilst the classifiers used within this study are all suitable for the data as it was presented to the e software, it may be that an alternative method of attribute selection or an alternative scale for the data would have improved the model performance but within the e confines of this work these potentially infinite transformations were not explored further.

4.3 Reflection on ANN techniques

Artificial Neural networks have been used in a number of fields of medicine with generally positive results and the outcome of this study correlates with this. The variability in the design of the networks made them very adaptive to the data cohort with the various combinations of hidden units and layers. Specific to colorectal cancer work has been done to assess the validity of outcome prediction for those with colorectal cancer based on pre, peri- and post-operative factors including histological staging[245]. This demonstrated a higher predictive accuracy with neural networks for both death and survival when compared to the prediction of clinicians. Similar studies have been performed using variants of Neural Networks such as the partial Logistic neural network (PLANN) [246] which allowed the creation of a web based survival prediction environment with the option of multiple online users.

The input data obtained from the patient questionnaires is quite complex in relation to the diagnosis related to the symptoms of the patients. While it appeared that the WEKA classifiers did not manage to find clarity within this a number of the neural networks evaluated managed to predict outcome to a high standard. The neural network performance was of a high standard and accuracy with a larger number of attributes for selection when compared with models containing fewer attributes. Whilst it is generally accepted that 10 times as many datasets are required as attributes, something that this study complied with it is plausible that had the dataset contained more patient episodes then model refinement could have been furthered.

4.4 Reflection on comparison model

The kind participation of two GP partners and colorectal consultants and their assessment of the test cases allowed the top performing models to be assessed with a 'virgin' set of data. When comparing the clinicians and GP's it was interesting to see the correlation in predictive accuracy between the clinical specialists, something that makes logical sense and is appropriate for their area of expertise.

One major drawback of this assessment of acumen is the lack of realism; primarily that an error in this model does not relate to a negative outcome in a patient. It is not realistic to suggest that such specificity would be a positive clinical attribute therefore the clinical index of suspicion that would instigate further appropriate tests is likely a lot lower than predicting the likelihood of a lower GI malignancy. In addition to this it is not only lower GI malignancy that is an important clinical finding; the majority of the other diagnoses seen within the cohort are in need of diagnosis and in some cases treatment.

Overall in the comparisons between the Neural network models, GP's and clinicians it was interesting to see just how accurate the Neural networks were in their levels of prediction. All of the neural networks performed better than their human counterparts in terms of percentage correct, sensitivity and specificity, PPV and NPV. This consistent level of performance may make the use of neural networks feasible as a screening tool in determining which patients should be 'fast tracked' and which should be seen on a slightly less urgent basis therefore not focusing resources inappropriately.

4.5 Overall assessment of Study

The study was able to compare a variety of KDD methods in the search for the most accurate model in terms of predicting the diagnosis of those referred as a 2ww patient. All of the models were based on the accuracy of the data provided in the patient's response to the questionnaires. While this information was transformed into a binary response it cannot be ignored that 'change in bowel habit' for example can mean one thing to one person and something completely different to another. Notwithstanding this fact, this is the same sort of clinical information that is provided by the patient to either their GP or Hospital specialist therefore remains the foundation for the basis of further clinical investigation.

Within the literature there are a few studies that have assessed prediction of colorectal cancer utilising both patient consultation questionnaires with scoring systems [169] and a smaller study from our unit that assessed the predictive capacity of neural networks for colorectal cancer found them to be of a higher predictive accuracy to clinicians although the training and validation sets were small compared to the number of variables assessed [277]. This study has corroborated the accuracy of neural networks at predicting those to be found with a colorectal malignancy compared to both alternative KDD classification methods and clinicians. This has been undertaken using robust methods of development, both of the neural networks and all appropriate KDD classifiers.

In terms of model development, a cohort of patient and their responses who all have a lower GI malignancy may have improved the accuracy of some of the KDD classifier models but it would not have been consistent with the environment that the model would be used in.

The use of alternative outcome measures in the model design process was an important part of the development. As already alluded to there is a large amount of pathology within patients referred via the 2ww pathway despite the main diagnosis only being found in 10% of cases. The ability to identify patients with a number of conditions and assess them on an urgent basis may be beneficial, not only in this dataset but in general for all patients referred from primary to secondary care. The best neural network model assessing outcome as Urgent / Non urgent had a sensitivity of 0.88 and specificity of 0.96. Once again the performance of the model deteriorated as the number of attributes used in its development was reduced.

4.6 Justification of methods

4.6.1

The 2ww referral pathway is the foundation for most urgent referrals from primary care with lower GI symptoms [139]. It was conceived in 2002 based on levels of evidence available at the time. These levels ranged between B (Fairly strong evidence) and D (Weak evidence) however there have been a number of subsequent studies undertaken to assess the accuracy of this referral pathway, all showing a similar low detection rate for colorectal cancer [162, 278-280].

Predictive accuracy is dependent on three primary components:

1. Predictive power of prognostic variables
2. Amount and quality of the data
3. Ability of method to capitalise on the prognostic indicators

These components require careful consideration prior to the conception of any predictive modelling task as they determine the suitable coding of variables, the collation of data, selection of models and based on the models how to measure model success. Within this study the majority of variables were Boolean in coding thus there was no requirement to collapse data into pre-determined categories as is commonly encountered.

Traditional statistical techniques for analysis of categorical data include logistical regression which makes linear assumptions between input variables and outcome. Such methods utilise the following function:

$$G = b + x1 + x2\ etc$$

The result of this function is then used to predict survival within this model.

Alternative methods in KDD include data mining methods have been established for many years and have been used in a wide variety of situations [281] [282-284]. They incorporate many techniques, including ANN and decision trees in the process of exploring data. They have the benefit of active learning from the input data and allow large data sets to be analysed to find a model that best fits the 'problem'

Medical problems are invariably complex and such tools have been used in various medical fields to improve rates of detection / prediction. Whilst there remains a conceptual issue with these methods when taken in conjunction with traditional statistical techniques they do enable the multidimensionality of the data to be assessed and thus may provide answers to complex problems that would not be available by more standard methods

4.6.2 Advantages of Data Mining Techniques

Predictive modelling within medicine is more commonly based on regression analysis, a more traditional statistical technique therefore, as outcome measures are categorical, binary logistical regression is frequently used. Whilst this method of modelling is widely accepted and can be performed with relative ease on numerous software packages, with statistical theory to validate the model fit, they depend on a linear relationship between input variables and outcome.

Data mining techniques differ in these assumptions and theoretically should offer advantages in complex modelling when compared to statistical approaches. They presume nonlinear relationships between variables and allow relationships between units to be arbitrary therefore permit the discovery of ‘rules’ that may not be apparent with more traditional methods. It is this ability of the KDD process that makes the technique worthy of assessment when modelling complex clinical outcomes.

4.6.3 Limitations of Data Mining techniques

The primary limitation with data mining techniques is the apparent lack of transparency within the model. It is conceptually difficult to gain insight into how the model uses the input data to derive an output value. Certain methods are less obscure than others, decision trees for example allowing a schematic flow diagram to illustrate the data pathway. Others however have a 'black box' approach, such as ANN whereby unless the user undertakes feature extraction, a complex and time consuming approach to deriving the intricacies of the model, the method used to apply weights to attributes remains unknown. This conceptual lack of clarity is likely the main reason for hesitancy in the use of these methods more frequently.

As with statistical methods the risk of 'over fit' is present in data mining methodology. There are a number of techniques that are used to avoid over fitting data to the model such as cross validation, bootstrapping and data splitting [217, 219, 285]. It is recognised and accepted that to accurately assess the predictive performance of the model a data set that was not part of the model-building process must be used. This, as is similar in regression analysis makes modelling small datasets difficult as over fitting can increase model error thus enforce erroneous conclusions.

4.7 Are KDD techniques viable in the identification of those with CRC?

Despite best efforts in early detection colorectal cancer remains difficult to diagnose based on clinical symptoms alone. This is likely attributable to numerous factors such as stage of disease, location of tumour and patients themselves to identify a few. Efforts are on-going to increase the detection rate of those with colorectal cancer at an earlier stage within the UK in the form of the national screening programme and FOB screening [148, 151]. Whilst very sensitive this compliance in the screening population is variable [151, 153] likely due to the method by which the individual provides the samples. In addition to the above, the use of flexible sigmoidoscopy in a mobile setting is being evaluated to optimise detection rates of colorectal cancer within the general population[286] [287].

Notwithstanding above, Colonoscopy remains the gold standard method of diagnosis [13, 288, 289] for colorectal cancer and it is not within the bounds of this study to compare KDD methods to colonoscopy, nor was it the aim to compare these techniques with screening tools. The aim was to optimise the referral pattern in those who attended their primary care physician with symptoms and were referred onto secondary care for further assessment, attempting to classify those who needed more urgent assessment and as such potentially assist in the more appropriate distribution of resources.

In this study KDD methods varied in their ability to predict patients with colorectal cancer. These ranged from the best model accurately predicting 95% of those within the dataset with sensitivity 0.63 and specificity 0.98.

In studies assessing prediction it is important to ensure the sample studied is sufficiently large to safeguard reliability. As such the ration of input variables to outcomes should be 10:1 [217] as failure to achieve this level has resulted in unstable models being created. In this study, all models explored had an appropriate ratio of input and output variables.

The use of KDD has a broad spectrum across all fields of medicine. The most commonly used method to date has been that of ANN with studies showing comparability, if not some degree of superiority to traditional techniques. [290] [291] [236] [292] [237]. The nonlinearity and ability of ANN to learn has made their use attractive when trying to stratify and predict outcomes in the field of medicine. Studies assessing outcomes of mortality and morbidity following cardiac surgery have been undertaken with positive results [242]

Alternative KDD methods used, specifically in the field of medicine include fuzzy logic classification systems such as PROAFTN [293] which has been applied to assist in diagnosing bladder tumours and acute leukaemia. Fuzzy KNN classifiers have been used and have been shown to produce a more robust model of prognostic markers than logistic regression and MLP's [294]. Fuzzy rule generation in conjunction with breast cancer datasets has been used with accuracy rates of 97% [295] [296].

The clinical environment in which a predictive system is used is the primary determinate of the model and its classification cut off point. In clinical settings such as this study's model the optimal system is one that has a small number of false positives and no false negatives. This will result in preference being given to model sensitivity at the cost of specificity. Whilst there is no theoretical guidance as to how the ideal cut off point in an ANN is chosen it may be possible to alter the number of cut off points in the ROC curve but studies looking at this have shown minimal gain [297]).

4.8 Limitations

The use of KDD is reliant upon the quality of information entered into the database for analysis. Whilst the data entry within this study was a direct reflection upon the answers given by patients regarding their symptoms it is feasible that the questionnaire may have been too complex. The initial questionnaire had been validated within a cohort of patients within the department however some additions were made prior to the distribution of the questionnaire for use within this study to try and increase the amount of data received. It is feasible that the addition of extra questions may have misled or confused those completing the questionnaire thus reducing its reproducibility. It is accepted that once any changes had been made to the questionnaire this should once again have been tested and validated on an independent cohort of patients both prior to and on attendance at a clinic to ensure that the answers were reproducible. Whilst this technique in itself may, due to human nature result in some anomalies it would allow the rigorous testing of the questionnaire and increase its validity within the setting of this study.

4.9 Conclusion

The complexity of medical diagnosis remains challenging both to the physician and computation models. Risk prediction remains central to a clinician's ability to successfully perform their duties, be it in a primary care setting, secondary or tertiary care. An array of tests and tools are at the disposal of those in a hospital setting, allowing the investigation of those deemed to be at increased risk of a condition. Clinicians use clinical evidence in conjunction with experience to initiate further investigations however there is variation in experience depending on the specialisation of the clinician.

This study has shown that the use of KDD tools as an adjuvant to clinical acumen can prove beneficial in identifying patients with lower GI pathology therefore expedite their diagnosis and treatment. While it would be ill-conceived to suggest that such computer models can replace physician-patient interaction further work assessing the feasibility of models such as the ones in this study directing patients 'straight to test' are worthy of consideration for both 2ww pathway patients and those referred in the low risk groups.

Appendix A

Bowel Symptom Questionnaire

We would be grateful if you could complete the following questions regarding symptoms you may have experienced recently.

Once completed please sign the bottom as evidence of consent as explained on the information sheet

1)Have you had any bleeding from your bottom	Yes ☐ No ☐
- Was this dark red	Yes ☐ No ☐
- Was this bright red	Yes ☐ No ☐
- Was it on your motion	Yes ☐ No ☐
- Was it on the toilet paper	Yes ☐ No ☐
- Was it mixed with your motion	Yes ☐ No ☐
2) Has this happened more than once in 6 weeks - If yes, how often?	Yes ☐ No ☐
3) Have you passed any mucus / slimy stuff from your bottom? - If yes, how often?	Yes ☐ No ☐
4) Have you passed any pus from your bottom in the past 6 weeks? - If yes, how often?	Yes ☐ No ☐
5) Has your bowel habit changed in the past few months?	Yes ☐ No ☐
- If Yes, how many times a day do you open them?	Yes ☐ No ☐
	Yes ☐ No ☐
- If No, has it changed in the last 12 months	Yes ☐ No ☐
- Are you more constipated	Yes ☐ No ☐
- Are your motions looser than normal	
- Have you had any diarrhoea	
6) Do you have to strain to open your bowels?	Yes ☐ No ☐

7) When you open your bowels do you feel as if you have completely emptied them?	Yes ☐ No ☐
8) Have you had any urgency when opening your bowels?	Yes ☐ No ☐
9) Do you have any pain when you open your bowels in the past 6 weeks?	Yes ☐ No ☐
10) Have you had any 'accidents' when opening your bowels recently?	Yes ☐ No ☐
11) Have you have any abdominal pain in the past 6 weeks?	Yes ☐ No ☐
12) Have you felt more tired than usual recently?	Yes ☐ No ☐
13) Have you recently found yourself short of breath doing activities that previously caused you no problems?	Yes ☐ No ☐
14) Do you get Short of Breath walking up stairs?	Yes ☐ No ☐
15) Has your weight been stable in the past 6 months? - Have you lost any weight recently? - Are your clothes looser fitting than before? - Have you gained any weight recently?	Yes ☐ No ☐ Yes ☐ No ☐ Yes ☐ No ☐ Yes ☐ No ☐
16) Has your appetite: Increased Decreased	 Yes ☐ No ☐ Yes ☐ No ☐
17) Do you take regular Aspirin	Yes ☐ No ☐
18) Do you take regular Painkillers	Yes ☐ No ☐
19) Have you ever had: - Polyps in your bowel - Bowel Cancer - Cancer elsewhere	 Yes ☐ No ☐ Yes ☐ No ☐ Yes ☐ No ☐

20) With reference to your immediate family (mother, father, brother and sister), have they ever had: - Polyps in the bowel - Bowel Cancer - Cancer elsewhere	 Yes ☐ No ☐ Yes ☐ No ☐ Yes ☐ No ☐
-If so who and at what age?	
21) With reference to other family members (aunts/uncles/cousins), have they ever had: - Polyps in the bowel - Bowel Cancer - Cancer elsewhere	 Yes ☐ No ☐ Yes ☐ No ☐ Yes ☐ No ☐
-If so who and at what age?	
22) Have you ever been diagnosed with Inflammatory Bowel Disease (Ulcerative Colitis / Crohns)	Yes ☐ No ☐
23) Has anyone in your family ever been diagnosed with Inflammatory bowel disease	Yes ☐ No ☐
-If yes, what relationship?	
24) Do you Smoke?	Yes ☐ No ☐
25) Have you ever Smoked?	Yes ☐ No ☐

References

1. GLOBOSCAN *Database*. 2002.

2. Baig, M.K. and C.G. Marks, *Referral guidelines for colorectal cancer: a threat or a challenge?* Hosp Med, 2000. **61**(7): p. 452-3.

3. Haenszel, W., *Mortality and morbidity statistics on all forms of cancer.* Acta - Unio Internationalis Contra Cancrum, 1961. **17**: p. 837-47.

4. Haenszel, W., *Incidence of and mortality from stomach cancer in the United States.* Acta - Unio Internationalis Contra Cancrum, 1961. **17**: p. 347-64.

5. Haenszel, W., *Cancer mortality among the foreign-born in the United States.* Journal of the National Cancer Institute, 1961. **26**: p. 37-132.

6. McMichael, A.J. and G.G. Giles, *Cancer in migrants to Australia: extending the descriptive epidemiological data.* Cancer research, 1988. **48**(3): p. 751-6.

7. Rickert, R.R., et al., *Adenomatous lesions of the large bowel: an autopsy survey.* Cancer, 1979. **43**(5): p. 1847-57.

8. Arminski, T.C. and D.W. McLean, *Incidence and Distribution of Adenomatous Polyps of the Colon and Rectum Based on 1,000 Autopsy Examinations.* Diseases of the colon and rectum, 1964. **7**: p. 249-61.

9. Williams, A.R., B.A. Balasooriya, and D.W. Day, *Polyps and cancer of the large bowel: a necropsy study in Liverpool.* Gut, 1982. **23**(10): p. 835-42.

10. Stryker, S.J., et al., *Natural history of untreated colonic polyps.* Gastroenterology, 1987. **93**(5): p. 1009-13.

11. Winawer, S.J. and A.G. Zauber, *Colorectal cancer screening: now is the time.* CMAJ : Canadian Medical Association journal = journal de l'Association medicale canadienne, 2000. **163**(5): p. 543-4; discussion 548.

12. Winawer, S.J., et al., *The National Polyp Study.* European journal of cancer prevention : the official journal of the European Cancer Prevention Organisation, 1993. **2 Suppl 2**: p. 83-7.

13. Winawer, S.J., et al., *Prevention of colorectal cancer by colonoscopic polypectomy. The National Polyp Study Workgroup.* The New England journal of medicine, 1993. **329**(27): p. 1977-81.

14. Henry, L.G., et al., *Risk of recurrence of colon polyps.* Annals of surgery, 1975. **182**(4): p. 511-5.

15. CROHN, B.B. and H. ROSENBERG, *The Sigmoidoscopic Picture of Chronic Ulcerative Colitis (Non-Specific).* The American Journal of the Medical Sciences, 1925. **170**(2): p. 220-227.

16. Munkholm, P., *Review article: the incidence and prevalence of colorectal cancer in inflammatory bowel disease.* Alimentary pharmacology & therapeutics, 2003. **18 Suppl 2**: p. 1-5.

17. Eaden, J.A., K.R. Abrams, and J.F. Mayberry, *The risk of colorectal cancer in ulcerative colitis: a meta-analysis.* Gut, 2001. **48**(4): p. 526-35.

18. Ekbom, A., et al., *Ulcerative colitis and colorectal cancer. A population-based study.* The New England journal of medicine, 1990. **323**(18): p. 1228-33.

19. Ekbom, A., et al., *Increased risk of large-bowel cancer in Crohn's disease with colonic involvement.* Lancet, 1990. **336**(8711): p. 357-9.

20. Terdiman, J.P., et al., *5-Aminosalicylic acid therapy and the risk of colorectal cancer among patients with inflammatory bowel disease.* Inflammatory bowel diseases, 2007. **13**(4): p. 367-71.

21. Muto, T., H.J. Bussey, and B.C. Morson, *The evolution of cancer of the colon and rectum.* Cancer, 1975. **36**(6): p. 2251-70.

22. Hoff, G. and M.H. Vatn, *Colonic adenoma: natural history.* Digestive diseases, 1991. **9**(2): p. 61-9.

23. Lynch, H.T. and A. de la Chapelle, *Hereditary colorectal cancer.* The New England journal of medicine, 2003. **348**(10): p. 919-32.

24. Lynch, H.T., et al., *Genetics, natural history, tumor spectrum, and pathology of hereditary nonpolyposis colorectal cancer: an updated review.* Gastroenterology, 1993. **104**(5): p. 1535-49.

25. Jess, T., et al., *Increased risk of intestinal cancer in Crohn's disease: a meta-analysis of population-based cohort studies.* The American journal of gastroenterology, 2005. **100**(12): p. 2724-9.

26. Larsson, S.C., N. Orsini, and A. Wolk, *Diabetes mellitus and risk of colorectal cancer: a meta-analysis.* Journal of the National Cancer Institute, 2005. **97**(22): p. 1679-87.

27. Ron, E., et al., *Acromegaly and gastrointestinal cancer.* Cancer, 1991. **68**(8): p. 1673-7.

28. Potter, J.D., *Colon cancer--do the nutritional epidemiology, the gut physiology and the molecular biology tell the same story?* The Journal of nutrition, 1993. **123**(2 Suppl): p. 418-23.

29. Steinmetz, K.A. and J.D. Potter, *Vegetables, fruit, and cancer prevention: a review.* Journal of the American Dietetic Association, 1996. **96**(10): p. 1027-39.

30. Research., W.C.R.F.A.I.f.C., ., and W.D. AICR, *Food, Nutrition, Physical Activity, and the Prevention of Cancer: a Global Perspective.* 2007.

31. Terry, P., et al., *Fruit, vegetables, dietary fiber, and risk of colorectal cancer.* Journal of the National Cancer Institute, 2001. **93**(7): p. 525-33.

32. Michels, K.B., et al., *Prospective study of fruit and vegetable consumption and incidence of colon and rectal cancers.* Journal of the National Cancer Institute, 2000. **92**(21): p. 1740-52.

33. Trock, B., E. Lanza, and P. Greenwald, *Dietary fiber, vegetables, and colon cancer: critical review and meta-analyses of the epidemiologic evidence.* Journal of the National Cancer Institute, 1990. **82**(8): p. 650-61.

34. Trock, B.J., E. Lanza, and P. Greenwald, *High fiber diet and colon cancer: a critical review.* Progress in clinical and biological research, 1990. **346**: p. 145-57.

35. Jacobs, E.J., et al., *Multivitamin use and colorectal cancer incidence in a US cohort: does timing matter?* American journal of epidemiology, 2003. **158**(7): p. 621-8.

36. Wu, K., et al., *A prospective study on supplemental vitamin e intake and risk of colon cancer in women and men.* Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology, 2002. **11**(11): p. 1298-304.

37. Connelly-Frost, A., et al., *Selenium, apoptosis, and colorectal adenomas.* Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology, 2006. **15**(3): p. 486-93.

38. Norat, T., et al., *Meat, fish, and colorectal cancer risk: the European Prospective Investigation into cancer and nutrition.* Journal of the National Cancer Institute, 2005. **97**(12): p. 906-16.

39. English, D.R., et al., *Red meat, chicken, and fish consumption and risk of colorectal cancer.* Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology, 2004. **13**(9): p. 1509-14.

40. Chao, A., et al., *Meat consumption and risk of colorectal cancer.* JAMA : the journal of the American Medical Association, 2005. **293**(2): p. 172-82.

41. Bingham, S.A., R. Hughes, and A.J. Cross, *Effect of white versus red meat on endogenous N-nitrosation in the human colon and further evidence of a dose response.* The Journal of nutrition, 2002. **132**(11 Suppl): p. 3522S-3525S.

42. Cross, A.J., J.R. Pollock, and S.A. Bingham, *Haem, not protein or inorganic iron, is responsible for endogenous intestinal N-nitrosation arising from red meat.* Cancer research, 2003. **63**(10): p. 2358-60.

43. Lewin, M.H., et al., *Red meat enhances the colonic formation of the DNA adduct O6-carboxymethyl guanine: implications for colorectal cancer risk.* Cancer research, 2006. **66**(3): p. 1859-65.

44. Kampman, E., et al., *Calcium, vitamin D, sunshine exposure, dairy products and colon cancer risk (United States).* Cancer causes & control : CCC, 2000. **11**(5): p. 459-66.

45. Baron, J.A., et al., *Calcium supplements for the prevention of colorectal adenomas. Calcium Polyp Prevention Study Group.* The New England journal of medicine, 1999. **340**(2): p. 101-7.

46. Flood, A., et al., *Calcium from diet and supplements is associated with reduced risk of colorectal cancer in a prospective cohort of women.* Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology, 2005. **14**(1): p. 126-32.

47. Slattery, M.L., et al., *Dietary calcium, vitamin D, VDR genotypes and colorectal cancer.* International journal of cancer. Journal international du cancer, 2004. **111**(5): p. 750-6.

48. Grau, M.V., et al., *Vitamin D, calcium supplementation, and colorectal adenomas: results of a randomized trial.* Journal of the National Cancer Institute, 2003. **95**(23): p. 1765-71.

49. Giovannucci, E., et al., *Physical activity, obesity, and risk for colon cancer and adenoma in men.* Annals of internal medicine, 1995. **122**(5): p. 327-34.

50. Slattery, M.L., *Physical activity and colorectal cancer.* Sports medicine, 2004. **34**(4): p. 239-52.

51. Samad, A.K., et al., *A meta-analysis of the association of physical activity with reduced risk of colorectal cancer.* Colorectal disease : the official journal of the Association of Coloproctology of Great Britain and Ireland, 2005. **7**(3): p. 204-13.

52. Slattery, M.L., *Diet, lifestyle, and colon cancer.* Seminars in gastrointestinal disease, 2000. **11**(3): p. 142-6.

53. Moore, L.L., et al., *BMI and waist circumference as predictors of lifetime colon cancer risk in Framingham Study adults.* International journal of obesity and related metabolic disorders : journal of the International Association for the Study of Obesity, 2004. **28**(4): p. 559-67.

54. Cho, E., et al., *Alcohol intake and colorectal cancer: a pooled analysis of 8 cohort studies.* Annals of internal medicine, 2004. **140**(8): p. 603-13.

55. Farinati, F., et al., *Effect of chronic ethanol consumption on activation of nitrosopyrrolidine to a mutagen by rat upper alimentary tract, lung, and hepatic tissue.* Drug metabolism and disposition: the biological fate of chemicals, 1985. **13**(2): p. 210-4.

56. Garro, A.J. and C.S. Lieber, *Alcohol and cancer.* Annual review of pharmacology and toxicology, 1990. **30**: p. 219-49.

57. Giovannucci, E., et al., *Alcohol, low-methionine--low-folate diets, and risk of colon cancer in men.* Journal of the National Cancer Institute, 1995. **87**(4): p. 265-73.

58. Terry, P., et al., *Long-term tobacco smoking and colorectal cancer in a prospective cohort study.* International journal of cancer. Journal international du cancer, 2001. **91**(4): p. 585-7.

59. Slattery, M.L., et al., *Associations between cigarette smoking, lifestyle factors, and microsatellite instability in colon tumors.* Journal of the National Cancer Institute, 2000. **92**(22): p. 1831-6.

60. Luchtenborg, M., et al., *Cigarette smoking and colorectal cancer: APC mutations, hMLH1 expression, and GSTM1 and GSTT1 polymorphisms.* American journal of epidemiology, 2005. **161**(9): p. 806-15.

61. Fijten, G.H., et al., *The incidence and outcome of rectal bleeding in general practice.* Family practice, 1993. **10**(3): p. 283-7.

62. Wauters, H., V. Van Casteren, and F. Buntinx, *Rectal bleeding and colorectal cancer in general practice: diagnostic study.* Bmj, 2000. **321**(7267): p. 998-9.

63. Thompson, et al., *Rectal bleeding in general and hospital practice; 'the tip of the iceberg'.* Colorectal Disease, 2000. **2**(5): p. 288-293.

64. Crosland, A. and R. Jones, *Rectal bleeding: prevalence and consultation behaviour.* BMJ, 1995. **311**(7003): p. 486-8.

65. Sandler, R.S., *Epidemiology of irritable bowel syndrome in the United States.* Gastroenterology, 1990. **99**(2): p. 409-15.

66. Everhart, J.E., et al., *A longitudinal survey of self-reported bowel habits in the United States.* Digestive diseases and sciences, 1989. **34**(8): p. 1153-62.

67. Thompson, M.R., et al., *Identifying and managing patients at low risk of bowel cancer in general practice.* Bmj, 2003. **327**(7409): p. 263-5.

68. Morrell, D.C. and C.J. Wale, *Symptoms perceived and recorded by patients.* The Journal of the Royal College of General Practitioners, 1976. **26**(167): p. 398-403.

69. Fijten, G.H., et al., *Predictive value of signs and symptoms for colorectal cancer in patients with rectal bleeding in general practice.* Family practice, 1995. **12**(3): p. 279-86.

70. Metcalf, J.V., et al., *Incidence and causes of rectal bleeding in general practice as detected by colonoscopy.* The British journal of general practice : the journal of the Royal College of General Practitioners, 1996. **46**(404): p. 161-4.

71. Mant, A., et al., *Rectal bleeding. Do other symptoms aid in diagnosis?* Diseases of the colon and rectum, 1989. **32**(3): p. 191-6.

72. Shallow, T.A., F.B. Wagner, Jr., and R.E. Colcher, *Clinical evaluation of 750 patients with colon cancer; diagnostic survey and follow-up covering a fifteen-year period.* Annals of surgery, 1955. **142**(2): p. 164-75.

73. Ellis, B.G. and M.R. Thompson, *Factors identifying higher risk rectal bleeding in general practice.* The British journal of general practice : the journal of the Royal College of General Practitioners, 2005. **55**(521): p. 949-55.

74. Calvey, H.D. and C.M. Castleden, *Gastrointestinal investigations for anaemia in the elderly: a prospective study.* Age and ageing, 1987. **16**(6): p. 399-404.

75. Rockey, D.C., *Gastrointestinal tract evaluation in patients with iron deficiency anemia.* Seminars in gastrointestinal disease, 1999. **10**(2): p. 53-64.

76. Rockey, D.C., *Occult gastrointestinal bleeding.* Gastroenterology clinics of North America, 2005. **34**(4): p. 699-718.

77. Rockey, D.C., *Lower gastrointestinal bleeding.* Gastroenterology, 2006. **130**(1): p. 165-71.

78. Rockey, D.C., *Occult and obscure gastrointestinal bleeding: causes and clinical management.* Nature reviews. Gastroenterology & hepatology, 2010. **7**(5): p. 265-79.

79. Kepczyk, T. and S.C. Kadakia, *Prospective evaluation of gastrointestinal tract in patients with iron-deficiency anemia.* Digestive diseases and sciences, 1995. **40**(6): p. 1283-9.

80. Mandel, J.S., et al., *The effect of fecal occult-blood screening on the incidence of colorectal cancer.* The New England journal of medicine, 2000. **343**(22): p. 1603-7.

81. Duthie, G.S., et al., *A UK training programme for nurse practitioner flexible sigmoidoscopy and a prospective evaluation of the practice of the first UK trained nurse flexible sigmoidoscopist.* Gut, 1998. **43**(5): p. 711-4.

82. Maslekar, S., et al., *Patient satisfaction with lower gastrointestinal endoscopy: doctors, nurse and nonmedical endoscopists.* Colorectal disease : the official journal of the Association of Coloproctology of Great Britain and Ireland, 2010. **12**(10): p. 1033-8.

83. Maslekar, S., et al., *Quality assurance in flexible sigmoidoscopy: medical and nonmedical endoscopists.* Surg Endosc, 2010. **24**(1): p. 89-93.

84. Pathmakanthan, S., et al., *Nurse endoscopists in United Kingdom health care: a survey of prevalence, skills and attitudes.* J Adv Nurs, 2001. **36**(5): p. 705-10.

85. Schoenfeld, P., et al., *Accuracy of polyp detection by gastroenterologists and nurse endoscopists during flexible sigmoidoscopy: a randomized trial.* Gastroenterology, 1999. **117**(2): p. 312-8.

86. Schoenfeld, P.S., et al., *Effectiveness and patient satisfaction with screening flexible sigmoidoscopy performed by registered nurses.* Gastrointest Endosc, 1999. **49**(2): p. 158-62.

87. Atkin, W.S., et al., *Uptake, yield of neoplasia, and adverse effects of flexible sigmoidoscopy screening.* Gut, 1998. **42**(4): p. 560-5.

88. Imperiale, T.F., et al., *Risk of advanced proximal neoplasms in asymptomatic adults according to the distal colorectal findings.* The New England journal of medicine, 2000. **343**(3): p. 169-74.

89. Levin, T.R., *Flexible sigmoidoscopy for colorectal cancer screening: valid approach or short-sighted?* Gastroenterology clinics of North America, 2002. **31**(4): p. 1015-29, vii.

90. Levin, T.R., et al., *Complications of screening flexible sigmoidoscopy.* Gastroenterology, 2002. **123**(6): p. 1786-92.

91. Levin, T.R. and A.M. Palitz, *Flexible sigmoidoscopy: an important screening option for average-risk individuals.* Gastrointestinal endoscopy clinics of North America, 2002. **12**(1): p. 23-40, vi.

92. Ahmad, N.A., et al., *Efficacy, safety, and clinical outcomes of endoscopic mucosal resection: a study of 101 cases.* Gastrointest Endosc, 2002. **55**(3): p. 390-6.

93. Conio, M., et al., *EMR of large sessile colorectal polyps.* Gastrointest Endosc, 2004. **60**(2): p. 234-41.

94. Jameel, J.K., et al., *Endoscopic mucosal resection (EMR) in the management of large colo-rectal polyps.* Colorectal disease : the official journal of the Association of Coloproctology of Great Britain and Ireland, 2006. **8**(6): p. 497-500.

95. Sano, Y., et al., *A newly developed bipolar-current needle-knife for endoscopic submucosal dissection of large colorectal tumors.* Endoscopy, 2006. **38 Suppl 2**: p. E95.

96. Fujishiro, M., et al., *Outcomes of endoscopic submucosal dissection for colorectal epithelial neoplasms in 200 consecutive cases.* Clin Gastroenterol Hepatol, 2007. **5**(6): p. 678-83; quiz 645.

97. Repici, A., et al., *Insulated-tip knife endoscopic mucosal resection of large colorectal polyps unsuitable for standard polypectomy.* Am J Gastroenterol, 2007. **102**(8): p. 1617-23.

98. Hurlstone, D.P., et al., *Salvage endoscopic submucosal dissection for residual or local recurrent intraepithelial neoplasia in the colorectum: a prospective analysis.* Colorectal disease : the official journal of the Association of Coloproctology of Great Britain and Ireland, 2008. **10**(9): p. 891-7.

99. Zhou, P., et al., *Endoscopic submucosal dissection for locally recurrent colorectal lesions after previous endoscopic mucosal resection.* Diseases of the colon and rectum, 2009. **52**(2): p. 305-10.

100. Rex, D.K., et al., *Colonoscopic miss rates of adenomas determined by back-to-back colonoscopies.* Gastroenterology, 1997. **112**(1): p. 24-8.

101. Winawer, S.J., et al., *A comparison of colonoscopy and double-contrast barium enema for surveillance after polypectomy. National Polyp Study Work Group.* The New England journal of medicine, 2000. **342**(24): p. 1766-72.

102. Kung, J.W., et al., *Colorectal cancer: screening double-contrast barium enema examination in average-risk adults older than 50 years.* Radiology, 2006. **240**(3): p. 725-35.

103. Toma, J., et al., *Rates of new or missed colorectal cancer after barium enema and their risk factors: a population-based study.* Am J Gastroenterol, 2008. **103**(12): p. 3142-8.

104. Pickhardt, P.J., et al., *Computed tomographic virtual colonoscopy to screen for colorectal neoplasia in asymptomatic adults.* The New England journal of medicine, 2003. **349**(23): p. 2191-200.

105. Pescatore, P., et al., *Diagnostic accuracy and interobserver agreement of CT colonography (virtual colonoscopy).* Gut, 2000. **47**(1): p. 126-30.

106. Cotton, P.B., et al., *Computed tomographic colonography (virtual colonoscopy): a multicenter comparison with standard colonoscopy for detection of colorectal neoplasia.* JAMA : the journal of the American Medical Association, 2004. **291**(14): p. 1713-9.

107. Arnesen, R.B., et al., *Missed lesions and false-positive findings on computed-tomographic colonography: a controlled prospective analysis.* Endoscopy, 2005. **37**(10): p. 937-44.

108. Rockey, D.C., et al., *Analysis of air contrast barium enema, computed tomographic colonography, and colonoscopy: prospective comparison.* Lancet, 2005. **365**(9456): p. 305-11.

109. Burling, D., et al., *Polyp measurement and size categorisation by CT colonography: effect of observer experience in a multi-centre setting.* Eur Radiol, 2006. **16**(8): p. 1737-44.

110. Burling, D., et al., *CT colonography interpretation times: effect of reader experience, fatigue, and scan findings in a multi-centre setting.* Eur Radiol, 2006. **16**(8): p. 1745-9.

111. Thomas, S., J. Atchley, and A. Higginson, *Audit of the introduction of CT colonography for detection of colorectal carcinoma in a non-academic environment and its implications for the national bowel cancer screening programme.* Clin Radiol, 2009. **64**(2): p. 142-7.

112. *AJCC Cancer Staging Manual*. 7th ed, ed. S.B.B. Edge, D.R.; Compton, C.C.; Fritz, A.G.; Greene, F.L.; Trotti, A.2010: Springer.

113. Astler, V.B. and F.A. Coller, *The prognostic significance of direct extension of carcinoma of the colon and rectum.* Annals of surgery, 1954. **139**(6): p. 846-52.

114. Dukes, C., *Histological Grading of Rectal Cancer: (Section of Pathology).* Proceedings of the Royal Society of Medicine, 1937. **30**(4): p. 371-6.

115. Quirke, P., et al., *Effect of the plane of surgery achieved on local recurrence in patients with operable rectal cancer: a prospective study using data from the MRC CR07 and NCIC-CTG CO16 randomised clinical trial.* Lancet, 2009. **373**(9666): p. 821-8.

116. Sebag-Montefiore, D., et al., *Preoperative radiotherapy versus selective postoperative chemoradiotherapy in patients with rectal cancer (MRC CR07 and NCIC-CTG C016): a multicentre, randomised trial.* Lancet, 2009. **373**(9666): p. 811-20.

117. Siegel, R., et al., *Preoperative short-course radiotherapy versus combined radiochemotherapy in locally advanced rectal cancer: a multi-centre prospectively randomised study of the Berlin Cancer Society.* BMC cancer, 2009. **9**: p. 50.

118. Jensen, L.H., et al., *Clinical outcome in 520 consecutive Danish rectal cancer patients treated with short course preoperative radiotherapy.* European journal of surgical oncology : the journal of the European Society of Surgical Oncology and the British Association of Surgical Oncology, 2010. **36**(3): p. 237-43.

119. Latkauskas, T., et al., *Initial results of a randomised controlled trial comparing clinical and pathological downstaging of rectal cancer after preoperative short-course radiotherapy or long term chemoradiotherapy both with delayed surgery.* Colorectal disease : the official journal of the Association of Coloproctology of Great Britain and Ireland, 2011.

120. Senapati, A., et al., *Low rates of local recurrence after surgical resection of rectal cancer suggest a selective policy for preoperative radiotherapy.* Colorectal disease : the official journal of the Association of Coloproctology of Great Britain and Ireland, 2011.

121. Hartley, A., et al., *Retrospective study of acute toxicity following short-course preoperative radiotherapy.* The British journal of surgery, 2002. **89**(7): p. 889-95.

122. Hazebroek, E.J., *COLOR: a randomized clinical trial comparing laparoscopic and open resection for colon cancer.* Surgical endoscopy, 2002. **16**(6): p. 949-53.

123. *A comparison of laparoscopically assisted and open colectomy for colon cancer.* The New England journal of medicine, 2004. **350**(20): p. 2050-9.

124. Guillou, P.J., et al., *Short-term endpoints of conventional versus laparoscopic-assisted surgery in patients with colorectal cancer (MRC CLASICC trial): multicentre, randomised controlled trial.* Lancet, 2005. **365**(9472): p. 1718-26.

125. Heald, R.J. and R.D. Ryall, *Recurrence and survival after total mesorectal excision for rectal cancer.* Lancet, 1986. **1**(8496): p. 1479-82.

126. Scott, N., et al., *Total mesorectal excision and local recurrence: a study of tumour spread in the mesorectum distal to rectal cancer.* The British journal of surgery, 1995. **82**(8): p. 1031-3.

127. Buess, G., et al., *Technique and results of transanal endoscopic microsurgery in early rectal cancer.* American journal of surgery, 1992. **163**(1): p. 63-9; discussion 69-70.

128. Steele, R.J., et al., *Transanal endoscopic microsurgery--initial experience from three centres in the United Kingdom.* The British journal of surgery, 1996. **83**(2): p. 207-10.

129. Bretagnol, F., et al., *Local excision of rectal tumours by transanal endoscopic microsurgery.* The British journal of surgery, 2007. **94**(5): p. 627-33.

130. Jemal, A., et al., *Cancer statistics, 2002.* CA: a cancer journal for clinicians, 2002. **52**(1): p. 23-47.

131. Andre, T., et al., *Oxaliplatin, fluorouracil, and leucovorin as adjuvant treatment for colon cancer.* The New England journal of medicine, 2004. **350**(23): p. 2343-51.

132. Hackett, T.P., N.H. Cassem, and J.W. Raker, *Patient delay in cancer.* The New England journal of medicine, 1973. **289**(1): p. 14-20.

133. Holliday, H.W. and J.D. Hardcastle, *Delay in diagnosis and treatment of symptomatic colorectal cancer.* Lancet, 1979. **1**(8111): p. 309-11.

134. Blakeborough, A., M.B. Sheridan, and A.H. Chapman, *Complications of barium enema examinations: a survey of UK Consultant Radiologists 1992 to 1994.* Clinical radiology, 1997. **52**(2): p. 142-8.

135. Waye, J.D., O. Kahn, and M.E. Auerbach, *Complications of colonoscopy and flexible sigmoidoscopy.* Gastrointestinal endoscopy clinics of North America, 1996. **6**(2): p. 343-77.

136. Talley, N.J. and M. Jones, *Self-reported rectal bleeding in a United States community: prevalence, risk factors, and health care seeking.* The American journal of gastroenterology, 1998. **93**(11): p. 2179-83.

137. Fijten, G.H., G.H. Blijham, and J.A. Knottnerus, *Occurrence and clinical significance of overt blood loss per rectum in the general population and in medical practice.* The British journal of general practice : the journal of the Royal College of General Practitioners, 1994. **44**(384): p. 320-5.

138. DOH, *Department of Health. Cancer Waiting Targets: a guide (Version 4)* 2005.

139. Thompson, M.R., *ACPGBI Referral guidelines for colorectal cancer.* Colorectal disease : the official journal of the Association of Coloproctology of Great Britain and Ireland, 2002. **4**(4): p. 287-297.

140. Hemingway, D.M., J. Jameson, and M.J. Kelly, *Straight to test: introduction of a city-wide protocol driven investigation of suspected colorectal cancer.* Colorectal Disease, 2006. **8**(4): p. 289-295.

141. Thorne, K., H.A. Hutchings, and G. Elwyn, BMC Health Services Research, 2006. **6**(1): p. 43.

142. Pearse IH, C., LH, *The Peckham Experiment; a Study of the Living Structure of Society,* 1985: Edinburgh & London

143. Wadsworth MEJ, B.W., Blaney R *Health and Sickness, the Choice of Treatment. Perception of Illness and Use of Services in an Urban Community*1971, London & Southampton: Tavistock Publications, Camelot Press Ltd.

144. DR, H., *The Symptom Iceberg: a Study of Community Health,* 1979, Routledge & Kegan Paul: London.

145. Kelly, S.B., et al., *Nurse specialist led flexible sigmoidoscopy in an outpatient setting.* Colorectal disease : the official journal of the Association of Coloproctology of Great Britain and Ireland, 2008. **10**(4): p. 390-3.

146. Maruthachalam, K., et al., *Evolution of the two-week rule pathway--direct access colonoscopy vs outpatient appointments: one year's experience and patient satisfaction survey.* Colorectal disease : the official journal of the Association of Coloproctology of Great Britain and Ireland, 2005. **7**(5): p. 480-5.

147. Corman, M.L., M.C. Veidenheimer, and J.A. Coller, *Colorectal carcinoma: a decade of experience at the Lahey Clinic.* Diseases of the colon and rectum, 1979. **22**(7): p. 477-9.

148. Paul Tappenden, S.E., Richard Nixon, Jim Chilcott, Hannah Sakai, Jon Karnon, *Cost-effectiveness, cost-utility and resource impact of alternative screening options for colorectal cancer*, 2004, ScHARR.

149. Mandel, J.S., et al., *Reducing mortality from colorectal cancer by screening for fecal occult blood. Minnesota Colon Cancer Control Study.* The New England journal of medicine, 1993. **328**(19): p. 1365-71.

150. Kronborg, O., et al., *Randomised study of screening for colorectal cancer with faecal-occult-blood test.* Lancet, 1996. **348**(9040): p. 1467-71.

151. Hardcastle, J.D., et al., *Randomised controlled trial of faecal-occult-blood screening for colorectal cancer.* Lancet, 1996. **348**(9040): p. 1472-7.

152. Hardcastle, J.D. and T.A. Justin, *Screening high-risk groups for colorectal neoplasia.* The American journal of gastroenterology, 1996. **91**(5): p. 850-2.

153. *Results of the first round of a demonstration pilot of screening for colorectal cancer in the United Kingdom.* BMJ, 2004. **329**(7458): p. 133.

154. Ahlquist, D.A., et al., *Accuracy of fecal occult blood screening for colorectal neoplasia. A prospective study using Hemoccult and HemoQuant tests.* JAMA : the journal of the American Medical Association, 1993. **269**(10): p. 1262-7.

155. Mandel, J.S., et al., *Sensitivity, specificity, and positive predictivity of the Hemoccult test in screening for colorectal cancers. The University of Minnesota's Colon Cancer Control Study.* Gastroenterology, 1989. **97**(3): p. 597-600.

156. Allison, J.E., R. Feldman, and I.S. Tekawa, *Hemoccult screening in detecting colorectal neoplasm: sensitivity, specificity, and predictive value. Long-term follow-up in a large group practice setting.* Annals of internal medicine, 1990. **112**(5): p. 328-33.

157. Allison, J.E., et al., *Improving the fecal occult-blood test.* The New England journal of medicine, 1996. **334**(24): p. 1607-8.

158. Uno, Y. and A. Munakata, *Endoscopic and histologic correlates of colorectal polyp bleeding.* Gastrointestinal endoscopy, 1995. **41**(5): p. 460-7.

159. Hanna, S.J., A. Muneer, and K.H. Khalil, *The 2-week wait for suspected cancer: time for a rethink?* International journal of clinical practice, 2005. **59**(11): p. 1334-9.

160. Chohan, D.P.K., et al., *How has the 'two-week wait' rule affected the presentation of colorectal cancer?* Colorectal Disease, 2005. **7**(5): p. 450-453.

161. Debnath, D., N. Dielehner, and K.A. Gunning, *Guidelines, compliance, and effectiveness: a 12 months' audit in an acute district general healthcare trust on the two week rule for suspected colorectal cancer.* Postgraduate medical journal, 2002. **78**(926): p. 748-51.

162. Eccersley, A.J., et al., *Referral guidelines for colorectal cancer--do they work?* Annals of the Royal College of Surgeons of England, 2003. **85**(2): p. 107-10.

163. Chohan, D.P., et al., *How has the 'two-week wait' rule affected the presentation of colorectal cancer?* Colorectal disease : the official journal of the Association of Coloproctology of Great Britain and Ireland, 2005. **7**(5): p. 450-3.

164. Flashman, K., *The Department of Health's "two week standard" for bowel cancer: is it working?* Gut, 2004. **53**(3): p. 387-391.

165. Cantillon, P. and R. Jones, *Does continuing medical education in general practice make a difference?* BMJ, 1999. **318**(7193): p. 1276-9.

166. Renehan, A.G., et al., *Impact on survival of intensive follow up after curative resection for colorectal cancer: systematic review and meta-analysis of randomised trials.* BMJ, 2002. **324**(7341): p. 813.

167. Walsh, S.R., et al., *Trends in colorectal cancer survival following the 2-week rule.* Colorectal Disease, 2007. **9**(3): p. 207-209.

168. Bevis, P.M., et al., *The association between referral source and stage of disease in patients with colorectal cancer.* Colorectal disease : the official

journal of the Association of Coloproctology of Great Britain and Ireland, 2008. **10**(1): p. 58-62.

169. Selvachandran, S., et al., *Prediction of colorectal cancer by a patient consultation questionnaire and scoring system: a prospective study.* The Lancet, 2002. **360**(9329): p. 278-283.

170. Rai, S., et al., *Assessment of a patient consultation questionnaire-based scoring system for stratification of outpatient risk of colorectal cancer.* The British journal of surgery, 2008. **95**(3): p. 369-74.

171. Bouckaert, G.H., J., *Performance and Performance Management*. *Handbook of Public Policy*, ed. J. B.G. Peter and Pierre2006, London: SAGE Publications.

172. Burke, W., *Organization Change – Theory and Practice*. 2nd ed2008, New York: SAGE Publication.

173. Hamlin, R.G., *Evidence based Policy and Performance Management. The American Review of*

Public Administration, 2007. ***37***(3): p. 255-277.

174. Krone, O., Syväjärvi, A. & Stenvall, J, *Knowledge Integration for Enterprise Resources Planning*

Application Design. Knowledge and Process Management, 2009. **16**(1): p. 1-12.

175. Fayyad, U.M., *Data mining and knowledge discovery: making sense out of data.* IEEE Expert Intelligent Systems & Their Applications, 1996. **11**(5): p. 20-23.

176. Chen, H.C., M, *Web Minind: Machine Learning for Web Applications*, in *Annual Review of Information Science and Technology*2004. p. 289-329.

177. Dunham, M., *Data Mining: Introductory and Advanced Analysis*2002, New Jersey: Prenhall.

178. Hearst, M. *Untangling text data mining*. in *ACL '99 Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*. 1999. Stroudsburg, PA, USA.

179. Chen, Z., *Data Mining and Uncertain Reasoning: An Integrated Approach*2001, New York: John Wiley& Sons.

180. Hayes-Roth, F. and N. Jacobstein, *The state of knowledge-based systems.* Commun. ACM, 1994. **37**(3): p. 26-39.

181. Simon, H., *Search and reasoning in problem solving.* Search and reasoning in problem solving, 1983.

182. McCulloch, W. and W. Pitts, *A logical calculus of the ideas immanent in nervous activity.* Bulletin of Mathematical Biology, 1943. **5**(4): p. 115-133.

183. Hebb, D., *The Organization of behavior*1949, New York: Wiley.

184. Rosenblatt, F., *The perceptron: a probalilistic model for information storage and organization in the brain.* Psycological Review, 1958. **65**(6): p. 386-408.

185. Widrow, B.A., JB, *Reliable, Trainable Networks for Computing and Control.* Aerospace Engineering 1962: p. 78-123.

186. Widrow, B.H., ME, *ADAPTIVE SWITCHING CIRCUITS.*1960: Wescon Convention Record.

187. Fukushima, K. and S. Miyake, *Neocognitron: A new algorithm for pattern recognition tolerant of deformations and shifts in position.* Pattern Recognition, 1982. **15**(6): p. 455-469.

188. Fukushima, K. and N. Wake, *Handwritten alphanumeric character recognition by the neocognitron.* IEEE transactions on neural networks / a publication of the IEEE Neural Networks Council, 1991. **2**(3): p. 355-65.

189. Werbos, P., *Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences*, 1974, Harvard University.

190. Le Cun, Y., *A Theoretical Framework for Back-Propagation.* 1988: p. 1-8.

191. Parker, D., *Learning Logic*, in *Invention Report*1982, Stanford UNiversity: Stanford. p. S64-81.

192. Rumelhart, D.E., G.E. Hinton, and R.J. Williams, *Learning representations by back-propagating errors.* Nature, 1986. **323**(6088): p. 533-536.

193. Rumelhart, D.E., McClelland, J L, *Parallel distributed processing: Explorations in the microstructures of cognition*1986, Cambridge, MA: MIT Press.

194. Anderson, J.A.S., Jack W.;Ritz, Stephen A.;Jones, Randall S, *Distinctive features, categorical perception, and probability learning: Some applications of a neural model.* Psychological Review, 1977. **84**(5): p. 413-451.

195. Hopfield, J.J., *Neural networks and physical systems with emergent collective computational abilities.* Proceedings of the National Academy of Sciences, 1982. **79**(8): p. 2554-2558.

196. Hartman, E.J., J.D. Keeler, and J.M. Kowalski, *Layered Neural Networks with Gaussian Hidden Units as Universal Approximations.* Neural Computation, 1990. **2**(2): p. 210-215.

197. Hornik, K., M. Stinchcombe, and H. White, *Multilayer feedforward networks are universal approximators.* Neural Networks, 1989. **2**(5): p. 359-366.

198. Jordan, M. *Attractor dynamics and parallelism in a connectionist sequential machine*. in *Proceedings of the Eighth Annual Conference of the Cognitive Science Society*. 1986. Englewood Cliffs, NJ: IEEE Tutorials.

199. Broomhead, D.L., D, *Multivariable Functional Interpolation and Adaptive Networks.* Complex Systems, 1988. **2**: p. 321-355.

200. Haykin, S.S., *Neural networks : a comprehensive foundation*1994, New York

Toronto: Macmillan ;

Maxwell Macmillan Canada ;

Maxwell Macmillan International. xix, 696 p.

201. Moody, J. and C.J. Darken, *Fast Learning in Networks of Locally-Tuned Processing Units.* Neural Computation, 1989. **1**(2): p. 281-294.

202. Kohonen, T., *Self-organized formation of topologically correct feature maps.* Biological Cybernetics, 1982. **43**(1): p. 59-69.

203. Patterson, D.W., *Artificial neural networks : theory and applications*1996, Singapore ; New York: Prentice Hall. xiv, 477 p.

204. McCullagh, P. and J.A. Nelder, *Generalized linear models*. 2nd ed. Monographs on statistics and applied probability1989, London ; New York: Chapman and Hall. xix, 511 p.

205. Auer, P., H. Burgsteiner, and W. Maass, *A learning rule for very simple universal approximators consisting of a single layer of perceptrons.* Neural networks : the official journal of the International Neural Network Society, 2008. **21**(5): p. 786-95.

206. Breiman, L., *Classification and regression trees*. Wadsworth statistics/probability series1984, Belmont, Calif.: Wadsworth International Group. x, 358 p.

207. Breiman, L., *The Little Bootstrap and Other Methods for Dimensionality Selection in Regression: X-Fixed Prediction Error*. Journal of the American Statistical Association, 1992. **87**(419): p. 738-754.

208. Quinlan, J.R. and E.B. Hunt, *A Formal Deductive Problem-Solving System.* J. ACM, 1968. **15**(4): p. 625-646.

209. Quinlan, J.R., *Induction of decision trees.* Machine Learning, 1986. **1**(1): p. 81-106.

210. Quinlan, J.R., *C4.5 : programs for machine learning*. Morgan Kaufmann series in machine learning1993, San Mateo, Calif.: Morgan Kaufmann Publishers. x, 302 p.

211. Duda, R.H., P, *Pattern classification and Scene Analysis*1973: Wiley.

212. Langley P, I.W., Thompson K. *An analysis of Bayesian classifiers*. in *Tenth national conference on arti cial intelligence*. 1992. AAAI Press amd MIT Press.

213. Vapnik, V.N., *Statistical learning theory*. Adaptive and learning systems for signal processing, communications, and control1998, New York: Wiley. xxiv, 736 p.

214. Yang, Y. and X. Liu, *A re-examination of text categorization methods.* Proceedings of SIGIR: International Conference on R&D in Information Retrieval, 1999. **22**: p. 42-49.

215. Holland, R.R., *Decision Tables.* JAMA: The Journal of the American Medical Association, 1975. **233**(5): p. 455-457.

216. L.A, Z., *Fuzzy sets.* Information and Control, 1965. **8**(3): p. 338-353.

217. R, K. *A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection.* in *International Joint Conference on Artificial Intelligence (IJCAI)*. 1995. Montreal, Canada: Morgan-Kaufmann.

218. H. M. Finucan, R.F.G.a.M.S., *Moments Without Tears in Simple Random Sampling from a Finite Population.* Biometrika, 1974. **61**(1): p. 151-154.

219. Efron, B. and R.J. Tibshirani, *An Introduction to the Bootstrap (Chapman & Hall/CRC Monographs on Statistics & Applied Probability)*1994: Chapman and Hall/CRC.

220. Efron, B., *Estimating the Error Rate of a Prediction Rule: Improvement on Cross-Validation.* Journal of the American Statistical Association, 1983. **78**(382): p. 316-331.

221. Efron, B., *The Jackknife, the Bootstrap, and Other Resampling Plans (CBMS-NSF Regional Conference Series in Applied Mathematics)*1987: Society for Industrial Mathematics.

222. Drew, P.J. and J.R. Monson, *Artificial neural networks.* Surgery, 2000. **127**(1): p. 3-11.

223. Kandaswamy, A., et al., *Neural classification of lung sounds using wavelet coefficients.* Computers in Biology and Medicine, 2004. **34**(6): p. 523-537.

224. Tourassi, G.D., et al., *A neural network approach to breast cancer diagnosis as a constraint satisfaction problem.* Medical physics, 2001. **28**(5): p. 804-11.

225. Tourassi, G.D., et al., *Application of the mutual information criterion for feature selection in computer-aided diagnosis.* Medical physics, 2001. **28**(12): p. 2394-402.

226. Baxt, W.G., *Application of artificial neural networks to clinical medicine.* Lancet, 1995. **346**(8983): p. 1135-8.

227. Baxt, W.G., *Use of an artificial neural network for the diagnosis of myocardial infarction.* Annals of internal medicine, 1991. **115**(11): p. 843-8.

228. Baxt, W.G. and J. Skora, *Prospective validation of artificial neural network trained to identify acute myocardial infarction.* Lancet, 1996. **347**(8993): p. 12-5.

229. Needham, C.J., et al., *Predicting the effect of missense mutations on protein function: analysis with Bayesian networks.* BMC Bioinformatics, 2006. **7**: p. 405.

230. Luk, J., et al., *Artificial neural networks and decision tree model analysis of liver cancer proteomes.* Biochemical and Biophysical Research Communications, 2007. **361**(1): p. 68-73.

231. Horowitz, N., et al., *Applying Data Mining Techniques in the Development of a Diagnostics Questionnaire for GERD.* Digestive Diseases and Sciences, 2007. **52**(8): p. 1871-1878.

232. Mofidi, R., et al., *Identification of severe acute pancreatitis using an artificial neural network.* Surgery, 2007. **141**(1): p. 59-66.

233. Liew, P., et al., *Comparison of artificial neural networks with logistic regression in prediction of gallbladder disease among obese patients.* Digestive and Liver Disease, 2007. **39**(4): p. 356-362.

234. Patil, S., et al., *Neural network in the clinical diagnosis of acute pulmonary embolism.* Chest, 1993. **104**(6): p. 1685-9.

235. Saftoiu, A., et al., *Neural network analysis of dynamic sequences of EUS elastography used for the differential diagnosis of chronic pancreatitis and pancreatic cancer.* Gastrointestinal Endoscopy, 2008. **68**(6): p. 1086-94.

236. Wu, E.J., et al., *Artificial neural network: border detection in echocardiography.* Medical & biological engineering & computing, 2008. **46**(9): p. 841-8.

237. Harrison, R. and R. Kennedy, *Artificial Neural Network Models for Prediction of Acute Coronary Syndromes Using Clinical Data From the Time of Presentation.* Annals of Emergency Medicine, 2005. **46**(5): p. 431-439.

238. Cucchetti, A., et al., *Artificial neural network is superior to MELD in predicting mortality of patients with end-stage liver disease.* Gut, 2007. **56**(2): p. 253-258.

239. Daskalakis, A., et al., *Design of a multi-classifier system for discriminating benign from malignant thyroid nodules using routinely H&E-stained cytological images.* Computers in Biology and Medicine, 2008. **38**(2): p. 196-203.

240. Mat-Isa, N.A., M.Y. Mashor, and N.H. Othman, *An automated cervical pre-cancerous diagnostic system.* Artificial Intelligence in Medicine, 2008. **42**(1): p. 1-11.

241. van Gerven, M.A., B.G. Taal, and P.J. Lucas, *Dynamic Bayesian networks as prognostic models for clinical patient management.* Journal of Biomedical Informatics, 2008. **41**(4): p. 515-29.

242. Peng, S.Y. and S.K. Peng, *Predicting adverse outcomes of cardiac surgery with the application of artificial neural networks.* Anaesthesia, 2008. **63**(7): p. 705-13.

243. Barbini, E., et al., *A comparative analysis of predictive models of morbidity in intensive care unit after cardiac surgery – Part I: model planning.* BMC Medical Informatics and Decision Making, 2007. **7**(1): p. 35.

244. Rowan, M., et al., *The use of artificial neural networks to stratify the length of stay of cardiac patients based on preoperative and initial postoperative factors.* Artificial Intelligence in Medicine, 2007. **40**(3): p. 211-221.

245. Bottaci, L., et al., *Artificial neural networks applied to outcome prediction for colorectal cancer patients in separate institutions.* The Lancet, 1997. **350**(9076): p. 469-472.

246. Dolgobrodov, S.G., et al., *Artificial neural network: predicted vs observed survival in patients with colonic cancer.* Diseases of the colon and rectum, 2007. **50**(2): p. 184-91.

247. Lin, C.S., et al., *Predicting hypotensive episodes during spinal anesthesia with the application of artificial neural networks.* Computer Methods and Programs in Biomedicine, 2008. **92**(2): p. 193-7.

248. Lisboa and P., *A Bayesian neural network approach for modelling censored data with an application to prognosis after surgery for breast cancer.* Artificial Intelligence in Medicine, 2003. **28**(1): p. 1-25.

249. Lisboa, P.J., et al., *Time-to-event analysis with artificial neural networks: an integrated analytical and rule-based study for breast cancer.* Neural networks : the official journal of the International Neural Network Society, 2008. **21**(2-3): p. 414-26.

250. Wu, Y., et al., *Artificial neural networks in mammography: application to decision making in the diagnosis of breast cancer.* Radiology, 1993. **187**(1): p. 81-7.

251. Jarman, I.H., et al., *An integrated framework for risk profiling of breast cancer patients following surgery.* Artificial Intelligence in Medicine, 2008. **42**(3): p. 165-88.

252. Mofidi, R., et al., *Prediction of survival from carcinoma of oesophagus and oesophago-gastric junction following surgical resection using an artificial neural network☆.* European Journal of Surgical Oncology, 2006. **32**(5): p. 533-539.

253. Copeland, G.P., D. Jones, and M. Walters, *POSSUM: a scoring system for surgical audit.* The British journal of surgery, 1991. **78**(3): p. 355-60.

254. Prytherch, D.R., et al., *POSSUM and Portsmouth POSSUM for predicting mortality. Physiological and Operative Severity Score for the enUmeration of Mortality and morbidity.* The British journal of surgery, 1998. **85**(9): p. 1217-20.

255. Tekkis, P.P., et al., *Risk-adjusted prediction of operative mortality in oesophagogastric surgery with O-POSSUM.* The British journal of surgery, 2004. **91**(3): p. 288-95.

256. Tekkis, P.P., et al., *Development of a dedicated risk-adjustment scoring system for colorectal surgery (colorectal POSSUM).* The British journal of surgery, 2004. **91**(9): p. 1174-82.

257. Hart, A.W., J, *Connectionist mdels in medicine:an investigation of their potential*, in *"nd European Conference on Artificial Intelligence in Medicine*, J.C. Hunter, J Wyatt, J, Editor 1989, Springer: Heidberg. p. 115-124.

258. Baxt, W.G., *Use of an Artificial Neural Network for Data Analysis in Clinical Decision-Making: The Diagnosis of Acute Coronary Occlusion.* Neural Computation, 1990. **2**(4): p. 480-489.

259. Goldman, L., et al., *A computer protocol to predict myocardial infarction in emergency department patients with chest pain.* The New England journal of medicine, 1988. **318**(13): p. 797-803.

260. Boon, M.E. and L.P. Kok, *Neural network processing can provide means to catch errors that slip through human screening of pap smears.* Diagnostic cytopathology, 1993. **9**(4): p. 411-6.

261. Burke, H.B., D.B. Rosen, and P.H. Goodman. *Comparing artificial neural networks to other statistical methods for medical outcome prediction*. in *Neural Networks, 1994. IEEE World Congress on Computational Intelligence., 1994 IEEE International Conference on*. 1994.

262. Fraser HS, K.R., Ross P, Harrison R, *A Comparison Of Radial Basis Functions And Back-Propagation In The Diagnosis Of Myocardial Infarction*, in *International Conference of Expert Systems and Neural Networks in Medicine*1994: Plymouth.

263. Fricker, J., *Artificial neural networks improve diagnosis of acute myocardial infarction.* The Lancet, 1997. **350**(9082): p. 935.

264. Chu, A., et al., *A decision support system to facilitate management of patients with acute gastrointestinal bleeding.* Artificial Intelligence in Medicine, 2008. **42**(3): p. 247-259.

265. WEKA, *Waikato Environment for Knowledge Analysis*, University of Waikato.

266. Berry, W.F., S, *Multiple regression in practice*. Quantative applications in the social sciences1985: SAGE.

267. Studenmund, A.H., Cassidy, H.J, *Using econometrics: A practical guide*1987, Boston: Little Brown.

268. Beart, R.W., et al., *Management and survival of patients with adenocarcinoma of the colon and rectum: a national survey of the Commission on Cancer.* Journal of the American College of Surgeons, 1995. **181**(3): p. 225-36.

269. Ferraris, R., et al., *Predictive value of rectal bleeding for distal colonic neoplastic lesions in a screened population.* European journal of cancer, 2004. **40**(2): p. 245-52.

270. Helfand, M., et al., *History of visible rectal bleeding in a primary care population. Initial assessment and 10-year follow-up.* JAMA : the journal of the American Medical Association, 1997. **277**(1): p. 44-8.

271. Leung, E., et al., *The effectiveness of the '2-week wait' referral service for colorectal cancer.* International journal of clinical practice, 2010. **64**(12): p. 1671-4.

272. Smith, R.A., et al., *Outcomes in 2748 patients referred to a colorectal two-week rule clinic.* Colorectal Disease, 2007. **9**(4): p. 340-343.

273. Cook, I.J., et al., *Gastrointestinal investigation of iron deficiency anaemia.* British medical journal, 1986. **292**(6532): p. 1380-2.

274. Hardwick, R.H. and C.P. Armstrong, *Synchronous upper and lower gastrointestinal endoscopy is an effective method of investigating iron-deficiency anaemia.* The British journal of surgery, 1997. **84**(12): p. 1725-8.

275. Zuckerman, G. and J. Benitez, *A prospective study of bidirectional endoscopy (colonoscopy and upper endoscopy) in the evaluation of patients with occult gastrointestinal bleeding.* The American journal of gastroenterology, 1992. **87**(1): p. 62-6.

276. James, M.W., et al., *Risk factors for gastrointestinal malignancy in patients with iron-deficiency anaemia.* European journal of gastroenterology & hepatology, 2005. **17**(11): p. 1197-203.

277. Maslekar, S., et al., *Artificial neural networks to predict presence of significant pathology in patients presenting to routine colorectal clinics.* Colorectal disease : the official journal of the Association of Coloproctology of Great Britain and Ireland, 2010. **12**(12): p. 1254-9.

278. Davies, R.J., et al., *A prospective study to assess the implementation of a fast-track system to meet the two-week target for colorectal cancer in Somerset.* Colorectal disease : the official journal of the Association of Coloproctology of Great Britain and Ireland, 2002. **4**(1): p. 28-30.

279. Mahon, C.C., et al., *Preliminary evaluation of United Kingdom National Referral Guidelines for lower gastrointestinal tract cancer.* Colorectal disease : the official journal of the Association of Coloproctology of Great Britain and Ireland, 2002. **4**(2): p. 111-114.

280. Walsh, S., et al., *The fourteen-day rule and colorectal cancer.* Annals of the Royal College of Surgeons of England, 2002. **84**(6): p. 386-8.

281. Lundin, M., et al., *Artificial neural networks applied to survival prediction in breast cancer.* Oncology, 1999. **57**(4): p. 281-6.

282. Hunt, D.L., et al., *Effects of computer-based clinical decision support systems on physician performance and patient outcomes: a systematic review.* JAMA : the journal of the American Medical Association, 1998. **280**(15): p. 1339-46.

283. Huo, Z., M.L. Giger, and C.E. Metz, *Effect of dominant features on neural network performance in the classification of mammographic lesions.* Physics in medicine and biology, 1999. **44**(10): p. 2579-95.

284. Qureshi, K.N., et al., *Neural network analysis of clinicopathological and molecular markers in bladder cancer.* The Journal of urology, 2000. **163**(2): p. 630-3.

285. Witten I, F.E., *Data Mining: Practical machine learning tools and techniques*2005, San Francisco: Elseiver.

286. Atkin, W.S., et al., *Once-only flexible sigmoidoscopy screening in prevention of colorectal cancer: a multicentre randomised controlled trial.* The Lancet, 2010. **375**(9726): p. 1624-1633.

287. Weissfeld, J.L., et al., *Flexible Sigmoidoscopy in the PLCO Cancer Screening Trial: Results From the Baseline Screening Examination of a Randomized Trial.* Journal of the National Cancer Institute, 2005. **97**(13): p. 989-997.

288. Lieberman, D.A., et al., *Use of colonoscopy to screen asymptomatic adults for colorectal cancer. Veterans Affairs Cooperative Study Group 380.* The New England journal of medicine, 2000. **343**(3): p. 162-8.

289. Schoenfeld, P., et al., *Colonoscopic screening of average-risk women for colorectal neoplasia.* The New England journal of medicine, 2005. **352**(20): p. 2061-8.

290. Mango L J. Tjon R, H.J., *Computer assisted Pap Smear screening using neural networks.* World Congress on Neural Networks: 1994 International Neural Network Society, 1994. **1**: p. 84-89.

291. Dolgobrodov, S.G., et al., *Artificial Neural Network: Predicted vs. Observed Survival in Patients with Colonic Cancer.* Diseases of the Colon & Rectum, 2006. **50**(2): p. 184-191.

292. Lin, C., et al., *Predicting hypotensive episodes during spinal anesthesia with the application of artificial neural networks.* Computer Methods and Programs in Biomedicine, 2008. **92**(2): p. 193-197.

293. Belacel, N. and M.R. Boulassel, *Multicriteria fuzzy assignment method: a useful tool to assist medical diagnosis.* Artificial intelligence in medicine, 2001. **21**(1-3): p. 201-7.

294. Seker, H., et al., *A fuzzy logic based-method for prognostic decision making in breast and prostate cancers.* IEEE transactions on information technology in biomedicine : a publication of the IEEE Engineering in Medicine and Biology Society, 2003. **7**(2): p. 114-22.

295. Kim MW, R.J., *Optimized fuzzy classification using genetic algorithm*, in *Lecture notes in artificial intelligence*,, J.Y. Wang L, Editor 2005, Springer: Berlin. p. 392-401.

296. Sivasankar, E.R., R.S, *Knowledge discovery in medical datasets using a Fuzzy Logic rule based classifier.* Electronic Computer Technology (ICECT), 2010 International Conference, 2010: p. 208 - 213.

297. Rowan, M., et al., *The use of artificial neural networks to stratify the length of stay of cardiac patients based on preoperative and initial postoperative factors.* Artificial Intelligence in Medicine, 2007. **40**(3): p. 211-21.