University of Cincinnati

Date: 3/20/2014

I, Boyang Bian , hereby submit this original work as part of the requirements for the degree of Doctor of Philosophy in Pharmaceutical Sciences/Biopharmaceutics.

It is entitled:

Exploring and Developing Algorithm of Predicting Advanced Cancer Stage of Colorectal Cancer Based on Medical Claim Database

Student's name: **Boyang Bian**

This work and its defense approved by:

Committee chair: Jianfei Guo, Ph.D.

Committee member: Jane Pruemer, Pharm.D.

Committee member: Christina Kelton, Ph.D.

Committee member: Wei Pan, Ph.D.

Committee member: Patricia Wigle, Pharm.D.

9253

Exploring and Developing Algorithm of Predicting Advanced Cancer Stage of Colorectal Cancer Based on Medical Claim Database

A dissertation submitted to the
Graduate School
of University of Cincinnati
in partial fulfillment of the
requirement of the degree of

Doctor of Philosophy

in the Division of Pharmacy Practice and Administrative Sciences
of the James Winkle College of Pharmacy

By
Boyang Bian

March 2014

Dissertation Committee Members

Jeff J. Guo, Ph.D. (Chair)

Christina M.L. Kelton, Ph.D.

Wei Pan, Ph.D.

Jane M. Pruemer, Pharm. D.

Patricia R. Wigle, Pharm. D.

Abstract

Background:

Colorectal cancer (CRC) is a type of cancer which develops from uncontrolled cell growth in the colon or rectum. It is the third most commonly diagnosed cancer in males and the second in females. In epidemiologic research for CRC, advanced cancer stage is an important factor for determining disease development and treatment patterns. However, this variable is not available because medical claims databases is retrospective and only original built for financial analysis only. Algorithms to predict advanced CRC stage were developed based on the existing medical information in claims database.

Method:

Study cohorts were identified from the Surveillance Epidemiology and End Results (SEER)-Medicare database. Two algorithms were constructed based on covariates obtained from the database for different study periods, including demographic, treatment pattern variables. The training set was used to derive predictive equations by using logistic regression model, then applied to validation set for evaluating the predictive characteristics (sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV)). The developed algorithm were applied to MarketScan® Commercial Claims and Encounters Database and tested the predictive values.

Results:

The algorithm of predicting advanced CRC stage in 1999 to 2003 achieved sensitivity 50.3% and specificity 95.0%, PPV 66.78% and NPV 90.58% while the equation distinguishing CRC stage IV in 2004 to 2007 achieved sensitivity 56.8%, specificity 95.3%, PPV 71.86% and NPV 91.19%. All algorithms made better predictive values than the single ICD-9 metastatic diagnosis as the predictor. Then the algorithm for 1999 to 2003 was applied to MarketScan database. 9484 patients were predicted as non-advanced CRC group while 1097 patients were assigned to advanced CRC group.

Conclusion

Claims-based algorithms were developed to predict advanced cancer stage. These algorithms were shown to be successful in the recent study period due to the inclusion of new biologic agents, which were utilized in advanced cancer treatment. This predictive algorithm may be applied in claims database and generate cancer stage information, which can assist with epidemiologic study of patients with CRC.

ACKNOWLEGMENTS

I would like to dedicate this work to my parents in China, for their love and support for the journey of my life. May my father rest in peace.

My deepest heartfelt thank you is given to my advisor, Dr. Jeff J. Guo, who has been a great mentor and role model for me. His enthusiasm in academic research and the strong motivation has encouraged me during my whole graduate study time.

It is impossible to express my appreciation to Dr. Christina ML. Kelton, who has been a remarkably valuable asset to my research in graduate study time and a truly friend. With her support and guidance, each year I am able to make solid step forward.

Great appreciation is extended to Dr. Patricia R. Wigle, Dr. Jane M. Pruemer, Dr. Wei Pan, for their time, attention and continuing guidance, I appreciate their knowledge and commentary for this dissertation research.

I would like to give my exceptional thanks to my wife Ying Xia, for her assistance, encouragement during last five year.

Table of Contents

List of Tables

List of Figures

Introduction

Background

Colorectal cancer (CRC) is a type of cancer which develops from uncontrolled cell growth in the colon or rectum (part of the large intestine). As the third most commonly diagnosed cancer in males and the second in females, approximately 1.2 million new colorectal cancer cases and an estimated 608,700 deaths occurred worldwide in 2008 (Jemal et al., 2011). It also causes about a half million deaths annually around the world and has a higher prevalence rate in developed countries than developing countries (Merika, Saif, Katz, Syrigos, & Morse, 2010). In 2007, colorectal cancer was the third leading cause of cancer mortality for males and the fourth leading cause for females in America (27,125 cases and 26,461 cases, respectively). The age-adjusted death rate was 16.9 people per 100,000 in the US (Xu, 2009).

In 2008, colorectal cancer had the second highest economic impact of all cancers, which resulted in $99 billion in direct and indirect costs globally. In high-income countries, which includes US, the estimated Disability-Adjusted Life Years (DALYs) lost for colorectal cancer was 2,117,900. (AmericanCancerSociety, 2010) In the United States, the estimated expenditures for colorectal cancer were $5.3 billion per year, including direct and indirect costs together. (Jansman, Postma, & Brouwers, 2007)

Cancer stage is a clinically useful classification scheme to encompass the attributes of the tumor that define its behavior. In pharmacoepidemiologic and health-outcomes studies, cancer stage is an important predictor of outcome. Surprisingly, the need of cancer stage predicting algorithms for claims database have been ignored for a long time. Limited success in predicting

advanced cancer stage from claims databases thus far have led to the theory that greater success is currently achievable because of the new pharmacotherapies available to treat advanced-stage cancer.

Definition and Pathology of Colorectal Cancer

CRC is a type of cancer in which malignant tumors arise from the inner wall of the colon or rectum. The symptoms of CRC can vary depending on the location of the lesion. Right-sided tumors normally present with symptoms like anemia, abdominal pain, or a change in bowel habits. If the lesion is left-sided, the most common presenting symptoms are a change in bowel habits, rectal blood loss, and abdominal pain.(Wayne, Cath, & Pamies, 1995) Most CRC cases can be attributed to sporadic factors (88-94%). A smaller number of CRC cases are hereditary.

The large intestine in an average adult is about 1.5 meters and extends from the terminal ileum to the anal canal. It consists of three major parts: cecum, colon (including ascending, transverse, descending, and sigmoid colon), and rectum. As the last part of the digestive system in the human body, the ascending colon is connected to the ileum by a large tube-like section of bowel called the cecum. It extracts water and salt from solid wastes before they are excreted by the body. (Potter, 1999) Food waste passes through the ascending, transverse, and descending colon, then remains in the sigmoid colon until it is ready to be excreted from the body. The rectum is located after the sigmoid colon and extends 13 to 15 cm to the anus. It is the temporary storage site for feces prior to discharge. (Greene & American Joint Committee on Cancer., 2006b)

Colorectal carcinoma is a multiple step progression of genetic mutation and phenotypic alternations, which eventually leads to uncontrolled cell growth, proliferation and tumor growth. Causes of genetic tumor genesis events for CRC include gene mutations, epigenetic silencing of gene transcription, loss of heterozygosity and gene amplification. (Potter, 1999) Genetic changes for CRC include oncogenes activation, tumor suppressor genes inactivation, and defects in MMR genes, especially the K-*ras* and N-*ras* genes. The *ras* gene family is responsible for encoding protein which is in charge of the transmission of the nucleus growth. (Arends, 2000) Activation of *ras* leads to a constitutive activity of protein, resulting in continuous stimulus of cell proliferation and other activities that promote carcinogenesis. Inactivation of tumor suppressor genes may assist or accelerate transformation of normal cells to cancer cells.

Epidemiology

CRC is the third most common cancer worldwide. Approximately 1.2 million new cancer cases and an estimated 608,700 deaths occurred in 2008. (Jemal et al., 2011) It also caused about half a million deaths annually around the world and has a higher prevalence rate in developed countries compared to developing countries. (Merika et al., 2010) The geographical variations in both incidence and mortality rate are significant. The highest rates are estimated in Australia/New Zealand (39 per million), Western Europe (33.1) North America (30.1), Eastern Asia (18) and more recently in Japan. The highest incidence rate of CRC is estimated in the Czech Republic (43). The lowest incidence rates are estimated in Africa (3.6 except South Africa) and South-Central Asia (4.5). The highest mortality rates in both sexes are estimated in Central Europe (20.3 for male patients, 12.1 for female patients) and the lowest in Middle Africa (3.5

and 2.7 respectively). (Miladinov-Mikov, 2010) (Seen in Figure 1) These geographic differences appear to be attributable to differences in dietary and environmental exposures that are imposed upon a background of genetically determined susceptibility. In the US, colorectal cancer was the third leading cause of cancer mortality for males and the fourth leading cause for females in 2007. The age-adjusted death rate is 16.9 people per 100,000 in the US.(Xu JQ, 2010) The incidence rate of CRC declined by 2.9% annually from 1998 which has been attributed to increasing CRC screening and therapeutic interventions. The 5-year survival rate for patients diagnosed with CRC is approximately 60%; however, survival improves substantially if the cancer is diagnosed while it is still localized (74% for stage I and 67% for stage IIA). Unfortunately, approximately 20% of CRC patients who do receive screening may be diagnosed in the later or even metastatic stage.

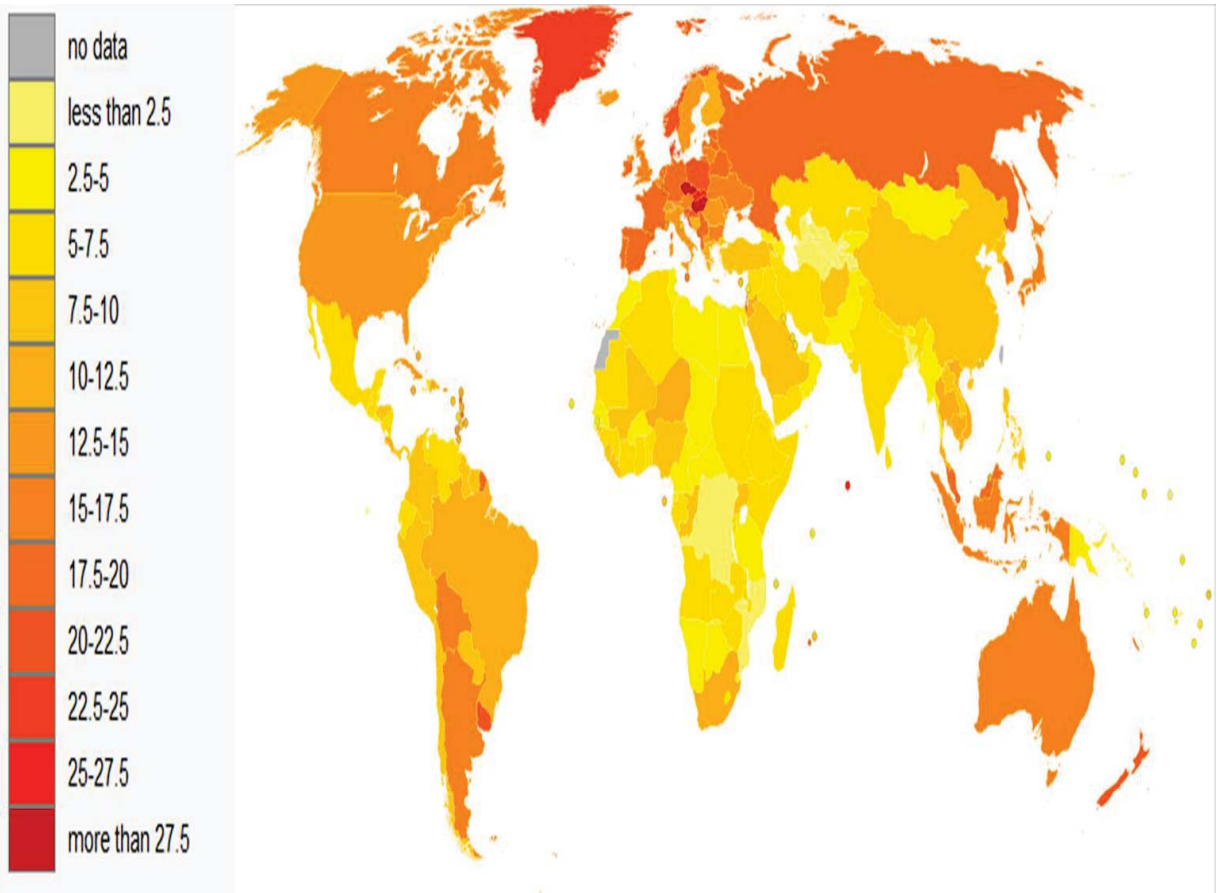


Figure 1. Age-standardized death from colorectal cancer per 100,000 inhabitants in 2004(WHO, 2010)

Economic Impact

CRC has the second highest economic impact of all cancers, resulting in $99 billion in direct and indirect costs worldwide in 2008. In high-income countries (which includes US), the estimated DALYs lost for colorectal cancer was 2,117,900. (AmericanCancerSociety, 2010) In the United States, the estimated expenditures on colorectal cancer were $5.3 billion per year. Based on the Surveillance, Epidemiology, and End Results (SEER) Medicare data, the total respective estimated direct cost over a 25 year period for patients is $59,919 for male and $59,438 for female (year 1984-1994 value). (Etzioni, Ramsey, Berry, & Brown, 2001) The nonmedical costs of CRC are another part of its socioeconomic impact. According to Yabroff, et al., (Yabroff, Warren, Knopf, Davis, & Brown, 2005) the patient time costs for treating CRC were $4,592 in the initial first 12 months of cancer care, $2,788 in the last 12 months of the disease's terminal phase, and $25 per month in the continuing treatment phase. In summary, the estimated lifetime cost of disease management for a colorectal cancer patient is close to $100,000 based on North American Data (mainly US and Canada). (Jansman et al., 2007)

Risk Factors of Colorectal Cancer

A great number of studies show several risk factors are associated with the development of colorectal cancer, including dietary and nutrition factors, lifestyle patterns, certain clinical comorbid conditions, and genetic susceptibilities.

Dietary factors. Composed of remnants of plant cells which cannot be processed by the human digestive system, dietary fiber is thought to protect against cancer and postulated to protect colonic cell exposure from certain carcinogens. (Asano & McLeod, 2002)

On the other side, studies suggest the association between fat and higher colorectal cancer risk, especially red and processed meat. Higher fat intake is related to an increased risk of colorectal cancer even after adjusting for fruit and vegetable consumption and other relevant factors. (Jarvinen, Knekt, Hakulinen, Rissanen, & Heliovaara, 2001)

Lifestyle Factors. Alcohol and tobacco consumption increases the risk of colorectal cancer in both incidence and mortality. This finding has been observed for both men and women in different observational studies. (Bagnardi, Blangiardo, La Vecchia, & Corrao, 2001; Chao et al., 2000; Liang, Chen, & Giovannucci, 2009; Moskal, Norat, Ferrari, & Riboli, 2007)

Evidence from observational and intervention studies suggests that aspirin and other non-steroidal anti-inflammatory drugs (NSAIDs) protect against the probability of developing colorectal cancer. Regular use of aspirin and other NSAIDs are associated with significant reduction in the risk of colorectal cancer in patients. (Dubé et al., 2007; Rostom et al., 2007)

Independent of different levels of physical activity, exogenous hormone use is also associated with an elevated risk of colon or rectal cancer.

Clinical and Generic Risk Factors. Patients with chronic ulcerative colitis and colonic Crohn's disease have an increased risk of colorectal cancer compared with the general population. If the lesion involves the entire large intestine, the risk is 5 to 10 times higher than average.

Patients who have familial colon cancer represent the least-understood pattern of colorectal cancer. Approximately twenty percent of these patients have family history of colorectal cancer. The most common hereditary colorectal cancers resulting from specific germ

line mutation are familial adenomatous polyposis (FAP) and hereditary nonpolyposis colorectal cancer (HNPCC). The risk of developing colorectal cancer for individuals with untreated FAP is virtually 100%; most will develop colorectal cancer when the patients are approaching 40 to 50 years of age. HNPCC (also called Lynch syndrome) is an autosomal dominant inherited syndrome that accounts for up to 5% of colon cancer cases. Multiple generations within a family are affected and colorectal cancer often develops early in life, with a mean age at the time of diagnosis of 45 years.

Disease Prevention

Currently, cancer prevention strategies can be classified as either primary or secondary. The aim of primary prevention is to prevent the development of colorectal cancer in at-risk populations, while secondary strategies are more focused on avoiding malignancy progression in patients who have already demonstrated an initial disease diagnosis. The list of prevention strategies for colorectal cancer is found in Table 1.

Diet. Although certain diets are commended for CRC high risk people, no consistent research data shows that increasing dietary fiber and/or decreasing dietary fat might reduce the risk of CRC. Further investigation of the role of high fiber diets or the use of fiber supplements is needed. Vitamin D, and folate are reported to have an inverse relationship with colorectal risk in observational studies. (Lamprecht & Lipkin, 2003)

Medication. The most widely used medication for preventing CRC is nonsteroidal anti-inflammatory medications like aspirin. Other NSAIDs are also reported to have preventative effects for reducing incident cases of colorectal cancer.(Rostom et al., 2007) However, according to the US Preventive Services Task Force's clinical guideline, (USPreventiveServicesTaskForce, 2007) the potential harm related to the use of these drugs outweighs the benefits for prevention in the general population.

Aspirin's effect on primary or secondary prevention for CRC is still controversial. Regular use of aspirin, especially in high doses, could reduce the incidence of CRC by 22%, while the risk of bleeding complications also increases with higher doses. Celecoxib, a COX-2 inhibitor, showed reduced sporadic adenoma formation by more than 30% compared to placebo. However, the risk of cardiovascular adverse events was also increased in the patients who received celecoxib.

Other agents like calcium for chemoprevention therapy have also showed positive effects. (Baron et al., 1999; Wactawski-Wende et al., 2006)

Surgical Prevention. Surgical resection is still an option for high risk individuals to prevent colon cancer development. Although NSAIDs could potential reduce CRC development, the effect is incomplete and cannot replace surgical resection for high risk patients, like those with HNPCC. Currently, certatin procedures including Fecal occult blood test (FOBTs), Double-contrast barium enema (DCBE), Fecal immunochemical tests (FITs), and flexible sigmoidoscopy are commended for average risk population. For higher risk group, colonoscopic polypetomy or

removal of polyps during screening colonscopies is considered the standard of care for CRC prevention.

Table 1. Prevention Strategies for Colorectal Cancer

Primary Prevention Strategies	
Diet	High-fiber diet supplementation
	Dietary fat reduction
	Vitamin D, folate
Chemoprevention	Nonsteroidal anti-inflammatory medications (asprine, non-asprine NSAIDs or COX-2 inhibitors)
	Calcium
	Selenium
	Estrogens
	Ursodeoxycholic acid
	Eflornithine
	Curcumin
Secondary Prevention Strategies	
Average risk population	Fecal occult blood test (FOBTs)
	Fecal immunochemical tests (FITs)
	Flexible sigmoidoscopy
	Double-contrast barium enema (DCBE)
	Colonoscopy
Increased risk population	Colonoscopy (time to begin with various year based on different procedures)
High risk population	Early surveillance with endoscopy and genetic testing
	Colonoscopy and genetic testing
	Colonoscopy biopsies for dysolasia

Screening

The screening tests for CRC can be classified as structure (luminal) tests and fecal-based/stool tests. Both strategies can detect early cancer as well as adenomatous polyps. (Burt, 2010)

The structure test uses imaging technology for polyp screening. The best tools for luminal screening tests are colonoscopy, sigmoidoscopy, and CT colography. Colonoscopy is the most complete screening procedure for examining the entire large intestine and removing polyps at certain session, and is also considered to be the "gold standard" (Hewett, Kahi, & Rex, 2010) for assessing the accuracy of other screening methods. Although no randomized control trials directly confirm mortality reduction by colonoscopy, some observational epidemiological studies (Citarda, Tomaselli, Capocaccia, Barcherini, & Crespi, 2001; Muller & Sonnenberg, 1995; Winawer et al., 1993) suggest colonoscopies have a significant impact for CRC, which has been estimated to be more than a 50% reduction in incidence. Sigmoidoscopy (or flexible sigmoidoscopy) is another significant tool for reducing CRC mortality risk. Evidence from randomized clinical trials shows sigmoidoscopy reduces CRC incidence and mortality for individuals who used this screening tool (33% and 43% , respectively, compared with no screening group. (Atkin et al., 2010)

Fecal tests are designed for detecting cancer risk in stool samples, particularly occult blood (FOBT) or exfoliated DNA alternation (stool DNA test). Compared with structure tests, these tests are noninvasive and do not require bowel clearance. Although they are able to demonstrate efficacy in early detection of CRC risk, only limited information exists for their role in detecting precancerous polyps. Direct evidence from randomized clinical trials shows FOBT

reduces 13-year cumulative mortality rate by 33% compared with the unscreened group. (Mandel et al., 1993) Another observational study indicates the reduction of CRC mortality by using FOBT was 18% after adjusting for the non-compliance rate. (Scholefield, Moss, Mangham, Whynes, & Hardcastle, 2011) For those who are unwilling or unable to perform colonoscopy screening, there is evidence suggesting a stool DNA test may provide valuable noninvasive screening results. However (since it has not been approved in US), this screening tool is not considered a first-line screening tool in the current clinical guidelines.

Medical Treatment

Based on the National Comprehensive Cancer Network (NCCN) clinical practice guidelines, treatment goals for cancer of the colon or rectum are based on the stage of disease. Stage I, II and III are considered potentially curable, but need to managed in the context of micrometastases, which may be present. (Engstrom et al., 2009a, 2009b) Based on the number and site(s) of metastases, approximately 20% to 30% of patients with resectable metastases may be cured. Most patients with stage IV disease are not considered curable, but treatments for controlling metastatic disease and palliating symptoms exist, such as avoiding disease-related complications and prolonging survival.

For the patients whose cancer is considered curable, surgical resection of the primary tumor(s) is the mainstay component of treatment. Depending on the stage of disease and the site of the tumor, further adjuvant chemotherapy and/or radiation therapy after surgery of primary tumor(s) may be another appropriate option. For those who develop resectable metastatic disease, systemic chemotherapy is the standard treatment procedure and is more desirable than surgery.

Radiation therapy could also be helpful for disease palliation when the tumor is localized or chemotherapy is not effective anymore.

Surgery. For operable CRC, surgical approaches generally include complete removal (resection) of the tumor along with the marginal tumor-free bowel and a lymphadenectomy in the nearby area. Regional rectal cancer surgery procedures depend on the region of tumor involvement. For patients with lesions in the middle to upper rectum, a low anterior resection is the primary procedure of choice. An abdominoperineal resection is the procedure for patients with lesions in the lower rectum if either the amount of unaffected bowel is insufficiently far enough away from the tumor or too close to areas that cannot perform an anastomosis. Depending on the type and extent of procedure/surgery for CRC, the associated mortality rate is approximately 2%, while the morbidity rate is from 8% to 15%.

Surgery for metastatic stage CRC is more complicated. Depending on the extent and site of metastatic disease, complete resection of discrete hepatic, pulmonary, abdominal, or even brain metastases may be needed. It may offer patients the opportunity to experience extended disease free survival (DFS) time. Since 25% of patients present with hepatic metastases at diagnosis or 60% of them will develop during the course of the disease, hepatic-limited resection must be done in a timely manner. The survival rate is significantly favorable compared to the control group. (Gill, Blackstock, & Goldberg, 2007) However, two-thirds of patients who receive hepatic metastases resection will develop a recurrence, which is why post-surgery adjuvant chemotherapy is needed for these patients.

Adjuvant Therapy Procedures for Operable Disease Stage. For operable disease (stages I, II and III), adjuvant therapy for CRC will be administrated to patients for the purpose of eliminating residual micrometastatic disease after complete surgical removal of the tumor, therefore, decreasing the probability of tumor recurrence as well as improving survival rates. For stage I, adjuvant therapy may not be necessary since most patients are cured by surgery in this stage.(DeVita, Lawrence, & Rosenberg, 2011) Adjuvant therapies have not shown better treatment results in stage II patients unless they are at high-risk for relapsing with inadequate lymph node or other clinical symptoms. However, radiation is still necessary for stage II rectal cancer for controlling the marginal areas, which are difficult to resect. It is essential to administer adjuvant therapy to stage III patients (for both radiation and chemotherapy), since regional node involvement makes these patients have a high risk for recurrence and five year mortality. In this group, adjuvant therapy could significantly decrease the risk of cancer relapse and death.

Fluorouracil (5-FU) has been the most widely used chemotherapy medication for the adjuvant treatment of CRC. The combination treatment of 5-FU plus leucovorin (LV) in adjuvant therapy has shown substantial improvements in response rates compared to 5-FU monotherapy in several large randomized trials in stage II or III CRC patients. (Haller et al., 2005; IMPACT, 1995; Wolmark et al., 1999) In addition, Oxaliplatin combination therapy regimens have been proven to reduce the risk of cancer relapse and increase 3-year DFS as compared to 5-FU plus LV alone. (André et al., 2004) The NCCN guidelines recommend Oxaliplatin-based treatment as an option for stage III colon cancer patients who can tolerate combination therapy. (Engstrom et al., 2009a, 2009b) Other new chemotherapy agents and chemotherapy regimens are constantly being investigated in an attempt to improve upon the response and safety of fluorouracil plus LV

in the adjuvant therapy for early stage patients. With the success of cetuximab and bevacizumab in the metastatic setting, most current adjuvant trials are evaluating monoclonal antibodies in combination with the previously mentioned regimens for stage III patients.

Chemotherapy For Metastatic CRC. Most metastatic colorectal cancers are incurable and treatment goals are to reduce patient symptoms, improve quality of life, and extend survival. The common regimens for metastatic disease consist of *FOLFOX* (oxaliplatin plus 5-FU and LV), *FOLFIRI* (irinotecan plus 5-FU and LV), bevacizumab plus 5-FU or LV or *FOLFOX* or *FOLFIRI, CapOx* (capecitabine plus oxaliplatin) or capecitabine alone, 5-FU plus LV alone. (Engstrom et al., 2009b) Two meta-analyses (D. J. Jonker, Maroun, & Kocha, 2000; Simmonds, 2000) have estimated the magnitude of benefit and harm associated with palliative chemotherapy for metastatic colorectal cancer and the results from both studies suggest chemotherapy is beneficial in terms of palliation and improved overall survival (OS) in patients with metastatic colorectal cancer.

5-FU-based Regimens

Similar to adjuvant chemotherapy, 5-FU is a first-line chemotherapeutic option for metastatic colorectal cancer, although some studies consider IV bolus 5-FU ineffective for advanced stage CRC.(Sobrero, Aschele, & Bertino, 1997) Different regimens have been developed to extend the duration of drug effect, but also decrease the toxicity of the treatments. Based on one meta-analysis, 5-FU based regimens showed significant yet marginal benefits on both response rate and overall survival. ("Efficacy of intravenous continuous infusion of

fluorouracil compared with bolus administration in advanced colorectal cancer. Meta-analysis Group In Cancer," 1998; Sobrero et al., 1997) Mayo Clinic conducted several studies (Buroker et al., 1994; de Gramont et al., 1997) to assess the treatment effect of 5-FU based regimens, the results show no statistical difference in response rate, median survival or palliative effects. Meanwhile some toxic effects like leukopenia or stomatitis were caused by certain regimens and require hospitalization to manage these toxicities. With the incorporation of new chemotherapeutic agents, there is the potential for better efficacy and lower toxicity in the management of metastatic colorectal cancer.

***FOLFOX* (5-FU and LV plus oxaliplatin)**

FOLFOX is recommended by NCCN for first-line chemotherapy regimens for advanced (or metastatic) CRC. On January 9, 2004, the U.S. Food and Drug Administration approved oxaliplatin for injection (Eloxatin™, a trademark of Sanofi-Synthelabo Inc.), for use in combination with infusional5-FU and LV for the initial treatment of advanced colorectal cancer. (NationalCancerInstitute, 2010). Unlike cisplatin, the DNA damage induced by oxaliplatin may not be recognized by DNA MMR complex, therefore it could achieve better treatment outcomes for CRC triggered by HNPCC. The oxaliplatin and 5-FU based regimen has been recommended by NCCN as the first-line chemotherapy for metastatic CRC and shows higher response rates as well as increased progression-free survival (PFS). One meta-analysis demonstrated (Simmonds, 2000) a significant improvement in tumor response (50.7% vs 22.3%) and PFS (median: 9months vs 6.2 months) in the FOLFOX group compared to 5-FU monotherapy. However, this result was not statistically significant but did improve patient quality of life.

Oxaliplatin adverse events include renal toxicity, nausea and vomiting, as well as neuropathies, both acute and persistent. Acute neuropathies occur in approximately 90% of patients while persistent ones are cumulative adverse effects and seen mostly in all patients who respond to treatment. (Goldberg et al., 2004; Grothey, 2003)

FOLFIRI (5-FU and LV plus irinotecan)

FOLFIRI is another chemotherapy regimen recommended in NCCN for advanced CRC treatment. Several investigations (Colucci et al., 2005; Petrelli et al., 2013)have been done for assessing if the regimen can provide additional survival time for late stage CRC patients and other clinical outcomes. The results suggest the *FOLFIRI* group achieved better outcomes in response rate, median time –to-event disease progression and OS compared to 5-FU plus LV alone. The addition of irinotecan to 5-FU and LV doesn't increase or decrease the quality of life for end stage patients. Certain adverse effects which may cause the regimen reduction or discontinuation were also observed. Neutropenia was the most common one causing dose reduction or discontinuation. Other observed adverse effects include diarrhea, nausea and vomiting, asthenia, and abdominal pain.

Capecitabine

On June 15, 2005, the U.S. Food and Drug Administration approved capecitabine (Xeloda®), an oral, tumor-selective fluoropyrimidine carbamate, as a single-agent adjuvant treatment for stage III colon cancer patients who have undergone complete resection of the primary tumor and later been used in metastatic CRC. The convenient administrative method and

different toxicity profile makes it a great alternative and combination to 5-FU in the treatment of metastatic disease. Twelves (Twelves, 2002) pooled the two studies comparing oral capecitabine with 5-FU regimens for advanced CRC. In these studies,1207 patients were randomized to oral capecitabine or 5-FU plus LV and the results showed capecitabine was more favorable compared to Mayo Clinic regimen. Normally, infusional 5-FU is considered to be superior to bolus administration, and oral capecitabine may be easier to use. Also irinotecan and oxaliplatin have been combined with capecitabine and data show this incorporation will be safe and effective in the initial treatment of metastatic colorectal cancer. (Koopman et al.) The current FDA-approved indication for capecitabine in metastatic colon cancer is when therapy with a fluoropyrimidine alone is desired. Replacement of 5-FU and LV with capecitabine in other regimens is not currently approved (Mayer, 2007)

Biologic Therapy Agents

Bevacizumab is a humanized monoclonal antibody which inhibits vascular endothelial growth factor A (VEGF-A). (Los, Roodhart, & Voest, 2007) It was approved by FDA as an initial treatment for patients with certain metastatic cancers in 2004, including CRC. Several studies have shown significant benefits gained when compared to chemotherapy alone. Hurwitz H, et al (Hurwitz et al., 2004) conducted a randomized clinical trial to compare bevacizumab in combination with FOLFIRI in metastatic CRC patients. The addition of bevacizumab to normal chemotherapeutic regimena showed an increase in response rate (44.9% vs. 34.7%), median survival (20.3% vs 15.6%) and PFS (10.6 vs 6.24 months) compared to FOLFIRI. The combination of bevacizumab with FOLFOX was also assessed by Saltz, et al (Saltz et al., 2011)

and Hochster HS, et al (Hochster et al., 2008). Their results demonstrate significant benefits (response rate, PFS and OS) in the bevacizumab group.

Cetuximab is a chimeric (mouse and human recombinant) monoclonal antibody which can directly inhibit epidermal growth factor receptor (EGFR). It binds to EGFR and turns off the uncontrolled growth in cancers with EGFR mutations. Cetuximab was approved by FDA in 2004 for treatment of EGFR-expressing, recurrent metastatic colorectal carcinoma in patients who are intolerant to irinotecan-based chemotherapy.(NationalCancerInstitute, 2011) It is recommended by NCCN as a second-line therapy agent for metastatic CRC. Research data suggest that cetuximab can be beneficial as an addition to oxaliplatin-based regimens. (Venook, 2006) As a single agent, cetuximab is associated with a 23% increasing in OS compared to supportive care. (Derek J. Jonker et al., 2007)

Cancer Staging Classification

Cancer stage is a clinically useful classification scheme to encompass the attributes of the tumor that define its behavior. It is based on the premise that cancers of the same anatomic site and histology share similar patterns of growth and similar outcomes. Normally, the cancer stage is defined by the TNM system which is accepted by the International Union Against Cancer (UICC) and American Joint Committee on Cancer (AJCC).

The TNM system is based on the extent of the tumor (T), the extent of spread to the lymph nodes (N), and the presence of distant metastasis (M). (Greene & American Joint Committee on Cancer., 2006b) A number for each letter indicates the size or extent of the tumor and the extent of cancer spread. Staging of cancer is the most important factor for predicting the

patient's survival rate, and cancer treatment is primarily determined by staging. Thus, staging does not change with progress of the disease as it is used to assess prognosis.

The colorectal cancer TNM classification is more detailed and precise than other identification system (like Dukes' system) in the prognostic subgroups. This staging classification applies to all carcinomas arising from the colon or rectum. The broader stage of a cancer is usually quoted as a number I, II, III, IV derived from the TNM value grouped by prognosis; a higher number indicates a more advanced cancer and likely a worse outcome. For CRC, the stage IV is defined as advanced cancer stage since the survival rate is significantly different than other stages.

Advanced disease stage of cancer includes regional metastatic disease and distant metastatic disease. Regional metastatic disease means the cancer cells from the original site penetrate or infiltrate in the tissues and form new tumor(s) in the adjacent site in the same region. Distant metastasis means the cancer cells penetrate the walls of lymphatic or blood vessels and finally form new tumor(s) and/or lymph node(s) in another site. The expression of advanced disease stage has been correlated with a poor prognosis based on the AJCC cancer staging manual (6th edition). Stage IV CRC has a lower 5-year survival rate compared with other less severity stages (8.1% for stage IV vs >44.3% for other stages). (Greene & American Joint Committee on Cancer., 2006b)

Purpose of Study

The objective of this study is to explore a certain registry database for additional information that might lead to a superior algorithm, relative to those developed earlier, for

identifying advanced cancer stage from claims data. We focus on colorectal cancer, although it is our hope, as well, to be able to shed some light on other cancers.

The objectives of this research are as follows:

1. Explore the SEER-Medicare database for the periods 1999-2003 and 2004-2007 with the goal of developing algorithms to predict advanced CRC stage;
2. Compare, using the standard measures of sensitivity, specificity, PPV, and negative predictive value, the best possible predictions for the earlier time period with those of the later time period; and
3. Understand the added value of any predictors used in the best algorithms which could provide as much generalizability to our results as possible in case future researchers work with databases with different types of predictor variables available.

Hypothesis of Study

Because of the importance of cancer stage, especially advanced- versus early-stage cancer, as the most clinically meaningful cancer-patient stratification, researchers relying on administrative databases for epidemiological or outcomes research are interested in identification of stage from information available in the data.

Based on existing literature, it would be tempting to give up on finding an algorithm to predict cancer stage from a claims database. However, there is a reason to believe that current advances in pharmacotherapy in the treatment of advanced-stage cancer may facilitate the development of better algorithms to predict cancer stage. For example, the drugs capecitabine

(Xeloda®) and bevacizumab (Avastin®) were introduced to the market in 2004 and are now widely used as first-line treatment or in combination with other pharmacologic agents for patients with advanced-stage cancer.

With the advancement of cancer treatment, especially late-stage treatment, it is our objective to explore the SEER-Medicare database for additional information that might lead to a superior algorithm, relative to those developed earlier, for identifying advanced cancer stage from claims data. We focus on colorectal cancer, although it is our hope, as well, to be able to shed some light on other cancers. Although it was not sure whether we would be able to improve on earlier results and consider our study exploratory in nature, we are hopeful that the effort would worthwhile in order to facilitate future pharmacoepidemiologic research in oncology in which stratification by disease stage could prove helpful.

Our primary research hypothesis is the following:

By identifying the relatively new biologic therapy agents in the 2004-2007 Medicare claims data, we will be able to develop a superior algorithm for identifying late-stage cancer than the algorithms developed using data from 1999-2003.

Significance

In pharmacoepidemiologic and health-outcomes studies, cancer stage is an important predictor of outcome. It could be employed as covariates or inclusion and exclusion criteria. However, normally, the cancer stage is defined by the TNM system.

Surprisingly, the need of cancer stage predicting algorithms for claims database have been ignored for a long time. Limited success in predicting advanced cancer stage from claims databases thus far have led to the theory that greater success is currently achievable because of the new pharmacotherapies available to treat advanced-stage cancer. Because of the importance of cancer stage in pharmacoepidemiologic and health-outcomes studies, it is important to keep working on algorithm development using the widely used database for oncologic studies, SEER-Medicare, available for that purpose. On the other side, the economic impact brought by the cost for CRC treatment, especially after new generation of biologic therapeutic agents, keeps adding great burdens in our society and different medical care payers.

Our study is significant for the following reasons:

- The very specific cancer stage algorithm for colorectal cancer has never been developed before.
- No claim-base algorithm study has been done for particular age group.
- The contribution for this study is to generate valuable cancer stage information in claim database and it could help other pharmacoepidemiologic studies to identify specific patients in different cancer stages.
- This study could also benefit insurance companies to apply the similar methods on different cancers and create unique cancer staging variable in claim database for managing the premium level on different cancer patients.

Literature Review

Cancer Staging

Purpose of Cancer Stage. To determine the extent of disease is the goal of staging examinations. A useful classification of disease could help a physician develop treatment options and estimate the prognosis of disease. In terms of oncology, cancer staging allows an oncologist to develop the best treatment strategy and predict patient's survival. Traditionally, Dukes classification (Dukes & Bussey, 1941) was used as the staging criteria for CRC. Now, AJCC cancer staging classification is widely used to determine oncology treatments and in research. Three significant elements of cancer, local tumor growth (T), spread to regional lymph nodes (N) and metastasis statues (M), are used to indicate the extent of disease at a particular time when the symptoms occur in a clinical examination. The mixture of the T, N, and M classifications into stage groupings is a method of designating the anatomic extent of a cancer and is related to the natural history of the particular type of cancer. It is intended to provide a means by which this information can readily be communicated to others, to assist in therapeutic decisions, and to help estimate prognosis. Eventually, it provides a mechanism for comparing similar groups of patients when evaluating different potential therapies. The significance of criteria for defining extent of disease differs from sites of tumor and histologic types. So, the T, N, M classification should be defined for each anatomic site to make the scheme valid. In addition to anatomic extent, the histologic type and histologic grade of the tumor may be important to determine the classification for staging.

The staging of cancer is used to analyze and compare groups of patients. It is preferable to achieve accurate information for the anatomic extent of the disease for each site, because the

precise clinical description and pathologic classification of malignant cancer may serve a number of related objectives: (1) to select primary and adjuvant therapy, (2) to estimate prognosis of diseases, (3) to assist the evaluation of treatment outcomes, (4) to facilitate of the exchange of information among healthcare institutions, and (5) to contribute to the continuing research of human cancers.

General Rule for Cancer Staging. The TNM system is a classification of the pathological extent of cancer and is based on three components:

- T the extent of the primary tumor
- N the presence and extent of regional lymph node metastasis
- M the presence of distant metastatic disease

The use of detail subsets system of the TNM components specifies the progressive extent of malignant statues.

Primary tumor: T0, T1, T2, T3, T4

T0: No primary tumor

Tis: Carcinoma *in situ*

T1-T4: Increasing size or local extent of primary tumor (tumor extension to lymph node is classified as lymph node metastasis.)

Regional Lymph Nodes: N0, N1, N2, N3

N0: No regional lymph node metastasis

N1-N3: Increasing involvement of regional lymph nodes (Not regional lymph node in any lymph system is classified as a distant metastatic disease.)

Distant Metastasis: M0, M1

M0: No distant metastasis

M1: Distant metastasis (either clinical or pathologic disease)

TNM Definition and Staging Classification for CRC. The colorectal cancer TNM classification is more detailed and precise than other identification system (like Dukes system) in the prognostic subgroups. TNM system is based on tumor invasion depth into the intestine's wall or extension to adjacent structures (T), the number of involved regional lymph nodes (N), and the presence or absence of distant metastasis (M). This staging classification applies to all carcinomas arising from the colon or rectum. Table 2 summarizes the detailed definition in AJCC TNM system. (Greene & American Joint Committee on Cancer., 2006a) Table 3 shows the staging grouping assignment based on TNM classification. (Greene & American Joint Committee on Cancer., 2006a)

CRC Prognosis by Stage. Based on the classification scheme of the AJCC staging, CRC has four categories (stage I, II, III, IV) based on the TNM system. The survival rate is significantly different among different stages. According to stages defined by the AJCC sixth edition system, the 5-year survival rate for patients diagnosed with advanced-stage CRC is no more than 8.1%, as opposed to a 44.3% or greater 5-year survival rate for patients in the earlier

stages of the disease. Table 4 shows the different stages survival data for colon and rectum cancer from SEER. (Greene & American Joint Committee on Cancer., 2006a)

Table 2. AJCC TNM Staging TNM for Colorectal Cancer

Criteria	Classifications	Definitions
Primary Tumor (T)		
	T0	No evidence of primary tumor
	Tis	Carcinoma *in situ*: intraepithelial or invasion of lamina propria
	T1	Tumor invades submucosa
	T2	Tumor invades muscularis propria
	T3	Tumor invades through the muscularis propria into the subserosa, or into nonperitonealized pericolic or perirectal tissues
	T4	Tumor directly invades other organs or structures, and/or perforates visceral peritoneum
Regional Lymph Nodes (N)		
	N0	No regional lymph node metastasis
	N1	Metastasis in 1 to 3 regional lymph nodes
	N2	Metastasis in 4 or more regional lymph nodes
Distant Metastasis (M)		
	M0	No distant metastasis
	M1	Distant metastasis

Table 3. AJCC TNM Classification for Colorectal Cancer

Stage	T	N	M
0	Tis	N0	M0
I	T1	N0	M0
	T2	N0	M0
IIa	T3	N0	M0
IIb	T4	N0	M0
IIIa	T1-T2	N1	M0
IIIb	T3-T4	N1	M0
IIIc	Any T	N2	M0
IV	Any T	Any N	M1

Table 4 Survival Rates of Colon and Rectum Cancer by Stage in SEER Registry

Cancer Stage	5-years Survival Rate	
	Colon	Rectum
I	74%	74%
II a	67%	65%
II b	59%	52%
II c	37%	32%
III a	73%	74%
III b	46%	45%
III c	28%	33%
IV	6%	6%

Medical Treatment in Advanced Stage Colorectal Cancer

In the last decade, some achievements have been made in metastatic CRC treatment. Surgery and radiation therapy are still used to manage isolated tumors in different sites. Chemotherapy, on the other side, is becoming most useful for patients who have disseminated or have unresectable metastatic disease.

Surgery. According to data from SEER registry(SEER, 2012), 19% of colorectal patients are diagnosed with stage IV disease. Complete surgical resection of colorectal related hepatic, pulmonary, abdominal, or brain metastases is critical and will prolong the DFS experience for advanced stage patients. Among those patients, 80-90% of them have unresectable metastatic liver disease. It has been estimated that over 50% of CRC mortality cases have hepatic metastases and the liver metastases are the cause of death in the majority of those patients. (Foster, 1984) Studies show surgical removal of colorectal liver metastases is a possible cure and the 5-year disease-free survival rates following by the procedure are approximately 20%. (Choti et al., 2002; Pawlik et al., 2005) Since two-thirds of the patients undergoing metastases resection will have cancer relapse, post-surgery therapy (e.g. adjuvant chemotherapy) should be taken to improve long-term outcomes for those patients.

However, for some patients who cannot perform resection due to metastases lesion site(s) or comorbidity, tumor ablation therapy may be the best option for them. However, the treatment effect may not as good as resection. A series of observational studies compared the effects of radiofrequency ablation (RFA) and liver resection in the treatment of liver metastases.(Gleisner et al., 2008; Hur et al., 2009; Reuter, Woodall, Scoggins, McMasters, & Martin, 2009) Most of

the results showed RFA is inferior to resection in recurrence rate as well as OS. NCCN clinical guideline(Engstrom et al., 2009a) concludes ablation should not be considered a substitute for resection in patients with resectable metastases.

Radiation Therapy. Radiation therapy is not normally used to treat advanced stage CRC but it may be used in certain circumstances. Symptom control is the primary goal for patients with advanced CRC. Radiation therapy accompanied with chemotherapy, is frequently used in the adjuvant or neoadjuvant setting for the treatment of rectal cancers, whereas chemotherapy alone is more common for the adjuvant and neoadjuvant treatment of colon cancers.

Chemotherapy. Chemotherapy is recommended after surgery for treating micrometastatic disease and prolonging the DFS. Current management of disseminated metastatic CRC uses various medication regimens, either single agents or in combination. 5-FU-based regimens, FOLFOX, FOLFIRI, capecitabine, irintotecan, bevacizumab, cetuximab and panitumumab are widely used in the treatment of stage IV CRC. The decision of therapy choice should be based on the goals of treatment, toxicity profile of regimens, as well as the previous treatment type and timing.

For patients with metastatic disease who qualify for intensive initial therapy, one of the five chemotherapy regimens could be selected: FOLFOX (including FOLFOX4 or mFOLFOX6), FOLFIRI, CAPOx, infusional 5-FU/LV or capecitabine, or FOLFOXRI. Biologic therapy agents could also be included as part of initial therapy. To define an appropriate

treatment plan, the site(s) of tumor involvement and history of prior chemotherapy need to be taken into consideration.

According to a phase III trial (Nordlinger et al., 2008) conducted by the European Organization for Research and Treatment of Cancer, FOLFOX treatment showed significant improvement in PFS for advanced stage patients. For resectable and resected patients, the absolute increase in rate of PFS in 3 years was 8.1% and 9.2%, respectively. The partial response rate after FOLFOX was 40% and mortality was less than 1% for both groups. Bevacizumab is considered in addition to FOLFOX for initial therapy for patients who suffer CRC characterized by the wild-type KRAS gene.

Evidence has shown comparable efficacy for FOLFOX and FOLFIRI. Tournigand et al. (Tournigand et al., 2004) assessed the interchangeability between FOLFOX and FOLFIRI regimens. Patients were randomly assigned to either the FOLFOX or FOLFIRI group at the beginning of treatment. At disease progression, two groups switched the chemotherapy agents. Finally, median survival was 21.5 months in FOLFIRI initial treatment group versus 20.6 months in FOLFOX initial group (P=0.99). Similar results were also found for median PFS.

CAPOx, the combination of capecitabine and oxaliplatin, is another first-line regimen for advanced stage CRC patients. Cassidy et al. (Cassidy et al., 2008) conducted a randomized trial to compare CAPOx with FOLFOX4 for treating advanced CRC. In a total of 2,034 patients, the two groups demonstrated similar efficacy for median PFS (8.0 months in CAPOx group vs. 8.5 months in FOLFOX4 group) and in OS (19.8 months in CAPOx group vs. 19.6 months in FOLFOX4 group). Neither result showed statistical significance. This study concluded CAPOx

shown no statistical significance to FOLFOX4 and could be used as first-line treatment of metastatic disease.

Infusional 5-FU/LV is an alternative treatment recommendation for patients who have impaired tolerance to other aggressive therapies. After this less intensive initial treatment, metastatic patients with no improvement should receive best supportive care. Furthermore, Capecitabine could be added to the regimen as an option for initial therapy.

FOLFOXIRI is recommended by NCCN clinical guideline (Engstrom et al., 2009a, 2009b) as a category 2B therapy regimen for unresectable metastatic patients. This regimen should be used without the addition of a biologic agent since the efficacy and safety data for the combination is insufficient. According to two randomized trials conducted by Falcone et al (Falcone et al., 2007) and Souglakos et al,(Souglakos et al., 2006), they observed better or similar results in PFS or OS in the FOLFOXIRI arm. However, studies also showed increased toxicity in the FOLFOXIRI regimen, which includeneurotoxicity, neutropenia, and diarrhea.

Bevacizumab is a humanized monoclonal antibody which inhibits vascular endothelial growth factor A (VEGF-A). (Los et al., 2007) It was approved by the FDA as initial treatment for patients with certain metastatic cancers in 2004, including CRC. Several studies show significant benefit compared to chemotherapy alone. Hurwitz H, et al (Hurwitz et al., 2004) conducted a randomized clinical trial to compare bevacizumab in combination with FOLFIRI in metastatic CRC patients. The addition of bevacizumab to normal chemotherapeutic regimens showed an increase in response rate (44.9% vs34.7%), median survival (20.3% vs 15.6%) and PFS (10.6 vs 6.24 months) compared to FOLFIRI. The combination of bevacizumab with

FOLFOX was also assessed by Saltz, et al (Saltz et al., 2011) and Hochster HS, et al (Hochster et al., 2008). The results also demonstrated significant benefit (response rate, PFS and OS).

Cetuximab is a chimeric (mouse and human recombinant) monoclonal antibody which can directly inhibit epidermal growth factor receptor (EGFR). Cetuximab binds to EGFR and turns off the uncontrolled growth in cancers with EGFR mutations. Cetuximab was approved by the FDA in 2004 for the treatment of EGFR-expressing, recurrent metastatic colorectal carcinoma in patients who are intolerant to irinotecan-based chemotherapy regimens.(NationalCancerInstitute, 2011) NCCN recommended it as a second-line therapy agent for metastatic CRC. Research data suggest that cetuximab can be beneficial as an addition to oxaliplatin-based regimens. (Venook, 2006) As a single agent, cetuximab is associated with a 23% increase in OS compared to supportive care. (Derek J. Jonker et al., 2007)

Cancer Staging Algorithms Research

Because of the importance of cancer stage, researchers relying on administrative databases for epidemiological or outcomes research are interested in identifying new methods of determining stage from available information in the database. International Classification of Diseases, 9th Revision, (ICD-9) codes, (Guo et al., 2006, 2007; Heaton et al., 2006) are used for developing inclusion and exclusion criteria and cohort selection in cancer patients for pharmacoepidemiologic database studies.(Carey et al., 2006; Iwashyna & Lamont, 2002; M. R. Smith et al., 2005) Several studies have demonstrated the reliability of ICD-9 codes for this purpose.(Freeman, Zhang, Freeman, & Goodwin, 2000; Nattinger, Laud, Bajorunaite, Sparapani, & Freeman, 2004; Warren, Feuer, Potosky, Riley, & Lynch, 1999) However, determination of

cancer stage is not so straightforward. Although there are some ICD-9 codes that are relevant to cancer stage, they are inadequate for the purpose of stage identification.

Cooper et al. Study. In 1999, Cooper et al (Cooper et al., 1999) used the SEER-Medicare claims database to evaluate the relative accuracy of ICD-9-based cancer stage identification for six commonly diagnosed cancers. All patients who were older than age 64 with incident cases of invasive breast, colorectal, endometrial, lung, pancreatic, and prostate cancer from 1984 to 1993 in SEER-Medicare database were selected. Cancer staging at diagnosis was included in SEER data and was coded by AJCC TNM system as well as the historic staging. This is considered the "gold standard" for future analysis. ICD-9 information was extracted for the cohort from Medicare Provider Analysis and Review files (MEDPAR) and outpatient files. For each patient, ICD-9 codes for primary and secondary diagnosis within three months of diagnosis date was searched. The predicted cancer stage was imputed for patients based on the ICD-9 codes. Finally, the sensitivity and positive predictive values (PPV) of predicted cancer stage were evaluated by using SEER stage with two-by-two table.

The results generated from 320,637 eligible cancer cases found that the identification method overestimated the localized tumors and insufficiently identified the distant-stage disease. In summary, ICD-9 codes are limited by misclassifying patients, especially the ones with distant stage cancers.

Thomas et al. Study. In 2002, Thomas et al (Thomas, Brooks, Mullins, Baquet, & Merchant, 2002) assessed the validity of ICD-9 codes on classifying disease stage in lung cancer patients. Medical records from 1996 to 1997, including a private insurance claims database and a

registry database for cancer pathologic records, were used in this study. All selected cases were cancer-free for six months prior to the beginning of study entry. The TNM staging information in this registry was considered as the "gold standard" and specifically converted to AJCC stage 0 to IV for lung cancer. In their definition, two groups were clustered: the localized stage (includes stage 0-I) and the advanced stage (stage II-IV).

A 77 patient cohort was generated from the database. Thirty out of 44 advanced-stage cancers (sensitivity of 68.2%) were classified correctly by using ICD-9 codes. This study still shows the ICD-9 codes are associated with underestimate in advanced cancer stage prediction.

Smith et al. Study. Relevant ICD-9 codes have proved insufficient to adequately determine cancer stage., Other researchers have attempted to develop more sophisticated algorithms. In 2010, Smith et al (G. Smith, Shih, Giordano, Smith, & Buchholz, 2010) developed a method to identify breast-cancer stage that was based on ICD-9 classification codes on medical procedures, clinic visits, medications, and demographic variables. These variables were shown to have a statistically significant relationship to cancer stage during algorithm development. They found that their algorithm outperformed the earlier ICD-9 method developed by Cooper et al. Prediction of distant disease using the Smith et al. method achieved a sensitivity of 81% (95% CI: 80%-84%) and a specificity of 89% (95% CI: 86%-89%).

Although Smith's study showed successful prediction for elderly patients with breast cancer. However, this algorithm could be improved by:

1) The positive predictive value (PPV) for Smith's algorithm was only 24%, and there were analytical issues that limit the usefulness of Smith's results. Examples are a

too restrictive set of predictive variables and an arbitrary cutoff for metastatic disease,

2) The predictors used in this algorithm did not follow the clinical guideline. Hormonal therapy was not included in the predicting variables. Also, the adjuvant therapies and combination regimens could be considered as combination variables and then applied to this algorithm;
3) After the calculation of predicted probability for different stages, the author used several arbitrarily selected cut-points to define the threshold of metastatic disease progress, instead of following the distribution of different cancer stage.
4) The method was limited to the SEER-Medicare population. The application for this methodology was not well addressed. It is unclear how to apply the method to a medical claim database, like MarketScan©.

Summary

Cancer stage is a clinically useful classification scheme to encompass the attributes of the tumor and define its behavior. It is based on the premise that cancers at the same anatomic site and with the same histology share similar patterns of growth and similar outcomes. Normally, the cancer stage is defined by the TNM system which is accepted by UICC and AJCC. The TNM system is based on the extent of the tumor (T), the extent of spread to the lymph nodes (N), and the presence of distant metastasis (M). Unlike some diseases such as asthma, where the disease progress can be inferred from the medication treatment pattern, treatment for cancer patients may be the same although the patient is in different stages of the disease.

Advanced disease staging for cancer includes regional metastatic disease and distant metastatic disease. The expression of advanced disease stage has been correlated with a poor prognosis. Based on the AJCC cancer staging manual (6th edition), the 5-year survival rate for colorectal cancer (stage IV) is no more than 6%. It is significantly lower than other stages which have at least a 28% 5-year survival rate.

In epidemiologic studies, advanced cancer stage is an important predictor of outcome. Unfortunately, this type of data does not exist in a medical claims database. The need for cancer stage predicting algorithms from a claims database has been ignored for a long time. Although several studies have assessed the validity of clinical classification codes for predicting cancer stage or tried to develop an algorithm for particular cancers, none of them achieved acceptable results.

Therefore, the primary objective of this study is to construct a predictive model for CRC advanced stage disease and to apply the algorithm to a medical claim database.

Methods

Overview

This dissertation seeks to develop an algorithm for predicting advanced cancer staging for CRC by using existing variables in an established cancer research database and then apply the predictors to a medical claims database. This chapter reviews the methodology to be used to generate the algorithms from the files of SEER-Medicare database colorectal cancer patients from two separate study periods: 1999 to 2003 and 2004 to 2007.

This study has two major sections:

1) Algorithms development: A series of advanced cancer stage algorithms were generated from SEER-Medicare database based on the clinical guideline and expert opinions. Logistic regression models were constructed for each study period.
2) Algorithm predictors' application in medical claim database: The algorithms developed in the previous section were applied to a medical claims database, then the predicted cancer stage for all CRC patients existing in the database were calculated.

Data Source

For the two research arms in this study, two databases were utilized.

SEER-Medicare Database. In the algorithms development section, the SEER-Medicare linked database was used to develop algorithms to predict advanced cancer stage. The SEER-Medicare database links two population-based databases which both provide unique

medical records of cancer patients from selected areas since 1973. SEER collects demographic, cancer staging, and cause of death, survival rate information (survival status and survival time). Medicare databases are well-established for medical claims data for all Medicare beneficiaries. Combining these two medical sources of records creates a unique research database for epidemiological as well as health outcome research. SEER-Medicare contains patients with incidents of cancer at different sites, which account for up to 26% of the United States' population. The data files we used in this study were as follows (See figure 2):

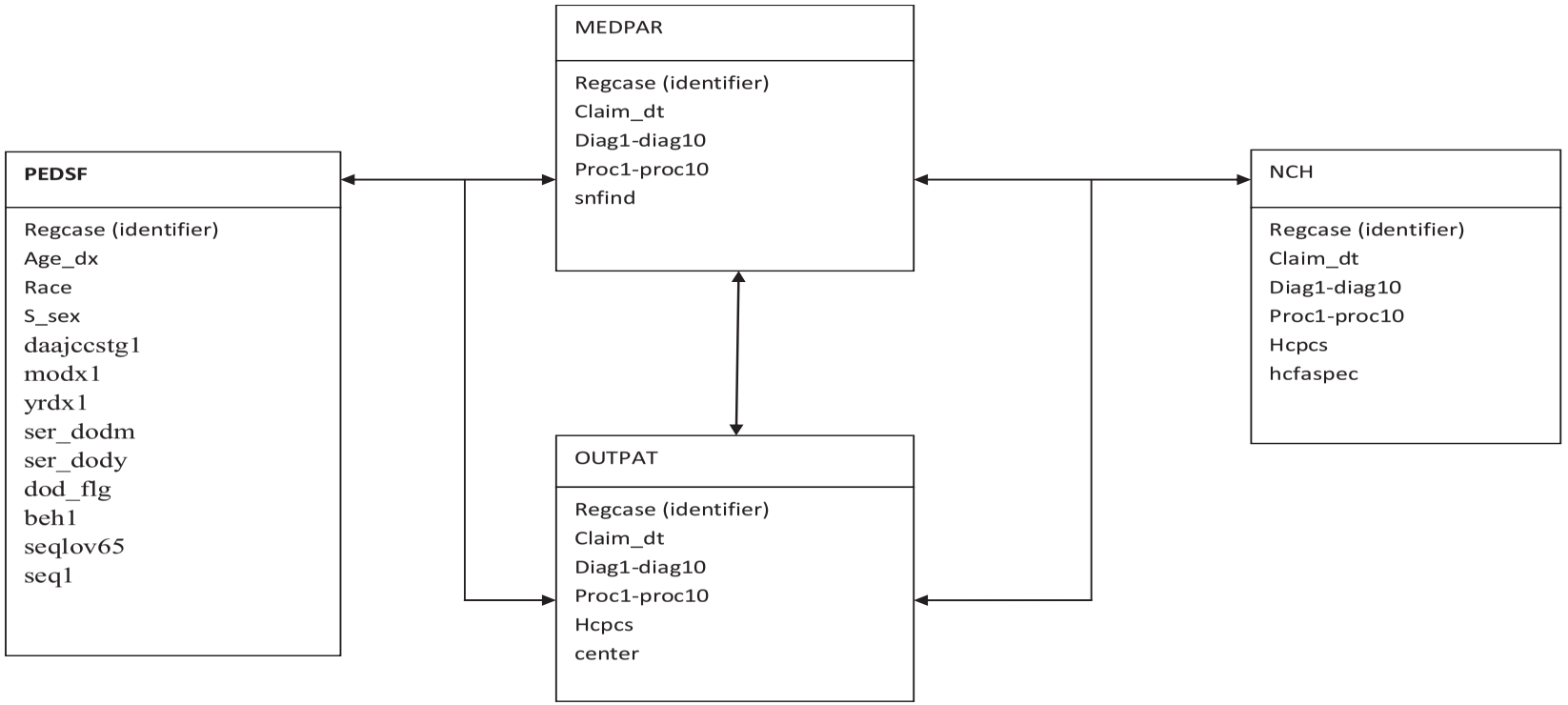
(1) Patient Entitlement and Diagnosis Summary File (PEDSF), which contains one record per person for individuals in the SEER database who have been matched with Medicare enrollment records;

(2) Medicare Provider Analysis and Review (MEDPAR), which includes all Part-A short-stay, long-stay, and skilled-nursing-facility (SNF) bills for each calendar year;

(3) Carrier claims (NCH), which contains claims from physicians and other non-institutional providers for each calendar year;

(4) Outpatient Claims (OUTPAT), which contains claims for each calendar year from institutional outpatient providers.

The medical information (including diagnosis codes, medication codes, claim date, and service type) were recorded in each file. All four files were linked by the common patient identifier (regcase) and pulled together to reconstruct the CRC treatment information for each individual.



Note:
Regcase: patient id age_dx: age at diagnosis race: race s_sex: gender daajccstg1: cancer stage modx1: primary diagnosis month yrdx1: primary diagnosis
yearser_dodm: SEER month of death ser_dody: SEER year of death dod_flag: decease flag beh1:ICD-O-3 behavior code seq1: Sequence Number
seq1ov65: sequence number for over 65 yrs old claim_dt: claim date diag: ICD-9 diagnosis code proc: ICD-9 procedure code hcpcs: HCPCS code
center: venue code snfind: MEDPAR Short Stay/Long Stay/SNF Indicator Code hcfaspec: HCFA provider code

Figure 2 Data Schema of SEER-Medicare Database

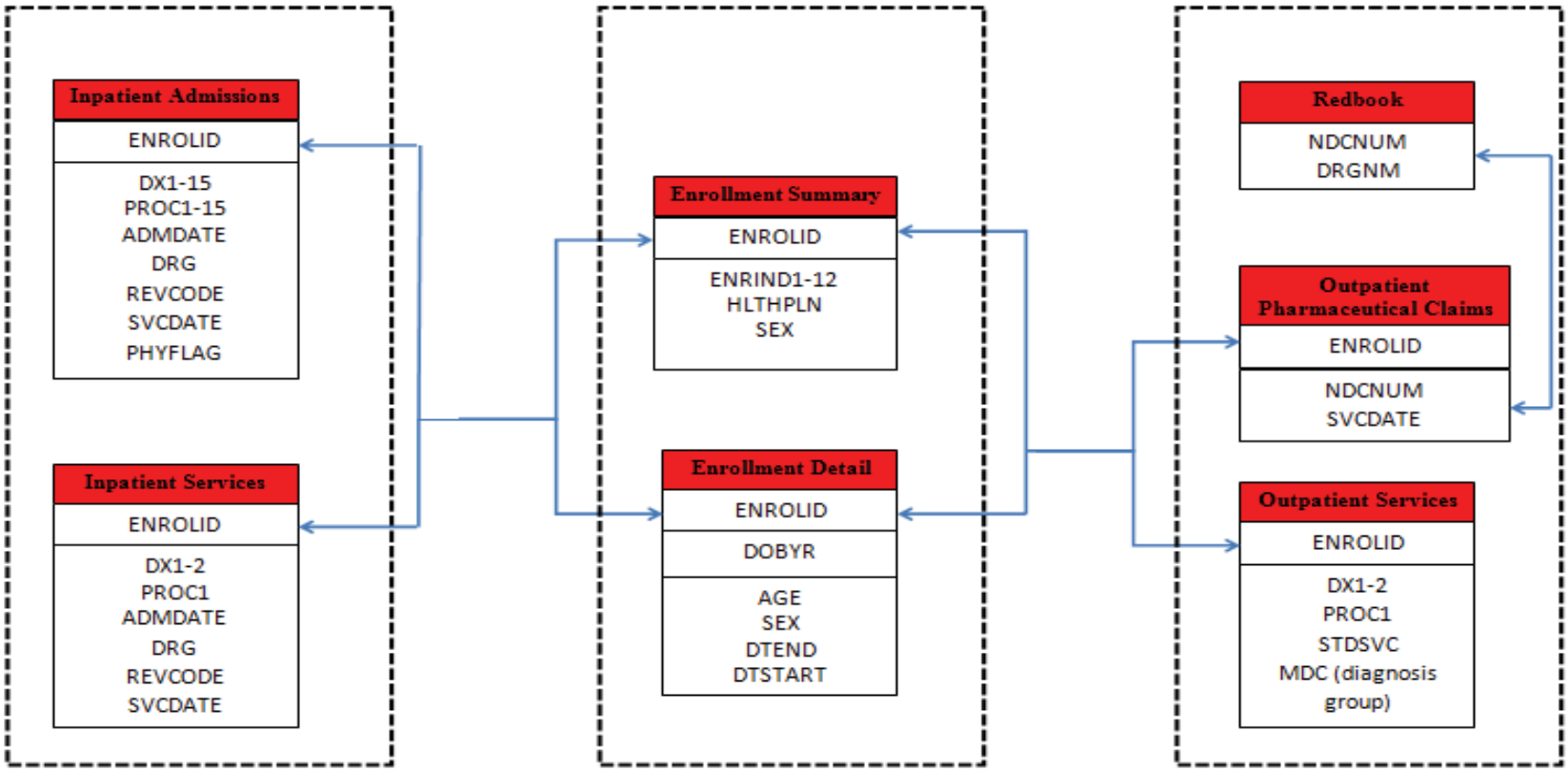
MarketScan® Commercial Claims and Encounters Database. After the final algorithms were generated from the SEER-Medicare database, they were applied to the MarketScan® Commercial Claims and Encounters Database. This private sector data originated from health insurance charge claims (both employer- and health plan sourced) in the US. It consists of service-level medical (inpatient and outpatient) and prescription claims for all insured individuals and their dependents. The detail of medication information, such as therapeutic class, manufacturer's average wholesale price (AWP) and generic identifier are added later. For the purpose of this study, the diagnostic and procedure information (which allows us to identify the different treatment for CRC), was extracted from this database. The data files we used are listed as follows:

1) The Inpatient Admissions files (IA), which contains the summarized information about hospital admission for individuals.
2) The Inpatient Services files (IS), which contains the individual facility and professional encounters and services that the inpatient admission record comprises.
3) The Outpatient Pharmaceutical Claims file (D) is available for a large portion of the individuals represented in the medical/surgical and populations tables.
4) The Outpatient Services File (O) contains encounters and claims for services that were rendered in a doctor's office, hospital outpatient facility, emergency room or other outpatient facility.
5) The Annual Enrollment Summary File (A), which contains a single record per person per period of continuous enrollment.

6) The Enrollment Detail File (T), which contains one record per person per month of enrollment for an individual enrollee regardless of whether or not any demographic values have changed from the previous month.

The Redbook® file, a supplement file in the MarketScan database, was also used to extract additional medication information for this study.

The medical information (including diagnosis codes, medication codes, claim date, and service type) were recorded in each file. All six main files (except redbook) were linked by the common patient identifier (enrolid) and pulled together to reconstruct the CRC treatment information for each individual.



Note:

EnrolID: patient id	dx: ICD-9 diagnosis code	proc: procedure code	admdate: administration date	drg: diagnosis related group code
Revcode: revenue code	svcdate: service date	phyflag: physician flag	enrind: enrollment indicator	hlthpln:health plan
sex: gender	dobyr: birth year	age: age	dtend: enrollment end date	dtstart: enrollment start date
ndcnum: NDC number	drgnm: drug name	stdsvc: service type	MDC:major diagnosis category	

Figure 3 Data Schema of MarketScan Commercial Claims and Encounters Database

Study Period

The study period is from 1999 to 2007. We intend to compare algorithm success (and failure) for the years 1999-2003 with that from 2003-2007. There was a change in cancer-stage coding between the AJCC's Cancer Staging Manual, 3rd and 6th Editions, but SEER-Medicare kept clear records during the transformation.

Due to availability issues, data from the MarketScan database used in this study is only from 2002 to 2003.

The Definition of Colorectal Cancer in SEER-Medicare Database

Based on the unique nature and design of the SEER-Medicare Database, instead of using the normal ICD 9 codes as the method of classifying colorectal cancer patients, SEER site codes were used as the definition for colorectal cancer. This site code is based on the primary site and ICD-O-3 morphology. For colorectal cancer, the following values of siterkm1were used to define CRC sites (siterkm1 is a variable recoded based on based on primary site and ICD-O-3 histology in order to make analyses of site/histology groups easier):

- 15: Cecum

- 16: Appendix
- 17: Ascending colon
- 18: Hepatic flexure
- 19: Transverse colon
- 20: Splenic flexure
- 21: Descending colon
- 22: Sigmoid colon
- 23: Large intestine, NOS
- 25: Recosigmoid junction
- 26: Rectum

Patient Selection

SEER-Medicare Database. Since the results in the pilot study showed the algorithm for all age groups cannot achieve better predictive results, patients in this study were included only if:

- Age is greater than 65 years-old
- Have been diagnosed with colorectal cancer between 1999 to 2007 as identified in SEER.
- CRC is the primary diagnosis for the patient
- They have malignant cases

They were excluded if they met any of the following criteria:

- Diagnosis month missing
- Patient's SEER historic stage is unclear, given as not applicable, occult, or unknown.
- Deceased or lost to follow-up in 6 months following primary cancer diagnosis
- Do not have continuous Medicare FFS coverage or had HMO enrollment from 6 months prior to 6 months after the diagnosis date

MarketScan Database. The inclusion criteria are the following:

- Diagnosed with colorectal cancer between 2002 and 2003 in MarketScan. The ICD-9 classification codes are the following: malignant neoplasm of hepatic flexure (153.0), malignant neoplasm of transverse colon (153.1), malignant neoplasm of descending colon (153.2), malignant neoplasm of sigmoid colon (153.3), malignant neoplasm of cecum (153.4), malignant neoplasm of appendix vermiformis (153.5), malignant neoplasm of ascending colon (153.6), malignant neoplasm of splenic flexure (153.7), malignant neoplasm of other specified sites of large intestine (153.8), malignant neoplasm of colon, unspecified site (153.9), malignant neoplasm of rectosigmoid junction (154.0), malignant neoplasm of rectum (154.1), malignant neoplasm of other sites of rectum, rectosigmoid junction, and anus (154.8).

The exclusion criteria are as follows:

- Diagnosis month missing
- Deceased or lost to follow-up in 6 months following cancer diagnosis.

- HMO enrollment from 6 months prior to 6 months after the diagnosis date

Theoretical Models

In the algorithm development section, a series of predictors were first extracted from the target database to generate the cancer staging predictive algorithms from different study periods. The factors of predictive models were selected based on the results of a literature review and clinical guideline review. Finally, the models were evaluated to determine the accuracy for predicting the advanced cancer stage of CRC. (Seen in Figure 4) After the algorithm predictors were developed, they were applied in a medical claims database and modified based on the variables' availability in that database. (Seen in Figure 5)

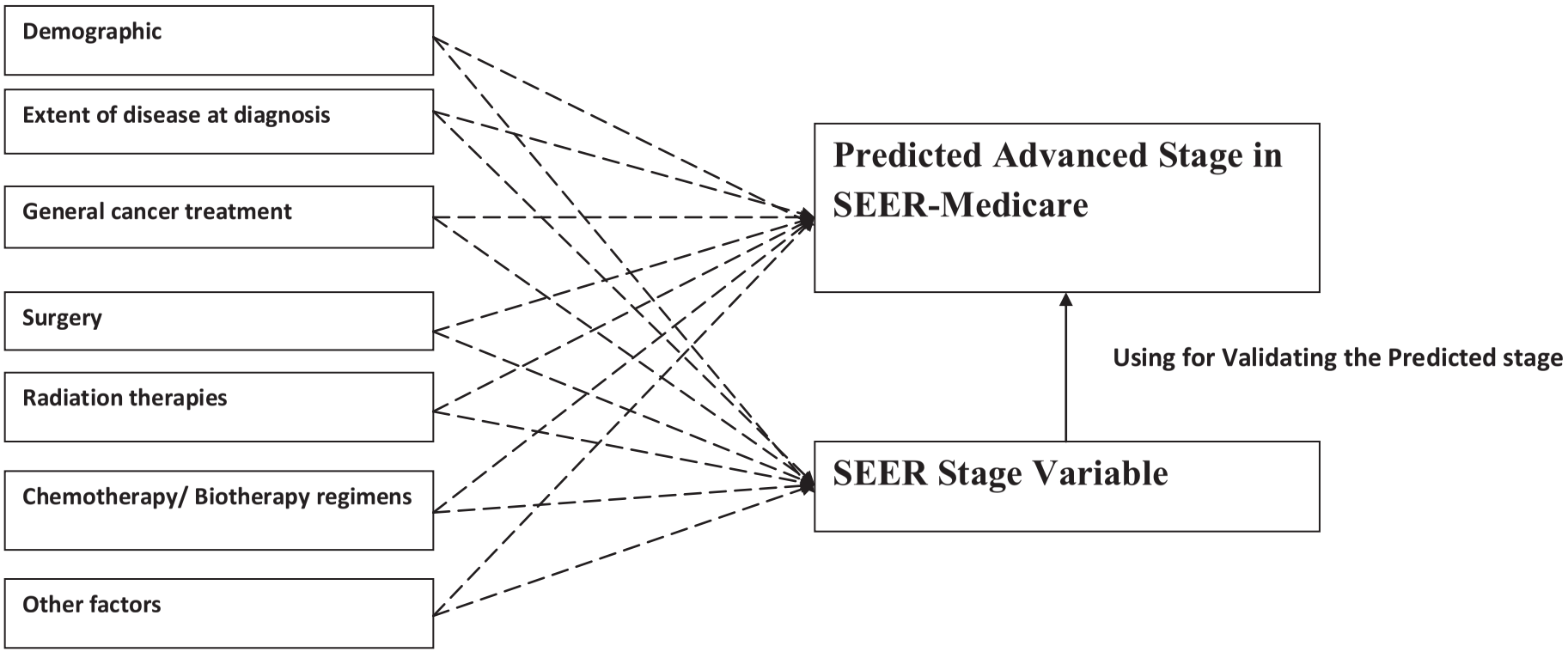


Figure 4 Theoretical Model for the Advanced Cancer Stage Predicting Method Used In SEER-Medicare Database

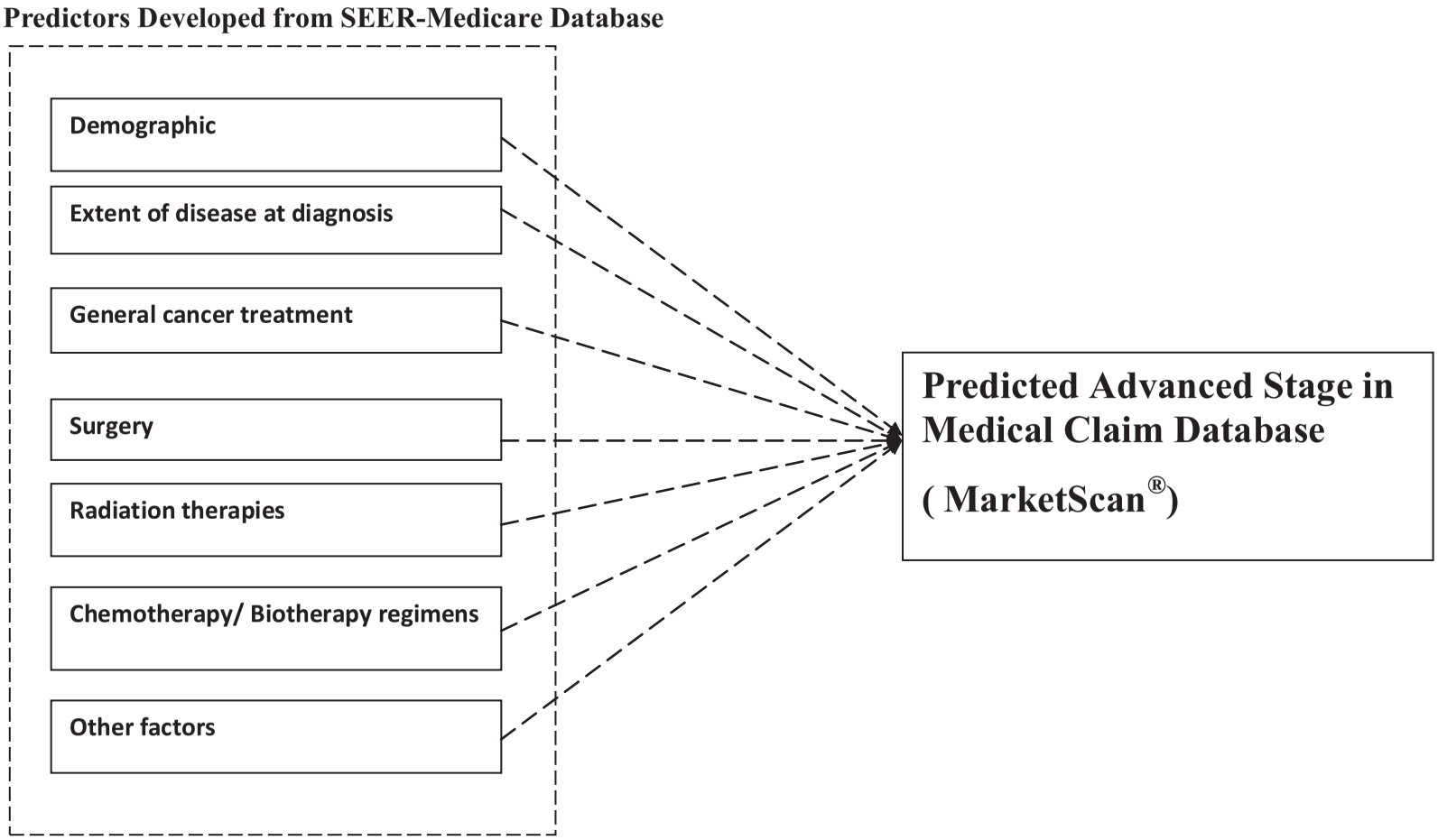


Figure 5 Theoretical Model for the Advanced Cancer Stage Predicting Method Used in Medical Claim Database

Independent Variables

The independent variables were chosen based on an extensive literature review, consultation with oncologists, and statistical significance during algorithm exploration. All clinical outcomes in the independent group were defined as dichotomous variables for at least one record of a certain diagnosis or procedure shown in a particular database.

The following factors have been identified to date:

(1) The demographics: age, gender. Race was not included in the algorithm since this variable is not available in MarketScan database;

(2) The extent of the disease at diagnosis, including lymph node involvement and metastasis (secondary malignant neoplasm) diagnosis;

(3) Cancer treatment condition, including number of visits to a surgeon, number of visits to a medical oncologist, number of visits to a radiation oncologist, and any hospital admission (Pollock & Vickers, 1998) or surgery; (DeVita, Hellman, & Rosenberg, 2005; Kahnamoui, Cadeddu, Farrokhyar, & Anvari, 2007)

(4) Radiation therapy; (DeVita et al., 2005; Kahnamoui et al., 2007)

(5) Chemotherapy or biotherapy; (DeVita et al., 2005)

(6) Screening tools, imaging tests, and disease history; (Davila et al., 2006; R. A. Smith, Cokkinides, & Eyre, 2007)

(7) Surgery. (Choti et al., 2002; Foster, 1984; Pawlik et al., 2005)

All the independent variables were identified from SEER-Medicare and MarketScan by using ICD-9, Current Procedural Terminology (CPT), Healthcare Common Procedure Coding System (HCPCS), revenue, and National Drug Code (NDC) codes. (Seen in Table 5, 6)

Table 5 Covariates and Claims Codes for Colorectal Cancer Stage

Predictor Variable	Time period searched	Source File	ICD-9 Diagnosis	ICD-9 Procedure	CPT/HCPCS	Revenue Center	HCFA code
Demographic							
Age at diagnosis*	At diagnosis	Pedsf					
Gender		pedsf					
Personal history							
Colorectal cancer	Previous diag	Male and female	V1011				
Ovarian cancer	Previous diag	Female	V1043				
Uterine cancer	Previous diag	Female	V1041				
Breast cancer	Previous diag	Female	V103				
Disease extension at diagnosis							
LN involvement	3 mos before and after diagnosis		1963				
Metastatic disease	3 mos before and after diagnosis		1962, 1965-6, 197, 1970, 1971, 1972, 1973, 1974, 1975,				

			1976, 1977, 1978, 198, 1980, 1981, 1982, 1983, 1984, 1985, 1986, 1987, 1988, 19881, 19882, 19889				
Cancer Treatment							
No. visits to surgeon *	In the year after diagnosis						02,14,19,20,23,24,28, 33,40,77,78,85,91
No. visits to medical oncologist *	In the year after diagnosis						90
No. visits to radiation oncologist *	In the year after diagnosis						92
Imaging (CT,MRI,PET, or bone scan)	In the 3 mos after diagnosis			8801, 8703, 8741, 8891, 8896, 8874, 9218, 9214	70450, 70460, 70470, 70551-3, 71250, 71260, 71270, 72192-4, 74150, 74160, 74170, 76700, 78315, 78320, 78812-6, G0213-5		

Radiation therapy	In the 6 mos after diagnosis		V580, V661, V671	9221-7, 9229	77401-77525, 77520,77523,77761-77799, G0256, G0261	0330, 0333	
LN dissection	3 mos before to 1 year after			4023, 4051, 8543, 8547	38740, 38745, 19162, 19200, 19220, 19240		
CEA (Elevated carcinoembryonic antigen)	In the 6 mos after diagnosis				79581,82387		
Biopsy (colon, rectum)	In the 6 mos after diagnosis			4525,4824	44025,44100,45100		
Polypectomy (colon, rectum)	In the 6 mos after diagnosis			4542,4836	45383-45385		
Transanal excision (Including Microsurgery)	In the 6 mos after diagnosis				45170		
Low anterior resection (LAR)	In the 6 mos after diagnosis				44139,44145, 44207,44213		
Total mesorectal excision (TME)	In the 6 mos after diagnosis				0184T		
AbdominAL perineal resection (APR)	In the 6 mos after diagnosis				45110		
Hepatic resection	In the 6 mos after diagnosis			5022,503	47380, 47382		

Pulmonary resection	In the 6 mos after diagnosis			3239,3249, 3250,3259	32480		
Chemotherapy (any agent)	In the 6 mos after diagnosis		V581, V662, V672	9925	96400-96549, J9000-J9999, Q0083-5	0331, 0332, 0335	
Preventive care							
FOBT	3 mos before and after diagnosis				G0107, 82270-82274		
Sigmoidoscopy	3 mos before and after diagnosis			4821-4825,4524,4542	45300,45303,45305,45307,45309,45327, 45330-4,45337-9,45340-2,45345, G0104		
Colonoscopy	3 mos before and after diagnosis			4521-4523,4525	G0105,G0121,45378, 45379-45392		
DCBE	3 mos before and after diagnosis			8764	74270,74280, G0106, G0120, G0122		
NSAIDs	3 mos before and after diagnosis		V5864				

* Treat as continuous variables.

Table 6 Chemotherapy Codes for Advanced Colorectal Cancer

Generic Name	Brand Name	NDC Number	HCPCS
Fluorouracil (5-FU)	Adrucil	00004150603, 00004170406, 00004170506, 00004190406, 00004197701, 00013102691 00013103691, 00013104694, 00013105694, 00023081030, 00023081230, 00066715030 00081039001, 00081039003, 10139006320, 10139006350, 38779002501, 38779002504 38779002505, 38779002509, 38779002510, 38779002525, 39769001210, 39769001240 39769001250, 39769001290, 49452317501, 49452317502, 49452317503, 49452317504 51309021720, 51309021750, 51309021798, 51432040910, 51432047010, 51552073301 51552073302, 51552073304, 51552073305, 51672406201, 51672406301, 51927108500 53258171003, 53258171100, 53443000207, 53905011110, 54569110000, 54569140600 54569156500, 54569156600, 54868095100, 54868095101, 54868545000, 58016201701 58016910601, 60346054881, 61703040932, 61703040953, 61703040967, 62991148601 62991148602, 62991148603, 63323011710, 63323011720, 63323011751, 63323011761 63370009515, 63370009525, 63370009535, 66530024940, 66758004401, 68682000431 68682008531	J9190
Oxaliplatin*	Eloxatin	00024059010, 00024059120, 00024059240, 00024059602, 00024059704	J9263
Capecitabin*	Xeloda	00004110020, 00004110051, 00004110116, 00004110150, 54569571700, 54868414300 54868414301, 54868414302, 54868414303, 54868526000, 54868526001, 54868526002 54868526003, 54868526004, 54868526005, 54868526006, 54868526007, 54868526008	J8520, J8521

		54868526009	
Irinotecan	Camptosar	00009752901, 00009752902, 00591318902, 00591318926, 00703443211, 00703443411 00703443491, 00703443711, 00781306672, 00781306675, 10019093401, 10019093402 10019093417, 10019093479, 10518010310, 10518010311, 18111000202, 18111000203 25021020002, 25021020005, 55390029501, 55390029601, 59762752901, 59762752902 61703034909, 61703034916, 61703034936, 61703034961, 61703034962, 63323019302 63323019305	J9206
Leucovorin		00005450164, 00005452564, 00005452583, 00005452590, 00005453640, 00544496130 00054449625, 00054449705, 00054449710, 00054449805, 00054449810, 00054449911 00054849619, 00054849706, 00054849806, 00054849906, 00074454102, 00074454104 00081063120, 00081063213, 00081063793, 00173063120, 00173063135, 00173063155 00173063225, 00173063893, 00182186901, 00182186917, 00182187024, 00205400451 00205470249, 00205533019, 00304191001, 00304217756, 00469013701, 00469013825 00469137030, 00469574030, 00517860525, 00536414807, 00536414904 00555048401,00555048402, 00555048527, 00603418321, 00603418435, 00641236441 00686514001, 00703513001, 00703513801, 00703514001, 00703514501, 00781122001 00781122031, 00781122263, 00839746306, 00839746406, 00839746408, 00839759630 00839759730, 00839759733, 00904231560, 00904231617, 11845046801, 11845046808 11845046925, 47679074901, 47679074927, 47679074935, 47679075049, 47679075097	J0641

		49452403601, 49452403602, 49452403603, 49452403604, 49884023701, 49884023711 49884023801, 49884023815, 51079058101, 51079058106, 51079058201, 51079058205 51309023302, 51309023405, 51309024110, 51309024220, 51309074112, 51309074114 51309074130, 51309074191, 51309074201, 51309074225, 51309074302, 51309074312 51309074324, 51432025103, 51432025130, 51432025525, 51432042110, 51432055710 51927269200, 53258137003, 54569264100, 54868331000, 54868331001, 54868331002 54868591500, 55390000901, 55390005110, 55390005210, 55390005301, 55390005401 58406062105, 58406062137, 58406062206, 58406062235, 58406062307, 58406062333 58406062462, 58406062467, 58406062668, 58406062674, 61703041050, 62701090030 62701090099, 62701090125, 63323071050, 63323071100, 66479024725, 99999317823 99999778801, 38779033700, 38779033703, 38779033704, 38779033706, 38779033709 38779033711, 38779033715, 38779033725, 38779033753, 55390081810, 55390082401 55390082501, 55390082601,00641236941, 00686024110, 00686058106,	
Bevacizumab*	Avastin	50242006001, 50242006002, 50242006101	J9035

Dependent Variable

According to the AJCC's Cancer Staging Manual, 3rd and 6th editions, advanced CRC stage is defined by stage IV. Over the study period, the SEER database coding changed. In 1999-2003, the variable **ajccstg1** in PEDSF was used to identify advanced stage at diagnosis. In 2004-2007, the variable **daajccstg1** in PEDSF was used.

In the MarketScan database, the predicted advanced cancer stage was calculated for every observation based on the coefficients generated by algorithms developed from SEER-Medicare. The cut-off criteria were based on the cut-off point from the original algorithm.

Statistical Analysis

Descriptive analysis such as mean, median, range and standard deviations were calculated for both dependent and independent variables in two databases.

Algorithms Development Section. We used a split sample approach to develop and validate our logistic models. Each model was derived from the "training set," which selected using simple random sampling without replacement and contained 50% of cohort population. Algorithms in each study period for predicting advanced cancer stage were constructed based on independent variables in SEER-Medicare. The logistic models of the associations between predictor covariates and the dichotomous outcomes (advanced stage versus non-advanced stage from SEER-Medicare) were developed.

The General Model is illustrated in the following:

$$Pr_{(Y=1)}=\frac{1}{1+e^{-f(z)}},\ f(z)=\beta_0+\beta_1*\ Demo_{ij}+\beta_2*\ ExtDis_{ij}+\beta_3*\ CnTrt_{ij}+\beta_4*Srgy_{ij}+\beta_5*RT_{ij}+\beta_6*Chemo_{ij}+\beta_7*Scrn_{ij}+\text{error}$$

Where

- Y is advanced stage from SEER stage variable (dichotomies variable, 1=advanced cancer, 0=non-advanced cancer stage);
- i is for a specific patient (samples from 1 to N);
- j is a particular factor in certain independent variable categories.
- β_1 to β_7 are the coefficients of those independent variables categories.
- Demo: Demographic

  ExtDis: Extent of disease

  CnTrt: Cancer treatment condition

  Srgy: Surgery

  RT: Radiation therapy

  Chemo: Chemotherapy

  Scrn: Screening tools

The proposed full regression equation for predicting advanced cancer stage during 1999 to 2003 is shown as follows:

logit [pr (y=1)] = β_0 + β_1 (age) + β_2 (gender) + β_3 (metastatic disease) + β_4 (LN involvement) + β_5 (Radiation therapy) + β_6 (imaging) + β_7 (chemotherapy (any agent)) + β_8 (hospitalization) +β_9 (history of tobacco use) + β_{10} (history of female breast cancer) + β_{11} (history of previous colorectal cancer) + β_{12} (history of ovarian cancer) + β_{13} (history of uterine cancer) +

β_{14} (no. visits of surgeon) +β_{15} (no. visits of medical oncologists) + β_{16} (no. visits of radiation oncologists)+ β_{17} (no. visits of hospital) + β_{18} (CEA) + β_{19} (Biopsy (colon, rectum)) + β_{20} (polyectomy) + β_{21} (transanal excision) + β_{22} (LAR) + β_{23} (APR)+ β_{24} (hepatic resection) + β_{25} (FOBT) + β_{26} (sigmoidoscopy) + β_{27} (colonoscopy) + β_{28} (DCBE) + β_{29} (NSAIDs)+ β_{30} (5-FU) + β_{31} (oxaliplatin) + β_{32} (irinotecan) + β_{33} (leucovorin) + β_{34} (FOLFOX)+ β_{35} (FOLFIRI)

where

- Y is advanced stage from SEER stage variable (dichotomies variable, 1=advanced cancer, 0=non-advanced cancer stage);
- β_1 to β_{35} are the coefficients of those independent variable categories.

The proposed full regression equation for predicting advanced cancer stage during 2004 to 2007 is shown as follows:

logit [pr (y=1)] = β_0 + β_1 (age) + β_2 (gender) + β_3 (metastatic disease) + β_4 (LN involvement) + β_5 (Radiation therapy) + β_6 (imaging) + β_7 (chemotherapy (any agent)) + β_8 (hospitalization) +β_9 (history of tobacco use) + β_{10} (history of female breast cancer) + β_{11} (history of previous colorectal cancer) + β_{12} (history of ovarian cancer) + β_{13} (history of uterine cancer) + β_{14} (no. visits of surgeon) +β_{15} (no. visits of medical oncologists) + β_{16} (no. visits of radiation oncologists)+ β_{17} (no. visits of hospital) + β_{18} (CEA) + β_{19} (Biopsy (colon, rectum)) + β_{20} (polyectomy) + β_{21} (transanal excision) + β_{22} (LAR) + β_{23} (APR)+ β_{24} (hepatic resection) + β_{25} (FOBT) + β_{26} (sigmoidoscopy) + β_{27} (colonoscopy) + β_{28} (DCBE) + β_{29} (NSAIDs)+ β_{30} (5-FU) + β_{31} (oxaliplatin) + β_{32} (irinotecan) + β_{33} (leucovorin) + β_{34} (FOLFOX)+ β_{35} (FOLFIRI)+ β_{35} (bevacizumab) + β_{36} (cetuximab) + β_{37} (capecitabine)

where

- Y is advanced stage from SEER stage variable (dichotomies variable, 1=advanced cancer, 0=non-advanced cancer stage);
- β_1 to β_{37} are the coefficients of those independent variable categories.

Logistic models were examined for any potential interaction effects among the independent variables for checking the multicollinearity issue by using collinearity matrix and tolerance.

Correlation matrix and tolerance were applied to all IVs in the algorithm development models in both study periods. Correlation matrix were constructed to check the multicollinearity assumption. Pearson correlation coefficients (r) were calculated for measuring how well the variables are related. Since the correlation matrix results may not be sufficient for the case that a group of variables may be highly interdependent instead of a pair of variables. So tolerance were calculated for preventing this based on regressing each variable on all the other explanatory variables, calculating the R^2 then subtracted by 1. Low tolerance corresponds with high multicollinearity. Any potential variables which violated the assumption by either the absolute value of the Pearson correlation coefficients was higher than 0.6 or tolerance was lower than 0.40 has lower tolerance than 0.40 or were removed from final models.

Backward selection method was applied to rule out covariates. Under this approach, the original model started with fitting a parsimonious model with all the variables of interest, and then the least significant variable was dropped to achieve better goodness of fit for the overall

model, so long as it was not significant at our chosen critical level. The progress of model re-fitting reduction was successively continued until the best goodness-of-fit was constructed.

After the algorithms were derived from the "training set," all parameters associated with each predictive factor estimated from the derivation dataset were applied to the other half of the cohort population ("validation set") to calculate the predictive probability of each patient having advanced stage cancer (stage IV disease). Since the algorithms were generated by a logistic regression model, the general formula of predictive probability is shown in the following:

$$\Pr_{(Y=1)} = \exp(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \cdots) \Big/ 1 + \exp(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \cdots)$$

where

- i is for a specific patient (samples from 1 to N);
- Pr(y=1) is predictive probability for advanced stage cancer patient i in the "validation set";
- X (k=1 to n) are the independent variables (predictors);
- β (k=1 to n) are the coefficients of those independent variables.

The "gold standard" for advanced stage was considered the SEER stage; the test stage was based on the calculated probability. (for example, for a probability cutpoint of α, patients were predicted to have stage IV disease if their calculated probability was $\geqslant \alpha$, and not to have stage IV disease if their calculated probability was $< \alpha$). The predictive probability that achieves the most correct predictive cases were used for the cut-off level in the next section. Sensitivity

measures the proportion of actual positives which are correctly identified as such, which in our case is the percentage of advanced cancer stage cases in the PEDSF file. (See Table 7) Specificity measures the proportion of negatives which are correctly identified, which is the percentage of non-advanced cancer stage cases. The positive predictive values (PPVs) and negative predictive values (NPVs) were calculated to evaluate the accuracy of the algorithm.

Two sets of the predictive values (sensitivity, specificity, PPV and NPV) were compared between the final algorithm for 1999-2003 and that for 2004-2007.

Table 7 Two by Two table for Predicted Advanced Cancer Stage Validation

		Predicted advanced cancer stage		
		Positive	Negative	
Advanced cancer stage determined by SEER stage variable ("gold standard")	Positive	TP	FN	Sensitivity=TP/(TP+FN)
	Negative	FP	TN	Specificity=TN/(FP+TN)
		PPV=TP/(TP+FP)	NPV=TN/(FN+TN)	

Note: TP=True positive FP=False positive
FN=False negative TN=True negative
PPV= Positive predictive value
NPV=Negative predictive value

Algorithm predictors' application in medical claims database section. Due to limited data availability, only 2002 to 2003 MarketScan commercial data was applied for the algorithms. The predictors generated from SEER-Medicare were applied in the commercial claims database (MarketScan) and this data was used to calculate the cancer stage predictive probability.

The parameter associated with each predictor estimated from the previous section was applied to each patient in a commercial medical claims database to calculate each patient's predictive probability of having advanced disease by using the following equation:

$$\boldsymbol{P'_{i\,pred}} = \exp(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \cdots) \Big/ 1 + \exp(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \cdots)$$

where

- i is for a specific patient (samples from 1 to N);
- $\boldsymbol{P'_{i\,pred}}$ are the predicted probability for a specific patients i in MarketScan database;
- X (k=1 to n) are the independent variables (predictors);
- β (k=1 to n) are the coefficients of those independent variables.

The threshold for advanced cancer stage was determined by the cut-off point generated from the previous section.

Results

Algorithms Development in SEER-Medicare Database

Patient Selection. From 1999 to 2007, we identified 212,345 SEER participants who had at least one diagnosis of colorectal cancer. Among those patients, 168,667 patients had at least one primary diagnosis. As detailed in Figure 6, we further excluded patients based on the exclusion criteria: we excluded 19,803 patients whose age were less than 65 years-old during the study period; 6,354 non-malignant cases were excluded; 480 cases were excluded due to death or loss of follow-up within 6 months following cancer diagnosis; and 14,523 patients who did not have continuous Medicare fee-for-service coverage or had HMO coverage from 6 months prior to 6 months after their diagnosis date since the claims information may not be complete.

After all patient selection criteria were applied, it yielded a final sample size of 127,507 patients in our study. Among this final sample, 70,264 patients had their first diagnosis during 1999 to 2003 and 57,243 patients had the first diagnosis during 2004 to 2007. In the study period, 50 percent of the cohort were randomly selected as the algorithm development data set and the other 50 percent were used as the algorithm validation data set.

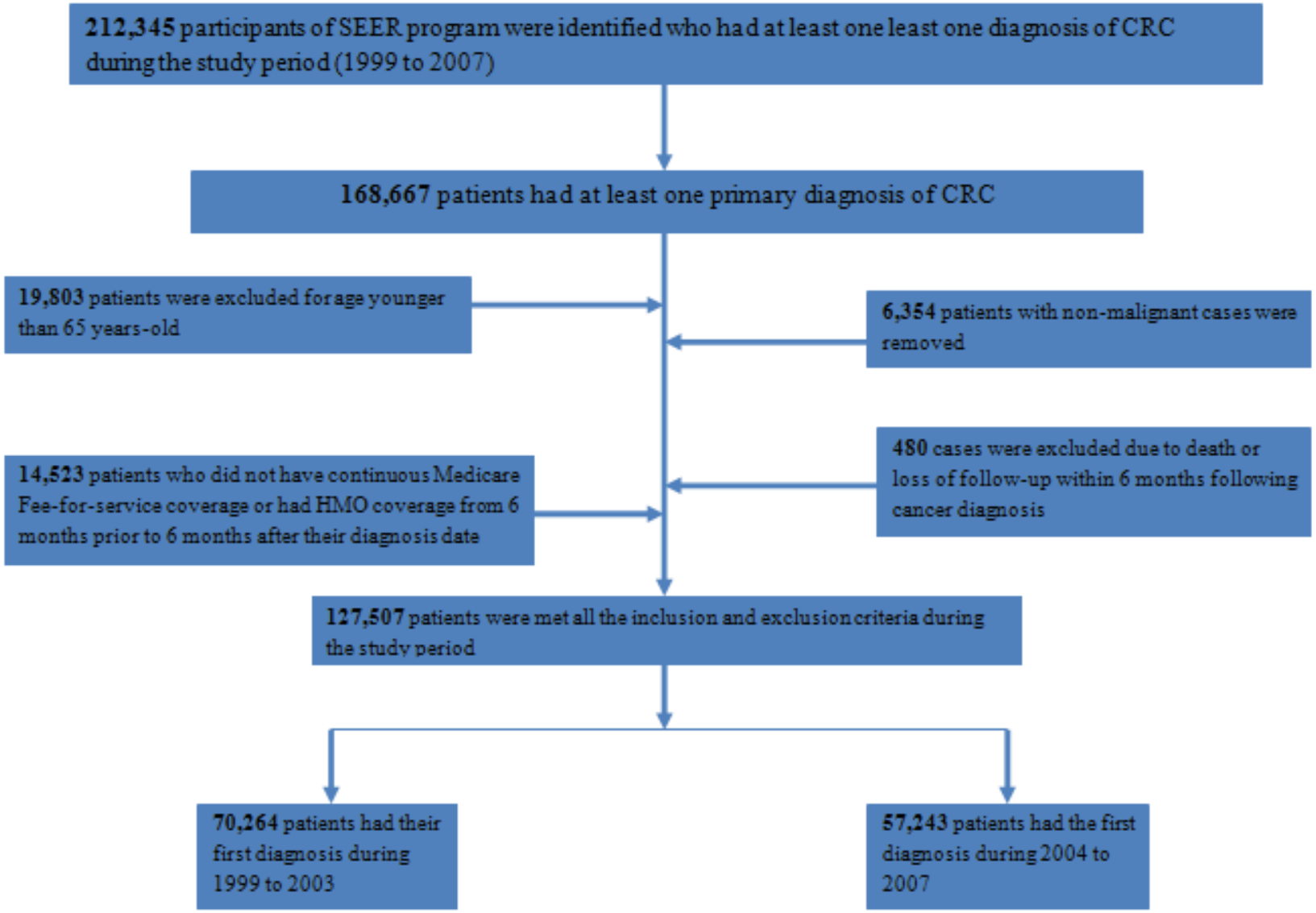


Figure 6 Patient Selection Flowchart for SEER-Medicare

Description of Independent Variables. The descriptive analysis of all independent variables' are the following:

Demographic variables

In all 127,507 patients, the mean age was 77.21 years old (SD 7.68, range 65 to 114). 45.51 percent (58,027) of the final population are male and 54.49 percent are female.

All results of the four sub data set (the algorithm development set and validation set of 1999 to 2003 and 2004 to 2007) are showed in table 8. T-tests were calculated for continuous variables and chi-square tests were calculated for categorical variables.

Table 8 Description of Demographic Independent Variables in sub-groups

Variables		Development Set			Validation Set			
		Mean (or Frequency)	SD (or pertentage)	Range	Mean (or Frequency)	SD (or pertentage)	Range	P Value
Study Period: 1999 to 2003								
Age at first CRC Diagnosis		77.24	7.57	65 - 104	77.22	7.61	65 - 105	0.77
Gender (%)	Male	15963	45.44		15831	45.06		0.32
	Female	19169	54.56		19301	54.94		
Study Period: 2004 to 2007								
Age at first CRC Diagnosis		77.14	7.77	65 - 106	77.24	7.8	65 - 114	0.15
Gender (%)	Male	13118	45.83		13115	45.82		0.98
	Female	15504	54.17		15506	54.18		

Cancer Treatment

Within six months after the primary colon or rectum cancer diagnosis, the total number of visits to cancer surgeon was 691,736. Of those patients, the mean number of visits was 5.42 (SD 8.19, range 0-244). 351,865 visits to medical oncologists were found for all eligible patients. The mean number of visits for medical oncologists was 2.76 (SD 12.95, range 0-549). The total number of visits to radiologists was 99,931. The mean number of radiation treatment visits was 0.78 (SD 3.24, range 0-165).

Radiation therapy

Within six months after the first diagnosis, 5,589 patients had at least one radiation therapy which accounts for 4.38% of the study population.

Chemotherapy and biologic therapy

For all eligible CRC patients, National Drug Codes (NDC) or clinical codes (e.g. HCPCS or CPT) were extracted from prescription records as well as outpatient drug records and inpatient services records.

22,748 patients had 5-FU prescriptions, which accounted for 17.84 percent of the whole study population. 7,934 patients received irinotecan in the 6 months after diagnosis (~6.22 percent of the final sample). The total number of leucovorin prescriptions for these patients was 8,718 or about 16.53 percent of the population. 9,935 patients had bevacizumab or bevacizumab combination in the 6 months after diagnosis, which is about 4.61 percent of the whole study

group. Lastly, 1.84 percent of the study population had cetuximab in the 6 months after diagnosis, which was 2,343 patients.

Screening tools, imaging tests and disease history

14.18% of the study population who had at least one record for FOBT after 6 months horizon before and after the primary diagnosis were found in the database, which accounts for 18,084 patients. The number of patients who had at least one sigmoidoscopy procedure was 24,429, about 19.16% of the study population. The number of patients who had at least one colonscopy was 69,125, which accounts for 54.21% of total patients. The number of patients who had at least one DCBE was 18,084, which accounts for 14.18%.

All types of imaging procedures were taken into account for this study, 56,778 patients (44.53 percent of the whole study population) were found that had at least one imaging procedure including CT, PET, MRI or bone scan.

For the aspect of disease history, 10,914 patients have a record indicating a history of tobacco use. The number of patients who had history of previous ovarian cancer or uterine cancer were 299 and 149, respectively.

Surgery

All medical records within 6 months after the first CRC diagnosis were extracted for CRC related surgery procedures. 76 patients were found that had liver resection within 6 months after the primary diagnosis. The number of patients who had at least one colon or rectum biopsy

procedure was 11,647. 2250 patients had at least one polypectomy on either colon or rectum. Transanal excision (including microsurgery) procedure was found in 150 patients' clinical records. The number of patients who had at least one low anterior resection (LAR) procedure was 1412. 183 patients who had at least one abdominal perineal resection (APR) procedure were found in clinical records.

Table 9 displays the summary of all baseline characteristics difference of CRC treatment patterns. T-tests were calculated for continuous variables and Chi-square tests were calculated for categorical variables.

Table 9 Description of Dependent Variable of CRC Treatment Patterns

Variables	Development Set			Validation Set			P Value
	Mean (or Frequency)	SD (or pertentage)	Range	Mean (or Frequency)	SD (or pertentage)	Range	
Study Period: 1999 to 2003							
Cancer Surgeon (visits)	5.83	9.05	0 - 178	5.82	8.88	0 - 242	0.91
Medical Oncologists (visits)	2.91	13.5	0 - 343	2.82	13.82	0 - 549	0.39
Radiologist (visits)	0.77	3.09	0 - 106	0.78	3.28	0 - 165	0.73
Radiation Therapy (%)	1449	2.06		1390	1.98		0.26
5-FU (%)	6089	9.69		6655	9.47		0.14
Irinotecan (%)	2490	3.54		2329	3.31		0.06
leucovorin	6281	8.94		6064	8.63		0.31
FOBT (%)	5647	8.04		5799	8.25		0.12
Sigmoidoscopy (%)	8419	11.98		8370	11.91		0.66
Colonscopy (%)	18953	26.97		19032	27.09		0.55
DCBE (%)	2973	4.23		2939	4.18		0.64
Imaging Procedures (%)	14215	20.23		14120	20.10		0.47
Tobacco Use History (%)	2721	3.87		2682	3.82		0.58
Ovarian Cancer History (%)	79	0.11		83	0.12		0.75
Uterine cancer History (%)	45	0.06		31	0.04		0.11
Liver Resection (%)	10	0.01		20	0.03		0.06
Biopsy (%)	4878	6.94		4967	7.07		0.33
Polypetomy (%)	5568	7.92		5619	8.00		0.60
Transanal excision (%)	283	0.40		304	0.43		0.38

LAR (%)	2087	2.97		2082	2.96		0.94
APR (%)	238	0.34		219	0.31		0.37
Study Period: 2004 to 2007							
Cancer Surgeon (visits)	4.92	7.11	0 - 142	4.94	7.08	0 - 128	0.82
Medical Oncologists (visits)	2.63	12.09	0 - 440	2.63	12.09	0 - 238	0.99
Radiologist (visits)	0.8	3.28	0 - 136	0.78	54.94	0 - 132	0.54
Radiation Therapy (%)	1382	2.41		1368	2.39		0.78
5-FU (%)	4654	8.13		4630	8.09		0.79
Irinotecan (%)	1612	2.82		1503	2.63		0.12
Leucovorin (%)	4358	7.61		4375	7.64		0.84
Bevacizumab (including combination) (%)	3509	2.11		3489	2.54		0.85
Cetuximab (%)	774	1.35		752	1.31		0.57
FOBT (%)	3369	5.89		3269	5.71		0.19
Sigmoidoscopy (%)	3817	6.67		3823	6.68		0.94
Colonscopy (%)	15527	27.12		15613	27.27		0.47
DCBE (%)	1643	2.87		1706	2.98		0.26
Imaging Procedures (%)	14227	24.85		14216	24.83		0.93
Tobacco Use History (%)	2802	4.89		2709	4.73		0.19
Ovarian Cancer History (%)	69	0.12		68	0.12		0.93
Uterine cancer History (%)	40	0.07		33	0.06		0.41
Liver Resection (%)	25	0.04		21	0.04		0.55
Biopsy (%)	888	1.55		914	1.60		0.53
Polypetomy (%)	4969	8.68		4948	8.64		0.81
Transanal excision (%)	289	0.50		303	0.53		0.56
LAR (%)	2478	4.33		2374	4.15		0.12
APR (%)	167	0.29		140	0.24		0.12

Description of Independent Variable

Among all CRC final population from 1999 to 2007, 21,644 patients had at least one advanced disease diagnosis, which takes 16.97 of the total population. Table 10 shows the baseline characteristics of the development sets and the validation sets in 1999 to 2003 and 2004 to 2007.

Table 10 Description of Independent Variable of CRC Advanced Disease Statue

CRC Advanced Disease	Development Set (counts %)	Validation Set (counts %)	P Value
1999 to 2003	**5830 (8.3%)**	**5754 (8.19%)**	**0.44**
2004 to 2007	**5024 (8.78%)**	**5036 (8.8%)**	**0.89**

Model Assumption Examination and Limitations Evaluation

Since the logistic regression technique were used for developing the algorithms, we need to make sure the analysis is valid. Therefore, making sure the assumptions of logistic regression was met is the first priority.

Unlike ordinary linear regression where several assumptions must be met before applying the method, the assumption of multicollinearity for IVs still needs to be met for logistic regression. If two or more independent variables appear to be determined highly correlated with another variable in the model, this model suffers multicolliearity, which is hard to abtain good estimates of their unique effects on the DVs. Although it doesn't bias the coefficient estimations, multicollinearity could mostly make the estimations unstable (have less effects).

Multicollinearity diagnosis for the algorithm development sub-group from 1999 to 2003

Both the correlation matrix and tolerance confirmed the multicollinearity assumption of logistic regression was met for the algorithm development sub dataset. All IVs Pearson correlation coefficients met the criteria and tolerance values were greater than 0.40.

Table 11 Correlation Matrix of Independent Variables of the Algorithm Development Sub-group: From 1999 to 2003

part 1:

Pearson Correlation Coefficients (r) Prob > |r| under H0: Rho=0								
	agedx	**cmet**	**crt**	**cimg**	**cchemo**	**cinpadm**	**ctbchis**	**cbchis**
agedx	1	-0.01121	-0.06277	0.03416	-0.15661	-0.1187	-0.12994	-0.02174
		0.0357	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
cmet	-0.01121	1	0.02929	0.30753	0.19418	0.10047	0.02287	-0.0147
	0.0357		<.0001	<.0001	<.0001	<.0001	<.0001	0.0059
crt	-0.06277	0.02929	1	0.16761	0.20546	0.07047	0.34527	-0.00454
	<.0001	<.0001		<.0001	<.0001	<.0001	<.0001	0.3945
cimg	0.03416	0.30753	0.16761	1	0.22864	0.22462	0.10892	0.01449
	<.0001	<.0001	<.0001		<.0001	<.0001	<.0001	0.0066
cchemo	-0.15661	0.19418	0.20546	0.22864	1	0.16073	0.12984	0.02637
	<.0001	<.0001	<.0001	<.0001		<.0001	<.0001	<.0001
cinpadm	-0.1187	0.10047	0.07047	0.22462	0.16073	1	0.09886	0.02226
	<.0001	<.0001	<.0001	<.0001	<.0001		<.0001	<.0001
ctbchis	-0.12994	0.02287	0.34527	0.10892	0.12984	0.09886	1	0.00972
	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001		0.0685
cbchis	-0.02174	-0.0147	-0.00454	0.01449	0.02637	0.02226	0.00972	1
	<.0001	0.0059	0.3945	0.0066	<.0001	<.0001	0.0685	
cCOLOhis	-0.03731	-0.00204	0.02056	0.02285	0.02301	0.0403	0.05522	0.00868
	<.0001	0.7022	0.0001	<.0001	<.0001	<.0001	<.0001	0.1036
cOVChis	-0.01239	0.00221	0.00224	0.01841	0.01029	0.00871	0.01097	0.06561

	0.0202	0.6793	0.6744	0.0006	0.0539	0.1024	0.0397	<.0001
cUTNhis	-0.0106	0.00344	-0.00343	0.00128	-0.00093	0.00346	0.00451	0.01348
	0.0469	0.5195	0.5208	0.8097	0.861	0.5165	0.398	0.0115
totsurg	-0.14089	-0.06923	0.10697	0.13668	0.11385	0.20633	0.13033	0.05267
	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
totmed	-0.111	0.03849	0.06253	0.07908	0.19773	0.102	0.09261	0.02977
	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
totinpadm	-0.20487	-0.01586	0.08888	0.14498	0.23606	0.45167	0.12509	0.05495
	<.0001	0.003	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
cbio	-0.02833	0.04458	0.01896	0.17916	0.05447	0.12512	0.0364	0.03254
	<.0001	<.0001	0.0004	<.0001	<.0001	<.0001	<.0001	<.0001
cpolytm	-0.04683	-0.01941	0.02679	0.16707	0.04064	0.12302	0.04396	0.01495
	<.0001	0.0003	<.0001	<.0001	<.0001	<.0001	<.0001	0.0051
ctran	0.00524	-0.02326	0.04376	0.03276	-0.01751	0.00171	0.03345	0.00494
	0.3259	<.0001	<.0001	<.0001	0.001	0.7482	<.0001	0.3544
clar	0.00343	0.07029	-0.00307	0.15838	0.08255	0.09068	0.00872	0.00037
	0.5202	<.0001	0.5644	<.0001	<.0001	<.0001	0.1021	0.9451
capr	0.00827	0.02054	0.0422	0.06978	0.02304	0.02824	0.03449	0.00687
	0.121	0.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.1979
cliver	-0.00165	0.03974	0.00499	0.01015	0.01458	0.01322	0.00774	-0.0016
	0.7567	<.0001	0.3501	0.057	0.0063	0.0132	0.1471	0.7636
cfobt	-0.05169	-0.04512	0.0164	0.07376	0.06467	0.12105	0.03411	0.04634
	<.0001	<.0001	0.0021	<.0001	<.0001	<.0001	<.0001	<.0001
csigscopy	-0.09205	-0.00777	0.15987	0.18593	0.08185	0.17367	0.13469	0.02724
	<.0001	0.1454	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
ccolonscopy	-0.0892	0.03069	0.08279	0.33338	0.16344	0.31514	0.10876	0.05032
	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001

cdcbe	-0.00137	0.06069	0.05265	0.14546	0.07139	0.09448	0.05002	0.00905
	0.7974	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.0898
cnsaids	-0.00977	-0.01421	-0.00985	-0.00241	0.01029	0.00624	0.00648	0.00186
	0.0671	0.0077	0.065	0.6521	0.0539	0.2423	0.2247	0.7275
cfu5	-0.22465	0.14427	0.19886	0.22564	0.58017	0.21407	0.18982	0.03357
	<.0001	<.0001	<.0001	<.0001	0.0687	<.0001	<.0001	<.0001
ccapbin	-0.01777	0.02395	0.00965	0.01213	0.02926	0.02248	0.02312	-0.0038
	0.0009	<.0001	0.0704	0.023	<.0001	<.0001	<.0001	0.4763

Part 2:

Pearson Correlation Coefficients									
Prob > |r| under H0: Rho=0									
	cCOLOhis	**cOVChis**	**cUTNhis**	**totsurg**	**totmed**	**totinpadm**	**cbio**	**cpolytm**	**ctran**
agedx	-0.03731	-0.01239	-0.0106	-0.14089	-0.111	-0.20487	-0.02833	-0.04683	0.00524
	<.0001	0.0202	0.0469	<.0001	<.0001	<.0001	<.0001	<.0001	0.3259
cmet	-0.00204	0.00221	0.00344	-0.06923	0.03849	-0.01586	0.04458	-0.01941	-0.02326
	0.7022	0.6793	0.5195	<.0001	<.0001	0.003	<.0001	0.0003	<.0001
crt	0.02056	0.00224	-0.00343	0.10697	0.06253	0.08888	0.01896	0.02679	0.04376
	0.0001	0.6744	0.5208	<.0001	<.0001	<.0001	0.0004	<.0001	<.0001
cimg	0.02285	0.01841	0.00128	0.13668	0.07908	0.14498	0.17916	0.16707	0.03276
	<.0001	0.0006	0.8097	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
cchemo	0.02301	0.01029	-0.00093	0.11385	0.19773	0.23606	0.05447	0.04064	-0.01751
	<.0001	0.0539	0.861	<.0001	<.0001	<.0001	<.0001	<.0001	0.001
cinpadm	0.0403	0.00871	0.00346	0.20633	0.102	0.45167	0.12512	0.12302	0.00171
	<.0001	0.1024	0.5165	<.0001	<.0001	<.0001	<.0001	<.0001	0.7482
ctbchis	0.05522	0.01097	0.00451	0.13033	0.09261	0.12509	0.0364	0.04396	0.03345
	<.0001	0.0397	0.398	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
cbchis	0.00868	0.06561	0.01348	0.05267	0.02977	0.05495	0.03254	0.01495	0.00494
	0.1036	<.0001	0.0115	<.0001	<.0001	<.0001	<.0001	0.0051	0.3544
cCOLOhis	1	0.0026	0.00585	0.06523	0.06592	0.0982	0.01664	0.01717	0.00645
		0.6257	0.2729	<.0001	<.0001	<.0001	0.0018	0.0013	0.2268
cOVChis	0.0026	1	0.0319	0.03195	0.03293	0.04573	0.01396	0.00408	-0.00428
	0.6257		<.0001	<.0001	<.0001	<.0001	0.0089	0.4445	0.4226
cUTNhis	0.00585	0.0319	1	0.00884	0.00019	0.00233	0.00173	0.00189	-0.00323
	0.2729	<.0001		0.0976	0.9721	0.6623	0.7457	0.7229	0.5453

totsurg	0.06523	0.03195	0.00884	1	0.15061	0.42387	0.1223	0.13393	0.04993
	<.0001	<.0001	0.0976		<.0001	<.0001	<.0001	<.0001	<.0001
totmed	0.06592	0.03293	0.00019	0.15061	1	0.32231	0.03997	0.01466	0.00241
	<.0001	<.0001	0.9721	<.0001		<.0001	<.0001	0.006	0.6512
totinpadm	0.0982	0.04573	0.00233	0.42387	0.32231	1	0.11454	0.12139	0.01017
	<.0001	<.0001	0.6623	<.0001	<.0001		<.0001	<.0001	0.0565
cbio	0.01664	0.01396	0.00173	0.1223	0.03997	0.11454	1	0.18729	-0.00211
	0.0018	0.0089	0.7457	<.0001	<.0001	<.0001		<.0001	0.6922
cpolytm	0.01717	0.00408	0.00189	0.13393	0.01466	0.12139	0.18729	1	0.05768
	0.0013	0.4445	0.7229	<.0001	0.006	<.0001	<.0001		<.0001
ctran	0.00645	-0.00428	-0.00323	0.04993	0.00241	0.01017	-0.00211	0.05768	1
	0.2268	0.4226	0.5453	<.0001	0.6512	0.0565	0.6922	<.0001	
clar	0.01047	-0.00176	0.0011	0.05518	0.02747	0.07201	0.04361	0.08746	-0.00918
	0.0497	0.7413	0.8366	<.0001	<.0001	<.0001	<.0001	<.0001	0.0855
capr	-0.0034	-0.00392	-0.00296	0.01793	0.01865	0.00871	-0.00406	0.02022	0.00032
	0.5242	0.4624	0.5793	0.0008	0.0005	0.1026	0.4467	0.0002	0.9519
cliver	-0.0015	-0.0008	-0.0006	-0.00062	-0.00339	0.00806	-0.00678	0.00192	-0.00152
	0.7788	0.8807	0.9098	0.9074	0.5256	0.1307	0.2041	0.7192	0.7756
cfobt	0.01829	0.03322	0.00816	0.22703	0.05394	0.21626	0.08358	0.10674	0.01692
	0.0006	<.0001	0.1261	<.0001	<.0001	<.0001	<.0001	<.0001	0.0015
csigscopy	0.03488	0.01417	0.006	0.25872	0.06459	0.19841	0.23585	0.42465	0.09637
	<.0001	0.0079	0.2611	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
ccolonscopy	0.03994	0.01613	0.01073	0.36647	0.10685	0.3169	0.36703	0.39236	0.05387
	<.0001	0.0025	0.0442	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
cdcbe	0.00897	0.01148	0.00913	0.11408	0.04049	0.07938	0.02786	0.01255	0.0012
	0.0928	0.0315	0.0871	<.0001	<.0001	<.0001	<.0001	0.0187	0.8216
cnsaids	0.0026	-0.00225	-0.0017	0.04496	0.00041	0.02421	0.00527	0.00573	0.03606

	0.6257	0.6727	0.75	<.0001	0.938	<.0001	0.3235	0.2832	<.0001
cfu5	0.09454	0.01777	0.00056	0.20492	0.31614	0.3872	0.09862	0.04434	-0.00955
	<.0001	0.0009	0.9163	<.0001	<.0001	<.0001	<.0001	<.0001	0.0736
ccapbin	-0.00355	0.01317	-0.00143	0.01311	0.02529	0.04116	0.01284	0.0022	-0.0036
	0.5059	0.0136	0.7885	0.014	<.0001	<.0001	0.0161	0.6805	0.4998

Part 3:

Pearson Correlation Coefficients										
Prob > |r| under H0: Rho=0										
	clar	**capr**	**cliver**	**cfobt**	**csigscopy**	**ccolonscopy**	**cdcbe**	**cnsaids**	**cfu5**	**ccapbin**
agedx	0.00343	0.00827	-0.00165	-0.05169	-0.09205	-0.0892	-0.00137	-0.00977	-0.22465	-0.01777
	0.5202	0.121	0.7567	<.0001	<.0001	<.0001	0.7974	0.0671	<.0001	0.0009
cmet	0.07029	0.02054	0.03974	-0.04512	-0.00777	0.03069	0.06069	-0.01421	0.14427	0.02395
	<.0001	0.0001	<.0001	<.0001	0.1454	<.0001	<.0001	0.0077	<.0001	<.0001
crt	-0.00307	0.0422	0.00499	0.0164	0.15987	0.08279	0.05265	-0.00985	0.19886	0.00965
	0.5644	<.0001	0.3501	0.0021	<.0001	<.0001	<.0001	0.065	<.0001	0.0704
cimg	0.15838	0.06978	0.01015	0.07376	0.18593	0.33338	0.14546	-0.00241	0.22564	0.01213
	<.0001	<.0001	0.057	<.0001	<.0001	<.0001	<.0001	0.6521	<.0001	0.023
cchemo	0.08255	0.02304	0.01458	0.06467	0.08185	0.16344	0.07139	0.01029	0.48017	0.02926
	<.0001	<.0001	0.0063	<.0001	<.0001	<.0001	<.0001	0.0539	0.4526	<.0001
cinpadm	0.09068	0.02824	0.01322	0.12105	0.17367	0.31514	0.09448	0.00624	0.21407	0.02248
	<.0001	<.0001	0.0132	<.0001	<.0001	<.0001	<.0001	0.2423	<.0001	<.0001
ctbchis	0.00872	0.03449	0.00774	0.03411	0.13469	0.10876	0.05002	0.00648	0.18982	0.02312
	0.1021	<.0001	0.1471	<.0001	<.0001	<.0001	<.0001	0.2247	<.0001	<.0001
cbchis	0.00037	0.00687	-0.0016	0.04634	0.02724	0.05032	0.00905	0.00186	0.03357	-0.0038
	0.9451	0.1979	0.7636	<.0001	<.0001	<.0001	0.0898	0.7275	<.0001	0.4763
cCOLOhis	0.01047	-0.0034	-0.0015	0.01829	0.03488	0.03994	0.00897	0.0026	0.09454	-0.00355
	0.0497	0.5242	0.7788	0.0006	<.0001	<.0001	0.0928	0.6257	<.0001	0.5059
cOVChis	-0.00176	-0.00392	-0.0008	0.03322	0.01417	0.01613	0.01148	-0.00225	0.01777	0.01317
	0.7413	0.4624	0.8807	<.0001	0.0079	0.0025	0.0315	0.6727	0.0009	0.0136
cUTNhis	0.0011	-0.00296	-0.0006	0.00816	0.006	0.01073	0.00913	-0.0017	0.00056	-0.00143
	0.8366	0.5793	0.9098	0.1261	0.2611	0.0442	0.0871	0.75	0.9163	0.7885

totsurg	0.05518	0.01793	-0.00062	0.22703	0.25872	0.36647	0.11408	0.04496	0.20492	0.01311
	<.0001	0.0008	0.9074	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.014
totmed	0.02747	0.01865	-0.00339	0.05394	0.06459	0.10685	0.04049	0.00041	0.31614	0.02529
	<.0001	0.0005	0.5256	<.0001	<.0001	<.0001	<.0001	0.938	<.0001	<.0001
totinpadm	0.07201	0.00871	0.00806	0.21626	0.19841	0.3169	0.07938	0.02421	0.3872	0.04116
	<.0001	0.1026	0.1307	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
cbio	0.04361	-0.00406	-0.00678	0.08358	0.23585	0.36703	0.02786	0.00527	0.09862	0.01284
	<.0001	0.4467	0.2041	<.0001	<.0001	<.0001	<.0001	0.3235	<.0001	0.0161
cpolytm	0.08746	0.02022	0.00192	0.10674	0.42465	0.39236	0.01255	0.00573	0.04434	0.0022
	<.0001	0.0002	0.7192	<.0001	<.0001	<.0001	0.0187	0.2832	<.0001	0.6805
ctran	-0.00918	0.00032	-0.00152	0.01692	0.09637	0.05387	0.0012	0.03606	-0.00955	-0.0036
	0.0855	0.9519	0.7756	0.0015	<.0001	<.0001	0.8216	<.0001	0.0736	0.4998
clar	1	-0.01341	-0.00424	0.04083	0.11252	0.14232	0.07805	-0.0043	0.08271	0.00203
		0.0119	0.4267	<.0001	<.0001	<.0001	<.0001	0.4199	<.0001	0.7032
capr	-0.01341	1	-0.00139	0.00165	0.04712	0.04289	-0.00142	-0.00392	0.02886	0.0054
	0.0119		0.7939	0.7574	<.0001	<.0001	0.7899	0.4624	<.0001	0.3116
cliver	-0.00424	-0.00139	1	-0.00279	0.00634	-0.00811	0.00699	-0.0008	0.0088	-0.00067
	0.4267	0.7939		0.601	0.2348	0.1286	0.1898	0.8807	0.099	0.8994
cfobt	0.04083	0.00165	-0.00279	1	0.13648	0.25352	0.05378	0.01522	0.0899	0
	<.0001	0.7574	0.601		<.0001	<.0001	<.0001	0.0043	<.0001	0.9996
csigscopy	0.11252	0.04712	0.00634	0.13648	1	0.3742	0.11465	0.01136	0.12151	0.01601
	<.0001	<.0001	0.2348	<.0001		<.0001	<.0001	0.0333	<.0001	0.0027
ccolonscopy	0.14232	0.04289	-0.00811	0.25352	0.3742	1	0.09872	0.02698	0.21709	0.01115
	<.0001	<.0001	0.1286	<.0001	<.0001		<.0001	<.0001	<.0001	0.0366
cdcbe	0.07805	-0.00142	0.00699	0.05378	0.11465	0.09872	1	0.01148	0.09749	0.00836
	<.0001	0.7899	0.1898	<.0001	<.0001	<.0001		0.0315	<.0001	0.1171
cnsaids	-0.0043	-0.00392	-0.0008	0.01522	0.01136	0.02698	0.01148	1	0.00409	-0.0019

	0.4199	0.4624	0.8807	0.0043	0.0333	<.0001	0.0315		0.4436	0.7222
cfu5	0.08271	0.02886	0.0088	0.0899	0.12151	0.21709	0.09749	0.00409	1	0.03637
	<.0001	<.0001	0.099	<.0001	<.0001	<.0001	<.0001	0.4436		<.0001
ccapbin	0.00203	0.0054	-0.00067	0	0.01601	0.01115	0.00836	-0.0019	0.03637	1
	0.7032	0.3116	0.8994	0.9996	0.0027	0.0366	0.1171	0.7222	<.0001	

Note:

agedx: Age at the first diagnosis cmet: Metastatic disease crt : Radiation therapy cimg: Imaging (CT,MRI,PET, or bone scan)

cchemo: Any chemotherapy agents cinpadm: Any inpatient stay ctbchis: History of tobacco use cbchis: History of breast cancer

cCOLOhis: History of CRC cOVChis: History of ovarian cancer cUTNhis: History of Uterine cancer totsurg: No. visit to Surgeon

totmed: No. visits to medical oncologist totinpadm: No. visits to hospital cbio: Biopsy (colon, rectum) cpolytm: Polypectomy (colon, rectum)

ctran: Transanal excision (Including Microsurgery) clar: Low anterior resection (LAR) capr: Abdominal perineal resection (APR) cliver: Hepatic resection

cfobt: Fecal occult blood test (FOBT) csigscopy: Sigmoidoscopy ccolonscopy: Colonoscopy cdcbe: DCBE

cnsaids: NSAIDS cfu5: FU-5 ccapbin: Capecitabin

*: Independent Variable's Pearson Coefficient is not met the criteria.

Table 12 Multicollinearity Diagnosis of the Algorithm Development Sub-group: From 1999 to 2003

Parameter Estimates		
Variable	**Tolerance**	**Variance Inflation**
Intercept	.	0
Age at the first diagnosis	0.90756	1.10185
Metastatic disease	0.85577	1.16854
Radiation therapy	0.8269	1.20934
Imaging (CT,MRI,PET, or bone scan)	0.73377	1.36283
Any chemotherapy agents	0.51592	1.9383
Any inpatient administration	0.73827	1.35452
History of tobacco use	0.84726	1.18027
History of breast cancer (female)	0.98931	1.01081
History of Colorectal cancer	0.98128	1.01907
History of ovarian cancer	0.99125	1.00882
History of Uterine cancer	0.9984	1.00161
No. visits to surgeon	0.72378	1.38164
No. visits to medical oncologist	0.84658	1.18122
No. of hospital admission	0.57249	1.74676
Biopsy (colon, rectum)	0.8446	1.184
Polypectomy (colon, rectum)	0.74433	1.3435
Transanal excision (including Microsurgery)	0.98334	1.01695
Low anterior resection (LAR)	0.95286	1.04947
Abdominal perineal resection (APR)	0.99002	1.01008
Hepatic resection	0.99771	1.00229
Fecal occult blood test (FOBT)	0.90032	1.11072
Sigmoidoscopy	0.71041	1.40765
Colonoscopy	0.57844	1.72879
DCBE	0.95404	1.04818
NSAIDS	0.99567	1.00435
FU_5	0.44899	2.22721
Capecitabin	0.99645	1.00356

Multicollinearity diagnosis for the algorithm development sub-group from 2004 to 2007

Both the correlation matrix and tolerance confirmed the multicollinearity assumption of logistic regression was met for the algorithm development sub dataset. All IVs Pearson correlation coefficients met the criteria and tolerance values were greater than 0.40. Therefore, 5-FU, Oxaliplatin, Irinotecan, Leucovorin, FOLFOX, FOLFIRI were removed from final model.

Table 13 Correlation Matrix of Independent Variables of the Algorithm Development Sub-group: From 2004 to 2007

Part 1

Pearson Correlation Coefficients Prob > |r| under H0: Rho=0										
	agedx	**cmet**	**crt**	**cimg**	**cchemo**	**cinpadm**	**ctbchis**	**cbchis**	**cCOLOhis**	**cOVChis**
agedx	1	-0.01162	-0.0713	0.05645	-0.16424	-0.1539	-0.11374	0.00698	-0.02126	-0.00614
		0.0492	<.0001	<.0001	<.0001	<.0001	<.0001	0.2376	0.0003	0.2987
cmet	-0.01162	1	0.01039	0.24942	0.20119	0.0774	0.02734	0.00175	0.02891	0.01398
	0.0492		0.0788	<.0001	<.0001	<.0001	<.0001	0.7675	<.0001	0.018
crt	-0.0713	0.01039	1	0.18647	0.20319	0.06568	0.3141	0.00828	0.01758	0.00887
	<.0001	0.0788		<.0001	<.0001	<.0001	<.0001	0.1614	0.0029	0.1335
cimg	0.05645	0.24942	0.18647	1	0.21061	0.16876	0.10822	0.02257	0.03121	0.00955
	<.0001	<.0001	<.0001		<.0001	<.0001	<.0001	0.0001	<.0001	0.1062
cchemo	-0.16424	0.20119	0.20319	0.21061	1	0.13015	0.09182	0.02241	0.05023	0.00648
	<.0001	<.0001	<.0001	<.0001		<.0001	<.0001	0.0002	<.0001	0.2727
cinpadm	-0.1539	0.0774	0.06568	0.16876	0.13015	1	0.07609	0.02384	0.01942	0.00406
	<.0001	<.0001	<.0001	<.0001	<.0001		<.0001	<.0001	0.001	0.4918
ctbchis	-0.11374	0.02734	0.3141	0.10822	0.09182	0.07609	1	0.00345	0.037	0.00059
	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001		0.5589	<.0001	0.9208
cbchis	0.00698	0.00175	0.00828	0.02257	0.02241	0.02384	0.00345	1	0.01353	0.04474
	0.2376	0.7675	0.1614	0.0001	0.0002	<.0001	0.5589		0.0221	<.0001
cCOLOhis	-0.02126	0.02891	0.01758	0.03121	0.05023	0.01942	0.037	0.01353	1	-0.00388

	0.0003	<.0001	0.0029	<.0001	<.0001	0.001	<.0001	0.0221		0.5118
cOVChis	-0.00614	0.01398	0.00887	0.00955	0.00648	0.00406	0.00059	0.04474	-0.00388	1
	0.2987	0.018	0.1335	0.1062	0.2727	0.4918	0.9208	<.0001	0.5118	
cUTNhis	0.00207	0.00441	-0.00843	0.00022	0.0025	0.0134	-0.00603	0.01818	-0.00295	0.07444
	0.7261	0.4557	0.154	0.9704	0.6726	0.0233	0.3077	0.0021	0.6176	<.0001
totsurg	-0.09934	-0.02924	0.14488	0.26306	0.15744	0.15539	0.12268	0.03312	0.04094	0.01748
	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.0031
totmed	-0.10063	0.08228	0.07097	0.10197	0.2271	0.08168	0.05315	0.02937	0.06289	-0.00375
	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.5263
totinpadm	-0.18099	0.06819	0.13068	0.23001	0.33506	0.44849	0.12185	0.03986	0.07821	0.01266
	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.0323
cbio	0.00297	0.03679	0.01422	0.02201	0.01195	0.07728	0.0075	-0.00184	-0.00126	-0.00469
	0.6158	<.0001	0.0162	0.0002	0.0433	<.0001	0.2042	0.7559	0.8308	0.4278
cpolytm	-0.03316	-0.02934	0.01638	0.18874	0.00976	0.09383	0.04114	-0.00002	0.0062	0.00004
	<.0001	<.0001	0.0056	<.0001	0.0986	<.0001	<.0001	0.9971	0.294	0.9947
ctran	-0.00499	-0.02553	0.04898	0.03449	-0.01624	-0.0008	0.02318	0.01118	0.00541	0.00929
	0.3989	<.0001	<.0001	<.0001	0.006	0.8919	<.0001	0.0585	0.3604	0.1161
clar	-0.00406	0.01822	-0.0253	0.17798	0.04136	0.09043	0.00352	0.00291	0.00107	0.02287
	0.4921	0.0021	<.0001	<.0001	<.0001	<.0001	0.5519	0.6227	0.8562	0.0001
capr	-0.00355	-0.00202	0.04052	0.05412	0.0068	0.01455	0.03033	-0.00673	0.00566	0.00559
	0.5477	0.7324	<.0001	<.0001	0.2499	0.0138	<.0001	0.2551	0.3383	0.3445
cliver	-0.0173	0.07146	-0.00114	0.01318	0.00149	0.01793	-0.00576	-0.0026	0.01275	-0.00145
	0.0034	<.0001	0.8467	0.0257	0.8009	0.0024	0.3298	0.6605	0.031	0.8058
cfobt	-0.01474	-0.02603	0.01433	0.12455	0.06544	0.06583	0.01466	0.0177	0.01129	0.013

	0.0127	<.0001	0.0154	<.0001	<.0001	<.0001	0.0131	0.0028	0.0561	0.0279
csigscopy	-0.03808	0.00561	0.21703	0.16458	0.07216	0.09879	0.09729	0.01509	0.02018	0.00377
	<.0001	0.3427	<.0001	<.0001	<.0001	<.0001	<.0001	0.0107	0.0006	0.5238
ccolonscopy	-0.04341	0.02971	0.0881	0.41827	0.14699	0.2351	0.09037	0.02913	0.02414	0.02226
	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.0002
cdcbe	0.00244	0.03721	0.07754	0.14159	0.05212	0.05476	0.05365	0.0197	0.00927	0.00625
	0.6803	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.0009	0.1167	0.2907
cfu5	-0.21977	0.1605	0.19403	0.21885	0.51957	0.16608	0.13936	0.02216	0.07515	0.00344
	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.0002	<.0001	0.5609
ccapbin	-0.01152	0.02308	0.02397	0.02581	0.03831	0.02333	0.01067	-0.00348	0.01937	-0.00195
	0.0513	<.0001	<.0001	<.0001	<.0001	<.0001	0.0711	0.5555	0.0011	0.7414
coxal	-0.2157	0.1731	0.09599	0.17256	0.4696	0.1481	0.09084	0.01126	0.0652	-0.00312
	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.0568	<.0001	0.5972
cirin	-0.14737	0.2116	0.04536	0.11267	0.30431	0.09581	0.08373	-0.00232	0.09478	-0.00892
	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.6947	<.0001	0.1313
cleuc	-0.21059	0.16425	0.13726	0.21002	0.58929	0.16243	0.11662	0.0197	0.07697	0.00098
	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.0009	<.0001	0.8684
cbeva	-0.12639	0.23159	0.03893	0.13253	0.31418	0.10001	0.07351	0.00021	0.10176	-0.00274
	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.9711	<.0001	0.6428
ccetu	-0.10141	0.16823	0.02073	0.07766	0.21192	0.06811	0.05598	-0.00228	0.08576	0.00498
	<.0001	<.0001	0.0005	<.0001	<.0001	<.0001	<.0001	0.6998	<.0001	0.3994
cfolfox	-0.19923	0.15929	0.08224	0.15868	0.4438	0.13299	0.08512	0.01281	0.06117	-0.0025
	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.0302	<.0001	0.6722
cfolfiri	-0.13771	0.19176	0.03849	0.10599	0.29596	0.08572	0.078	-0.00416	0.09475	-0.00747
	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.4817	<.0001	0.2062

Part 2

Pearson Correlation Coefficients											
Prob > |r| under H0: Rho=0											
	cUTNhis	**totsurg**	**totmed**	**totinpadm**	**cbio**	**cpolytm**	**ctran**	**clar**	**capr**	**cliver**	**cfobt**
agedx	0.00207	-0.09934	-0.10063	-0.18099	0.00297	-0.03316	-0.00499	-0.00406	-0.00355	-0.0173	-0.01474
	0.7261	<.0001	<.0001	<.0001	0.6158	<.0001	0.3989	0.4921	0.5477	0.0034	0.0127
cmet	0.00441	-0.02924	0.08228	0.06819	0.03679	-0.02934	-0.02553	0.01822	-0.00202	0.07146	-0.02603
	0.4557	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.0021	0.7324	<.0001	<.0001
crt	-0.00843	0.14488	0.07097	0.13068	0.01422	0.01638	0.04898	-0.0253	0.04052	-0.00114	0.01433
	0.154	<.0001	<.0001	<.0001	0.0162	0.0056	<.0001	<.0001	<.0001	0.8467	0.0154
cimg	0.00022	0.26306	0.10197	0.23001	0.02201	0.18874	0.03449	0.17798	0.05412	0.01318	0.12455
	0.9704	<.0001	<.0001	<.0001	0.0002	<.0001	<.0001	<.0001	<.0001	0.0257	<.0001
cchemo	0.0025	0.15744	0.2271	0.33506	0.01195	0.00976	-0.01624	0.04136	0.0068	0.00149	0.06544
	0.6726	<.0001	<.0001	<.0001	0.0433	0.0986	0.006	<.0001	0.2499	0.8009	<.0001
cinpadm	0.0134	0.15539	0.08168	0.44849	0.07728	0.09383	-0.0008	0.09043	0.01455	0.01793	0.06583
	0.0233	<.0001	<.0001	<.0001	<.0001	<.0001	0.8919	<.0001	0.0138	0.0024	<.0001
ctbchis	-0.00603	0.12268	0.05315	0.12185	0.0075	0.04114	0.02318	0.00352	0.03033	-0.00576	0.01466
	0.3077	<.0001	<.0001	<.0001	0.2042	<.0001	<.0001	0.5519	<.0001	0.3298	0.0131
cbchis	0.01818	0.03312	0.02937	0.03986	-0.00184	-0.00002	0.01118	0.00291	-0.00673	-0.0026	0.0177
	0.0021	<.0001	<.0001	<.0001	0.7559	0.9971	0.0585	0.6227	0.2551	0.6605	0.0028
cCOLOhis	-0.00295	0.04094	0.06289	0.07821	-0.00126	0.0062	0.00541	0.00107	0.00566	0.01275	0.01129
	0.6176	<.0001	<.0001	<.0001	0.8308	0.294	0.3604	0.8562	0.3383	0.031	0.0561
cOVChis	0.07444	0.01748	-0.00375	0.01266	-0.00469	0.00004	0.00929	0.02287	0.00559	-0.00145	0.013

	<.0001	0.0031	0.5263	0.0323	0.4278	0.9947	0.1161	0.0001	0.3445	0.8058	0.0279
cUTNhis	1	0.01568	0.00516	0.00597	0.00409	-0.0048	-0.00378	-0.00487	-0.00287	-0.00111	0.00085
		0.008	0.3825	0.3128	0.4886	0.4167	0.5227	0.4104	0.6278	0.8516	0.8861
totsurg	0.01568	1	0.14887	0.40324	0.00527	0.15554	0.05402	0.13654	0.04725	-0.00301	0.18838
	0.008		<.0001	<.0001	0.3724	<.0001	<.0001	<.0001	<.0001	0.6111	<.0001
totmed	0.00516	0.14887	1	0.31393	-0.01031	0.00775	-0.00241	0.01874	0.01379	0.00873	0.05311
	0.3825	<.0001		<.0001	0.0811	0.19	0.6831	0.0015	0.0196	0.1399	<.0001
totinpadm	0.00597	0.40324	0.31393	1	-0.00599	0.10972	0.00424	0.10843	0.0247	0.00839	0.17751
	0.3128	<.0001	<.0001		0.3107	<.0001	0.4737	<.0001	<.0001	0.1556	<.0001
cbio	0.00409	0.00527	-0.01031	-0.00599	1	0.04247	0.00611	-0.01066	0.00481	0.00835	0.01093
	0.4886	0.3724	0.0811	0.3107		<.0001	0.3009	0.0713	0.4156	0.1577	0.0645
cpolytm	-0.0048	0.15554	0.00775	0.10972	0.04247	1	0.05428	0.10555	0.02908	-0.00731	0.08791
	0.4167	<.0001	0.19	<.0001	<.0001		<.0001	<.0001	<.0001	0.2164	<.0001
ctran	-0.00378	0.05402	-0.00241	0.00424	0.00611	0.05428	1	-0.02115	0.01521	-0.00299	0.02059
	0.5227	<.0001	0.6831	0.4737	0.3009	<.0001		0.0003	0.0101	0.6134	0.0005
clar	-0.00487	0.13654	0.01874	0.10843	-0.01066	0.10555	-0.02115	1	-0.01869	0.00351	0.06219
	0.4104	<.0001	0.0015	<.0001	0.0713	<.0001	0.0003		0.0016	0.5522	<.0001
capr	-0.00287	0.04725	0.01379	0.0247	0.00481	0.02908	0.01521	-0.01869	1	-0.00227	-0.00805
	0.6278	<.0001	0.0196	<.0001	0.4156	<.0001	0.0101	0.0016		0.7016	0.1731
cliver	-0.00111	-0.00301	0.00873	0.00839	0.00835	-0.00731	-0.00299	0.00351	-0.00227	1	-0.00713
	0.8516	0.6111	0.1399	0.1556	0.1577	0.2164	0.6134	0.5522	0.7016		0.2278
cfobt	0.00085	0.18838	0.05311	0.17751	0.01093	0.08791	0.02059	0.06219	-0.00805	-0.00713	1
	0.8861	<.0001	<.0001	<.0001	0.0645	<.0001	0.0005	<.0001	0.1731	0.2278	

csigscopy	-0.00092	0.19015	0.03939	0.11399	0.10703	0.15746	0.14439	0.13249	0.02797	0.00232	0.06975
	0.8763	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.6951	<.0001
ccolonscopy	0.01183	0.37093	0.1006	0.29699	0.14977	0.40981	0.05347	0.18797	0.0501	-0.00133	0.20899
	0.0454	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.8214	<.0001
cdcbe	-0.00521	0.13214	0.0483	0.08581	-0.00258	-0.00644	-0.0129	0.06022	-0.01102	0.00796	0.03244
	0.378	<.0001	<.0001	<.0001	0.6629	0.276	0.029	<.0001	0.0623	0.1783	<.0001
cfu5	-0.00128	0.25082	0.32496	0.48767	0.00033	0.0265	-0.0123	0.064	0.02964	0.01582	0.08291
	0.8289	<.0001	<.0001	<.0001	0.9551	<.0001	0.0374	<.0001	<.0001	0.0075	<.0001
ccapbin	-0.00148	0.01768	0.01725	0.03757	-0.00201	0.00277	-0.00401	0.0066	-0.00304	-0.00117	0.00466
	0.8017	0.0028	0.0035	<.0001	0.7332	0.64	0.4978	0.2643	0.607	0.8427	0.4304
coxal	0.00224	0.19157	0.31196	0.44513	-0.02327	0.00971	-0.01676	0.05232	0.01071	0.02045	0.06149
	0.7053	<.0001	<.0001	<.0001	<.0001	0.1003	0.0046	<.0001	0.07	0.0005	<.0001
cirin	-0.00508	0.11543	0.26908	0.40245	-0.01137	-0.00354	-0.01255	0.00509	0.00715	0.02356	0.02269
	0.3899	<.0001	<.0001	<.0001	0.0543	0.5489	0.0338	0.3895	0.2263	<.0001	0.0001
cleuc	-0.00284	0.23625	0.32758	0.48474	-0.00124	0.0245	-0.0214	0.06698	0.02499	0.0171	0.08571
	0.6311	<.0001	<.0001	<.0001	0.8341	<.0001	0.0003	<.0001	<.0001	0.0038	<.0001
cbeva	0.00398	0.1308	0.27177	0.38632	-0.01813	0.00716	-0.00935	0.0223	0.01045	0.02368	0.03722
	0.5008	<.0001	<.0001	<.0001	0.0022	0.226	0.1139	0.0002	0.0771	<.0001	<.0001
ccetu	-0.00624	0.07072	0.22017	0.32437	-0.0112	-0.0076	-0.00822	0.00382	0.0042	0.01694	0.01998
	0.2914	<.0001	<.0001	<.0001	0.0582	0.1983	0.1644	0.5179	0.4777	0.0041	0.0007
cfolfox	0.00176	0.17734	0.30727	0.43747	-0.0208	0.00942	-0.01748	0.0494	0.00897	0.02449	0.05492
	0.7664	<.0001	<.0001	<.0001	0.0004	0.1109	0.0031	<.0001	0.129	<.0001	<.0001
cfolfiri	-0.00826	0.1116	0.26574	0.39594	-0.01177	0.00084	-0.01069	0.00989	0.00922	0.02155	0.01922

	0.1622	<.0001	<.0001	<.0001	0.0465	0.8864	0.0706	0.0942	0.1186	0.0003	0.0011

Part 3:

Pearson Correlation Coefficients								
Prob > |r| under H0: Rho=0								
	csigscopy	ccolonscopy	cdcbe	cfu5	ccapbin	coxal	cirin	cleuc
agedx	-0.03808	-0.04341	0.00244	-0.21977	-0.01152	-0.2157	-0.14737	-0.21059
	<.0001	<.0001	0.6803	<.0001	0.0513	<.0001	<.0001	<.0001
cmet	0.00561	0.02971	0.03721	0.1605	0.02308	0.1731	0.2116	0.16425
	0.3427	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
crt	0.21703	0.0881	0.07754	0.19403	0.02397	0.09599	0.04536	0.13726
	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
cimg	0.16458	0.41827	0.14159	0.21885	0.02581	0.17256	0.11267	0.21002
	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
cchemo	0.07216	0.14699	0.05212	0.5857	0.03831	0.4696	0.30431	0.58929
	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
cinpadm	0.09879	0.2351	0.05476	0.16608	0.02333	0.1481	0.09581	0.16243
	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
ctbchis	0.09729	0.09037	0.05365	0.13936	0.01067	0.09084	0.08373	0.11662
	<.0001	<.0001	<.0001	<.0001	0.0711	<.0001	<.0001	<.0001
cbchis	0.01509	0.02913	0.0197	0.02216	-0.00348	0.01126	-0.00232	0.0197
	0.0107	<.0001	0.0009	0.0002	0.5555	0.0568	0.6947	0.0009
cCOLOhis	0.02018	0.02414	0.00927	0.07515	0.01937	0.0652	0.09478	0.07697
	0.0006	<.0001	0.1167	<.0001	0.0011	<.0001	<.0001	<.0001
cOVChis	0.00377	0.02226	0.00625	0.00344	-0.00195	-0.00312	-0.00892	0.00098
	0.5238	0.0002	0.2907	0.5609	0.7414	0.5972	0.1313	0.8684
cUTNhis	-0.00092	0.01183	-	-0.00128	-	0.00224	-	-0.00284

			0.00521		0.00148		0.00508	
	0.8763	0.0454	0.378	0.8289	0.8017	0.7053	0.3899	0.6311
totsurg	0.19015	0.37093	0.13214	0.25082	0.01768	0.19157	0.11543	0.23625
	<.0001	<.0001	<.0001	<.0001	0.0028	<.0001	<.0001	<.0001
totmed	0.03939	0.1006	0.0483	0.32496	0.01725	0.31196	0.26908	0.32758
	<.0001	<.0001	<.0001	<.0001	0.0035	<.0001	<.0001	<.0001
totinpadm	0.11399	0.29699	0.08581	0.48767	0.03757	0.44513	0.40245	0.48474
	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
cbio	0.10703	0.14977	-0.00258	0.00033	-0.00201	-0.02327	-0.01137	-0.00124
	<.0001	<.0001	0.6629	0.9551	0.7332	<.0001	0.0543	0.8341
cpolytm	0.15746	0.40981	-0.00644	0.0265	0.00277	0.00971	-0.00354	0.0245
	<.0001	<.0001	0.276	<.0001	0.64	0.1003	0.5489	<.0001
ctran	0.14439	0.05347	-0.0129	-0.0123	-0.00401	-0.01676	-0.01255	-0.0214
	<.0001	<.0001	0.029	0.0374	0.4978	0.0046	0.0338	0.0003
clar	0.13249	0.18797	0.06022	0.064	0.0066	0.05232	0.00509	0.06698
	<.0001	<.0001	<.0001	<.0001	0.2643	<.0001	0.3895	<.0001
capr	0.02797	0.0501	-0.01102	0.02964	-0.00304	0.01071	0.00715	0.02499
	<.0001	<.0001	0.0623	<.0001	0.607	0.07	0.2263	<.0001
cliver	0.00232	-0.00133	0.00796	0.01582	-0.00117	0.02045	0.02356	0.0171
	0.6951	0.8214	0.1783	0.0075	0.8427	0.0005	<.0001	0.0038
cfobt	0.06975	0.20899	0.03244	0.08291	0.00466	0.06149	0.02269	0.08571
	<.0001	<.0001	<.0001	<.0001	0.4304	<.0001	0.0001	<.0001
csigscopy	1	0.15375	0.12454	0.09869	0.00518	0.04871	0.04192	0.07661
		<.0001	<.0001	<.0001	0.3804	<.0001	<.0001	<.0001
ccolonscopy	0.15375	1	0.08101	0.19866	0.01343	0.1351	0.05278	0.19322

	<.0001		<.0001	<.0001	0.0231	<.0001	<.0001	<.0001
cdcbe	0.12454	0.08101	1	0.06751	-0.00979	0.05448	0.03092	0.06431
	<.0001	<.0001		<.0001	0.0976	<.0001	<.0001	<.0001
cfu5	0.09869	0.19866	0.06751	1	0.0088	0.43829	0.45665	0.44779
	<.0001	<.0001	<.0001		0.1365	<.0001	<.0001	<.0001
ccapbin	0.00518	0.01343	-0.00979	0.0088	1	0.04631	0.03621	0.01018
	0.3804	0.0231	0.0976	0.1365		<.0001	<.0001	0.085
coxal	0.04871	0.1351	0.05448	0.63829*	0.04631	1	0.46524	0.64755*
	<.0001	<.0001	<.0001	<.0001	<.0001		<.0001	<.0001
cirin	0.04192	0.05278	0.03092	0.45665	0.03621	0.46524	1	0.46382
	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001		<.0001
cleuc	0.07661	0.19322	0.06431	0.94779*	0.01018	0.64755*	0.46382	1
	<.0001	<.0001	<.0001	<.0001	0.085	<.0001	<.0001	
cbeva	0.03154	0.07317	0.03767	0.48618	0.04675	0.52865	0.60355	0.49126
	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
ccetu	0.02014	0.02426	0.01442	0.29428	0.05319	0.36825	0.59179	0.29146
	0.0007	<.0001	0.0147	<.0001	<.0001	<.0001	<.0001	<.0001
cfolfox	0.04371	0.13214	0.0477	0.69959*	0.01283	0.89263*	0.44417	0.72741*
	<.0001	<.0001	<.0001	<.0001	0.03	<.0001	<.0001	<.0001
cfolfiri	0.03978	0.06093	0.02753	0.50117	0.03311	0.44265	0.90398	0.52109
	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001

Part 4

Pearson Correlation Coefficients				
Prob > |r| under H0: Rho=0				
	cbeva	ccetu	cfolfox	cfolfiri
agedx	-0.12639	-0.10141	-0.19923	-0.13771
	<.0001	<.0001	<.0001	<.0001
cmet	0.23159	0.16823	0.15929	0.19176
	<.0001	<.0001	<.0001	<.0001
crt	0.03893	0.02073	0.08224	0.03849
	<.0001	0.0005	<.0001	<.0001
cimg	0.13253	0.07766	0.15868	0.10599
	<.0001	<.0001	<.0001	<.0001
cchemo	0.31418	0.21192	0.4438	0.29596
	<.0001	<.0001	<.0001	<.0001
cinpadm	0.10001	0.06811	0.13299	0.08572
	<.0001	<.0001	<.0001	<.0001
ctbchis	0.07351	0.05598	0.08512	0.078
	<.0001	<.0001	<.0001	<.0001
cbchis	0.00021	-0.00228	0.01281	-0.00416
	0.9711	0.6998	0.0302	0.4817
cCOLOhis	0.10176	0.08576	0.06117	0.09475
	<.0001	<.0001	<.0001	<.0001
cOVChis	-0.00274	0.00498	-0.0025	-0.00747
	0.6428	0.3994	0.6722	0.2062
cUTNhis	0.00398	-	0.00176	-0.00826

		0.00624		
	0.5008	0.2914	0.7664	0.1622
totsurg	0.1308	0.07072	0.17734	0.1116
	<.0001	<.0001	<.0001	<.0001
totmed	0.27177	0.22017	0.30727	0.26574
	<.0001	<.0001	<.0001	<.0001
totinpadm	0.38632	0.32437	0.43747	0.39594
	<.0001	<.0001	<.0001	<.0001
cbio	-0.01813	-0.0112	-0.0208	-0.01177
	0.0022	0.0582	0.0004	0.0465
cpolytm	0.00716	-0.0076	0.00942	0.00084
	0.226	0.1983	0.1109	0.8864
ctran	-0.00935	-0.00822	-0.01748	-0.01069
	0.1139	0.1644	0.0031	0.0706
clar	0.0223	0.00382	0.0494	0.00989
	0.0002	0.5179	<.0001	0.0942
capr	0.01045	0.0042	0.00897	0.00922
	0.0771	0.4777	0.129	0.1186
cliver	0.02368	0.01694	0.02449	0.02155
	<.0001	0.0041	<.0001	0.0003
cfobt	0.03722	0.01998	0.05492	0.01922
	<.0001	0.0007	<.0001	0.0011
csigscopy	0.03154	0.02014	0.04371	0.03978
	<.0001	0.0007	<.0001	<.0001
ccolonscopy	0.07317	0.02426	0.13214	0.06093
	<.0001	<.0001	<.0001	<.0001
cdcbe	0.03767	0.01442	0.0477	0.02753

	<.0001	0.0147	<.0001	<.0001
cfu5	0.48618	0.29428	0.49959	0.50117
	<.0001	<.0001	<.0001	<.0001
ccapbin	0.04675	0.05319	0.01283	0.03311
	<.0001	<.0001	0.03	<.0001
coxal	0.52865	0.36825	0.89263*	0.44265
	<.0001	<.0001	<.0001	<.0001
cirin	0.60355	0.59179	0.44417	0.90398*
	<.0001	<.0001	<.0001	<.0001
cleuc	0.49126	0.29146	0.72741*	0.52109
	<.0001	<.0001	<.0001	<.0001
cbeva	1	0.42128	0.49897	0.57715
		<.0001	<.0001	<.0001
ccetu	0.42128	1	0.33498	0.5135
	<.0001		<.0001	<.0001
cfolfox	0.49897	0.33498	1	0.50066
	<.0001	<.0001		<.0001
cfolfiri	0.57715	0.5135	0.50066	1
	<.0001	<.0001	<.0001	

Note:

agedx: Age at the first diagnosis cmet: Metastatic disease crt : Radiation therapy cimg: Imaging (CT,MRI,PET, or bone scan)

cchemo: Any chemotherapy agents cinpadm: Any inpatient stay ctbchis: History of tobacco use cbchis: History of breast cancer

cCOLOhis: History of CRC cOVChis: History of ovarian cancer cUTNhis: History of Uterine cancer totsurg: No. visit to Surgeon

totmed: No. visits to medical oncologist totinpadm: No. visits to hospital cbio: Biopsy (colon, rectum) cpolytm: Polypectomy (colon, rectum)

ctran: Transanal excision (Including Microsurgery) clar: Low anterior resection (LAR) capr: Abdominal perineal resection (APR) cliver: Hepatic resection

cfobt: Fecal occult blood test (FOBT)	csigscopy: Sigmoidoscopy	ccolonscopy: Colonoscopy	cdcbe: DCBE
cnsaids: NSAIDS	cfu5: FU-5	ccapbin: Capecitabin	coxal: Oxaliplatin
cirin: Irinotecan	cleuc: Leucovorin	cbeva: Bevacizumab	ccetu:Cetuximab
cfolfox:FOLFOX	cfolfiri:FOLFIRI		

*: Independent Variable's Pearson Coefficient is not met the criteria.

Table 14 Multicollinearity Diagnosis of the Algorithm Development Sub-group: From 2004 to 2007

Parameter Estimates		
Variable	**Tolerance**	**Variance Inflation**
Intercept	.	0
Age at the first diagnosis	0.89697	1.11487
Metastatic disease	0.84344	1.18562
Radiation therapy	0.79596	1.25634
Imaging (CT,MRI,PET, or bone scan)	0.69388	1.44117
Any chemotherapy agents	0.57249	1.74675
Any inpatient stay	0.74807	1.33677
History of tobacco use	0.87983	1.13658
History of breast cancer	0.99344	1.0066
History of CRC	0.98309	1.0172
History of ovarian cancer	0.99032	1.00977
History of Uterine cancer	0.99299	1.00706
No. visit to Surgeon	0.71831	1.39215
No. visits to medical oncologist	0.83736	1.19423
No. visits to hospital	0.48133	2.07757
Biopsy (colon, rectum)	0.9517	1.05075
Polypectomy (colon, rectum)	0.80988	1.23475
Transanal excision (Including Microsurgery)	0.97046	1.03044
Low anterior resection (LAR)	0.92555	1.08044
Abdominal perineal resection (APR)	0.99111	1.00897
Hepatic resection	0.99341	1.00664
Fecal occult blood test (FOBT)	0.92688	1.07888
Sigmoidoscopy	0.85972	1.16317
Colonoscopy	0.60331	1.65751
DCBE	0.95441	1.04776
5-FU	0.99724	1.00277
NSAID	0.09131	10.95151
Capecitabin	0.8355	1.15688
Oxaliplatin	0.16076	6.22031
Irinotecan	0.13725	7.28608
Leucovorin	0.0892	11.21045
Bevacizumab	0.85353	1.16021

Cetuximab	0.96248	1.03561
FOLFOX	0.14171	7.05689
FOLFIRI	0.14627	6.83678

Algorithms Development. Logistic regression were constructed to develop the algorithms for both the study periods, backward selection method was used to eliminate less significant covariates.

Candidate Covariates and Parameter Estimation of the Algorithm for 1999-2003

DCBE, total number of impatient stays, general chemotherapy indicator, and history of uterine cancer were removed from the final algorithm due to less statistical significance ($p>0.25$). The parameter estimate can be found in Table 15.

Table 15 Algorithm Parameter estimates for Predicting CRC Advanced Disease: 1999-2003

Model to Predict CRC Advanced Disease: 1999 to 2003		
Predictor Variable	**Parameter Estimate**	**P-Value**
Intercept	2.6285	0.0002
Age at the first diagnosis [a]	-0.0262	<.0001
Metastatic disease	-1.4492	<.0001
Radiation therapy	0.1179	0.0216
Imaging (CT,MRI,PET, or bone scan)	-0.2927	<.0001
Any inpatient administration	0.1959	<.0001
History of tobacco use	0.0911	0.0192
History of breast cancer (female)	0.3083	0.0229
History of Colorectal cancer	-0.246	0.0176
History of ovarian cancer	-0.3764	0.0387
No. visits to surgeon[a]	-0.1161	<.0001
No. visits to medical oncologist[a]	0.00534	<.0001
Biopsy (colon, rectum)	-0.1445	<.0001
Polypectomy (colon, rectum)	0.1738	<.0001
Transanal excision (Including Microsurgery)	0.4891	0.0129
Low anterior resection (LAR)	0.096	0.0161
Abdominal perineal resection (APR)	0.5581	<.0001
Hepatic resection	-1.0967	0.0454
Fecal occult blood test (FOBT)	0.2762	<.0001
Sigmoidoscopy	0.0843	0.0023
Colonoscopy	0.3457	<.0001
5-FU	-0.4199	<.0001
Capecitabin	-0.9	<.0001

Note: Model fit characteristics: R-square 0.2320, Max-rescaled R-square 0.3913

a: Entered as continuous variables.

Then classification table was computed to assess the performance of the algorithm for predicting the advanced disease during the study period, which in this case was 1999 to 2003. We tried to choose the best probability level based on the predicting values (including the sensitivity, specificity) from the classification table, which compared the counterpart results from using ICD-9 metastatic disease diagnosis as the predictor.

If ICD-9 metastatic disease diagnosis was applied as the single predictor for advanced disease, the sensitivity is 50.0%, the specificity is 92.0%, the PPV is 66.4% and NPV is 90.2%.

The algorithm for 1999 to 2003 could achieve slightly better predictive values (the sensitivity is 50.3%, the specificity is 95.0%, the PPV is 68.3% and the NPV is 93.4%) when the predictive probability=0.30 (for detail check table 16 and figure 7). Although we may achieve better correct rate, but it made the sensitivity drop significantly.

Table 16 The classification Table of Predicting Algorithm of Advanced CRC for 1999 to 2003

Classification Table of Predicting Algorithm for 1999 to 2003			
Probability	Percentages (%)		
Level	Correct Rate	Sensitivity	Specificity
0.20	84.6	59.1	89.7
0.21	85.7	56.7	91.4
0.22	86.5	54.4	92.9
0.23	87.0	53.2	93.7
0.24	87.1	52.6	94.0
0.25	87.2	52.2	94.2
0.26	87.3	51.7	94.4
0.27	87.4	51.3	94.6
0.28	87.5	51.0	94.8
0.29	87.5	50.6	94.9
0.30	87.6	50.3	95.0
0.31	87.6	50.0	95.1
0.32	87.7	49.8	95.2
0.33	87.7	49.6	95.3
0.34	87.8	49.3	95.4
0.35	87.8	49.0	95.5
0.36	87.8	48.8	95.5
0.37	87.8	48.6	95.6
0.38	87.8	48.4	95.7
0.39	87.9	48.1	95.8

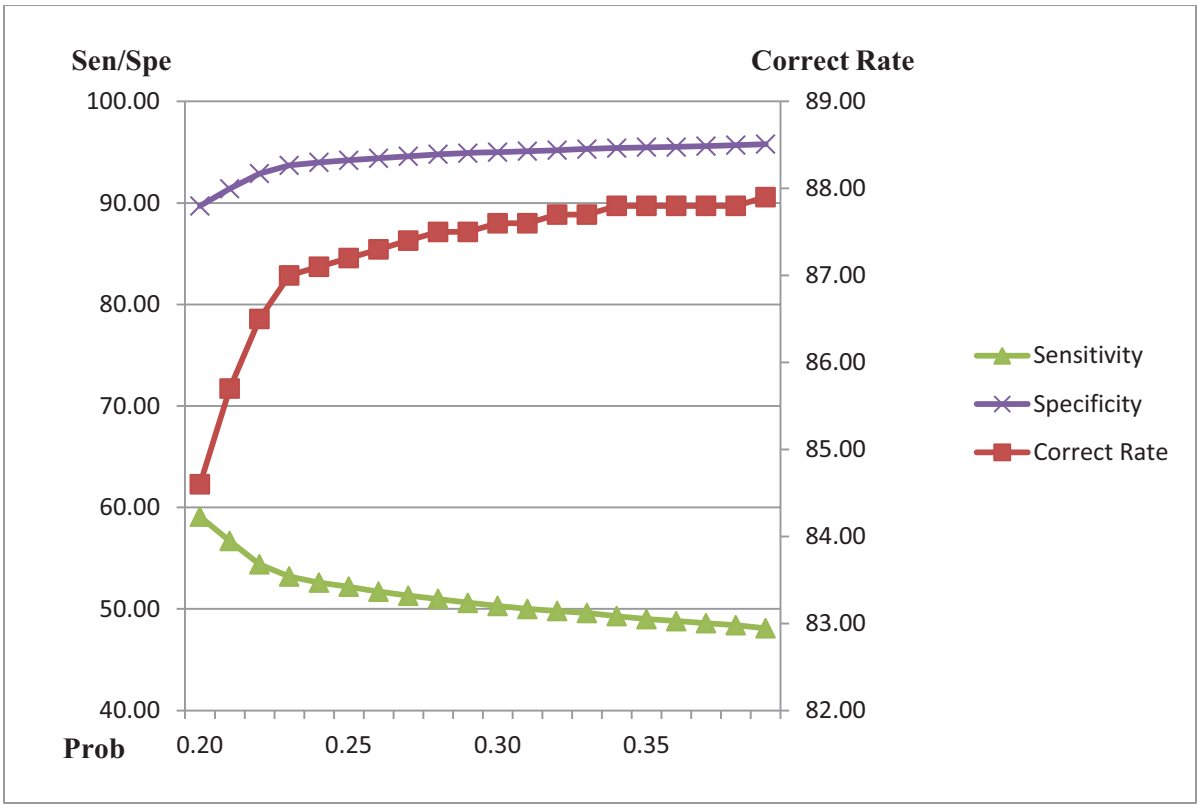


Figure 7 The Predictive Values Trend of Predicting Algorithm of Advanced CRC for 1999 to 2003

Candidate Covariates and Parameter Estimation of the Algorithm for 2004-2007

Total number of impatient stays, general radiation indicator, history of tobacco use, history of uterine cancer, history of colon cancer, and history of ovarian cancer were removed from the final algorithm due to less statistical significance ($p>0.25$).

The parameter estimation can be found in table 17.

Table 17 Algorithm Parameter estimates for Predicting CRC Advanced Disease: 2004 to 2007

Model to Predict CRC Advanced Disease: 2004 to 2007		
Predictor Variable	**Parameter Estimate [b]**	**P-Value**
Intercept	-0.0793	0.8609
Age at the first diagnosis [a]	-0.0121	<.0001
Metastatic disease	-1.6027	<.0001
Imaging (CT,MRI,PET, or bone scan)	-0.2276	<.0001
Any chemotherapy agents	-0.0819	0.0434
Any inpatient administration	0.16	<.0001
History of breast cancer (female)	0.2213	0.1144
No. visits to surgeon[a]	-0.0894	<.0001
No. visits to medical oncologist[a]	0.00319	0.0444
No. of hospital admission[a]	-0.0003	0.0021
Biopsy (colon, rectum)	-0.1506	0.0118
Polypectomy (colon, rectum)	0.1388	0.0001
Transanal excision (Including Microsurgery)	0.538	0.0022
Low anterior resection (LAR)	0.3285	<.0001
Abdominal perineal resection (APR)	0.4192	0.0229
Fecal occult blood test (FOBT)	0.3027	<.0001
Sigmoidoscopy	0.0776	0.0264
Colonoscopy	0.4227	<.0001
DCBE	-0.0834	0.0677
5-FU	-0.0761	0.0648
Capecitabin	0.3437	0.2004
Bevacizumab	-0.746	<.0001
Cetuximab	-0.6423	<.0001

Note: Model fit characteristics: R-square 0.2729, Max-rescaled R-square 0.4510

a: Entered as continuous variables.

The classification table was computed to assess the performance of the algorithm for predicting the advanced disease during the study period, which in this case was 2004 to 2007. The best probability level was chosen from the classification table based on the predicting values (including the sensitivity, specificity), then compared to the counterpart results from using ICD-9 metastatic disease diagnosis as the predictor.

ICD-9 metastatic disease diagnosis was applied as the single predictor for advanced disease, the sensitivity was 52.8%, the specificity was 93.2%, the PPV was 66.1% and the NPV was 90.4%. Compared with these results, we found when the predictive probability=0.33 (for detail check table 18 and figure 8), the algorithm could achieve slightly better predictive values (the sensitivity was 57.4%, the specificity was 96.1%, the PPV was 67.5% and the NPV was 93.4%). Although we may achieve better correct predictive rate, but it made the sensitivity drop significantly.

Table 18 The classification Table of Predicting Algorithm of Advanced CRC for 1999 to 2003

Classification Table of Predicting Algorithm for 1999 to 2003			
Probability	Percentages (%)		
Level	Correct Rate	Sensitivity	Specificity
0.20	87.3	60.8	94.9
0.21	87.6	60.5	95.0
0.22	87.8	60.0	95.1
0.23	87.9	59.6	95.2
0.24	88.0	59.1	95.3
0.25	88.1	58.8	95.4
0.26	88.2	58.7	95.5
0.27	88.2	58.4	95.6
0.28	88.3	58.3	95.7
0.29	88.3	58.2	95.8
0.30	88.4	58.0	95.9
0.31	88.4	57.7	96.0
0.32	88.4	57.5	96.1
0.33	88.4	57.4	96.2
0.34	88.5	57.1	96.3
0.35	88.5	56.9	96.4
0.36	88.6	56.6	96.5
0.37	88.6	56.3	96.6
0.38	88.6	56.1	96.7
0.39	88.6	55.9	96.8
0.40	88.6	55.5	96.9
0.41	88.6	55.3	97.0

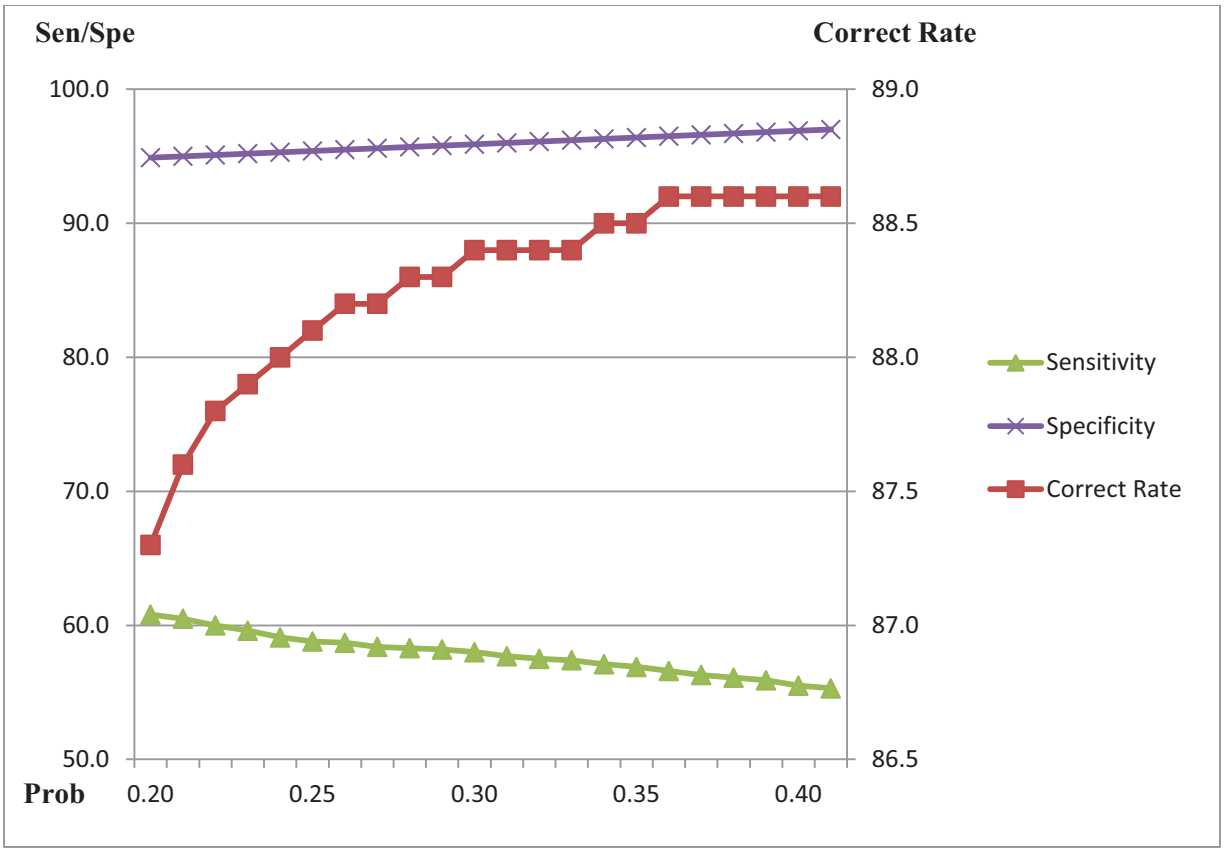


Figure 8 The Predictive Values Trend of Predicting Algorithm of Advanced CRC for 2004 to 2007

Algorithms Applications and Modification in MarketScan Database

Patient Selection. In the MarketScan® Commercial Claims and Encounters Database, 10,441,357 recipients were found in 2002 to 2003 files. Among those recipients, we identified 13,163 patients (raw diagnosis number) with a total 742,740 claim records who had at least one primary diagnosis for colon or rectum cancer in the database. The patients with CRC found in the -database included the following ICD-9 classification codes: malignant neoplasm of hepatic flexure (153.0), malignant neoplasm of transverse colon (153.1), malignant neoplasm of descending colon (153.2), malignant neoplasm of sigmoid colon (153.3), malignant neoplasm of cecum (153.4), malignant neoplasm of appendix vermiformis (153.5), malignant neoplasm of ascending colon (153.6), malignant neoplasm of splenic flexure (153.7), malignant neoplasm of other specified sites of large intestine (153.8), malignant neoplasm of colon, unspecified site (153.9), malignant neoplasm of rectosigmoid junction (154.0), malignant neoplasm of rectum (154.1), malignant neoplasm of other sites of rectum, rectosigmoid junction, and anus (154.8). (Detail is found in Table 19). Because of the insurance enrollment criteria, we excluded 2324 patients who didn't have 6 months before and after insurance coverage. We also excluded 236 patients for being lost to follow-up 6 months first CRC diagnosis. After all selection criteria were applied, there were 10,603 colorectal cancer patients included in this final study. (Seen in Figure 7)

Table 19 Number of Colorectal Cancer Patients by ICD-9 Classification in MarketScan Commercial Claims Database

ICD-9 Codes	Number of Patients	Percent (%)
153.0	261	1.98
153.1	298	2.26
153.2	286	2.17
153.3	1338	10.16
153.4	652	4.95
153.5	278	2.11
153.6	631	4.79
153.7	111	0.84
153.8	466	3.54
153.9	5907	44.88
154.0	0	0
154.1	2726	20.71
154.8	209	1.59

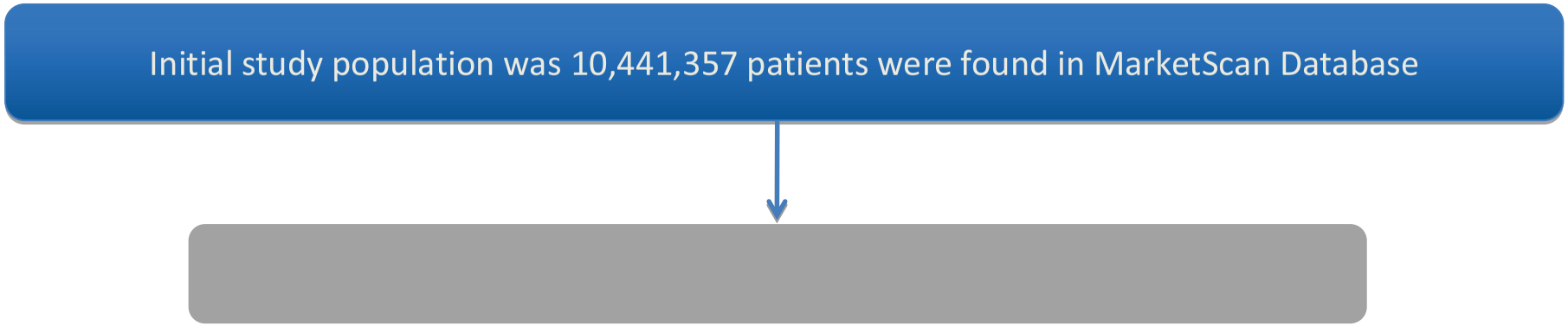


Figure 9 Patient Selection Flowchart for MarketScan Database

Description of Dependent Variables

Demographic

The mean of CRC patient population's age was 54.2 (median 56, SD 8.13, range 0-93). 22 percent of patients were 60 years old or older. 968 patients were 18 years old or younger. 5328 patients were males and 5275 were females.

The Extent of the Disease

During 3 month horizon before and after the first diagnosis, the number of patients who had at least one diagnosis for secondary metastatic disease was 1801. 71 patients had at least one lymph nodes involvement diagnosis.

Cancer Treatment

Within 6 months after the primary colon or rectum cancer diagnosis, the total number of visits to a cancer surgeon was 49,457. Of those patients, the mean number of visits was 4.66 (SD 6.63, range 0-64). 29,387 visits to medical oncologists were found for all eligible patients. The mean number of visits for medical oncologists was 2.77 (SD 12.04, range 0-163). The total number of visits to radiologists was 8,610. The mean number of radiation treatment visits was 0.81 (SD 3.21, range 0-68).

Radiation therapy

Within 6 months after the first diagnosis, 1,566 patients had at least one radiation therapy which accounts for 6.01% of the study population.

Chemotherapy and biologic therapy

For all eligible CRC patients, the prescription records were extracted from outpatient drug records as well as the inpatient services records, by using National Drug Codes (NDC) or clinical codes (e.g. HCPCS or CPT).

27,779 records were found in database for 5-FU prescriptions. The mean number of 5-FU use was 2.62 (SD 13.21). 178 patients had irinotecan in 6 months after diagnosis. The mean prescription use of irinotecan was 0.0167 (SD 2.27). The total number of leucovorin prescription for those patients was 8718. The mean prescription use of leucovorin was 0.822 (SD 13.1).

Screening tools, imaging tests and disease history

1298 patients who had at least one record for FOBT after 3 months horizon before and after the primary diagnosis were found in the database, it accounts for 12.25% of the study population. The number of patients who had sigmoidoscopy was 2778, about 21.48% of the study population. The number of patients who had at least one colonscopy was 6298, which accounts for 59.37% of total patients. The number of patients who had at least one DCBE was 736, which accounts for 6.94%. All types of imaging procedures were taken into account for this study, 5487 patients were found that had at least one imaging procedure including CT, PET, MRI or bone scan.

For the aspect of disease history, 93 patients have a record indicating the history of tobacco use, which was a surprisingly low number. The number of patients who had history of previous ovarian cancer, uterine cancer and breast cancer for female were 32, 15 and 64, respectively.

Surgery

All medical records within 6 months after the first CRC diagnosis were searched for the surgery procedures. 120 patients were found that had liver resection within 6 months after the primary diagnosis. The number of patients who had at least one colon or rectum biopsy procedure was 1766. 2250 patients had at least one polypectomy on either colon or rectum. Transanal excision (including microsurgery) procedure was found in 150 patients' clinical records. The number of patients who had at least one low anterior resection (LAR) procedure was 1412. 183 patients who had at least one abdominal perineal resection (APR) procedure were found in clinical records.

The SEER-Medicare Cancer Stage Algorithm Application. The algorithm developed from SEER-Medicare data during 1999 to 2003 was applied to all selected population from MarketScan. The probability of advanced CRC cancer was calculated for each individual. Then every patient was assigned to either advanced cancer group or non-advanced cancer group based on the best cut-point selected from the algorithm.

Since the best cut-point for the algorithm in 1999 to 2003 was 0.30, the participant whose predictive probability was lower than 0.30 was assigned to the non-advanced CRC group while the remaining was categorized as the advanced CRC group. Based on this principle, 9484 patients were predicted as non-advanced CRC group while 1097 patients were assigned to advanced CRC group. From the raw diagnosis number of CRC cancer patients, after compared with the SEER prevalence (44.19 per 100,000) (SEER, 2012) for CRC in 2002-2003, we found

the prevalence rate of CRC in MarketScan commercial database is much lower than the counterpart in SEER. Therefore we think the estimation of advanced disease for CRC may be higher in MarketScan population.

Due to the lack of cancer stage information in MarketScan commercial claim database, we cannot validate this result. This could be the new direction for the future study.

Discussion

Overview

This Chapter will discuss the results of the algorithms for predicting advanced CRC in different time periods and the application of those algorithms. The implications, limitations and future research will also be discussed.

The Improvement of the Cancer Predictive Algorithms

During 1999 to 2007, 212,345 SEER participants were identified for having at least one CRC diagnosis. Among those patients, 127,507 were selected in our cohort for developing the algorithm for advanced CRC stage.

When developing predictive algorithms for both study periods, we were able to achieve better predictive results than the single factor predictive model by using the ICD-9 metastatic disease code. The sensitivity was 57.4% and the specificity was 96.1% in the second algorithm for the later study period (2004 to 2007), compared with 52.8% and 93.2% for the single predictor model. The finding confirms our original hypothesis that due to the newly development cancer therapy (especially the introducing of biologic agents), superior predictive algorithm would be developed for the CRC cancer in the 2004 to 2007 Medicare claims data.

Overall, our predictive algorithms present an improvement based on the following characteristics:

Introducing the new biologic agents as predictors. The current advances in pharmacotherapy in the treatment of advanced-stage cancer specifically facilitate the development of better algorithms to predict cancer stage. Biological oncology agents provide a new path for treating colorectal cancer, especially for patients with advanced disease. Bevacizumab has been approved and several studies show significant benefit gained compared to chemotherapy alone. Hurwitz H, et al (Hurwitz et al., 2004) conducted a randomized clinical trial to compare bevacizumab in combination with FOLFIRI in metastatic CRC patients. The addition of bevacizumab to normal chemotherapeutic regimens showed an increase in response rate (44.9% vs34.7%), median survival (20.3% vs 15.6%) and PFS (10.6 vs 6.24 months) compared to FOLFIRI. The combination of bevacizumab with FOLFOX was also assessed by Saltz, et al (Saltz et al., 2011) and Hochster HS, et al (Hochster et al., 2008) and the results demonstrated significant benefits (response rate, PFS and OS) were received in combination treatment group. Research data also suggest that cetuximab can benefit oxaliplatin-based regimens. (Venook, 2006) As a single agent, cetuximab is associated with a 23% increase in OS compared to supportive care. (Derek J. Jonker et al., 2007) Both of these two biological agents serve as important factors in the second predictive algorithm from 2004 to 2007.

Introducing the treatment behaviors as predictors. As an important sector, treatment behaviors are well captured in a medical claims database. Moreover, this information is greatly affected by the process of cancer severity development. The number of specialist visits (cancer surgeon, medical oncologist, etc.) was the key predictors in both algorithms as well as other medical procedures.

The Implication of Findings

Cancer stage is a clinically useful classification scheme to encompass the attributes of the tumor that define its behavior. It is based on the premise that cancers of the same anatomic site and histology share similar patterns of growth and similar outcomes. Although previous research has met with, at best, limited success in predicting advanced cancer stage from claims databases. Because of the importance of cancer stage in pharmacoepidemiological and other health-outcomes studies, it is important to keep working on algorithm development using the best database, SEER-Medicare, available for that purpose. We were successful in the development of an algorithm using the very latest pharmacotherapy data, then we applied the algorithm to at least one other claims database to explore potential difficulties in the use of a database with different predictor variables.

Our algorithm can be applied as a great tool for assisting researchers for a series of epidemiological research questions in colorectal cancer fields and allows researchers to identify patients with advanced disease stage. The predicted cancer stage could serve as a covariate to address the treatment pattern or different treatment utilization or outcomes for the late stage colorectal cancer. With a probability cut point at 0.33, the algorithm for 2004 to 2007 would be highly sensitive and specific for identifying patients with advanced CRC. This selected predictive probability cut point could be used as a rule-out criteria. Also the highly NPV could lower the likelihood of misclassification bias in the sample selection. It will much easier for researchers to identify the advanced disease population and assess the medication effectiveness and adverse events, as well as monitor the quality of life for these particular patient group.

Also this algorithms can be served as a predictive tool for health insurance industry. Based on the predictive algorithm, health insurance company could estimate the population of advanced disease population for CRC and make certain decisions on manage care policy. Based on the estimation results from the predictive algorithm, health insurance company could follow up with those patients and analyze their treatment patterns and manage the certain prescriptions on the formulary.

Limitation

As proved with MarketScan commercial database, these predictive algorithms may not be directly applied to other populations. Since our study population was from Medicare (age was mainly greater than 65 years-old), there could be some major differences in the treatment patterns of those in the population. Therefore, the characteristic derived from the algorithm predictors may have underestimated or overestimated those factors, including chemotherapy, surgery procedure, and radiation utilization, compared with other populations.

There are other limitations associated with this study:

1. Ethnicity is not included in the predictive algorithm. Race is a very important predictor for cancer stage. The reason not included race into the predictive algorithm was to apply the data structure of MarketScan commercial database. Since the prevalence of CRC in different race group are quite vary, this factor should be reconsidered in the future.
2. The algorithms suffer with time-sensitive issue. The algorithms were developed based on the clinical guidelines and the newly available medication regimens

during the study periods. As the quickly development in oncology treatment, some treatment patterns may change, or applied to early stage CRC population or newly screening tools and medications may introduce, the algorithm may not able to apply to other time period.

3. The criteria of selecting the optimal predictive probability was based on the benchmark value of the single predictor (ICD-9 metastasis diagnosis code) since the specificity is very important for the advanced disease diagnosis. But this method may not be the best statistical solution for determining the optimal cut-off point. The optimal threshold for this particular situation could be determined the "costs" from perspective of patients themselves and societal perspective with the sensitivity and specificity.(Metz & Kronman, 1980; Zhou, Obuchowski, & McClish, 2008) However, such "costs" information were not available for CRC advanced disease diagnosis. This could be an interesting topic for future research work.
4. This study also has the potential for misclassification bias. For the reason of the misclassification error, there are several explanations. First, as the nature of this type of study, the medical claim database was originally designed for financial purpose but not for health research purpose. Although most medical claims contains up to more than 10 diagnosis variables, it is hard to capture all the medical diagnoses. Second, the metastatic (or advanced) patients may have secondary tumor, but the existing diagnosis code system (ICD-9 or even ICD-10) cannot distinguish each other. Both of the cancer would be listed as the records of metastatic disease, however it is hard to identify the original disease.

5. The accuracy of the diagnosis and other medical records in medical claim database cannot be verified due to lack of the ability to review the actual medical charts.
6. The adherence prescriptions (especially prescriptions for oral chemotherapeutic agents) was unknown because it cannot ensure that patients were actually taking medications correctly (or even taking them at all).
7. Since the chemotherapy can be administered in both outpatient and inpatient settings and these two systems don't share the same coding system, these chemotherapy agents may not be well captured in both settings.

Future Work

There are several fields that require future research.

First, in the second step of our study, we found the baseline of MarketScan for CRC patients were significantly different with SEER-Medicare. So the algorithm application for Marketscan cannot be validated. Other claims database should be considered, especially with the one which has an oncology cohort with a substantial number of population will be preferred.

Second, since biological agents have been widely used in clinical treatment for colorectal cancer as well as chemoprevention and early detection tools, those new innovations could change

the treatment pattern for the advanced CRC disease patients and need to be incorporated into the future algorithm development.

Third, with the application of ICD-10 code, the predicting results of metastatic disease using the diagnosis code could be changed. The impact of this change needs to be investigated and could improve the algorithm as well.

Fourth, with the development of the data mining technique, especially the newly text mining tools (such as SAS Enterprise Miner or SAS text miner), we also can try to develop the new algorithm based on certain patterns of series string of the CPT and/or HCPCS codes from oncology related claims.

Fifth, as mentioned in the limitation, the data of consequent "costs" of different predictive scenarios for advanced cancer were not available. The utility data could be a great way to assess this "costs". The QALY assessment for these population needs to conduct and could improve the understanding the patient expectation. For our study, this type data could improve the method of determine the best optimal cut-off point for predictive probability.

References

AmericanCancerSociety. (2010). The Global Economic Cost of Cancer. , from http://www.cancer.org/acs/groups/content/@internationalaffairs/documents/document/acspc-026203.pdf

André, T., Boni, C., Mounedji-Boudiaf, L., Navarro, M., Tabernero, J., Hickish, T., . . . de Gramont, A. (2004). Oxaliplatin, Fluorouracil, and Leucovorin as Adjuvant Treatment for Colon Cancer. *New England Journal of Medicine, 350*(23), 2343-2351. doi: doi:10.1056/NEJMoa032709

Arends, J. W. (2000). Molecular interactions in the Vogelstein model of colorectal carcinoma. *J Pathol, 190*(4), 412-416. doi: 10.1002/(sici)1096-9896(200003)190:4<412::aid-path533>3.0.co;2-p

Asano, T., & McLeod, R. S. (2002). Dietary fibre for the prevention of colorectal adenomas and carcinomas. [Review]. *Cochrane Database Syst Rev*(2), CD003430. doi: 10.1002/14651858.CD003430

Atkin, W. S., Edwards, R., Kralj-Hans, I., Wooldrage, K., Hart, A. R., Northover, J. M., . . . Cuzick, J. (2010). Once-only flexible sigmoidoscopy screening in prevention of colorectal cancer: a multicentre randomised controlled trial. *Lancet, 375*(9726), 1624-1633. doi: 10.1016/s0140-6736(10)60551-x

Bagnardi, V., Blangiardo, M., La Vecchia, C., & Corrao, G. (2001). A meta-analysis of alcohol drinking and cancer risk. [Meta-Analysis

Research Support, Non-U S Gov't]. *Br J Cancer, 85*(11), 1700-1705.

Baron, J. A., Beach, M., Mandel, J. S., van Stolk, R. U., Haile, R. W., Sandler, R. S., . . . Greenberg, E. R. (1999). Calcium Supplements for the Prevention of Colorectal Adenomas. *New England Journal of Medicine, 340*(2), 101-107. doi: doi:10.1056/NEJM199901143400204

Buroker, T. R., O'Connell, M. J., Wieand, H. S., Krook, J. E., Gerstner, J. B., Mailliard, J. A., . . . Gesme, D. H., Jr. (1994). Randomized comparison of two schedules of fluorouracil and leucovorin in the treatment of advanced colorectal cancer. *J Clin Oncol, 12*(1), 14-20.

Burt, R. W. (2010). Colorectal cancer screening. *Curr Opin Gastroenterol, 26*(5), 466-470. doi: 10.1097/MOG.0b013e32833d1733

Carey, L. A., Perou, C. M., Livasy, C. A., Dressler, L. G., Cowan, D., Conway, K., . . . Millikan, R. C. (2006). Race, breast cancer subtypes, and survival in the Carolina Breast Cancer Study. *JAMA, 295*(21), 2492-2502. doi: 10.1001/jama.295.21.2492

Cassidy, J., Clarke, S., Diaz-Rubio, E., Scheithauer, W., Figer, A., Wong, R., . . . Saltz, L. (2008). Randomized phase III study of capecitabine plus oxaliplatin compared with fluorouracil/folinic acid plus oxaliplatin as first-line therapy for metastatic colorectal cancer. *J Clin Oncol, 26*(12), 2006-2012. doi: 10.1200/jco.2007.14.9898

Chao, A., Thun, M. J., Jacobs, E. J., Henley, S. J., Rodriguez, C., & Calle, E. E. (2000). Cigarette Smoking and Colorectal Cancer Mortality in the Cancer Prevention Study II. *J Natl Cancer Inst, 92*(23), 1888-1896. doi: 10.1093/jnci/92.23.1888

Choti, M. A., Sitzmann, J. V., Tiburi, M. F., Sumetchotimetha, W., Rangsin, R., Schulick, R. D., . . . Cameron, J. L. (2002). Trends in long-term survival following liver resection for hepatic colorectal metastases. *Ann Surg, 235*(6), 759-766.

Citarda, F., Tomaselli, G., Capocaccia, R., Barcherini, S., & Crespi, M. (2001). Efficacy in standard clinical practice of colonoscopic polypectomy in reducing colorectal cancer incidence. *Gut, 48*(6), 812-815.

Colucci, G., Gebbia, V., Paoletti, G., Giuliani, F., Caruso, M., Gebbia, N., . . . Brunetti, C. (2005). Phase III randomized trial of FOLFIRI versus FOLFOX4 in the treatment of advanced colorectal cancer: a multicenter study of the Gruppo Oncologico Dell'Italia Meridionale. [Clinical Trial Clinical Trial, Phase III Comparative Study Multicenter Study Randomized Controlled Trial]. *J Clin Oncol, 23*(22), 4866-4875.

Cooper, G. S., Yuan, Z., Stange, K. C., Amini, S. B., Dennis, L. K., & Rimm, A. A. (1999). The Utility of Medicare Claims Data for Measuring Cancer Stage. *Medical Care, 37*(7), 706-711.

Davila, R. E., Rajan, E., Baron, T. H., Adler, D. G., Egan, J. V., Faigel, D. O., . . . Fanelli, R. D. (2006). ASGE guideline: colorectal cancer screening and surveillance. *Gastrointest Endosc, 63*(4), 546-557. doi: S0016-5107(06)00231-8 [pii]

10.1016/j.gie.2006.02.002

de Gramont, A., Bosset, J. F., Milan, C., Rougier, P., Bouche, O., Etienne, P. L., . . . Bedenne, L. (1997). Randomized trial comparing monthly low-dose leucovorin and fluorouracil bolus with bimonthly high-dose leucovorin and fluorouracil bolus plus continuous infusion for advanced colorectal cancer: a French intergroup study. *J Clin Oncol, 15*(2), 808-815.

DeVita, V. T., Hellman, S., & Rosenberg, S. A. (2005). *Cancer, principles & practice of oncology* (7th ed.). Philadelphia, PA: Lippincott Williams & Wilkins.

DeVita, V. T., Lawrence, T. S., & Rosenberg, S. A. (2011). *DeVita, Hellman, and Rosenberg's cancer : principles & practice of oncology* (9th ed.). Philadelphia: Wolters Kluwer Health/Lippincott Williams & Wilkins.

Dubé, C., Rostom, A., Lewin, G., Tsertsvadze, A., Barrowman, N., Code, C., . . . Moher, D. (2007). The Use of Aspirin for Primary Prevention of Colorectal Cancer: A Systematic Review Prepared for the U.S. Preventive Services Task Force. *Annals of Internal Medicine, 146*(5), 365-375.

Dukes, C. E., & Bussey, H. J. (1941). Venous Spread in Rectal Cancer: (Section of Proctology). *Proc R Soc Med, 34*(9), 571-573.

Efficacy of intravenous continuous infusion of fluorouracil compared with bolus administration in advanced colorectal cancer. Meta-analysis Group In Cancer. (1998). *J Clin Oncol, 16*(1), 301-308.

Engstrom, P. F., Arnoletti, J. P., Benson, A. B., Chen, Y.-J., Choti, M. A., Cooper, H. S., . . . Willett, C. (2009a). Colon Cancer. *Journal of the National Comprehensive Cancer Network, 7*(8), 778-831.

Engstrom, P. F., Arnoletti, J. P., Benson, A. B., Chen, Y.-J., Choti, M. A., Cooper, H. S., . . . Willett, C. (2009b). Rectal Cancer. *Journal of the National Comprehensive Cancer Network, 7*(8), 838-881.

Etzioni, R., Ramsey, S. D., Berry, K., & Brown, M. (2001). The impact of including future medical care costs when estimating the costs attributable to a disease: a colorectal cancer case study. *Health Econ, 10*(3), 245-256. doi: 10.1002/hec.580

Falcone, A., Ricci, S., Brunetti, I., Pfanner, E., Allegrini, G., Barbara, C., . . . Masi, G. (2007). Phase III trial of infusional fluorouracil, leucovorin, oxaliplatin, and irinotecan (FOLFOXIRI) compared with infusional fluorouracil, leucovorin, and irinotecan (FOLFIRI) as first-line treatment for metastatic colorectal cancer: the Gruppo Oncologico Nord Ovest. *J Clin Oncol, 25*(13), 1670-1676. doi: 10.1200/jco.2006.09.0928

Foster, J. H. (1984). Treatment of metastatic disease of the liver: a skeptic's view. *Semin Liver Dis, 4*(2), 170-179. doi: 10.1055/s-2008-1040656

Freeman, J. L., Zhang, D., Freeman, D. H., & Goodwin, J. S. (2000). An approach to identifying incident breast cancer cases using Medicare claims data. *J Clin Epidemiol, 53*(6), 605-614. doi: S0895-4356(99)00173-0 [pii]

Gill, S., Blackstock, A. W., & Goldberg, R. M. (2007). Colorectal cancer. *Mayo Clin Proc, 82*(1), 114-129. doi: 10.4065/82.1.114

Gleisner, A. L., Choti, M. A., Assumpcao, L., Nathan, H., Schulick, R. D., & Pawlik, T. M. (2008). Colorectal liver metastases: recurrence and survival following hepatic resection, radiofrequency ablation,

and combined resection-radiofrequency ablation. *Arch Surg, 143*(12), 1204-1212. doi: 10.1001/archsurg.143.12.1204

Goldberg, R. M., Sargent, D. J., Morton, R. F., Fuchs, C. S., Ramanathan, R. K., Williamson, S. K., . . . Alberts, S. R. (2004). A randomized controlled trial of fluorouracil plus leucovorin, irinotecan, and oxaliplatin combinations in patients with previously untreated metastatic colorectal cancer. *J Clin Oncol, 22*(1), 23-30. doi: 10.1200/jco.2004.09.046

Greene, F. L., & American Joint Committee on Cancer. (2006a). AJCC Cancer Staging Altas (pp. x, 353 p.). Retrieved from http://rave.ohiolink.edu/ebooks/ebc/9788847006942

Greene, F. L., & American Joint Committee on Cancer. (2006b). AJCC cancer staging atlas (pp. ix, 352 p.).

Grothey, A. (2003). Oxaliplatin-safety profile: neurotoxicity. *Semin Oncol, 30*(4 Suppl 15), 5-13.

Guo, J. J., Keck, P. E., Jr., Corey-Lisle, P. K., Li, H., Jiang, D., Jang, R., & L'Italien, G. J. (2006). Risk of diabetes mellitus associated with atypical antipsychotic use among patients with bipolar disorder: A retrospective, population-based, case-control study. *J Clin Psychiatry, 67*(7), 1055-1061.

Guo, J. J., Keck, P. E., Jr., Corey-Lisle, P. K., Li, H., Jiang, D., Jang, R., & L'Italien, G. J. (2007). Risk of diabetes mellitus associated with atypical antipsychotic use among Medicaid patients with bipolar disorder: a nested case-control study. *Pharmacotherapy, 27*(1), 27-35. doi: 10.1592/phco.27.1.27

Haller, D. G., Catalano, P. J., Macdonald, J. S., O'Rourke, M. A., Frontiera, M. S., Jackson, D. V., & Mayer, R. J. (2005). Phase III study of fluorouracil, leucovorin, and levamisole in high-risk stage II and III colon cancer: final report of Intergroup 0089. *J Clin Oncol, 23*(34), 8671-8678. doi: 10.1200/jco.2004.00.5686

Heaton, P. C., Guo, J. J., Hornung, R. W., Johnston, J. A., Jang, R., Moomaw, C. J., & Cluxton, R. J. (2006). Analysis of the effectiveness and cost benefit of leukotriene modifiers in adults with asthma in the Ohio Medicaid population. *J Manag Care Pharm, 12*(1), 33-42.

Hewett, D. G., Kahi, C. J., & Rex, D. K. (2010). Does Colonoscopy Work? *Journal of the National Comprehensive Cancer Network, 8*(1), 67-77.

Hochster, H. S., Hart, L. L., Ramanathan, R. K., Childs, B. H., Hainsworth, J. D., Cohn, A. L., . . . Hedrick, E. (2008). Safety and Efficacy of Oxaliplatin and Fluoropyrimidine Regimens With or Without Bevacizumab As First-Line Treatment of Metastatic Colorectal Cancer: Results of the TREE Study. *Journal of Clinical Oncology, 26*(21), 3523-3529. doi: 10.1200/jco.2007.15.4138

Hur, H., Ko, Y. T., Min, B. S., Kim, K. S., Choi, J. S., Sohn, S. K., . . . Kim, N. K. (2009). Comparative study of resection and radiofrequency ablation in the treatment of solitary colorectal liver metastases. *Am J Surg, 197*(6), 728-736. doi: 10.1016/j.amjsurg.2008.04.013

Hurwitz, H., Fehrenbacher, L., Novotny, W., Cartwright, T., Hainsworth, J., Heim, W., . . . Kabbinavar, F. (2004). Bevacizumab plus Irinotecan, Fluorouracil, and Leucovorin for Metastatic Colorectal Cancer. *New England Journal of Medicine, 350*(23), 2335-2342. doi: doi:10.1056/NEJMoa032691

IMPACT. (1995). Efficacy of adjuvant fluorouracil and folinic acid in colon cancer. International Multicentre Pooled Analysis of Colon Cancer Trials (IMPACT) investigators. *Lancet, 345*(8955), 939-944.

Iwashyna, T. J., & Lamont, E. B. (2002). Effectiveness of Adjuvant Fluorouracil in Clinical Practice: A Population-Based Cohort Study of Elderly Patients With Stage III Colon Cancer. *Journal of Clinical Oncology, 20*(19), 3992-3998. doi: 10.1200/jco.2002.03.083

Jansman, F. G. A., Postma, M. J., & Brouwers, J. R. B. J. (2007). Cost Considerations in the Treatment of Colorectal Cancer. *PharmacoEconomics, 25*(7), 537-562.

Jarvinen, R., Knekt, P., Hakulinen, T., Rissanen, H., & Heliovaara, M. (2001). Dietary fat, cholesterol and colorectal cancer in a prospective study. *Br J Cancer, 85*(3), 357-361.

Jemal, A., Bray, F., Center, M. M., Ferlay, J., Ward, E., & Forman, D. (2011). Global cancer statistics. *CA: A Cancer Journal for Clinicians, 61*(2), 69-90. doi: 10.3322/caac.20107

Jonker, D. J., Maroun, J. A., & Kocha, W. (2000). Survival benefit of chemotherapy in metastatic colorectal cancer: a meta-analysis of randomized controlled trials. *Br J Cancer, 82*(11), 1789-1794. doi: 10.1054/bjoc.1999.1254

Jonker, D. J., O'Callaghan, C. J., Karapetis, C. S., Zalcberg, J. R., Tu, D., Au, H.-J., . . . Moore, M. J. (2007). Cetuximab for the Treatment of Colorectal Cancer. *New England Journal of Medicine, 357*(20), 2040-2048. doi: doi:10.1056/NEJMoa071834

Kahnamoui, K., Cadeddu, M., Farrokhyar, F., & Anvari, M. (2007). Laparoscopic surgery for colon cancer: a systematic review. *Can J Surg, 50*(1), 48-57.

Koopman, M., Antonini, N. F., Douma, J., Wals, J., Honkoop, A. H., Erdkamp, F. L. G., . . . Punt, C. J. A. Sequential versus combination chemotherapy with capecitabine, irinotecan, and oxaliplatin in advanced colorectal cancer (CAIRO): a phase III randomised controlled trial. *The Lancet, 370*(9582), 135-142. doi: 10.1016/s0140-6736(07)61086-1

Lamprecht, S. A., & Lipkin, M. (2003). Chemoprevention of colon cancer by calcium, vitamin D and folate: molecular mechanisms. [10.1038/nrc1144]. *Nat Rev Cancer, 3*(8), 601-614.

Liang, P. S., Chen, T.-Y., & Giovannucci, E. (2009). Cigarette smoking and colorectal cancer incidence and mortality: Systematic review and meta-analysis. *International Journal of Cancer, 124*(10), 2406-2415. doi: 10.1002/ijc.24191

Los, M., Roodhart, J. M., & Voest, E. E. (2007). Target practice: lessons from phase III trials with bevacizumab and vatalanib in the treatment of advanced colorectal cancer. *Oncologist, 12*(4), 443-450. doi: 10.1634/theoncologist.12-4-443

Mandel, J. S., Bond, J. H., Church, T. R., Snover, D. C., Bradley, G. M., Schuman, L. M., & Ederer, F. (1993). Reducing mortality from colorectal cancer by screening for fecal occult blood. Minnesota Colon Cancer Control Study. *N Engl J Med, 328*(19), 1365-1371. doi: 10.1056/nejm199305133281901

Mayer, R. J. (2007). Should Capecitabine Replace Infusional Fluorouracil and Leucovorin When Combined With Oxaliplatin in Metastatic Colorectal Cancer? *Journal of Clinical Oncology, 25*(27), 4165-4167. doi: 10.1200/jco.2007.11.6582

Merika, E., Saif, M. W., Katz, A., Syrigos, K., & Morse, M. (2010). Review. Colon cancer vaccines: an update. [Review]. *In Vivo, 24*(5), 607-628.

Metz, C. E., & Kronman, H. B. (1980). Statistical significance tests for binormal ROC curves. *Journal of Mathematical Psychology, 22*(3), 218-243. doi: http://dx.doi.org/10.1016/0022-2496(80)90020-6

Miladinov-Mikov, M. (2010). Colorectal cancer epidemiology. *European journal of cancer (1990), 46*(4), 765.

Moskal, A., Norat, T., Ferrari, P., & Riboli, E. (2007). Alcohol intake and colorectal cancer risk: a dose-response meta-analysis of published cohort studies. [Meta-Analysis

Research Support, Non-U.S. Gov't]. *Int J Cancer, 120*(3), 664-671. doi: 10.1002/ijc.22299

Muller, A. D., & Sonnenberg, A. (1995). Prevention of colorectal cancer by flexible endoscopy and polypectomy. A case-control study of 32,702 veterans. *Ann Intern Med, 123*(12), 904-910.

NationalCancerInstitute. (2010). FDA Approval for Oxaliplatin Retrieved Apr 25, 2012, from http://www.cancer.gov/cancertopics/druginfo/fda-oxaliplatin

NationalCancerInstitute. (2011). FDA Approval for Cetuximab Retrieved Apr 23, 2012, from http://www.cancer.gov/cancertopics/druginfo/fda-cetuximab

Nattinger, A. B., Laud, P. W., Bajorunaite, R., Sparapani, R. A., & Freeman, J. L. (2004). An algorithm for the use of Medicare claims data to identify women with incident breast cancer. *Health Serv Res, 39*(6 Pt 1), 1733-1749. doi: HESR315 [pii]10.1111/j.1475-6773.2004.00315.x

Nordlinger, B., Sorbye, H., Glimelius, B., Poston, G. J., Schlag, P. M., Rougier, P., . . . Gruenberger, T. (2008). Perioperative chemotherapy with FOLFOX4 and surgery versus surgery alone for resectable liver metastases from colorectal cancer (EORTC Intergroup trial 40983): a randomised controlled trial. *Lancet, 371*(9617), 1007-1016. doi: 10.1016/s0140-6736(08)60455-9

Pawlik, T. M., Scoggins, C. R., Zorzi, D., Abdalla, E. K., Andres, A., Eng, C., . . . Vauthey, J. N. (2005). Effect of surgical margin status on survival and site of recurrence after hepatic resection for colorectal metastases. *Ann Surg, 241*(5), 715-722, discussion 722-714.

Petrelli, F., Borgonovo, K., Cabiddu, M., Ghilardi, M., Lonati, V., Maspero, F., . . . Barni, S. (2013). FOLFIRI-bevacizumab as first-line chemotherapy in 3500 patients with advanced colorectal cancer: a pooled analysis of 29 published trials. *Clin Colorectal Cancer, 12*(3), 145-151.

Pollock, A. M., & Vickers, N. (1998). Deprivation and emergency admissions for cancers of colorectum, lung, and breast in south east England: ecological study. *BMJ, 317*(7153), 245-252.

Potter, J. D. (1999). Colorectal cancer: molecules and populations. *J Natl Cancer Inst, 91*(11), 916-932.

Reuter, N. P., Woodall, C. E., Scoggins, C. R., McMasters, K. M., & Martin, R. C. (2009). Radiofrequency ablation vs. resection for hepatic colorectal metastasis: therapeutically equivalent? *J Gastrointest Surg, 13*(3), 486-491. doi: 10.1007/s11605-008-0727-0

Rostom, A., Dubé, C., Lewin, G., Tsertsvadze, A., Barrowman, N., Code, C., . . . Moher, D. (2007). Nonsteroidal Anti-inflammatory Drugs and Cyclooxygenase-2 Inhibitors for Primary Prevention of Colorectal Cancer: A Systematic Review Prepared for the U.S. Preventive Services Task Force. *Annals of Internal Medicine, 146*(5), 376-389.

Saltz, L., Badarinath, S., Dakhil, S., Bienvenu, B., Harker, W. G., Birchfield, G., . . . Cohn, A. (2011). Phase III Trial of Cetuximab, Bevacizumab, and 5-Fluorouracil/Leucovorin vs. FOLFOX-Bevacizumab in Colorectal Cancer. *Clin Colorectal Cancer*. doi: 10.1016/j.clcc.2011.05.006

Scholefield, J. H., Moss, S. M., Mangham, C. M., Whynes, D. K., & Hardcastle, J. D. (2011). Nottingham trial of faecal occult blood testing for colorectal cancer: a 20-year follow-up. *Gut*. doi: 10.1136/gutjnl-2011-300774

SEER. (2012). SEERDatabase: Incidence - SEER 9 Regs Research Data, Nov 2011 Sub (1973-2009) <Katrina/Rita Population Adjustment> - Linked To County Attributes - Total U.S., 1969-2010 Counties. In D. National Cancer Institute, Surveillance Research Program, Surveillance Systems Branch, (Ed.).

Simmonds, P. C. (2000). Palliative chemotherapy for advanced colorectal cancer: systematic review and meta-analysis. Colorectal Cancer Collaborative Group. *BMJ, 321*(7260), 531-535.

Smith, G., Shih, Y.-C., Giordano, S., Smith, B., & Buchholz, T. (2010). A method to predict breast cancer stage using Medicare claims. *Epidemiologic Perspectives & Innovations, 7*(1), 1.

Smith, M. R., Lee, W. C., Brandman, J., Wang, Q., Botteman, M., & Pashos, C. L. (2005). Gonadotropin-Releasing Hormone Agonists and Fracture Risk: A Claims-Based Cohort Study of Men With Nonmetastatic Prostate Cancer. *Journal of Clinical Oncology, 23*(31), 7897-7903. doi: 10.1200/jco.2004.00.6908

Smith, R. A., Cokkinides, V., & Eyre, H. J. (2007). Cancer screening in the United States, 2007: a review of current guidelines, practices, and prospects. *CA Cancer J Clin, 57*(2), 90-104. doi: 57/2/90 [pii]

Sobrero, A. F., Aschele, C., & Bertino, J. R. (1997). Fluorouracil in colorectal cancer--a tale of two drugs: implications for biochemical modulation. *J Clin Oncol, 15*(1), 368-381.

Souglakos, J., Androulakis, N., Syrigos, K., Polyzos, A., Ziras, N., Athanasiadis, A., . . . Georgoulias, V. (2006). FOLFOXIRI (folinic acid, 5-fluorouracil, oxaliplatin and irinotecan) vs FOLFIRI (folinic acid, 5-fluorouracil and irinotecan) as first-line treatment in metastatic colorectal cancer (MCC): a multicentre randomised phase III trial from the Hellenic Oncology Research Group (HORG). *Br J Cancer, 94*(6), 798-805. doi: 10.1038/sj.bjc.6603011

Thomas, S. K., Brooks, S. E., Mullins, C. D., Baquet, C. R., & Merchant, S. (2002). Use of ICD-9 coding as a proxy for stage of disease in lung cancer. *Pharmacoepidemiol Drug Saf, 11*(8), 709-713. doi: 10.1002/pds.759

Tournigand, C., Andre, T., Achille, E., Lledo, G., Flesh, M., Mery-Mignard, D., . . . de Gramont, A. (2004). FOLFIRI followed by FOLFOX6 or the reverse sequence in advanced colorectal cancer: a randomized GERCOR study. *J Clin Oncol, 22*(2), 229-237. doi: 10.1200/jco.2004.05.113

Twelves, C. (2002). Capecitabine as first-line treatment in colorectal cancer. Pooled data from two large, phase III trials. *Eur J Cancer, 38 Suppl 2*, 15-20.

USPreventiveServicesTaskForce. (2007). Routine aspirin or nonsteroidal anti-inflammatory drugs for the primary prevention of colorectal cancer: U.S. Preventive Services Task Force recommendation statement. *Ann Intern Med, 146*(5), 361-364.

Venook, A. (2006). Phase III study of irinotecan/5FU/LV (FOLFIRI) or oxaliplatin/5FU/LV (FOLFOX)±cetuximab for patients (pts) with untreated metastatic adenocarcinoma of the colon or rectum (MCRC): CALGB 80203 preliminary results. *Journal of Clinical Oncology, 24*(18s), 3509.

Wactawski-Wende, J., Kotchen, J. M., Anderson, G. L., Assaf, A. R., Brunner, R. L., O'Sullivan, M. J., . . . Manson, J. E. (2006). Calcium plus Vitamin D Supplementation and the Risk of Colorectal Cancer. *New England Journal of Medicine, 354*(7), 684-696. doi: doi:10.1056/NEJMoa055222

Warren, J. L., Feuer, E., Potosky, A. L., Riley, G. F., & Lynch, C. F. (1999). Use of Medicare hospital and physician data to assess breast cancer incidence. *Med Care, 37*(5), 445-456.

Wayne, M. S., Cath, A., & Pamies, R. J. (1995). Colorectal cancer. A practical review for the primary care physician. *Arch Fam Med, 4*(4), 357-366.

WHO. (2010). WHO Disease and injury country estimates.

Winawer, S. J., Zauber, A. G., Ho, M. N., O'Brien, M. J., Gottlieb, L. S., Sternberg, S. S., . . . et al. (1993). Prevention of colorectal cancer by colonoscopic polypectomy. The National Polyp Study Workgroup. *N Engl J Med, 329*(27), 1977-1981. doi: 10.1056/nejm199312303292701

Wolmark, N., Rockette, H., Mamounas, E., Jones, J., Wieand, S., Wickerham, D. L., . . . Fisher, B. (1999). Clinical trial to assess the relative efficacy of fluorouracil and leucovorin, fluorouracil and levamisole, and fluorouracil, leucovorin, and levamisole in patients with Dukes' B and C carcinoma of the colon: results from National Surgical Adjuvant Breast and Bowel Project C-04. *J Clin Oncol, 17*(11), 3553-3559.

Xu, J. (2009). Deaths: Data for 2007. *National vital statistics reports, 58*(1), 1.

Xu JQ, K. K., Murphy SL, Tejada-Vera B. . (2010). Deaths: Final data for 2007. National vital statistics reports (Vol. vol 58 no 19): National Center for Health Statistics..

Yabroff, K. R., Warren, J. L., Knopf, K., Davis, W. W., & Brown, M. L. (2005). Estimating patient time costs associated with colorectal cancer care. *Med Care, 43*(7), 640-648.

Zhou, X.-H., Obuchowski, N. A., & McClish, D. K. (2008). Measures of Diagnostic Accuracy *Statistical Methods in Diagnostic Medicine* (pp. 15-56): John Wiley & Sons, Inc.