University of Alberta

Urine Metabolomics and Colorectal Cancer Screening

by

Haili Wang

A thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of

Master of Science
in
Clinical Epidemiology

School of Public Health

Spring 2011
Edmonton, Alberta

Examining Committee

Dr. Richard Fedorak, Gastroenterology

Dr. Dan Schiller, Surgery

Dr. Clarence Wong, Gastroenterology

Dr. Karen Goodman, Gastroenterology & School of Public Health

Dr. Darryl Adamko, Pediatric Pulmonary Medicine

Dr. Dean Befus, Pulmonary Medicine (External Member)

Abstract

Colorectal cancer (CRC) is a major public health concern. The current population-based screening method used world-wide is fecal occult blood testing (FOBT), however this test has very low sensitivity for both colorectal cancer and adenomatous (pre-cancerous) polyps and is associated with low compliance. Metabolomics is a new field of science to study small molecules of metabolism and existing literature on metabolomics and CRC is limited. In this thesis, urine metabolomics has been shown to represent a novel, non-invasive, well-accepted screening tool for detecting CRC and adenomatous polyps with high sensitivity. The metabolomic fingerprint of CRC and that of adenomatous polyps have been explored to further understand metabolic changes in these disease states. After curative treatment of CRC, the CRC metabolomic fingerprint has been shown to remain.

Acknowledgements

This project would not have been possible without the help of numerous individuals. Special thanks to Dr. Richard Fedorak and the Clinical Investigator Program for the opportunity and financial support to work on this exciting research project. Mostly I appreciate having had the opportunity to be immersed in a special academic environment where I can bounce ideas off of colleagues and friends. Thanks to all the members of my thesis committee, namely Dr. Dan Schiller, Dr. Clarence Wong, Dr. Karen Goodman, Dr. Darryl Adamko, and Dr. Carolyn Slupsky (former member) for their guidance and advice throughout the last few years. The recruitment of the many controls and polyp patients was the combined efforts of the SCOPE study nurses. Identification of the colorectal cancer patients would not have been possible without the help of the general and colorectal surgeons and the gastroenterologists in Edmonton and Grande Prairie. Many thanks to Rae Foshaug and Victor Tso for processing and analyzing the urine samples. Also thanks to Ryan McKay and Deryck Webb from NANUC for their expertise in NMR, and David Chang from Chenomx for his advice on the analytical aspects of metabolomics.

Finally, I would like to thank my parents, sister, husband and daughter for their unconditional love, support and patience throughout all this.

Table of Contents

List of Tables

List of Figures or Illustrations

List of Abbreviations

AUC	Area Under the Curve
CRC	Colorectal Cancer
FOBT	Fecal Occult Blood Test
FIT	Fecal Immune Test
FTICR	Fourier Transform Ion Cyclotron Resonance
GC	Gas chromatography
HP	Histopathology
HPLC	High Performance Liquid Chromatography
LC	Liquid Chromatography
MS	Mass Spectrometry
NMR	Nuclear Magnetic Resonance
OPLS	Orthogonal Partial Least Squares
PCA	Principal Components Analysis
PLS-DA	Partial Least Squares - Discriminant Analysis
QUADAS	QUality Assessment of Diagnostic Accuracy Studies
ROC	Receiver Operating Characteristics
SCOPE	Stop COlorectal Cancer through Prevention and Education
TCA	Tricarboxylic Acid Cycle
TOFMS	Time-of-Flight Mass Spectrometry
VIP	Variable Importance Plot

1.0 Introduction

1.1 Aims and Outline of Thesis

Colorectal cancer is a major public health concern. The development of a more accurate, non-invasive, patient-accepted screening tool for colorectal cancer is much needed. The main goal of this thesis is to explore the potential of urine metabolomics as an effective diagnostic/screening tool for colorectal cancer and colonic adenoma.

The rest of **chapter 1** will highlight the Canadian public health perspective of CRC screening. In **chapter 2** the literature regarding metabolomics and colorectal cancer will be reviewed. **Chapter 3** will provide a detailed description of the methodology used in this study, including recruitment strategies, experimental design, as well as statistical analysis. The results of the colorectal cancer experiment will be described in **chapter 4** while the results of the colonic adenoma experiments will be described in **chapter 5**. **Chapter 6** will focus on the metabolomic fingerprint of postoperative colorectal cancer patients, and explores whether patients' metabolite profile changes after their cancer has been removed.

1.2 Colorectal Cancer Screening -- The Canadian Public Health Perspective

Colorectal cancer (CRC) is the third most common cancer and the second most frequent cause of cancer-related deaths in Canada. In 2006, an estimated 19,900

Canadians were diagnosed with CRC; 8500 died from the disease.[1] However, CRC is a disease that can be cured if identified early, and even preventable if found at the adenomatous polyp stage. The rest of this chapter will examine current screening guidelines and summarize evidence regarding adherence, but more importantly, the barriers to screening will be explored. Existing strategies to improve adherence from both the clinical and public health perspectives will be outlined.

1.2.1 Current Screening Modalities

Early detection and treatment of diseases has a potential to increase the lifespan of patients and decrease health care costs. Colorectal cancer is a suitable disease for screening since it is very common, it is serious with severe consequences, its treatment is more effective at an earlier stage, and it has a detectable preclinical phase (adenomatous polyps) that is fairly long and prevalent.[2] Current screening guidelines, for individuals with average risk (age 50 or older with no risk factors), recommended by the Canadian Association of Gastroenterology for colorectal cancer consist of one of the following: fecal occult blood testing (FOBT) every 2 years, flexible sigmoidoscopy every 5 years, flexible sigmoidoscopy combined with FOBT every 5 years, double contrast barium enema every 5 years, or colonoscopy every 10 years.[3] Each of these test modalities has limitations or potential risks associated with it. The most commonly used FOBT has been the guaiac-based test Hemoccult II. Patients are instructed to avoid consuming red meat, certain fruits and vegetables, and Vitamin C supplements for 3 days prior to

and during the stool collection in order to avoid false-positive and false-negative results. Although large population-based randomized control trials have demonstrated a survival benefit in patients who undergo annual or biannual screening with FOBT[4], the strict dietary measures and the handling of stool that are required may preclude widespread acceptance by the general population and in turn decrease compliance. Moreover, clinical studies using the Hemoccult test have estimated a low sensitivity for small cancers and polyps (26% for cancers and 13% for large adenomas).[5] Flexible sigmoidoscopy will miss proximal colonic lesions, while barium enema exposes patients to radiation and is less sensitive and specific than colonoscopy. Colonoscopy is the gold standard for screening and it can also be used for treatment of precancerous lesions, however it requires considerable resources and skilled personnel. Complications can result from sedation or the procedure itself. The risk of bleeding is approximately 1:100 and that of perforation is 1.3 per 1000 for diagnostic colonoscopy and 1.4 per 1000 for therapeutic.[6]

1.2.2 The Effectiveness of Screening

Current CRC screening strategies save lives. The largest FOBT trial to date was conducted in Nottingham, UK. From 1981 to 1991, this trial recruited 153,000 asymptomatic subjects between the ages of 45 and 74 and randomized them into control or intervention groups. The intervention group received a Hemoccult FOB test kit by mail every 2 years, which required self-collection of stool samples. Positive tests led to further investigations. Follow up continued for 4

more years after the end of the trial. This study detected a 15% reduction in colorectal cancer mortality in the intervention group compared to the control group, even though only 38% or those in the intervention group completed all the FOBTs.[7]

1.2.3 The Components of An Effective Screening Program

A screening program is much more than just a screening test, so when addressing the issue of effectiveness of screening, it is important to distinguish between test sensitivity and program sensitivity. Program sensitivity is the sensitivity achieved over time through serial testing in a program. [8] Test sensitivity can be improved with development of new and advanced technology such as Fecal Immune Testing (FIT), Stool DNA, and Urine Metabolomics etc.. Program sensitivity, however, is highly reliant on patient compliance especially since colorectal cancer screening requires repeat testing at regular intervals. The perfectly accurate and harmless test would still have minimal impact on disease prevention if barriers such as access, cost and awareness are not addressed. It is therefore important to examine what the current uptake rate is in terms of CRC screening, what barriers are preventing people from being screened and which populations should be targeted.

1.2.4 Screening Compliance

A population-based study done in Montreal in 2007 surveyed 17,498 subjects in four provinces (Ontario, NFL, Saskatchewan, and BC) and found that 70% of

respondents were not adherent to current CRC screening guidelines.[9] Similarly, a population-based study done in Ontario, using administrative data only, showed that of 982,443 screen-eligible men and women between the ages of 50-59, 79.5% did not have any tests to screen for CRC during a 6-yr follow-up. [10] In an Alberta study, only 14.3% of average risk adults (n=1,476) were up to date on CRC screening. [11]

1.2.5 Barriers to Effective Screening

A biopsychosocial framework can be used to better understand the disparities in adherence to CRC screening guidelines. Clinically, it was noted that people with health-care seeking tendencies, such as those who have a regular physician, those who get flu shots or those with a chronic condition, were more adherent to the guidelines. Psychologically, self-perceived stress was associated with increased adherence to screening, likely due to increased chances of receiving preventive health services. Since people who visit their physicians less frequently may be at risk for not receiving preventive healthcare, perhaps invitations for CRC screening should come from sources that are independent of physicians. While environmental factors such as the availability and access to screening services were not found to influence adherence[9], it has also been noted in the literature that socio-demographic factors associated with *increased* adherence include male sex[12], high-income level[13], and not working full-time[9]. Individuals who were born in Canada and were Caucasian were more likely to adhere to the guidelines[14]. The effect that socioeconomic status (SES) has on screening

behavior was studied by Whaynes et al., using the Nottingham trial data. The recruited subjects were divided into different socioeconomic groups based on the deprivation index of their general practitioners, determined by their postal code. Contrary to what was hypothesized, deprivation was not a significant factor in determining colorectal cancer prevalence. However, those with lower SES were less likely to accept the invitation to be screened, thus delaying the diagnosis and resulting in cancers diagnosed at a later stage, thus increasing mortality. Therefore, those with socioeconomic deprivation are disadvantaged by their lower participation rate.[13] These findings reinforced the need to tailor CRC screening to underserved groups.

Several US studies have also studied the barriers to CRC screening. Focus group interviews done at Harvard School of Public Health have identified three groups of factors that moderate perceived personal risk for colorectal cancer: knowledge factors, service system factors, and psychological or cognitive factors. Knowledge factors include unawareness of general prevalence of colorectal cancer (stark contrast to prostate or breast cancer), ignorance that women are just as likely to get colorectal cancer as men, misconception that family history was the single most important or even sole determinant of risk, and the assumption that risk is symptom-dependent. Service system factors include lack of preventive and screening information from providers and misconceptions about prior negative results and necessity for repeated screening. Psychological factors include concern with another health issue that offsets perceived risk for CRC,

disinclination to worry, and deference to authority of doctors who had not encouraged screening.[15]

A unique study by Klabunde et al. in 2005 looked at the barriers from the perspectives of both physicians and patients and compared the two groups. They used data from two large-scale surveys, namely the 1999-2000 Survey of Colorectal Cancer Screening Practices (n=1235) and the 2000 National Health Interview Survey (NHIS) (n=6497). Both primary care physicians and patients more often identified patient-related factors (fear of finding cancer, belief that screening isn't effective, embarrassment/anxiety about screening tests, lack of awareness of screening/CRC not perceived as a serious health threat) to be major barriers compared to system-related factors (screening costs too much/is not covered by insurance, physicians don't actively recommend screening to their patients, shortage of trained providers to conduct screening other than FOBT, shortage of trained providers to conduct follow-up with endoscopic procedures). The two groups also agreed on the fact that patients' lack of awareness of the need for screening and the lack of knowledge about CRC are important barriers.[16]

In summary, the main barriers for CRC screening are lack of awareness and misconceptions about the disease and lack of preventative information given to patients by their physicians. While cost and insurance coverage were shown to be barriers in some US studies[17], they were not shown to be barriers for screening in Canada due to our public healthcare system. The target populations should be

those that are underserved and those with low SES and the provider of information should not only be limited to physicians.

1.2.6 Overcoming Barriers to Screening

To overcome these barriers to screening, educational programs that address all these factors should be developed and implemented. The primary message should be that early detection and repeated screening for both men and women even in the absence of symptoms or family history is beneficial and important. Dissemination of knowledge is a key strategy to address some of these factors, but in addition, changes in clinical practice need to be made, such as integrating screening recommendations and follow-up as part of regular physical exams, and having active notifications of screening results.[15]

The method of recruitment for screening has traditionally been word of mouth from physician to patient or via post mail (as in some large studies). With the advancement of technology and the widespread use of electronic communication, some have hypothesized that perhaps email would be a more efficient way of notifying potential eligible screening subjects. However, a 2008 study in Houston, Texas showed implementing colon cancer screening through email over the Internet was no more effective than a mail-out reminder. The return rate for FOBT was only 25%.[18] This is perhaps because eligible screening subjects are those over the age of 50 and this portion of the population may not be as computer-literate as their younger counterparts.

Publishing guidelines on screening will help raise awareness, especially when there has been a change, as recommended by an expert panel. In 2001, the Canadian Task Force on Preventive Health Care revised its screening recommendations for CRC from its 1994 assessment that there was "inconclusive evidence to recommend screening in asymptomatic individuals over 40" (1994) to "there is good evidence to include annual or biennial screening with FOBT (grade A recommendation) and fair evidence to include flexible sigmoidoscopy (grade B recommendation)… for average risk individuals at least 50 years of age."[19] Subsequent to this publication, a study in the Canadian Journal of Surgery showed that the proportion of primary-care physicians that recommended CRC screening increased from 43% to 60%.[20]

However, guidelines for physicians will only solve part of the problem. There is a shortage of primary care physicians in Canada and resources will become even scarcer as the baby-boomers continue to age. It has also been shown that people who go to physicians regularly are not those at highest risk. Therefore, other healthcare workers in the field need to be involved in the promotion of and education on preventive health strategies. A recent Canadian study published in 2007 explored the idea for a workplace colorectal cancer-screening awareness program. In 2003, the Toronto Police Service partnered with Sunnybrook and Women's College Health Sciences Centre in Toronto to implement such a program. This program included first educating "trainers" with formal education

sessions performed by the head occupational health nurse using a 5-minute videotape containing information about CRC screening and testimonials of survivors of CRC. All trainers were also given ample opportunity for questions. The trainers then went on to educate the other members of his or her unit. Although 50% of members attended the program, only 13% completed the questionnaires and agreed to participate. This program identified that nearly 1/3 (298 out of 965) of subjects as having average or above-average risk for colorectal cancer, and would, therefore, benefit from screening.[1] Programs such as this would increase awareness through education and relieve some of the burden of addressing these potentially time-consuming issues from primary care physicians who are already over-worked. Contrary to the US where a 1995 survey of 1720 private-sector workplaces showed that 35.2% of workplaces with >750 employees have a cancer-screening program[21], there may be little financial incentive for employers in Canada to establish such programs due to the nature of our public healthcare system.

1.2.7 Alberta's Approach to Colorectal Cancer Screening

Similar to Ontario and Manitoba, Alberta is currently establishing a CRC screening program. In 2006, the Expert Working Group sponsored by Alberta Health and Wellness and the Alberta Cancer Board, recommended the adoption of population based screening for colorectal cancer for all Albertans aged 50–74 with annual fecal occult blood testing recommended as the primary screening tool for those at average risk. To meet this mandate, Capital Health (now part of

Alberta Health and Wellness), which served a population of almost 2 million people, established the Stop COlorectal cancer through Prevention and Education (SCOPE) program as a comprehensive, integrated, population-based screening program for residents of the Capital Health region. The program began as a pilot in early 2008. The program encompasses educational information, risk stratification as well as screening for both average and high-risk patients, and colonoscopy for those individuals who test positive by screening. Eligible subjects are currently referred to the program by their primary care physicians, but once the full program is launched, individuals will be able to access the program in several ways, including through HealthLink or through the website (http://www.capitalhealth.ca/EspeciallyFor/Scope/default.htm). Education sessions are held every month at the University of Alberta Hospital for eligible participants, where nurse navigators give a group session on colorectal cancer and screening. A website has been developed with helpful, easy-to-understand information regarding CRC and the SCOPE program. Paper-based education and information resources are available to individuals without access to the Internet. The pilot not only served to correct the potential problems before launching the full program, but it also provides a great opportunity for various research topics. Every subject in the pilot will be taking a fecal occult blood test (FOBT), two fecal immune tests (FIT), and providing a sample of their urine for metabolomic analysis, as well as undergoing a colonoscopy. This is a unique opportunity to compare the sensitivity and specificity of the fecal tests as well as the urine metabolomics, using colonoscopy as the gold standard, in the same individual.

Although this program is not set up specifically to target the under serviced groups, it does increase awareness through education and health promotion. Additionally, it changes the setting from a physician's office to an auditorium and increases efficiency by targeting a group of interested subjects rather than individuals.

1.2.8 Role of the Community

Health is typically thought of as a function of individual decisions, but the role of the community is vitally important. In an interview at the University of Toronto, Malcolm Glad well, the author of Outliers, illustrates this point with the story of Roseto. Roseto is a little town in the hills of Pennsylvania, which is a replica of a town in southern Italy established in the 1880's. This town is famous for the fact that the death rate is half that of the US average, despite the fact that everyone smoked, ate poorly, and was overweight. In this town of income equality and social equality, no one became rich but everyone "lived forever". Individually bad decisions were made, but as a community no one died. This is a subtle reminder that discussions regarding health should be started at the community level and worked backwards to the individuals.[22]

1.2.9 Conclusion

Colorectal cancer is a major public health concern in Canada. It is a serious but preventable condition, however, less than 30% of eligible individuals are currently being screened appropriately. There are biopsychosocial reasons for the

disparities in screening behaviors. Multiple studies have identified the barriers to screening mainly as lack of awareness and lack of prevention information given by physicians. Strategies to overcome these barriers include publishing clear guidelines, establishing education programs, and workforce-initiated screening programs. The Alberta SCOPE program is a local initiative to increase awareness of colorectal cancer and in turn improve population health through primary and secondary prevention. Developing more patient-friendly screening tests such as urine metabolomics could also improve patient compliance.

1.3 References

1. Bagai A, Parsons K, Malone B, Fantino J, Paszat L, Rabeneck L. Workplace colorectal cancer-screening awareness programs: An adjunct to primary care practice? Journal of Community Health 2007;32:157-167.
2. Aschengrau A, Seage III GR. Essentials of Epidemiology in Public Health. 2 ed: Jones and Bartlett Publishers, Inc, 2008:411.
3. Leddin D, Hunt R, Champion M, Cockeram A, Flock N, Gould M, Kim YI, Love J, Morgan D, Natsheh S, Sadowski D, Canadian Assoc Gastroenterology C. Canadian association of gastroenterology and the Canadian digestive health foundation: Guidelines on colon cancer screening. Canadian Journal of Gastroenterology 2004;18:93-99.
4. Mandel JS, Church TR, Ederer F, Bond JH. Colorectal cancer mortality: Effectiveness of biennial screening for fecal occult blood. Journal of the National Cancer Institute 1999;91:434-437.
5. Ahlquist DA. OCCULT BLOOD SCREENING - OBSTACLES TO EFFECTIVENESS. Cancer 1992;70:1259-1265.
6. Misra T, Lalor E, Fedorak RN. Endoscopic perforation rates at a Canadian university teaching hospital. Canadian Journal of Gastroenterology 2004;18:221-226.
7. Hardcastle JD, Chamberlain JO, Robinson MHE, Moss SM, Amar SS, Balfour TW, James PD, Mangham CM. Randomised controlled trial of faecal-occult-blood screening for colorectal cancer. Lancet 1996;348:1472-1477.
8. Levin B, Lieberman DA, McFarland B, Andrews KS, Brooks D, Bond J, Dash C, Giardiello FM, Glick S, Johnson D, Johnson CD, Levin TR, Pickhardt PJ, Rex DK, Smith RA, Thorson A, Winawer SJ, Ad ACSCC, Amer Coll Radiology Colon C. Screening and surveillance for the early detection of colorectal cancer and adenomatous polyps, 2008: A joint guideline from the American Cancer Society, the US Multi-Society Task Force on Colorectal Cancer, and the American College of Radiology. Gastroenterology 2008;134:1570-1595.
9. Sewitch MJ, Fournier C, Ciampi A, Dyachenko A. Adherence to colorectal cancer screening guidelines in Canada. Bmc Gastroenterology 2007;7.
10. Rabeneck L, Paszat LF. A population-based estimate of the extent of colorectal cancer screening in Ontario. American Journal of Gastroenterology 2004;99:1141-1144.
11. McGregor SE, Hilsden RJ, Li FX, Bryant HE, Murray A. Low uptake of colorectal cancer screening 3 yr after release of national recommendations for screening. American Journal of Gastroenterology 2007;102:1727-1735.
12. Brawarsky P, Brooks DR, Mucci LA. Correlates of colorectal cancer testing in Massachusetts men and women. Preventive Medicine 2003;36:659-668.
13. Whynes DK, Frew EJ, Manghan CM, Scholefield JH, Harldcastle JD. Colorectal cancer, screening and survival: the influence of socio-economic deprivation. Public Health 2003;117:389-395.
14. Swan J, Breen N, Coates RJ, Rimer BK, Lee NC. Progress in cancer screening practices in the United States - Results from the 2000 National Health Interview Survey. Cancer 2003;97:1528-1540.
15. Weitzman ER, Zapka J, Estabrook B, Goins KV. Risk and reluctance: Understanding impediments to colorectal cancer screening. Preventive Medicine 2001;32:502-513.
16. Klabunde CN, Vernon SW, Nadel MR, Breen N, Seeff LC, Brown ML. Barriers to colorectal cancer screening: A comparison of reports from primary care physicians and average-risk adults. Medical Care 2005;43:939-944.
17. Sarfaty M, Feng SB. Uptake of colorectal cancer screening in an uninsured population. Preventive Medicine 2005;41:703-706.
18. Chan ECY, Vernon SW. Implementing an intervention to promote colon cancer screening through e-mail over the internet - Lessons learned from a pilot study. Medical Care 2008;46:S117-S122.
19. Colorectal cancer screening - Recommendation statement from the Canadian Task Force on Preventive Health Care. Canadian Medical Association Journal 2001;165:206-208.

20. Asano TK, Toma D, Stern HS, McLeod RS. Current awareness in Canada of clinical practice guidelines for colorectal cancer screening. Canadian Journal of Surgery 2004;47:104-108.
21. CDC. Canadian Screening Offered by Worksites -- United States, 1992 and 1995: http://www.cdc.gov.login.ezproxy.library.ualberta.ca/mmwr/preview/mmwrhtml/00047655.htm. Accessed Nov 25, 2008
22. Interview with Malcolm Gladwell: http://media.rotman.utoronto.ca.login.ezproxy.library.ualberta.ca/vod?AdminView=yes&mediaid=1186. Accessed Dec. 1, 2008

2.0 Metabolomics and Detection of Colorectal Cancer in Humans – A Systematic Review

2.1 Abstract

Metabolomics represents one of the new "omics" sciences and capitalizes on the unique presence and concentration of small molecules in tissues and body fluids to construct a "fingerprint" that can be unique to the individual and, within that individual, unique to environmental influences, including health and disease states. As such, metabolomics has potential to serve an important role in diagnosis and management of human diseases. Colorectal cancer (CRC) is a major public health concern. Current population-based screening methods are suboptimal and whether metabolomics could represent a new tool of screening is under investigation. The purpose of this systematic review is to summarize existing literature on metabolomics and CRC, in terms of diagnostic accuracies and distinguishing metabolites. Eight studies are included[1-8]. A total of 12 metabolites (taurine, lactate, choline, inositol, glycine, phosphocholine, proline, phenylalanine, alanine, threonine, valine, and leucine) were found to be more prevalent in CRC and glucose was found to be in higher proportion in control specimens using tissue metabolomics. Serum and urine metabolomics identified several other differential metabolites between controls and CRC patients. This review highlights the novelty of the field of metabolomics in colorectal oncology.

2.2 Background

Colorectal cancer (CRC) is a leading cause of death in North America. Current non-invasive screening methods are suboptimal in sensitivity and have poor population compliance. Nuclear magnetic resonance (NMR)-based urine metabolomics is a highly novel assessment of urine-excreted small molecules that has potential to identify CRC and colonic polyps.[9] The potential for a simple single urine test to identify CRC and/or colonic polyps is exciting and of great interest to opinion leaders, politicians, scientists, physicians and patients given the world-wide interest and movement to population-based colon screening programs.

Metabolomics is an emerging field of research downstream from genomics, proteomics and transcriptomics and is a quantitative collection of low molecular weight compounds, such as metabolic substrates and products, lipids, small peptides, vitamins, and other protein cofactors[10], generated by metabolism. It is a precise, consistent, and quantitative method to examine and describe cellular growth, maintenance, and normal function.[11] It is currently being used as a mode of research in many disciplines of medicine, including psychiatry[12], obstetrics[13], gastroenterology[14], and oncology[15]. This technology is however fairly new and few human studies have been done to validate the results of existing cellular and animal studies, especially in the field of colorectal oncology. This systematic review aims to summarize the existing human literature on the diagnostic accuracies of metabolomics in the field of CRC.

Metabolomics can be performed on urine, serum, tissue, and less frequently, on fecal extracts, saliva and amniotic fluid. It is ideal for studying the effects that diseases and drugs have on the human body because it is downstream from transcriptome and proteome and thus the changes are amplified and are numerically more tractable. Also, the technology is generic such that a given metabolite is the same in every organism that contains it.[16]

The analytical techniques that make it possible to assay and quantitate components of the metabolome and to extract useful signatures from those data include liquid chromatography (LC) coupled with mass spectrometry (MS) and gas chromatography MS (GCMS) and NMR spectroscopy.[17] The outputs from MS or NMR are analyzed using multivariate analysis such as partial least squares discriminant analysis (PLS-DA) and principal component analysis (PCA).

2.3 Objectives

The primary objective of this systematic review was to summarize the reported diagnostic accuracies of serum, urine, and tissue metabolomics for detecting colorectal cancers in the adult population, using histopathology as the gold standard. The secondary objective was to summarize evidence of the most prevalent metabolites found in colorectal cancer, where studies have been carried out.

2.4 Methods

A protocol was prepared outlining the *a priori* design for this review, including criteria for considering studies for this review, search methods, data collection and analysis. In conjunction with a research librarian, a comprehensive, systematic literature search was performed. MeSH headings and keywords were used. Electronic searches through established databases [MEDLINE (Ovid) (1950 - Feb. 2009), EMBASE (Ovid) (1980 - Feb. 2009), PubMed (Sept. 2008-Feb. 2009), Cochrane Library (Issue 1, 2009), Scopus (Feb. 2009), Web of Science (1900 - Feb. 2009)], grey literature (i.e. literature not identifiable via conventional means) and conference proceedings were completed. No language restrictions were applied. Published papers and abstracts, as well as unpublished studies were included in the searches.

Broad screening of titles and abstracts were done by the author (HW). The inclusion/exclusion criteria used are outlined below.

Design: All randomized, quasi-randomized, non-randomized, retrospective and prospective cohort studies, and case-series were included. Single case reports and reviews were excluded.

Population: All studies involving human adult subjects (> 18 years of age) with primary CRC and with or without controls were included. For the serum and urine studies, the cancer patients had not commenced any medical or surgical treatment for their condition at the time of the index test. For the tissue studies,

the cancer patients had completed at least a biopsy or surgical treatment of the CRC at the time of the index test.

Index Tests: Serum, urine, or tissue metabolomics were the index tests of interest for this review.

Target Conditions: The target condition was primary CRC.

Reference Standard: The reference standard used to define the target condition was histopathological analysis of resected colorectal cancer specimens.

Outcomes: Studies were deemed suitable for inclusion if the sensitivity and specificity values of the index tests were available or derivable from the data reported in the primary studies or obtainable from the authors, or if there was information on occurrence of specific metabolites listed that distinguished CRC from controls.

Setting: Studies in any setting were included.

The assessment of the methodological quality of each included study was done by two independent reviewers (HW, VT), using the QUality Assessment of Diagnostic Accuracy Studies (QUADAS) tool,[18] which is a 14-question tool that evaluates spectrum bias, misclassification bias, disease progression bias, partial verification bias, differential verification bias, incorporation bias, review bias, and bias associated with study withdrawals and uninterpretable results. The QUADAS tool questions are included in table 2.1.

Table 2.1: The Quality Assessment of Diagnostic Accuracy Studies (QUADAS) tool

Item	Yes	No	Unclear
Was the spectrum of patients representative of the patients who will receive the test in practice?	( )	( )	( )
Were selection criteria clearly described?	( )	( )	( )
Is the reference standard likely to correctly classify the target condition?	( )	( )	( )
Is the time period between reference standard and index test short enough to be reasonably sure that the target condition did not change between the two tests?	( )	( )	( )
Did the whole sample of a random selection of the sample, receive verification using a reference standard of diagnosis?	( )	( )	( )
Did patients receive the same reference standard regardless of the index test results?	( )	( )	( )
Was the reference standard independent of the index test (i.e., the index text did not form part of the reference standard)?	( )	( )	( )
Was the execution of the index test described in sufficient detail to permit replication of the test?	( )	( )	( )
Was the execution of the reference standard described in sufficient detail to permit its replication?	( )	( )	( )
Were the index tests results interpreted without knowledge of the results of the reference standard?	( )	( )	( )
Were the reference standard results interpreted without knowledge of the results of the index test?	( )	( )	( )
Were the same clinical data available when test results were interpreted as would be available when the test is used in practice?	( )	( )	( )
Were uninterpretable/intermediate test results reported?	( )	( )	( )
Were withdrawals from the study explained?	( )	( )	( )

A standardized data extraction form was developed to collect the details of all included studies. Each of the two reviewers (HW, VT) independently extracted information from each study. Disagreements were resolved through discussion.

A qualitative synthesis of the results was done. Not enough homogenous data was available to perform a proper meta-analysis since only one of the five reviewed studies reported on sensitivity and specificity data. Sensitivity and subgroup analyses were planned but not done due to lack of data. Potential sources of heterogeneity are timing of index tests, threshold values for positive and negative diagnosis, expertise in the performance and interpretation of the

index tests, expertise in the interpretation of the reference standards. Subgroup analysis was planned for the different type of index test (urine, serum, tissue metabolomics) and the type of metabolite quantification – HPLC/MS vs. NMR. Publication bias was not assessed as the number of studies included was less than eight.

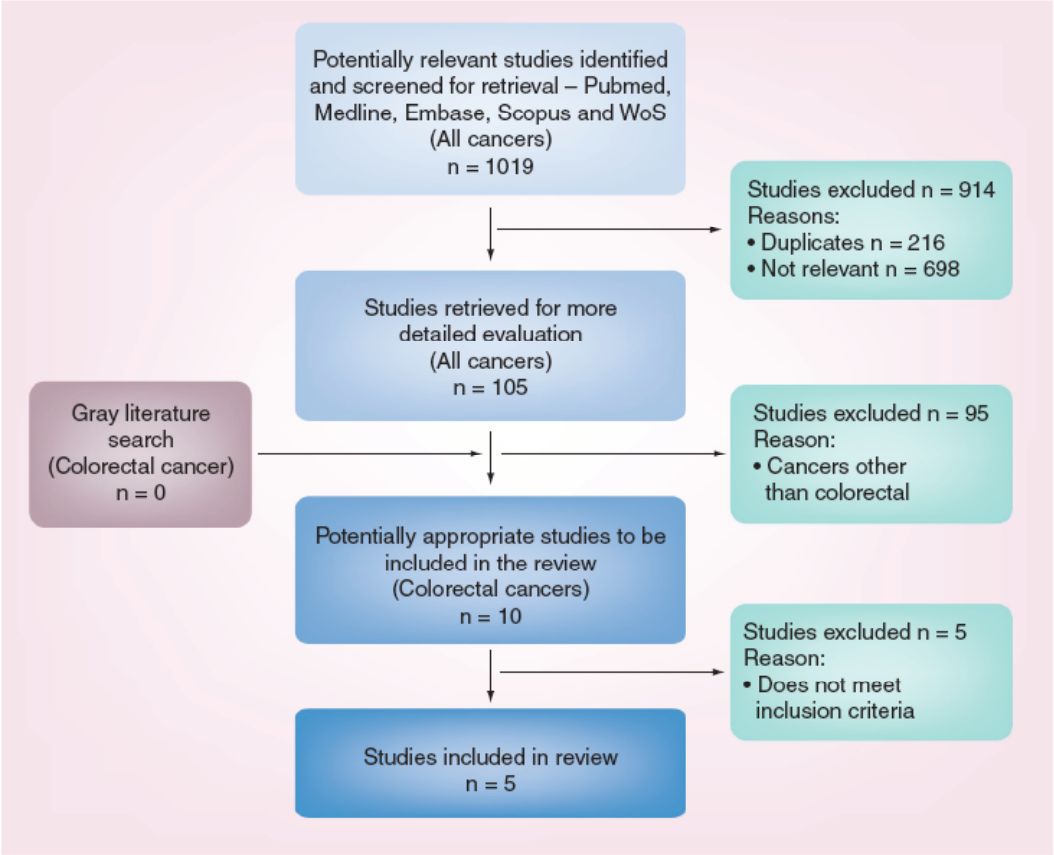
A table of the characteristics of included studies is shown in the Results section.

2.5 Results

2.5.1 Results of the Systematic Search

The electronic database searches resulted in 1019 items. After removing the 216 duplicate hits, 698 were excluded from broad screening of title and abstract because they did not fit the inclusion criteria for the review. Out of the 105 studies that were retrieved for more detailed evaluation, 95 were excluded because they were cancers other than colorectal. A grey literature search did not reveal any additional useful studies. Ten CRC papers were retrieved in full and reviewed for relevance in detail by the two reviewers (HW, VT) independently. Five were excluded for various reasons listed in table 2.2. In the end, five studies are included in this review. Figure 2.1 summarizes the trial flow.

Figure 2.1: Flow of studies through the selection process



2.5.2 Description of Included Studies

There were five studies that satisfied the criteria for inclusion in this review. Table 2.3 lists the characteristics of the included studies[1-5]. All studies were case series published in English consisting of 30 to 84 samples in 15 to 44 patients. Patient demographics were not reported in one study[4] and the setting was not reported in another[5]. Four out of five studies[2-5] used tissue metabolomics as the index test and one[1] used urine. Since one study[2] used both NMR and GC/MS techniques, there were in total three NMR studies[2, 3, 5], and three MS studies[1, 2, 4].

Only one study[4] had sensitivity and specificity results, but all had distinguishing metabolites.

Table 2.2: Characteristics of included studies

Author Year Country	Design	Setting	Age	Patients (n)	Samples (n)	Cancer	Index Test	Analytical technique	Reference standard	Sens Spec	Distinguishing metabolites	PCA/ PLS results
Ma 2008 China	CS	Tertiary	64±10 53-72	33	33	CRC	Urine	UPLC/MS	HP	N/A	Y	Y
Chan 2009 Singapore	CS	Tertiary	67±13	31	63	CRC	Tissue	NMR GC/MS	HP	N/A	Y	Y
Piotto 2008 France	CS	Tertiary	68±12 45-90	44	84	CRC	Tissue	NMR	HP	N/A	Y	Y
Denkert 2008 Germany	CS	Tertiary	?	27	45	Colon	Tissue	GC/MS	HP	95% 95%	Y	Y
Lean 1993 Australia	CS	?	37-82	15	30	CRC	Tissue	NMR	HP	N/A	Y	N

CRC: Colorectal cancer; CS: Case Series; GC: Gas chromatography; HP: Histopathology; MS: Mass spectroscopy; NMR: Nuclear Magnetic Resonance; PCA: Principal component analysis; PLS: Partial least squares; Sens: sensitivity; Spec: specificity; UPLC: Ultra-high performance liquid chromatography.

2.5.3 Summary of Included Studies

Ma et al. 2008[1]

This Chinese study used UPLC/MS to examine the urine samples of 24 CRC patients both before and after their cancer operations, and that of 9 controls. They noted that when compared to the healthy controls, preoperative CRC patients had significantly increased levels of low-molecular weight compounds 283 and 234 ($p<0.05$), and these compounds decreased significantly after the operation. Using PLS-DA analysis, the study demonstrated a clear and significant separation between preoperative, post-operative CRC patients and healthy controls. The authors concluded that once these pilot results are tested in a larger population, a urine test has the potential to identify affected patients. The limitation of this study was that the names of the distinguishing compounds were not identified, only the molecular weights were presented.

Chan et al. 2009[2]

This study from Singapore looked at 31 colon cancer tissue samples from 31 CRC patients and compared them to 32 matched segments of normal mucosa, from the same patient, 5-10cm away from the cancer. This study is unique in that it used both NMR and GC/MS techniques to analyze the samples. There were distinguishing metabolites ($p <0.05$ for all except glycine ($p=0.1751$) and phosphoethanolamine ($p=0.0541$)) identified with each technique, which are listed in Table 4. PLS-DA plots could clearly distinguish cancer versus normal tissues in both NMR and MS, and could even distinguish colon cancer versus rectal

cancer when NMR was used. The limitations of this study were the inadequate clinical information provided for each patient, and not all samples were analyzed using both techniques.

Piotto et al. 2008[3]

A total of 84 tissue samples consisting of cancer and normal mucosa from 44 patients were studied in this French paper, but it was unclear how many were from each group. NMR spectra were obtained from the tissue samples. After removing 12 outliers, PLS-DA model obtained for the remaining 72 samples demonstrated very clear separation for the cancer versus healthy biopsies. Distinguishing metabolites were also stated (statistically significant, but no p-values specified). The authors of the study then proceeded to build a second PLS-DA model consisting of only the distinguishing metabolites using the first 50 biopsies (27 cancer, 23 controls). The remaining 22 samples were subjected to a blind classification process, where visual inspection showed that the cancerous and healthy samples were classified in the correct region. This is the only study where the index test was explicitly stated to be blinded from the results of the reference standard. The limitation of this study, again, was that not enough clinical information about the patients was provided, such as co-morbidities and location of tumor.

Denkert et al. 2008[4]

This German study used GC/MS to examine 45 tissue samples (27 colon cancer, 18 normal). In total, 82 distinguishing metabolites were identified ($p< 0.01$), some not named. This is the only study to state a sensitivity and specificity value for the index test for detection of CRC. Sensitivity and specificity were both approximately 95%. PCA was able to separate CRC from normal tissues. The authors concluded that metabolic signatures, as well as individual metabolites can be detected from fresh-frozen tumor tissue of CRC and that these alterations can be linked to relevant biochemical pathways. Again, minimal clinical information regarding the patients was given.

Lean et al. 1993[5]

This Australian study used NMR tissue metabolomics to study 30 colonic samples of CRC and matched normals in 15 patients. Distinguishing metabolites were listed. Since this is an older study, there was no mention of PLS or PCA plots. The unique part of this study was that it was able to identify 6 samples in the 15 'normal' group as abnormal, indicating that NMR was able to identify abnormal colorectal mucosa, which is not morphologically manifest. A limitation was the small amount of clinical information provided.

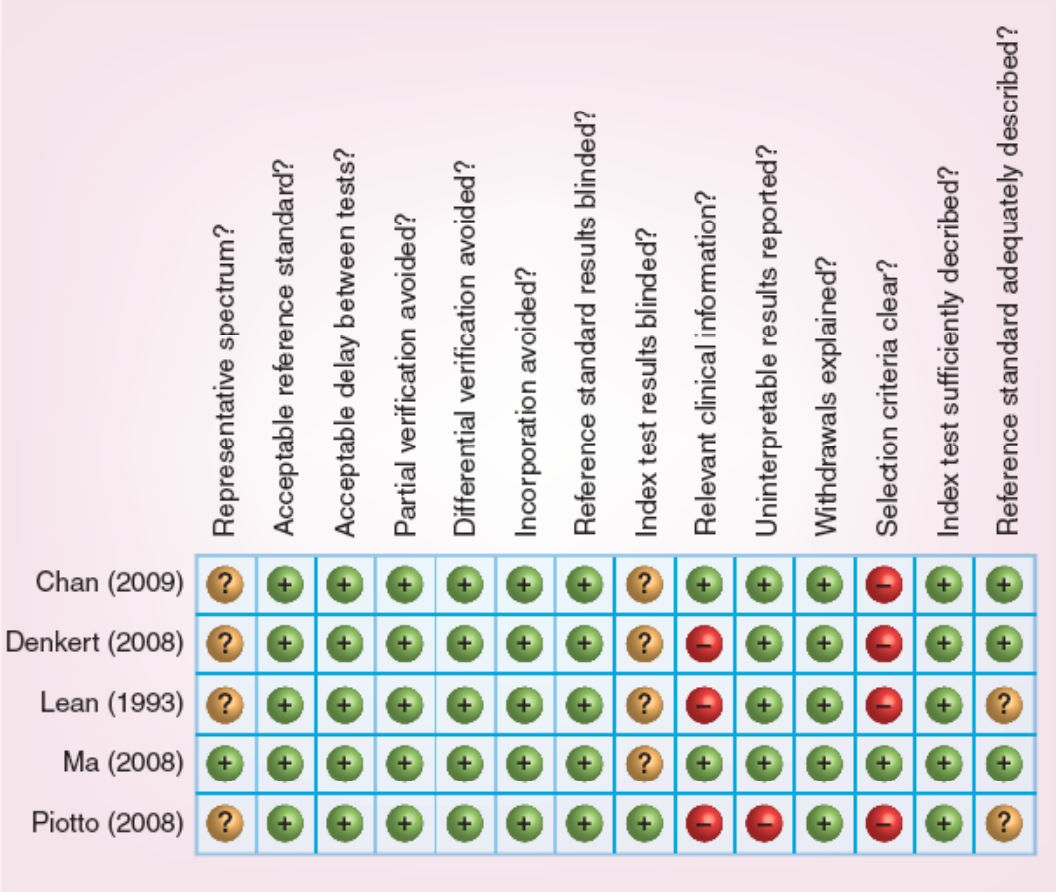
2.5.4 Methodological Quality of Included Studies

The quality assessment results for the individual studies are shown in figure 2.2. All but one study had a clear description of the demographic, clinical features and

the inclusion and exclusion criteria of the population studied, thus the generalizability of the results may be compromised. The reference standard for all the studies is histopathology; currently this is the gold standard for CRC diagnosis. For the tissue studies, the samples used for the index test and that used for the reference standard were obtained at the same time. For the urine study, the urine was obtained from the patients at 7 a.m. on the morning of their surgery. All of the study group received confirmation of the diagnosis by histopathology, that is, the results of the index test did not influence the decision to perform the reference standard, so partial verification bias was avoided. Since all patients received the same reference standard regardless of the results of the index test, differential verification bias was also avoided. Incorporation bias was not applicable as the reference standard was completely independent of the index test. The index tests were all described in sufficient detail to permit replication of the test. The results of the reference standard was interpreted without the knowledge of the result of the index test, however, it was unclear whether the index test results were interpreted without the knowledge of the results of the reference standard in all but one study. Relevant clinical data was not available in three of the five included studies since the location of the CRC was not provided. There were no withdrawals from the studies. Uninterpretable and/or intermediate results were accounted for except in one study[3], where 12 outliers were removed from analysis, without a detailed explanation as to why they were excluded.

Overall, the studies were strong in the technical aspects of test description, but were weak in the amount of clinical information provided and small number of subjects.

Figure 2.2: Risk of bias summary: review authors' judgments about each risk of bias item for each included study. +: Yes (high quality); - : No (low quality); ?: Unclear.



2.5.5 Individual Metabolite Analysis

All five studies listed metabolites that could distinguish CRC patients from controls. The results from each study are summarized in table 2.3. In tissue

metabolomics, the metabolites that appeared in more than one study are denoted with a "†" and those that appeared in more than two studies are denoted with a "‡". The ones that were most often found to be in higher proportion in CRC specimens are taurine, lactate, choline, inositol, glycine, phosphocholine, proline, phenylalanine, alanine, threonine, valine, and leucine. The metabolite most often found to be in higher proportion in normal control specimens was glucose. All recent studies showed that CRC can be clearly distinguished from normal controls on PCA or PLS-DA plots.

2.5.6 Update to Review

As the field of metabolomics is expanding very quickly, an updated literature review done in June 2010 revealed three other papers that would fit the criteria for this review[6-8], two on serum and one on urine. The results of these studies are also summarized in table 2.3.

Qiu *et al*. have recently published two studies, one on serum and one on urinary metabolite profiling of colorectal cancer. Both studies had good separation between CRC patients and healthy controls on orthogonal partial least squares (OPLS) plots but neither had sensitivity and specificity data. In the serum study[6], using gas and liquid chromatography coupled with time-of-flight mass spectrometry (GC-TOFMS) and ultra performance liquid chromatography-quadrupole time-of-flight mass spectrometry (UPLC-QTOFMS), respectively, 33 differential metabolites were identified, five of which were found using both

techniques, namely increased levels of pyruvate and lactate, and decreased levels of tryptophan, tyrosine, and uridine ($p<0.05$) in cancer patients. Oleamide was the most depleted serum metabolite and pyruvate was the metabolite most increased. In the urine study[7], using GC-MS, 16 differential metabolites were identified ($p< 0.05$ unless otherwise specified) including decreased levels of succinate, isocitrate, citrate, 3-methyl-histidine ($p=0.0582$), histidine ($p=0.0601$) and increased levels of 5-hydroxytryptophan, 5-hydroxyindoleacetate, tryptophan, glutamate, 5-oxoproline, N-acetyl-aspartate, p-cresol ($p=0.0961$), 2-hydroxyhippurate, phenylacetate ($p=0.0875$), phenylacetylglutamine, and p-hydroxyphenylacetate in cancer patients. In both of these studies, adequate clinical information was provided, but the index tests were not carried out without the knowledge of the reference standard.

Ritchie *et al.*[8] used Fourier transform ion cyclotron resonance mass spectrometry (FTICR-MS), LC-MS and NMR technologies and found significantly reduced levels of serum hydroxylated polyunsaturated ultra long-chain fatty acids in CRC patients. The authors were also able to validate their results in five independent study populations of CRC patients and controls and reported an average area under the curve (AUC) of 0.91 +/- 0.04, which translates into approximately 75% sensitivity and 90% specificity. Clinical data was provided in this study but it was unclear whether the index test was performed without the knowledge of the reference test.

Table 2.3: Summary of distinguishing metabolites identified in each reviewed study

Study	Technique	Distinguishing metabolites	
		Decreased in CRC	*Increased in CRC*
Tissue metabolites			
Chan (2009)	NMR	Lipids PEG Glucose‡	Choline-containing compounds† Taurine‡ Scyllo-inositol† Glycine† PEG Phosphoethanolamine Lactate‡ Phosphocholine†
	MS	Fumarate Malate D-mannose D-galactose D-glucose‡ 1-hexadecanol Arachidonic acid	Lactate‡ Phosphate L-glycine† 2-hydroxy-3-methylvalerate L-proline† L-phenylalanine† Fatty acids Uridine 11,14-eicosadienoic acid 11-eicosenoic acid 1-O-heptadecylglycerol 1-monooleoylglycerol Propyl octadecanoate Cholesterol
Piotto (2008)	NMR	Myo-inositol β-glucose‡	Taurine‡ Glutamate Aspartate Lactate‡
Denkert (2008)	GC/MS	Oleic acid N-acetylglycine Inositol stereoisomer Galactonate γ-lactone	Alanine† Methionine Hypoxanthine Cysteine Proline† Phenylalanine† Threonine† Uracil Isoleucine Leucine† Valine†
Lean (1993)	NMR		Choline† Phosphoryl-choline† Glycerol phosphoryl-choline Inositol† Taurine‡ Fucose Alanine† Glutamic acid/glutathione Histidine Leucine† Lysine Threonine† Valine†

†Metabolites that appeared in more than one tissue metabolomic study.
‡Metabolites that appeared in three different tissue metabolomic studies.
CRC: Colorectal cancer; FTICR: Fourier transform ion cyclotron resonance mass spectrometry; GC: Gas chromatography; LC: Liquid chromatography; LMW: Low molecular weight; MS: Mass spectroscopy; NMR: Nuclear magnetic resonance; PEG: Polyethylene glycol; QTOFMS: Quadrupole time-of-flight mass spectrometry; TOFMS: Time-of-flight mass spectrometry; UPLC: Ultraperformance liquid chromatography.

Table 2.3: Summary of distinguishing metabolites identified in each reviewed study. (Con't)

Study	Technique	Distinguishing metabolites	
		Decreased in CRC	*Increased in CRC*
Serum metabolomics			
Qiu (2009)	GC-TOFMS	Urea Valine Leucine Proline Threonine Threonic acid 4-hydroxyproline Citrulline 2-piperidinecarboxylic acid Ornithine Hippurate Lysine Tyrosine Tryptophan Oleamide Uridine	Pyruvate Lactate 2-hydroxybutanoic acid 3-hydroxybutanoic acid Malic acid Oleic acid
	UPLC-QTOFMS	Tyrosine Uridine Phenylalanine Tryptophan Myristic acid Palmitic acid Nervonic acid Arginine Glutamic acid Nicotinamide Dopamine	Glycerol phosphate Pyruvic acid Lactate Carnitine
Ritchie (2010)	FTICR-MS, LC-MS NMR	Hydroxylated polyunsaturated ultra long-chain fatty acids	
Urine metabolomics			
Ma (2008)	UPLC/MS		LMW 283 LMW 294
Qiu (2010)	GC-MS	Succinate Isocitrate Citrate 3-methyl-histidine Histidine	5-hydroxytryptophan 5-hydroxyindoleacetate Tryptophan Glutamate 5-oxoproline N-acetyl-aspartate p-cresol 2-hydroxyhippurate Phenylacetate Phenylacetylglutamine p-hydroxyphenylacetate

[†]Metabolites that appeared in more than one tissue metabolomic study.
[‡]Metabolites that appeared in three different tissue metabolomic studies.
CRC: Colorectal cancer; FTICR: Fourier transform ion cyclotron resonance mass spectrometry; GC: Gas chromatography; LC: Liquid chromatography; LMW: Low molecular weight; MS: Mass spectroscopy; NMR: Nuclear magnetic resonance; PEG: Polyethylene glycol; QTOFMS: Quadrupole time-of-flight mass spectrometry; TOFMS: Time-of-flight mass spectrometry; UPLC: Ultraperformance liquid chromatography.

2.6 Discussion

2.6.1 Summary of the Human Metabolomic Studies in CRC

A comprehensive literature search for studies related to urine, serum, and tissue metabolomics and the detection of primary colorectal cancer in adult humans was

completed. Up to June 2010 there were eight studies that met the inclusion/exclusion criteria established for this review. In total, there were two urine, two serum and four tissue metabolomic studies reviewed.

The main objective of this review was to summarize the diagnostic accuracies of serum, urine, and tissue metabolomics for detecting CRC in the adult population, using histopathology as the gold standard. However, since only two studies had sensitivity and specificity reported, this objective was not satisfactorily achieved.

Instead, the existing publications focused on the differentiation of CRC specimens from normal tissue based on the differences in metabolites detected. Therefore, we were able to achieve our secondary objective in summarizing the most prevalent metabolites found in CRC. In tissue, the two most distinguishing metabolites for CRC are increased levels of taurine and lactate, while secondarily important metabolites appear to be increased levels of choline, inositol, glycine, phosphocholine, proline, phenylalanine, alanine, threonine, valine, and leucine, and decreased glucose. In serum, the studies so far have shown increased levels of pyruvate and lactate, and decreased levels of tryptophan, tyrosine, uridine, and hydoxylated, polyunsaturated ultra long-chain fatty acids in patients with CRC. In urine, increased levels of 5-hydroxytryptophan, 5-hydroxyindoleacetate, tryptophan, glutamate, 5-oxoproline, N-acetyl-aspartate, p-cresol, 2-hydroxyhippurate, phenylacetate, phenylacetylglutamine, p-hydroxyphenylacetate, low molecular weight compounds 283 and 294 and

decreased levels of succinate, isocitrate, citrate, 3-methyl-histidine, and histidine are found in patients with CRC.

While the aforementioned studies focused on metabolomics in urine, tissue, or serum, recent publications have shown promising results with fecal water extracts[19, 20]. One half of the studies included in this review used colonic tissue to determine metabolomics. While tissue metabolomics provides promising results, urine metabolomics represents a much less invasive method of testing compared to tissue or serum metabolomics and would be a much easier platform to use for a CRC screening tool. Besides, tissue metabolomics does not offer any obvious advantages over histological analysis. So far, only two studies[1, 7] utilized urine as the platform for study. Not only was urine metabolomics able to show the differences between cancer and healthy subjects, both studies that utilized urine metabolomics were also able to demonstrate a change in the metabolic profiling after the cancer was surgically removed.

Taken together, the results of the various differential metabolites identified in these studies provide a number of speculated alterations in biochemical pathways in CRC. In tissue-based metabolomics, there is an up-regulation of amino acids likely reflecting cellular needs for higher turnover of structural proteins.[4] The higher level of uridine in CRC is thought to be associated with the higher propagation rate of the tumor cells. Lower levels of malate and fumarate in CRC are thought to be related to the higher metabolic rate of the tumors. Finally,

decreased levels of lipids in CRC specimens is perhaps due to increased utilization of lipids from increased membrane biosynthesis for cell propagation.[2] Lean et al. stated that the higher levels of choline and phosphocholine seen in CRC could be accounted for with increased phospholipid synthesis (necessary for membrane turnover) in tumor cells.[5] In serum-based metabolomics, the increase in pyruvate and lactate levels in CRC patients is reflective of altered glycolysis. Alterations in arginine and proline metabolism, fatty acid metabolism and oleamide metabolism are thought to account for the other differential metabolites.[6] In urine-based metabolomics, down-regulation of the tricarboxylic acid (TCA) cycle, up-regulation of tryptophan metabolism and altered gut microflora metabolism is suggested in patients with CRC. Abnormal glutamate and histamine metabolism may also play a role in CRC.[7]

There are a few metabolites that are particularly intriguing. Tryptophan was decreased in the serum[6] and increased in the urine[7] of patients with CRC, whereas uridine, lysine, proline, and threonine were decreased in serum[6] and increased in tissue[2, 5] and histidine was decreased in urine[7] but increased in tissue[5] metabolomics of patients with CRC. These findings suggest that serum, urine and tissue metabolism are intimately correlated and that factors such as renal or hepatic clearance may play a role in the metabolites identifiable by various testing processes. This implies that, in order to secure a complete picture of the metabolome and identify the alterations in the biochemical pathways of a specific

condition, one needs to perform serum, urine, and tissue metabolomics of the same patient and controls simultaneously, using the same techniques.

While there are differences in opinion as to the best analytical technique used to study metabolomics, several recent reviews have noted that a combination of techniques may be needed to provide a complete metabolome picture.[21-23] Nevertheless, it has been demonstrated that both NMR and MS are able to yield PCA/PLS plots that can clearly distinguish CRC from normal controls. In 2009, Chan observed that NMR was superior to MS in distinguishing colon cancer from rectal cancer[2]. NMR is also faster to do, whereas mass spectroscopy requires chromatography (liquid or gas) to separate out the metabolites first, which is an additional expensive and time-consuming step.

2.6.2 Strengths and Weaknesses of the Review

This systematic review is the first to summarize existing published literature on the role of metabolomics in detecting CRC in humans. While there are additional publications on breast[24-26] and prostate[27-29] cancers, there are only few publications on CRC. In fact, the majority of the papers included in this review were published within the last 2 years. In completing this review, we were able to highlight those metabolites that are currently found to be increased and/or decreased in patients with CRC compared to those patients without CRC. Nevertheless, it remains too early to project the true diagnostic or prognostic accuracies of metabolomics in CRC.

Given the lack of homogenous data, a meta-analysis of the diagnostic accuracies and summary diagnostics could not be completed. Investigation of heterogeneity and publication bias was also not assessed owing to insufficient data. The quality of studies was very good in describing details of the index tests, and the more recent studies provided sufficient clinical data. However, for the most part, it was unclear whether the index tests were interpreted without the knowledge of the result from the reference standard. This is a very important bias to address in studies of diagnostic and prognostic tests.

2.6.3 Applicability of Findings to Clinical Practice and Policy

This review represents the first step in determining whether there is a role for using metabolomics in the diagnosis and prognosis of CRC, and eventually as a pre-cancerous screening tool to detect adenomatous polyps as the precursor to the development of CRC. Metabolite identification unique to CRC will help scientists to be more focused and accurate in testing specimens in the future. Once the metabolomic "fingerprint" of CRC is firmly established, the next step would be to test the accuracy of this "fingerprint" and these metabolites in a prospective blinded study against the reference standard.[9] Other factors that could influence metabolomics such as diet, drugs, stress, microbiota etc. all need to be explored as well. Success with metabolomics as a diagnostic and prognostic tool is likely to fundamentally change the physicians' approach to health care.

2.7 Conclusions

Clinically, metabolomics has the potential to become a tool for diagnosing CRC and as an extension urine or serum metabolomics may represent a new, and less invasive method of screening for CRC. The low adherence to current CRC screening guidelines[30-32] is multi-factorial, but having a less invasive and more accessible test for screening will certainly improve compliance and improve public health through primary and secondary prevention. While it is encouraging to see the growth of metabolomics in colorectal oncology, appropriately powered, blinded, prospective and clinically validated serum and urine metabolomic studies are needed.

A version of this chapter (2.0) has been published.
Wang, Tso, Slupsky, Fedorak 2010. Future Oncology. 2010 6(9),1395-1406.

2.8 References

1. Ma Y-L, Qin H-L, Liu W-J, Peng J-Y, Huang L, Zhao X-P, Cheng Y-Y. Ultra-High Performance Liquid Chromatography-Mass Spectrometry for the Metabolomic Analysis of Urine in Colorectal Cancer. Digestive Diseases and Sciences 2009;54:2655-2662.
2. Chan ECY, Koh PK, Mal M, Cheah PY, Eu KW, Backshall A, Cavill R, Nicholson JK, Keun HC. Metabolic Profiling of Human Colorectal Cancer Using High-Resolution Magic Angle Spinning Nuclear Magnetic Resonance (HR-MAS NMR) Spectroscopy and Gas Chromatography Mass Spectrometry (GC/MS). Journal of Proteome Research 2009;8:352-361.
3. Piotto M, Moussallieh FM, Dillmann B, Imperiale A, Neuville A, Brigand C, Bellocq JP, Elbayed K, Namer IJ. Metabolic characterization of primary human colorectal cancers using high resolution magic angle spinning H-1 magnetic resonance spectroscopy. Metabolomics 2009;5:292-301.
4. Denkert C, Budczies J, Weichert W, Wohlgemuth G, Scholz M, Kind T, Niesporek S, Noske A, Buckendahl A, Dietel M, Fiehn O. Metabolite profiling of human colon carcinoma - deregulation of TCA cycle and amino acid turnover. Molecular Cancer 2008;7.
5. Lean CL, Newland RC, Ende DA, Bokey EL, Smith ICP, Mountford CE. ASSESSMENT OF HUMAN COLORECTAL BIOPSIES BY H-1 MRS - CORRELATION WITH HISTOPATHOLOGY. Magnetic Resonance in Medicine 1993;30:525-533.
6. Qiu YP, Cai GX, Su MM, Chen TL, Zheng XJ, Xu Y, Ni Y, Zhao AH, Xu LX, Cai SJ, Jia W. Serum Metabolite Profiling of Human Colorectal Cancer Using GC-TOFMS and UPLC-QTOFMS. Journal of Proteome Research 2009;8:4844-4850.
7. Qiu YP, Cai GX, Su MM, Chen TL, Liu YM, Xu Y, Ni Y, Zhao AH, Cai SJ, Xu LX, Jia W. Urinary Metabonomic Study on Colorectal Cancer. Journal of Proteome Research 2010;9:1627-1634.
8. Ritchie SA, Ahiahonu PWK, Jayasinghe D, Heath D, Liu J, Lu YS, Jin W, Kavianpour A, Yamazaki Y, Khan AM, Hossain M, Su-Myat KK, Wood PL, Krenitsky K, Takemasa I, Miyake M, Sekimoto M, Monden M, Matsubara H, Nomura F, Goodenowe DB. Reduced levels of hydroxylated, polyunsaturated ultra long-chain fatty acids in the serum of colorectal cancer patients: implications for early screening and detection. Bmc Medicine 2010;8.
9. Wang H, Slupsky CM, Wong CK, Schiller DE, Fedorak RN. Is there a role for using NMR urine metabolomics as a new method of screening for colorectal cancer? Volume 23 (Suppl. A): Canadian Journal of Gastroenterology, 2009.
10. Claudino WM, Quattrone A, Biganzoli L, Pestrin M, Bertini I, Di Leo A. Metabolomics: Available results, current research projects in breast cancer, and future applications. Journal of Clinical Oncology 2007;25:2840-2846.
11. Goodacre R, Vaidyanathan S, Dunn WB, Harrigan GG, Kell DB. Metabolomics by numbers: acquiring and understanding global metabolite data. Trends in Biotechnology 2004;22:245-252.
12. Kaddurah-Daouk R, McEvoy J, Baillie RA, Lee D, Yao JK, Doraiswamy PM, Krishnan KRR. Metabolomic mapping of atypical antipsychotic effects in schizophrenia. Molecular Psychiatry 2007;12:934-945.
13. Broadhurst DI, Brown M, Dunn WB, Carroll KB, Redman CW, Baker PN, Kell DB, Kenny LC. The application of UPLC/LTQ-Orbitrap mass Spectrometry to human metabolite biomarker discovery. A case study: Pre-eclampsia. Reproductive Sciences 2008;15:524.
14. Siffedeen JS, Rankin KN, Fu H, Dieleman LA, Slupsky C, Fedorak R. Metabolomics in inflammatory bowel disease: Differentiating patients with or without IBD. Gastroenterology 2008;134:A203-A203.
15. Kind T, Tolstikov V, Fiehn O, Weiss RH. A comprehensive urinary metabolomic approach for identifying kidney cancer. Analytical Biochemistry 2007;363:185-195.

16. Kell DB. Metabolomic biomarkers: search, discovery and validation. Expert Rev Mol Diagn 2007;7:329-33.
17. Kaddurah-Daouk R, Kristal B, Weinshilboum R. Metabolomics: A global biochemical approach to drug response and disease. Annual Review of Pharmacology and Toxicology 2008:653-683.
18. Whiting P, Rutjes AWS, Reitsma JB, Bossuyt PMM, Kleijnen J. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. BMC Med Res Methodol 2003;3:25.
19. Monleon D, Morales JM, Barrasa A, Lopez JA, Vazquez C, Celda B. Metabolite profiling of fecal water extracts from human colorectal cancer. Nmr in Biomedicine 2009;22:342-348.
20. Bezabeh T, Somorjai R, Dolenko B, Bryskina N, Levin B, Bernstein CN, Jeyarajah E, Steinhart AH, Rubin DT, Smith ICP. Detecting colorectal cancer by H-1 magnetic resonance spectroscopy of fecal extracts. Nmr in Biomedicine 2009;22:593-600.
21. Dunn WB, Ellis DI. Metabolomics: Current analytical platforms and methodologies. Trac-Trends in Analytical Chemistry 2005;24:285-294.
22. Lenz EM, Wilson ID. Analytical strategies in metabonomics. Journal of Proteome Research 2007;6:443-458.
23. Robertson DG. Metabonomics in toxicology: A review. Toxicological Sciences 2005;85:809-822.
24. Bathen TF, Jensen LR, Sitter B, Fjoesne HE, Halgunset J, Axelson DE, Gribbestad IS, Lundgren S. MR-determined metabolic phenotype of breast cancer in prediction of lymphatic spread, grade, and hormone status. Breast Cancer Research and Treatment 2007;104:181-189.
25. Bullinger D, Frohlich H, Klaus F, Neubauer H, Frickenschmidt A, Henneges C, Zell A, Laufer S, Gleiter CH, Liebich H, Kammerer B. Bioinformatical evaluation of modified nucleosides as biomedical markers in diagnosis of breast cancer. Analytica Chimica Acta 2008;618:29-34.
26. Frickenschmidt A, Frohlich H, Bullinger D, Zell A, Laufer S, Gleiter CH, Liebich H, Kammerer B. Metabonomics in cancer diagnosis: mass spectrometry-based profiling of urinary nucleosides from breast cancer patients. Biomarkers 2008;13:435-449.
27. Whitehead TL, Kieber-Emmons T. Applying in vitro NMR spectroscopy and H-1 NMR metabonomics to breast cancer characterization and detection. Progress in Nuclear Magnetic Resonance Spectroscopy 2005;47:165-174.
28. Burns MA, He WL, Wu CL, Cheng LL. Quantitative pathology in tissue MR spectroscopy based human prostate metabolomics. Technology in Cancer Research & Treatment 2004;3:591-598.
29. Cheng LL, Burns MA, Taylor JL, He WL, Halpern EF, McDougal WS, Wu CL. Metabolic characterization of human prostate cancer with tissue magnetic resonance spectroscopy. Cancer Research 2005;65:3030-3034.
30. Sewitch MJ, Fournier C, Ciampi A, Dyachenko A. Adherence to colorectal cancer screening guidelines in Canada. Bmc Gastroenterology 2007;7.
31. Rabeneck L, Paszat LF. A population-based estimate of the extent of colorectal cancer screening in Ontario. American Journal of Gastroenterology 2004;99:1141-1144.
32. McGregor SE, Hilsden RJ, Li FX, Bryant HE, Murray A. Low uptake of colorectal cancer screening 3 yr after release of national recommendations for screening. American Journal of Gastroenterology 2007;102:1727-1735.

3.0 Methodology

This chapter gives an overview of the methods used for this study including recruitment, samples handling, NMR and analysis. Specific details such as number of patients for the study are stated in subsequent chapters.

3.1 Recruitment and Urine Sample Collection

Normal, Adenoma, Hyperplastic Polyp Subjects: Study subjects who had normal findings, adenomatous polyps, or hyperplastic polyps on their screening colonoscopy were recruited from the SCOPE (Stop COlorectal cancer through Prevention and Education) pilot study – a population based study of over 1000 asymptomatic patients who are at average or high risk for colorectal cancer. The SCOPE pilot study was carried out between April 2008 and October 2009.

The SCOPE study population consisted of patients on wait lists referred for elective colonoscopy for screening purposes to the gastroenterologists participating in the SCOPE program. Once identified, the patients received a phone call from the SCOPE pilot nurse navigator. The nurse reviewed the inclusion and exclusion criteria, and outlined the SCOPE Pilot and its objectives. After patient questions were answered, the nurse requested verbal consent to participate in the SCOPE Pilot study. Formal written consent was also obtained at a later date to ensure participants understood the colonoscopy procedure and the risks associated with the procedure, including the risk of perforation, bleeding, infection, subsequent surgery, missed diagnosis of adenoma or cancer, and

reactions to the medications used. Participants also had the opportunity to ask questions about their participation in the Pilot with a registered nurse at the information session and about the colonoscopy with the gastroenterologist prior to the procedure.

Both average risk individuals and those with a family history of colon cancer or polyps were included. Inclusion criteria for average risk were as follows: asymptomatic, 50-75 years of age and no personal or family history of colorectal cancer or polyps. Participants at increased risk for colorectal cancer were included if they were 40-75 years of age with known personal or significant family history of colorectal cancer or polyps. Potential study participants were excluded if they were under 40 or over 75 years of age, unable to understand or sign the informed consent, or had a recent history of visible hematochezia or inflammatory bowel disease. Participants with significant co-morbidities were also excluded.[1]

Subjects participating in the SCOPE pilot study were each given a unique study identification number when they attended the information session given by the study nurses on colon cancer screening. The midstream urine sample was collected at the end of the education session, in the mid-afternoon; subjects were in their normal state, i.e. they did not have any diet modifications. In addition to providing a urine sample, all patients completed a medical questionnaire, had a FOBT (fecal occult blood test) and FIT (fecal immune test), prior to their colonoscopy as part of the SCOPE study. The study urine sample container was

pre-filled with 6 drops of dried sodium azide (27.3mg/mL) to prevent any bacterial growth in the urine while it was waiting to be frozen.

Cancer Patients: All newly diagnosed colorectal cancer patients presenting to general surgeons' offices or the endoscopy suite in Edmonton and Grande Prairie between Oct 2008 and June 2010 were screened for this study. Eligible patients were identified by the general surgeon or gastroenterologist as those who have not had any treatment for the colorectal cancer. The potential participant had an opportunity to review the informed consent and ask questions. Those that chose to participate signed the consent form and were assigned unique study identification numbers. The study medical questionnaire was filled out either in the surgeons' offices or in the endoscopy suite. A urine sample was collected from each patient *before* his/her surgery or neoadjuvant treatment in the pre-operative admission clinic (normal state) or in the endoscopy suite (may be fasting) using the study container pre-coated with sodium azide. Patients were excluded if they had already undergone medical (chemotherapy), radiation, or definitive surgical treatment for the CRC.

Post-op Patients: Those CRC patients that had curative treatment (surgery, or surgery with adjuvant treatment) were contacted again between 3 months to 1 year after their treatment. They were asked to provide another urine sample in their normal state. This sample was analyzed and compared to their pre-treatment sample to see if their metabolomic fingerprint for colorectal cancer disappeared after treatment.

In all cases, subjects were also excluded if they were anuric, oliguric, had end-stage renal failure, or were on hemo or peritoneal dialysis. Ethics approval for this study was obtained from the Health Research Ethics Board at the University of Alberta.

Urine was chosen as the biofluid for this trial on the basis of existing literature[2, 3] suggesting that metabolomic profile of CRC can be identified through urine. A urine sample is less invasive and more patient-friendly to obtain than a blood and stool sample, respectively. These two qualities alone would make urine a more compliant test and as stated previously, increasing patient compliance is a very important factor in improving screening rates.

3.2 Nuclear Magnetic Resonance (NMR)

As mentioned in the chapter 2, NMR and mass spectrometry (MS) are the two most used methods of quantifying the metabolites in the field of metabolomics. NMR was chosen for this project since the equipment is readily accessible, the expertise is available, and the process is faster than MS as it does not require the chromatography step. NMR is not only rapid, but requires minimal or no sample preparation, is non-destructive, robust, reproducible, quantitative, nonselective and cost-effective, however, it is not as sensitive as MS, which can also detect a wider range of metabolites.[4] The National High Field Nuclear Magnetic Resonance Centre (NANUC) is located on the University of Alberta campus and

is home to Varian 500Hz, 600Hz, and 800Hz NMR spectrometers and is equipped with highly skilled and knowledgeable personnel. For this project, the 600 Hz spectrometer was used since it was the most cost-effective. In June 2009, NANUC acquired a Varian 768 AS sample handling robot, so since that time, the urine samples were run with this robot (*i.e.* automated mode).

Generating the Spectra

The basis of NMR is that many atomic nuclei have an intrinsic spin (or angular momentum). When placed in a magnetic field, these nuclei absorb energy from any applied electromagnetic pulse and radiate this energy back out. The NMR apparatus is a series of coils that produces a static magnetic field and smaller induced excitation fields. The receiver picks up the induced voltage from the precessing spins and this generates a signal that decays with time, or a FID (free induction decay). Fourier transformation of the FID generates a more familiar spectrum. The width of the peak is related to the rate of decay of the signal – the faster the decay, the broader the peak.[5, 6]

Shimming

Signals decay faster when the magnetic field is inhomogeneous, *i.e.* the field varies from place to place across the sample, so atomic nuclei in different positions may resonate at different frequencies, such that the individual spins spread out in different directions and the total signal is therefore smaller, and the peak is broader. Shimming is used to adjust and eliminate the inhomogeneities in

a magnetic field and thus increase the resolution of the signal and increase the signal to noise ratio. Active shimming uses coils with adjustable current and passive shimming uses steel pieces that get magnetized from the permanent or superconducting magnet. The additional magnetic fields from the coil or effects of the steel add to the overall magnetic filed of the superconducting magnet in such a way that the total field becomes more homogenous. [5] In our experiments, active shimming is used.

Chemical shift

The total magnetic field experienced by a nucleus is affected by its electronic environment, *i.e.* neighbouring atoms, bond lengths, and angles between bonds, and this is reflected in the spin energy levels (and resonance frequencies). The variations of NMR frequencies of the same kind of nucleus, due to variations in the electron distribution, is called chemical shift.[5, 6] Chemical shift is the reason why we can identify metabolites based on the resonance frequencies of nuclei.

Water Suppression

Biological samples are in aqueous solution and therefore the signal from hydrogen atoms in water (55M) is on the order of 10,000 times greater than the metabolite resonances (<= 1mM). Suppression of the solvent signal is therefore desirable to obtain information regarding the chemical makeup of the sample. The transmitter offset (tof) is the exact position of the carrier frequency relative to the lock frequency. The correct tof gives the lowest intensity of the water peak. The

saturation frequency is the position of the water saturation pulse.[7] The tof and saturation frequency are set to the same value in our experiments.

Pulse Width, Gain, Sweep Width

Pulse width represents the amount of energy used to excite the nucleus from the resting state into the excited state. This needs to be optimized so that the nucleus is maximally excited but not too much so that it ends back in the resting state.[8]

Gain is the amplification of the signal. If the gain is set too high, then the most intense signal in the spectrum can overload the electronics necessary to observe the signal. Too low a gain setting can mean that weak signals are not sufficiently amplified and get lost in the noise.

Sweep width is the range of frequencies observed in a given spectrum. For Varian spectrometers a sweep width of 10,000 Hz would mean that we observe signals ±5000 Hz from the carrier position.[8]

3.3 Urine Processing & Preparation

3.3.1 Urine storage & processing

The urine samples were labeled with a four-digit study identification number and no patient information, thus the group assignment remained blinded for the sample processing and analyzing steps. All urine samples were frozen at -80°C within 24 hours of collection and if the samples could not be frozen immediately,

they were stored at 4°C within 4 hours of collection. On the day of processing, the samples were thawed and divided into four x 1mL aliquots. After 50 μL of sodium azide (27.3mg/mL) was added to each sample to prevent bacterial growth, the samples were frozen again at -80°C until the day prior to NMR acquisition.

3.3.2 Sample Preparation – day before NMR

The day prior to NMR acquisition, the urine samples were thawed at room temperature in the biohood. For the non-automated (manual) NMR acquisition, 585μL of each sample was diluted (1:10) with 65μL of internal standard consisting of 5 mM sodium 2,2-dimethyl-2-silapentane-5-sulfonate (DSS), 100 mM imidazole, 0.2% sodium azide in 99% D2O (Chenomx Inc., Edmonton, AB) to achieve a total volume of 650μL and stored at 4°C. For the automated (robotic) NMR acquisition, 675μL of each sample was diluted (1:10) with 75μL of the same Chenomx internal standard to achieve a total volume of 750μL and stored at 4°C.

3.3.3 Sample Preparation – day of NMR

On the day of NMR acquisition, the pH of each sample was measured. Various concentrations of HCl and NaOH were added to the samples to achieve a pH between 6.7 and 6.8 to minimize chemical exchange as the chemical shift will change with pH. For the non-automated samples, an aliquot of 600μL of the samples were placed in 5 mm NMR tubes and capped; for the automated samples, 700μL were used.

3.4 NMR Acquisition

3.4.1 Manual/Non-automated Mode

One-dimensional nuclear magnetic resonance spectra were acquired using an Oxford 600Hz NMR spectrometer with a Varian VNMRS two channel console and running VNMRJ software version 2.2C on a RHEL 4 host computer in the NANUC. Before samples were inserted into the spectrometer, the outside of the tubes were cleaned with ethanol and Kimwipes® to remove any debris or oils from handling. Samples (600 uL) were set to a depth of 66 mm in the depth gauge and then inserted into the spectrometer. All samples were run at a sweep width (sw) of 7225.43 Hz and a gain of 18. The saturation frequency (sfrq), transmitter offset (tof) and pulse width (pw) were all individually calibrated at the start of each day. The tof typically ranged from (-213 to -215 Hz) and the pw ranged from 6 to 8 μs. Shims were optimized until an acceptable line width value was obtained at relative peak heights of: 50% (< 1.0 Hz), 0.55% (< 12.0 Hz), and 0.11% (< 20.0 Hz) were achieved. Finally, during post-processing, zero filling was used to increase the actual acquired data points to the next largest factor of 2, and no weighting functions were applied.

We utilized the first increment of a 2D-^{1}H, ^{1}H-NOESY pulse sequence for the acquisition of 1H-NMR data and for suppressing the solvent signal. Experiments contained a 100 ms mixing time along with a 990 ms pre-saturation (~80 Hz gammaB_1). Spectra were collected at 25°C through a total of 32 scans over a

period of 3.5 min; a total recycle delay of 5 s was also used (*i.e.* 1 s recovery delay/saturation and a 4 s acquisition).

3.4.2 Automated/Robotic Mode

Automated runs followed exactly the same experimental parameters used in the manual mode with the exception of an additional 30 s of equilibration time in the NMR to allow the sample to equilibrate to 25°C. All sample handling was done with a Varian 768 AS sample handling robot. The first sample of the batch was manually shimmed to satisfactory line width values and subsequent samples were automatically shimmed. Any spectra that did not meet acceptable line height values were discarded and the sample was re run.

3.5 Post NMR Acquisition

3.5.1 pH Recheck

After the spectra were obtained, the samples were removed from NMR tubes with glass Pasteur pipettes and transferred into eppendorf tubes. The pH of each sample was then rechecked to ensure that the pH had not shifted a significant amount. This was recorded and could be referenced if a particular sample produced an unexpected spectrum. Samples were stored in the -80°C freezer.

3.5.2 Cleaning the NMR Tubes

The NMR tubes were first filled with bleach followed by soapy water, alcoholic KOH (120 g/L) and concentrated HCl (360 g/L). Between each wash solution,

the tubes were rinsed out five times with double distilled H_2O. After the tubes were clean, they were inverted on an NMR rack and allowed to air dry for at least 48 hours prior to the next use.

3.6 Summary of Sampling Methods

- Prepare urine containers coated with 6 drops of sodium azide (27.3mg/mL)
- Collect midstream urine samples from desired study populations
- Store urine sample in -80°C freezer within 24 hours of collection; if the samples cannot be frozen immediately, then store at 4°C within 4 hours of collection
- Sample processing
    - Thaw samples and aliquot four x 1mL samples into eppendorf tubes along with 50 μL of sodium azide (27.3mg/mL) and freeze at -80°C
- Sample preparation
    - Day prior to NMR
        - Thaw samples and take 585μL of each sample and dilute (1:10) with 65μL of internal standard consisting of 5 mM sodium 2,2-dimethyl-2-silapentane-5-sulfonate (DSS), 100 mM imidazole, 0.2% sodium azide in 99% D2O (Chenomx Inc., Edmonton, AB) to achieve a total volume of 650μL (manual NMR acquisition mode) *OR* 675μL of each sample and dilute (1:10) with 75μL of the Chenomx internal standard to achieve a total volume of 750μL (automated NMR acquisition mode)

        - Store at 4°C overnight
    - Day of NMR
        - pH each sample and add HCl or NaOH to achieve pH between 6.7 an d 6.8
        - Aliquot 600μL (manual mode) *OR* 700μL (automated mode) of the urine samples into 5 mm NMR tubes and cap
- NMR Acquisition
    - Oxford 600Hz NMR spectrometer with a Varian VNMRS two channel console and running VNMRJ software version 2.2C on a RHEL 4 host computer
    - Calibrate saturation frequency(-213 to -215 Hz), transmitter offset (-213 to -215 Hz) and pulse width (6 to 8 μs) at the start of each day
    - Sweep width = 7225.43 Hz and a gain =18
    - Clean NMR tubes with Kimwipes®
    - Set depth gauge to 66mm and insert samples into NMR magnet
    - Optimize shims at relative peak heights of: 50% (< 1.0 Hz), 0.55% (< 12.0 Hz), and 0.11% (< 20.0 Hz)
- Post NMR Acquisition
    - Recheck sample pH and record
    - Clean NMR tubes with bleach, soapy water, alcoholic KOH (120 g/L) and concentrated HCl (360 g/L); between each wash solution, rinse tubes five times with double distilled water

3.7 Metabolite Analysis

3.7.1 Metabolite quantification

Once the spectra were acquired, quantification of metabolites was done using the targeted profiling technique as implemented in Chenomx NMRSuite v7.0 (Chenomx, Inc. Edmonton, Canada), which compares the integral of a known reference signal (in this case DSS) with signals derived from a library of compounds to determine metabolite concentration relative to the reference signal. The quantification process was done by one individual and verified by a second individual to optimize accuracy. The spectral analyses were also spot checked by a third individual. Over 240 metabolites were considered and 72 were found to be significant, that is, the spectral peaks of 72 metabolites in the compound library were identified in the spectra of the study samples.

3.7.2 Normalization

Since hydration states of individuals can be different, the measured metabolite concentrations were normalized to account for the different dilutions of the urine samples. Traditionally, creatinine-normalization is done[9, 10]. However by doing this, creatinine is eliminated from the list of metabolites that could potentially contribute to the separation of normal vs. cancer/adenoma in the multivariate analysis. In fact, when the raw metabolite concentrations were used to generate an orthogonal partial least squares (OPLS) or partial least squares-discriminant analysis (PLS-DA) model of normal versus cancer, creatinine was within the top 10 metabolites that most contribute to the separation of normal and cancer.

Creatinine-normalization would have falsely eliminated this important metabolite. Also, creatinine normalization is only useful if it can be assumed that the kinetics of excretion of the metabolites studied is similar to creatinine, which is not the case for this study.[10, 11]

Since dilute urine is a reflection of more water content in the urine, and thus proportionately lower concentration of *all* the metabolites in the urine, and vice versa for a concentrated urine, it would make more sense to do total normalization; *i.e.* dividing each metabolite concentration by the total metabolite concentration. However, since there are high concentrations of urea in the samples, if total normalization was done, it would resemble urea normalization and thus could falsely nullify the contributions of urea to the model. Therefore, the concentrations were normalized to the total measured concentration of all metabolites minus the measured concentration of urea, i.e. [metabolite]/([total]-[urea]). This method of normalization is superior to creatinine-normalization because it essentially normalizes to the concentration of 68 metabolites rather than one (creatinine). This normalization method was also used in Slupsky et al..[12] Table 3.1 gives an example of the R^2Y (model's fit of data) and Q^2 (model's predictability of data in 7-fold cross-validation) values of the normal vs. CRC OPLS model built using each type of normalization (more on R^2Y and Q^2 later). It can be seen that normalizing to (total-urea) and log transformation gave the best R^2Y (0.478) and a relatively high Q^2 (0.355). Normalization was done using Microsoft Excel v. 11.3.3.

3.7.3 Transformation

Log transformation was done to account for the non-normal distributive nature of the concentrations in the SIMCA-P+ v12.0.1 (Umetrics, Umea, Sweden) program. The formula for log transformation is 10Log(C1*X+C2). For those concentrations that have a value of 0, C2 is made to equal 0.5 to overcome the issue of log (0) = ∝.

Table 3.1: Model characteristics of normal vs. CRC OPLS model built using each type of normalization, with and without log transformation

	Prior to Log Transformation		**After Log Transformation**	
	R^2Y	Q^2	R^2Y	Q^2
Raw Concentration	0.318	0.261	0.470	0.418
Creatinine Normalization	0.376	0.270	0.410	0.314
Total Normalization	0.414	0.353	0.415	0.355
Total-urea Normalization	0.468	0.342	0.478	0.355

3.7.4 Metabolite Selection

Finally, those metabolites that are not products of normal human metabolism, i.e. xenobiotics, such as ibuprofen and salicylurate, were excluded. The internal standard DSS was also excluded. Therefore, of the 72 metabolites, 69 were included in the data analysis.

3.7.5 Statistical Analysis

SIMCA-P+ v12.0.1 (Umetrics, Umea, Sweden) was used to perform the projection-based methods including principal component analysis (PCA), partial

least squares discriminant analysis (PLS-DA), and orthogonal partial least squares (OPLS). These methods convert the multi-dimensional data down to a more manageable 2 or 3 main components based on variance. Projection based models are conceptually very different than traditional regression models with independent predictor variables. They are able to handle many, incomplete, and correlated predictor variables in a simple and straightforward way.[13]

Principal Component Analysis (PCA)

A PCA model is unsupervised and provides a summary, or overview, of all observations or samples demonstrating groupings, trends, and outliers. PCA makes it possible to extract and display systematic variation in the data.[13] Each PCA model is generated based on the direction in the data demonstrating the highest variation, i.e. gender, age, diet, lifestyle, genes, unknown factors, etc. which might be distinctly different from the direction separating the classes.[13, 14]

Partial Least Squares (PLS)

Conventional PLS is used where a quantitative relationship exists between two data tables X & Y; it uses X to construct a model of Y, where the objective is to predict Y from the X for new samples in the prediction set. Systematic variation may reside in X which is not linearly correlated with Y – such variability in X is called Y-orthogonal variation. Although Y-orthogonal variation in X does not affect the predictive power of a PLS model, it may lead to some pitfalls regarding interpretation and has potentially major implications in selection of metabolite

biomarkers, i.e. positive correlation patterns can be interpreted as negligible or negative. The score-loading plot based on the PLS model is perturbed by the presence of Y-orthogonal variation in X.[13, 15]

Orthogonal Partial Least Squares (OPLS)

OPLS is an extension to the supervised PLS regression method with an integrated Orthogonal signal correction (OSC) filter, which removes the uncorrelated signals resulting in information of the within-class variation. The OPLS method is designed to handle variation in X that is orthogonal to Y. OPLS separates the systematic variation in X into two parts, one that is linearly related (and therefore predictive) to Y and one that is orthogonal to Y. The predictive variation of Y in X is modeled by the predictive components. The variation in X which is orthogonal to Y is modeled by the orthogonal components. This partitioning of the X-data provides improved model transparency and interpretability, but does not change the predictive power. OPLS is recommended to obtain a clearer and more straightforward interpretation. It can also provide an understanding of the interclass variation.[13-15]

3.7.6 Model Characteristics

The quality of a model is represented by R^2 and Q^2. R^2 is the percent of variation of the training set – X with PCA and Y with PLS – explained by the model. It is a measure of fit, i.e. how well the model fits the data. Q^2 is the percent of variation of the training set – X with PCA and Y with PLS – predicted by the model

according to cross validation. It indicates how well the model predicts new data in 7-fold cross validation. The range for these parameters is 0 to 1, where 1 indicates a perfect fit. A large R^2Y(close to 1) is a necessary condition for a good model and a large Q^2Y ($Q^2Y > 0.5$) indicates good predictivity.[15]

3.7.7 Fitting the Models

All models are auto-fitted using SIMCA-P+ v12.0.1 (Umetrics, Umea, Sweden). The number of components is thus determined by the software. As the model parameters are optimized, R^2 and Q^2 initially follow the same upward trend from 0 to 1. However as the models start to overfit, the trajectories diverge, R^2 toward 1 and Q^2 falling back toward 0. It is assumed that the model will have achieved its optimal predictive powers, and thus generalize well, at the initial point of divergence.[16] Auto-fitting in SIMCA is programmed to generate the number of components that results in the largest R^2 and Q^2 values.

3.7.8 The Plots

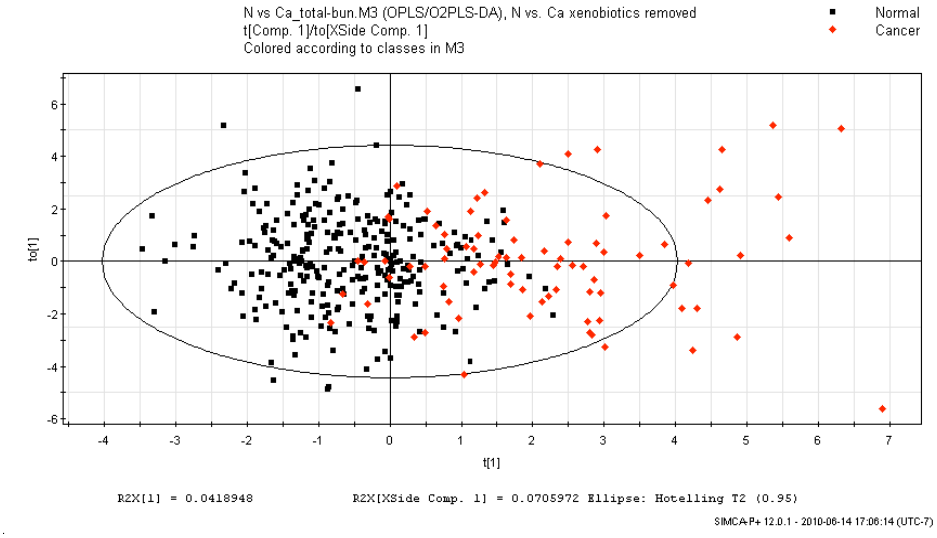
The statistical model can be graphically represented in a variety of methods, namely the scatter plot, loadings plot, variable importance plot (VIP), coefficient plot, observed versus predicted plot.

Scatter Plot

The most visual way to look at the model is a two-dimensional or three-dimensional scatter plot where the scores of the two groups of subjects are plotted

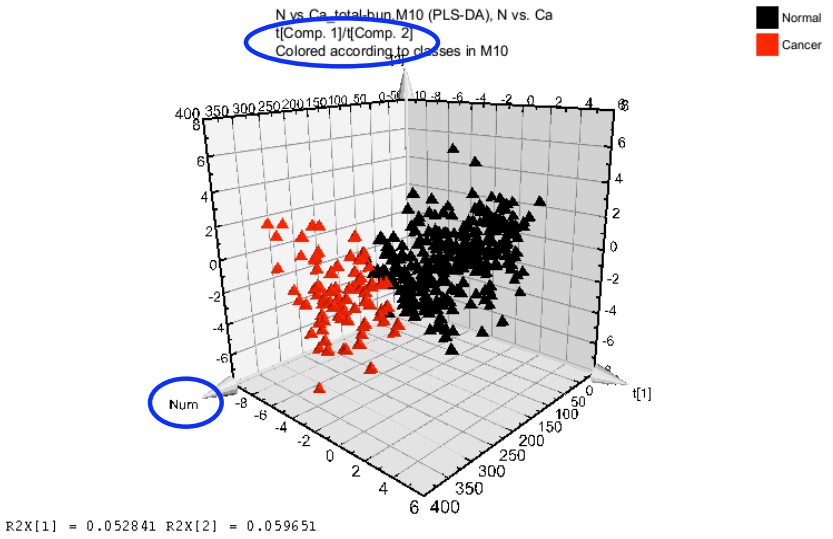
and represented by different colors. This plot reveals groups, trends, outliers, and similarities. An example is shown as figure 3.1.

Figure 3.1: An example of a scatter plot – normal (black squares) vs. CRC (red diamonds) model



A three dimensional plot of the PLS model is only possible when there are three components. If a three-dimensional plot is generated with only two components in the model, then the third dimension is by default separating based on the sample number (Num). Hence, one can be falsely led to thinking that there is a separation between the two groups. An example is shown as figure 3.2. One can always manually add another component in order to demonstrate the model in 3D but this is at the expense of lower Q^2 values.

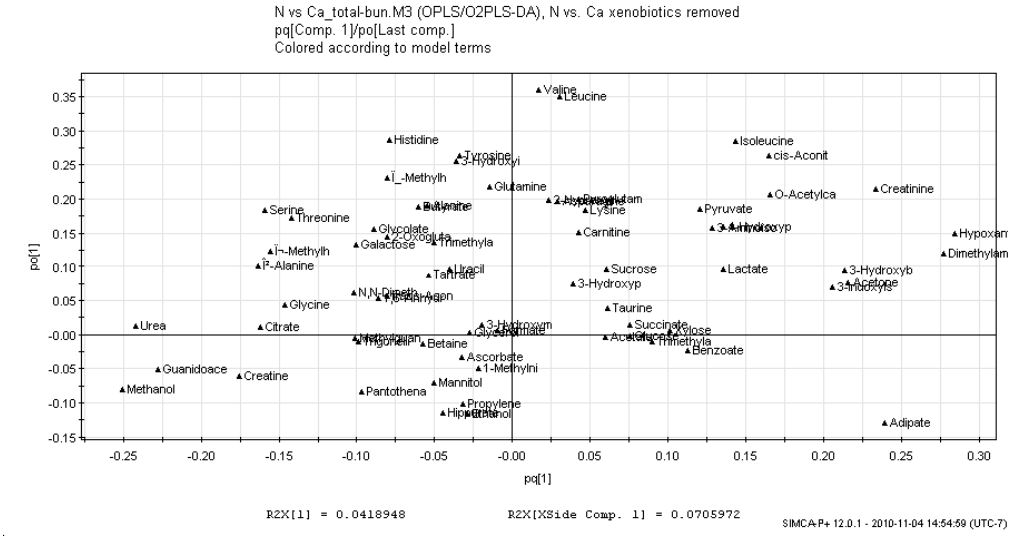
Figure 3.2: An example of a three dimensional scatter plot – normal (black pyramids) vs. CRC (red pyramids) model



Loading Scatter Plot

The loading scatter plot displays the correlation or importance of the x-variables in driving a particular group to the place that they are on the scatter plot. For example, in the figure 3.3, urea and methanol are variables that play a strong role in driving the normal group to the left side of the scatter plot while hypoxanthine and dimethylamine are metabolites that drive the cancer group to the right side.

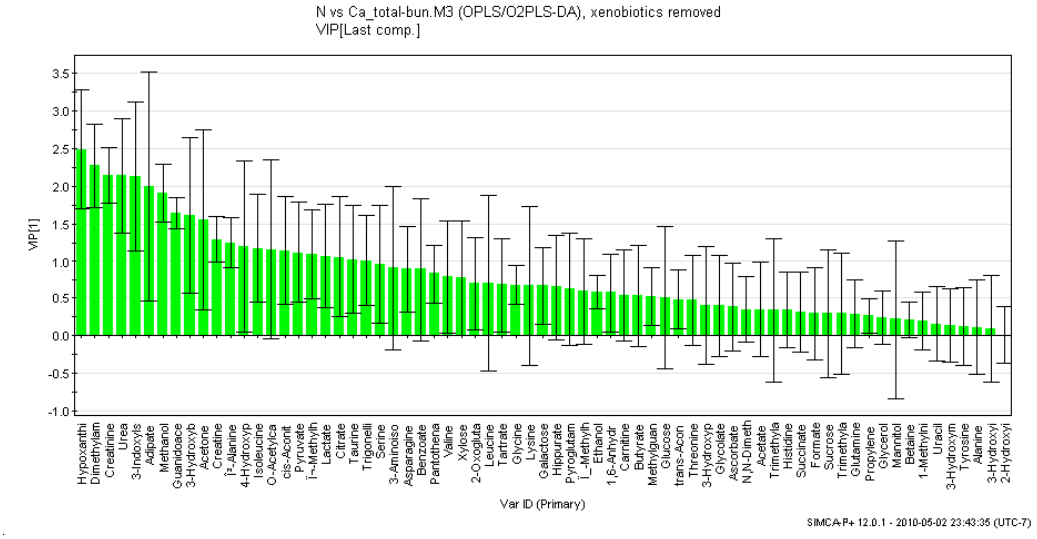
Figure 3.3: An example of a loading scatter plot – normal vs. CRC model



Variable Importance Plot (VIP)

The variable importance plot shows which metabolites are most contributing to the separation of the two groups in a weighted fashion. The VIP score is an absolute value representing the importance that each metabolite has on the separation between the two classes; *i.e.* metabolite impact score.

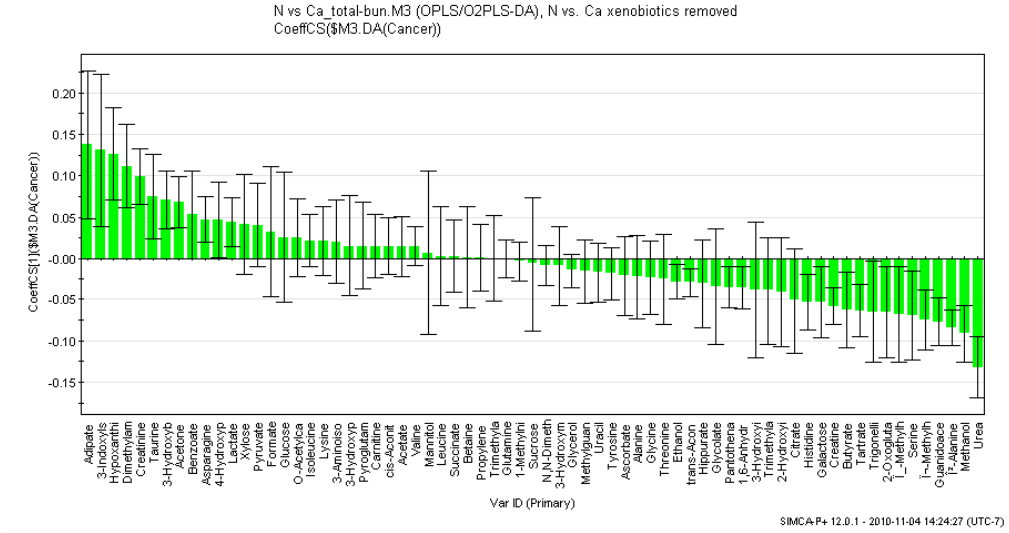
Figure 3.4: An example of a VIP plot – normal vs. CRC model



Coefficient Plot

The coefficient plot also demonstrates the most contributing metabolites in a weighted manner but separates the two groups, i.e. those that are on the left side of the plot are the metabolites that are higher in concentration in the cancer group and those that are on the right side are those that are higher in the normal group.

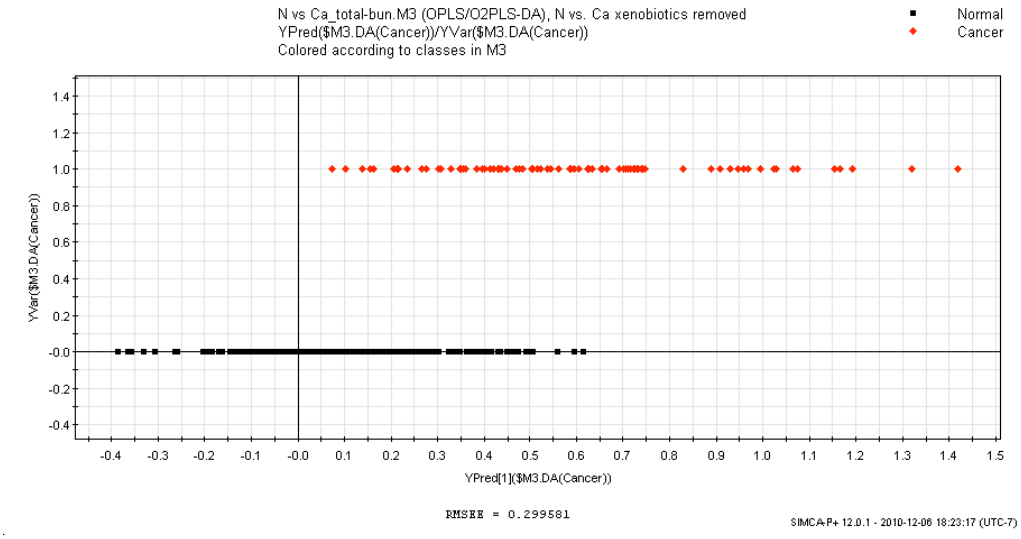
Figure 3.5: An example of a coefficient plot – normal vs. CRC model



Observed vs. Predicted Plot

The observed vs. predicted plot displays the observed values vs. the fitted or predicted values for each subject. This plot allows us to determine the true positives, false positives, true negatives, and false negatives and calculate sensitivity and specificity with a range of cutoffs (see next section).

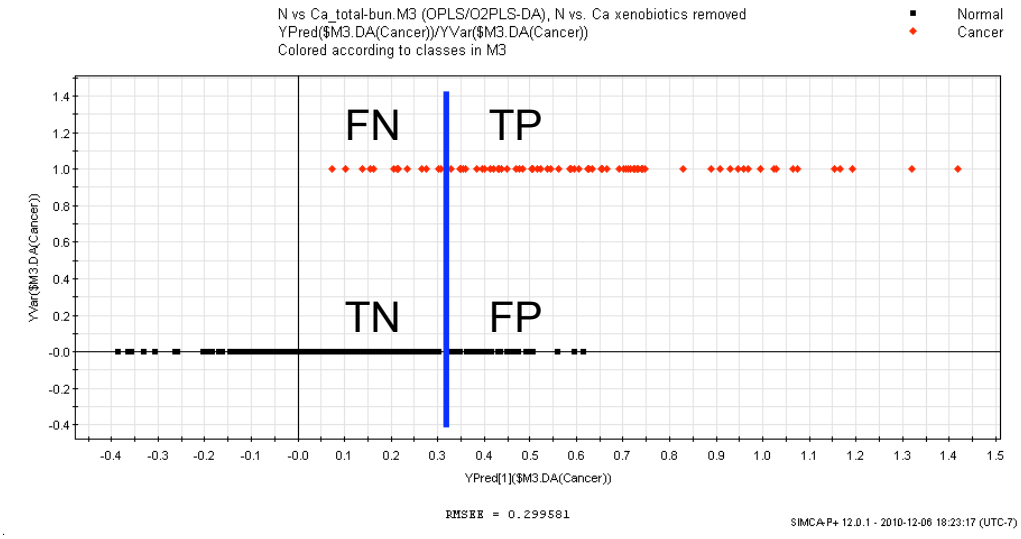
Figure 3.6: An example of an observed vs. predicted plot – normal (black squares) vs. CRC (red diamonds) model



3.7.9 Sensitivity & Specificity Calculations

To generate sensitivity and specificity data, arbitrary cutoff points for the predicted value (YPred) were chosen where the two groups overlapped on the Observed vs. Predicted plot. Those red diamond (cancer) dots that are to the left of the cutoff are false negatives (FN), while those to the right are the true positive (TP). Those black square (normal) dots to the left of the cutoff are the true negatives (TN) and those that are to the right of the cutoff are the false positives (FP) (figure 3.7). Sensitivity and specificity are calculated using the formulas TP/(TP + FN) and TN/(TN+FP), respectively. With this data, a receiver operating characteristics (ROC) curve of sensitivity versus 1-specificity was plotted and area under the curve (AUC) was calculated. Stata/SE 10.1 (Stata Corporation, TX, USA) was used to compute this.

Figure 3.7: An example of an observed vs. predicted plot demonstrating how sensitivity and specificity values are calculated



3.8 Clinical Data Acquisition and Analysis

Relevant clinical information such as age, gender, family history, comorbidities, medications etc. was obtained from the study questionnaires and the patients' medical charts and recorded in Microsoft Access database (Access 2007). Histopathology and pathology results from colonoscopy and surgery were also obtained from patients' medical charts and used as the gold standard for calculating the test diagnostics.

Clinical information was analyzed to examine the effects of clinical variables on the metabolomic test results and specifically whether they contribute to discordant results. The data was also stratified using clinical information such as gender, family history etc. In addition, the profiles of the CRC samples were correlated

with operative and histological findings to determine whether cancer location or stage changed the metabolomic fingerprint.

3.9 Quality Assurance/Quality Control

There were various steps performed throughout the study to ensure accuracy and good quality. During sample collection, all study individuals were instructed on how to collect a midstream urine sample. All urine containers were inspected to ensure there were no cracks and that they were properly coated with Sodium Azide. All urine sample processing and pH adjusting were done by at least 2 individuals to ensure that samples were appropriately labeled and handled. Three additional aliquots of each sample were stored in case of mishaps or if the first sample could not be analyzed properly. The protocol used for this study was based on SOPs from Dr. Slupsky's lab as used in previous urine metabolomics and IBD experiments. A pilot study using the first 53 subjects from the SCOPE program was done to test and refine the protocols (results not shown).

In the analysis stages, all NMR fids were either manually run or checked (for the automated samples) to ensure the fids were done properly. The spectral analysis done by Chenomx was done by one individual, verified by a second individual and then spot checked by a third individual The same version of the Chenomx software and library of compounds were used throughout the analyses.

3.10 References

1. Wong CK, Fedorak RN, Prosser C, Stewart ME, Van Zanten S, Sadowski DC. A comparison of guaiac and immunochemical fecal occult blood tests to colonoscopy for the detection of colonic polyps and cancer: *submitted* 2010.
2. Ma Y-L, Qin H-L, Liu W-J, Peng J-Y, Huang L, Zhao X-P, Cheng Y-Y. Ultra-High Performance Liquid Chromatography-Mass Spectrometry for the Metabolomic Analysis of Urine in Colorectal Cancer. Digestive Diseases and Sciences 2009;54:2655-2662.
3. Qiu YP, Cai GX, Su MM, Chen TL, Liu YM, Xu Y, Ni Y, Zhao AH, Cai SJ, Xu LX, Jia W. Urinary Metabonomic Study on Colorectal Cancer. Journal of Proteome Research 2010;9:1627-1634.
4. Lenz EM, Wilson ID. Analytical strategies in metabonomics. Journal of Proteome Research 2007;6:443-458.
5. Keeler J. Understanding NMR Spectroscopy. John Wiley & Sons Ltd., 2005.
6. Callaghan P. Introductory NMR & MRI: www.magritek.com/videos.html. Accessed Jan 15, 2011
7. Teng Q. Structural Biology: Practical NMR Applications. Springer, 2005.
8. McKay R. Personal Communications, 2010.
9. Mutti A. Biological monitoring in occupational and environmental toxicology. Toxicology Letters 1999;108:77-89.
10. Viau C, Lafontaine M, Payan JP. Creatinine normalization in biological monitoring revisited: the case of 1-hydroxypyrene. International Archives of Occupational and Environmental Health 2004;77:177-185.
11. Heavner DL, Morgan WT, Sears SB, Richardson JD, Byrd GD, Ogden MW. Effect of creatinine and specific gravity normalization techniques on xenobiotic biomarkers in smokers' spot and 24-h urines. Journal of Pharmaceutical and Biomedical Analysis 2006;40:928-942.
12. Slupsky CM, Rankin KN, Wagner J, Fu H, Chang D, Weljie AM, Saude EJ, Lix B, Adamko DJ, Shah S, Greiner R, Sykes BD, Marrie TJ. Investigations of the effects of gender, diurnal variation, and age in human urinary metabolomic profiles. Analytical Chemistry 2007;79:6995-7004.
13. Trygg J, Holmes E, Lundstedt T. Chemometrics in metabonomics. Journal of Proteome Research 2007;6:469-479.
14. Bylesjo M, Rantalainen M, Cloarec O, Nicholson JK, Holmes E, Trygg J. OPLS discriminant analysis: combining the strengths of PLS-DA and SIMCA classification. Journal of Chemometrics 2006;20:341-351.
15. SIMCA-P+ v12.0.1 Analysis Advisor (Umetrics, Umea, Sweden).
16. Broadhurst DI, Kell DB. Statistical strategies for avoiding false discoveries in metabolomics and related experiments. Metabolomics 2006;2:171-196.

4.0 A Novel and Highly Sensitive Test For Detecting Colorectal Cancer Using Metabolomics From a Spot Urine Sample

4.1 Abstract

Background: Colorectal cancer (CRC) is the third leading cause of cancer-related death in the Western World. However, if identified early, CRC is curable. Current non-invasive fecal-based screening methods for CRC are cumbersome to do and have low sensitivity. The science of metabolomics is the systematic study of unique small molecule metabolite fingerprints in bio-fluids and/or tissues. The metabolome represents the collection of all metabolites in and organism which, in the human, are the end products of both somatic and bacterial cellular processes. The extension of metabolomic fingerprints to their examination in disease states and altered physiologic conditions represents the potential for a highly sensitive, non-invasive, novel screening tool for detecting CRC.

Aim: The aim of this chapter was to use metabolomics from a spot urine sample to develop a diagnostic test that would distinguish healthy subjects from patients with CRC. We achieved this aim by building and refining a metabolomics model that estimated the sensitivity and specificity of CRC relative to the gold standard of colonoscopy or the diagnostic modality that was used to establish the diagnosis of the cancer.

Methods: Urine samples were collected from 444 colonoscopy-negative normal subjects and 116 CRC patients and analyzed using an Oxford 600Hz nuclear magnetic resonance (NMR) spectrometer with a Varian VNMRS two-channel

console. The 1H NMR spectrum of each urine sample was analyzed using Chenomx NMRSuite v7.0 (Chenomx, Inc. Edmonton, Canada). The first 294 normal and 82 CRC samples were used as a training set to establish the diagnostic metabolomic model of normal vs. CRC using multivariate analysis with the aid of SIMCA-P+ v12.0.1 (Umetrics, Umea, Sweden) and STATA/SE 10.1 (TX, USA). The model was then validated with the remaining 150 normal and 34 CRC urine samples (the testing set).

Results: Using 69 metabolites, the normal and cancer groups could be separated with a two-component orthogonal partial least squares (OPLS) model with a R^2Y of 0.478 (model's fit of data), and a Q^2 of 0.355 (model's predictability of data in 7-fold cross-validation). Diagnostic accuracies were calculated using the predicted values from the model and a sensitivity and specificity of 92.7% and 71.8%, respectively, were achieved. A receiver operating characteristics (ROC) curve was generated and area under the curve (AUC) was calculated to be 0.931 (95% CI 0.902, 0.961). Validation of the model with 184 blinded samples resulted in sensitivity and specificity values of 85.3% and 52.7%, respectively, confirming the robustness of the model. Stratification by gender and family history of cancer resulted in sub models that had even higher diagnostic accuracies.

Conclusions: This is the largest reported study to demonstrate that NMR urine metabolomics, as a diagnostic test, has the ability to distinguish normal healthy subjects from CRC patients with substantially better accuracy than that of current

fecal-based tests. Urine metabolomics has the potential to become an accurate, non-invasive, and inexpensive screening tool for CRC.

4.2 Introduction

CRC is a major public health concern as it is a leading cause of morbidity and mortality. In 2010, the National Cancer Institute estimates that there were 102,900 new cases of colon cancer and 39,670 new cases of rectal cancer and 51,370 combined colon and rectal cancer deaths in the United States.[1] The Canadian Cancer Society estimates that in 2010 there were 22,500 Canadians diagnosed with CRC and 9,100 died of it.[2] Current non-invasive screening method for CRC is guaiac-based fecal occult blood test (FOBT), Hemoccult II, which is a 3-day sample collection test that has a sensitivity for detecting CRC of 25%-38%[3]. Newer fecal immunochemical tests (FIT) specifically bind to human hemoglobin, and thus have a higher sensitivity for CRC (61-91%). However many of the studies done for the FITs only provided estimates for sensitivity as patients with negative results underwent flexible sigmoidoscopy or registry follow-up only.[4] Colonoscopy is currently the gold standard for screening but it is costly and has defined, non-negligible morbidity and potential mortality associated with it. Metabolomics is an emerging field of research that quantitatively identifies low molecular weight compounds, such as metabolic substrates and products, lipids, small peptides, vitamins, and other protein cofactors, generated by metabolism. Urine metabolomics is being investigated for a potential role in screening tests for colorectal cancer and precancerous lesions.

This chapter focuses on the metabolomic fingerprint of CRC, which is established by comparing the urine metabolite profile of patients with CRC relative to controls that have had a normal colonoscopy (unblinded training set). This fingerprint will then be validated with a separate population of normal subjects and CRC patients (blinded testing set).

4.3 Objectives

- To identify the most influential metabolites that contribute to the separation, in the spot urine metabolomic fingerprint, between patients with CRC and those without CRC.
- In patients with CRC and those without CRC, using an un-blinded training set of the metabolomic fingerprint data defined above, to build and refine a model that would estimate the sensitivity and specificity of urine metabolomics in identifying CRC relative to the gold standard of colonoscopy or the diagnostic modality that was used to establish the diagnosis of the cancer.
- Using a blinded testing set of the metabolomic fingerprint data, to further confirm the statistical models and the sensitivity and specificity of the spot urine metabolomic fingerprint as a diagnostic test that would distinguish patients with CRC from those without CRC.

4.4 Materials and Methods

4.4.1 Recruitment and Sample Collection

Urine samples were collected from 444 colonoscopy-negative (normal) subjects, i.e. subjects who had completely normal colonoscopy findings, aged 39-76 through a regional population-based screening program called SCOPE (Stop COlorectal cancer through Prevention and Education) in Edmonton, Alberta between April 2008 and October 2009. The screening population consisted of average and high-risk individuals (personal or family history of CRC). All screening subjects had a colonoscopy as the gold standard of their diagnosis. The urine samples were collected prior to the colonoscopy.

Urine samples were also collected from 116 CRC patients through the practices of general surgeons and gastroenterologists in Edmonton and Grand Prairie, Alberta between October 2008 and June 2010. The urine sample of the CRC patients was collected prior to any surgical or neoadjuvant chemoradiation treatment. All cancer patients had a tissue diagnosis of colon or rectal cancer from biopsies obtained during colonoscopy or from resected pathological specimens. All removed colonic tissue was sent for histological analysis. Pathologists were blinded to the urine metabolomic results.

In all cases, clinical information such as demographics, family history, co-morbidities, smoking status, etc. was gathered in the form of a questionnaire and from patients' health records. Subjects were excluded if they were anuric,

oliguric, have end-stage renal failure, or on hemo or peritoneal dialysis. Ethics approval for this study was obtained from the Health Research Ethics Board at the University of Alberta.

4.4.2 Sample Analysis

All urine samples were stored at -80°C until they were ready to be analyzed. The day prior to NMR acquisition, each sample was thawed to room temperature and was diluted (1:10) with internal standard consisting of 5 mM sodium 2,2-dimethyl-2-silapentane-5-sulfonate (DSS), 100 mM imidazole, 0.2% sodium azide in 99% D2O. The samples were stored at 4°C overnight. On the day of NMR acquisition, each sampled was adjusted to a pH between 6.7 and 6.8 and aliquoted into 5mm NMR tubes. One-dimensional nuclear magnetic resonance spectra was acquired using an Oxford 600Hz NMR spectrometer with a Varian VNMRS two channel console and running VNMRJ software version 2.2C on a RHEL 4 host computer in the Canadian National High Field NMR Centre (NANUC), Edmonton, Alberta. All samples were run at a sweep width (sw) of 7225.43 Hz. The saturation frequency (sfrq), transmitter offset (tof) and pulse width (pw) were all individually calibrated at the start of each day. The tof typically ranged from (-213 to -215 Hz) and the pw ranged from 6 to 8 μs. Shims were optimized until an acceptable line width value was obtained at relative peak heights of: 50% (< 1.0 Hz), 0.55% (< 12.0 Hz), and 0.11% (< 20.0 Hz) were achieved. Water suppression was performed. Spectra were collected at 25°C through a total of 32 scans over a period of 3.5 min; a total recycle delay of 5 s

was also used (*i.e.* 1 s recovery delay/saturation and a 4 s acquisition). The ^{1}H NMR spectrum of each urine sample was analyzed and quantitated using the targeted profiling technique[5] as implemented in Chenomx NMRSuite v7.0 (Chenomx, Inc. Edmonton, Canada). The quantification process was done independently by two individuals and verified by a third individual to optimize accuracy. 294 metabolites were considered and 72 were found to be significant.

The spectral acquisition and quantification process were performed without the knowledge of the pathology results.

4.4.3 Data Analysis

The first 294 normal and 82 CRC samples were used as a training set to establish the diagnostic metabolomic model of normal vs. CRC using projection-based methods and logistic regression with the aid of SIMCA-P+ v12.0.1 (Umetrics, Umea, Sweden) and STATA/SE 10.1 (TX, USA). The metabolite concentrations were normalized (to total metabolite concentration except urea) to account for the dilutional differences in the urine samples. Log transformation was done to account for the non-normal distributive nature of the concentrations. Finally, those metabolites that are not products of normal human metabolism, i.e. xenobiotics, such as ibuprofen and salicylurate, were excluded. Diagnostic accuracies (such as sensitivity, specificity, and AUC for ROC curve) were calculated from the statistical model and the model was then validated with 184 blinded urine samples (150 normal and 34 CRC), the testing set. The clinical

characteristics of the two groups were compared and stratified models were built. Chi squared test was used to compare proportional outcomes and student's t-test was used to compare continuous outcomes.

4.5 Results

4.5.1 Training Set Subject Characteristics

Table 4.1 lists the demographics and clinical characteristics comparing the normal and the CRC groups in the training set. It is not surprising that the two groups are quite different from each other. There are more females in the normal group (60%) compared to the CRC group (43%), p=0.005. The average age for the cancer group (68.6±1.2 years) is approximately 13 years older than that of the normal group (55.3±0.5 years). There are more people in the normal group with positive family history of CRC (69% vs. 25%, $p<0.001$) or any cancer (92% vs. 79%, $p<0.001$), but more people in the CRC group who smoke (20% vs. 9%, p=0.010), have diabetes (23% vs. 5%, $p<0.001$) and have symptoms of gastrointestinal bleeding (60% vs. 2%, $p<0.001$) and altered bowel habits (52% vs. 4%, $p<0.001$).

Table 4.1: Patient characteristics

	NORMAL [N=294] N (%)	CRC [N=82] N (%)	p-VALUE
Male:Female	117:177	47:35	0.005*
Average age (years±SEM)	55.3±0.5	68.6±1.2	<0.001*
FHx of Colon or Rectal cancer	191 (69)	20 (25)	<0.001*
FHx of Any Cancer	230 (92)	60 (79)	<0.001*
Smoking	26 (9)	15 (20)	0.010*
Diabetes	14 (5)	19 (23)	<0.001*
GI Bleeding	5 (2)	49 (60)	<0.001*
Change in Bowel Habit	13 (4)	43 (52)	<0.001*

Note: Not all % are calculated with the denominator of the total in each group as some clinical information was missing or unknown. * $p \leq 0.05$

Within the CRC group, patients presented with different stages of cancer (figure 4.1) at various locations of the colon (figure 4.2), and with different pathological features (figure 4.3). The majority of the patients within this study (60%) presented with stage three or four CRC, that is, the lymph nodes were involved or there is presence of distal metastasis. Approximately one-third of the cancers were right sided and one-thirds were in the rectum. About 35% of the cancers had lymphocytic response, about 25% had lymphatic invasion, and less than 10% had vascular invasion or perineural invasion.

Figure 4.1: Colorectal cancer by stage

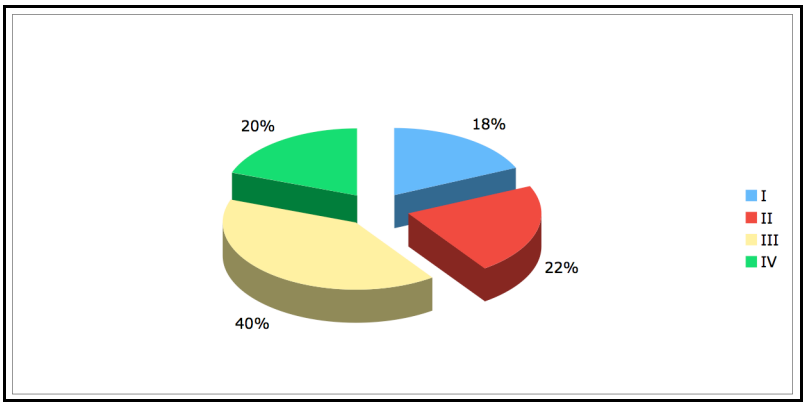


Figure 4.2: Colorectal cancer by location

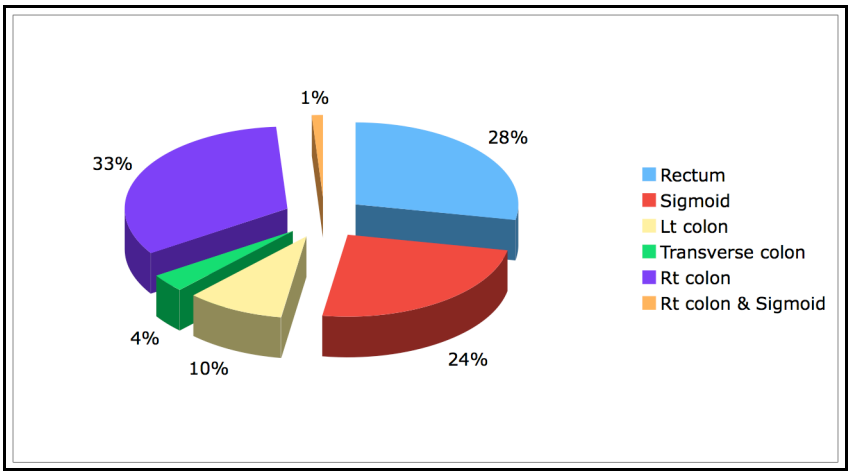
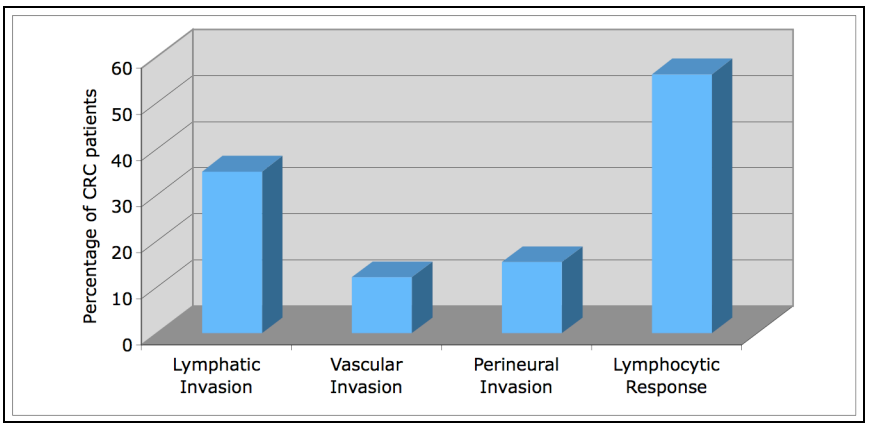


Figure 4.3: Pathological features of colorectal cancer specimens



4.5.2 Building the Models

Unsupervised principal component analysis (PCA) was unable to generate a statistically significant model to separate out the two groups. However, using two-component separation, a supervised orthogonal partial least squares (OPLS) model was built with R^2Y of 0.478, and Q^2 of 0.355. The OPLS scatter plot shown below (figure 4.4) illustrates the normal group in black squares and the cancer group in red diamonds. A crude exploratory data analysis shows that although there is a degree of overlap, it is clear that the two groups are showing up in different areas of the plot. A partial least squared discriminant analysis (PLS-DA) model was also built and an extra component was added to generate a 3-dimensional scatter plot of the same data (figure 4.5).

Figure 4.4: OPLS scatter plot of normal (black squares) vs. CRC (red diamonds)

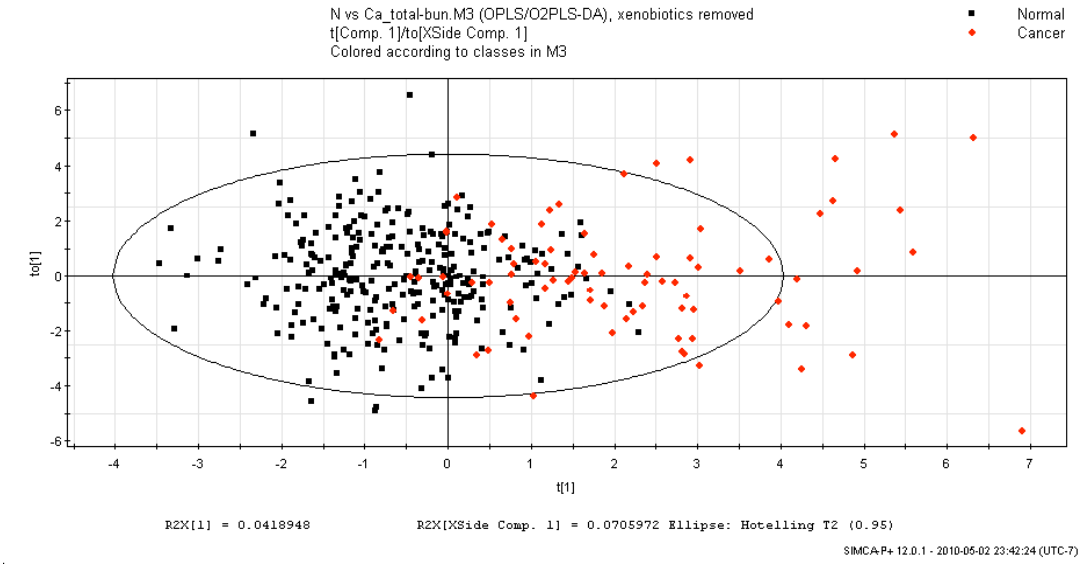
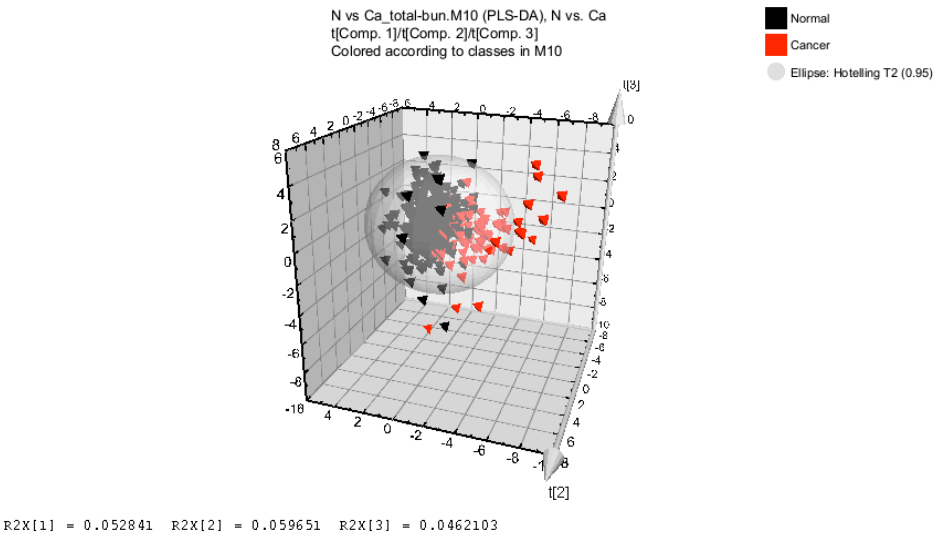


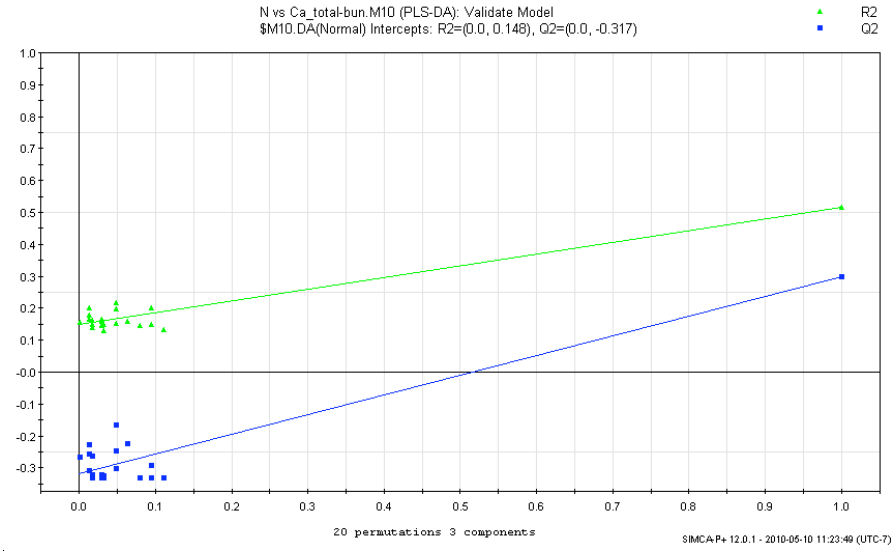
Figure 4.5: PLS-DA 3-D scatter plot of normal (black pyramids) vs. CRC (red pyramids)



To validate that this statistical model is not spurious and not overfit, permutation tests were used and a validation plot was generated using SIMCA (figure 4.6). Note that the validation plots can only be generated on the PLS-DA models. Permutation tests help assess over-fitting by randomly permuting class labels and refitting a new model with the same number of components as the original model. An over-fit model will have similar R^2 and Q^2 to that of the randomly permuted data. Well-fit models will have R^2 and Q^2 values that are always higher than that of the permuted data.[5] In this case, twenty models were generated based on the data where the order of the Y-observations has been randomly permuted while the X-matrix has been kept intact and the R^2 and Q^2 values for these models are shown on the left side of the validation plot. The R^2 and Q^2 values of the original normal vs. CRC model are shown far to the right. This validation plot shows that all the R^2 and Q^2 values for the randomly generated models are lower than the

original points to the right, that is, the goodness of fit of this model is better than 'random' and the model is not over-fit.

Figure 4.6: Validation plot



4.5.3 Diagnostic Accuracies

An Observed vs. Predicted plot (figure 4.7) was generated and a range of sensitivity and specificity values were calculated based on different cut-offs. The three representative pairs of diagnostic accuracies listed in table 4.2 were picked according to the following criteria: the highest sensitivity that results in a specificity of at least 50%, sensitivity and specificity that are similar in value and a pair in between the previous two. Receiver operating characteristics (ROC) curve is generated (figure 4.8) from the range of sensitivity and specificity values and AUC is calculated to be 0.9314 (95% CI 0.9017, 0.9611).

Figure 4.7: Observed vs. predicted plot of normal (black squares) vs. CRC (red diamonds) model

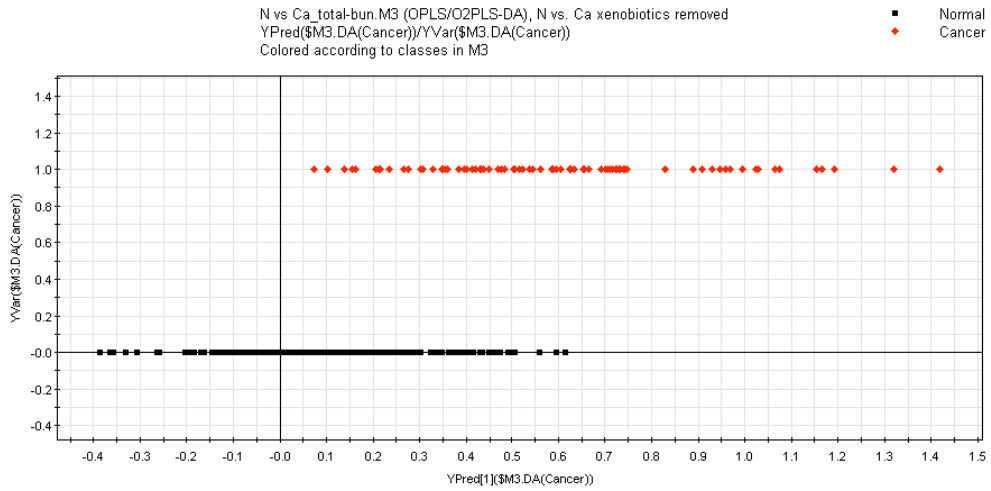
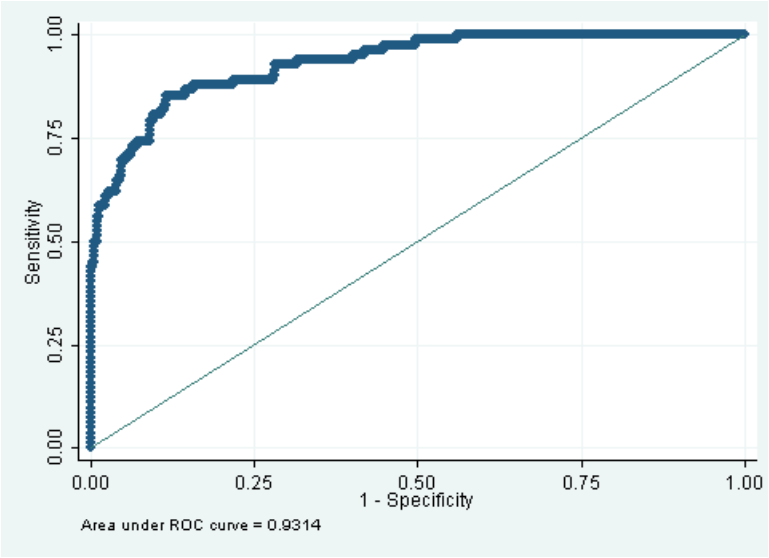


Table 4.2: Representative diagnostics and model characteristics for normal vs. CRC OPLS model

Cut off level	Sensitivity	Specificity	R^2Y	Q^2	AUC
0.101527	98.780%	50.340%	0.478	0.355	0.9314
0.212925	92.680%	71.770%			
0.279316	85.370%	85.370%			

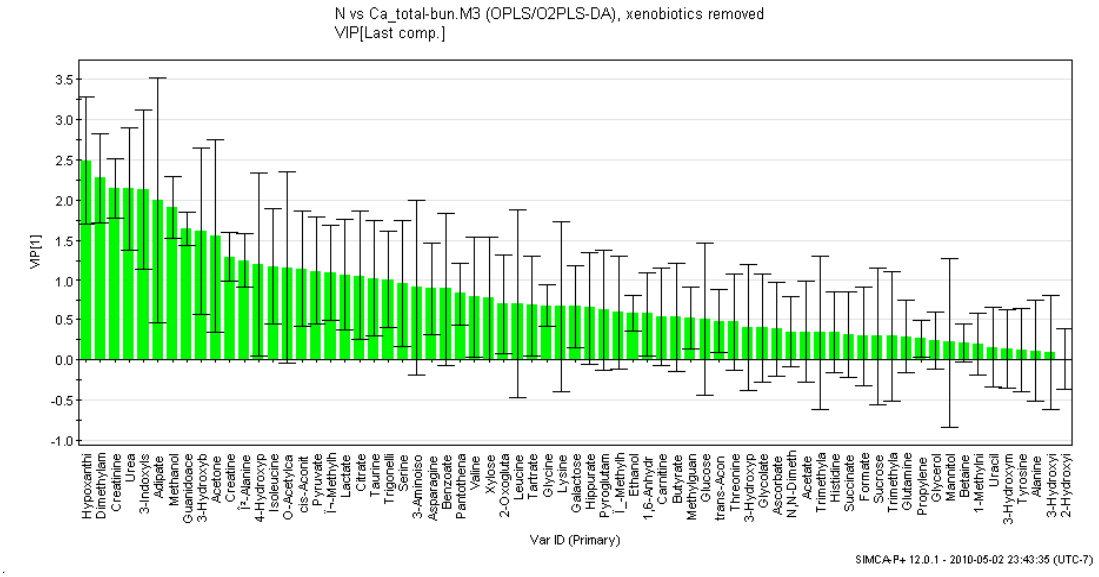
Figure 4.8: Receiver operating characteristics (ROC) curve for normal vs. CRC OPLS model



4.5.4 Metabolites

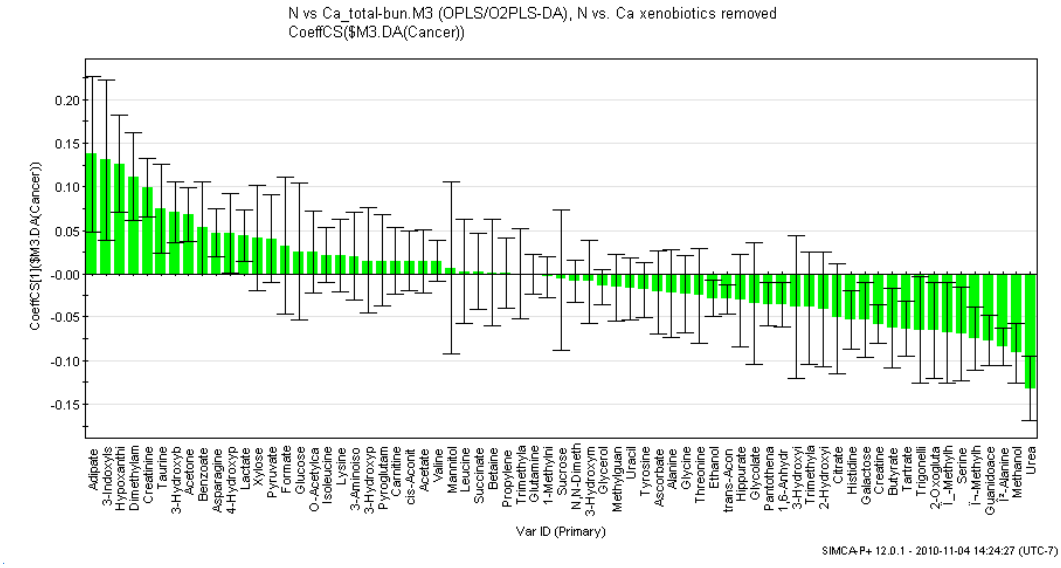
A Variable Importance Plot was generated to illustrate which metabolites contribute the most to the separation between normal and cancer.

Figure 4.9: Variable importance plot of normal vs. CRC OPLS model



The top 10 metabolites that contribute to the separation of normal and CRC (in order of importance) are: hypoxanthine, dimethylamine, creatinine, urea, 3-indoxylsulfate, adipate, methanol, guanidoacetate, 3-hydroxybutyrate, and acetone. Specifically, the top 5 metabolites that are higher in concentration in the cancer samples are adipate, 3-indoxylsulfate, hypoxanthine, dimethylamine, and creatinine; and those that are higher in normal samples are urea, methanol, β-alanine, π-methylhistidine, and serine. This is shown by the coefficient plot below.

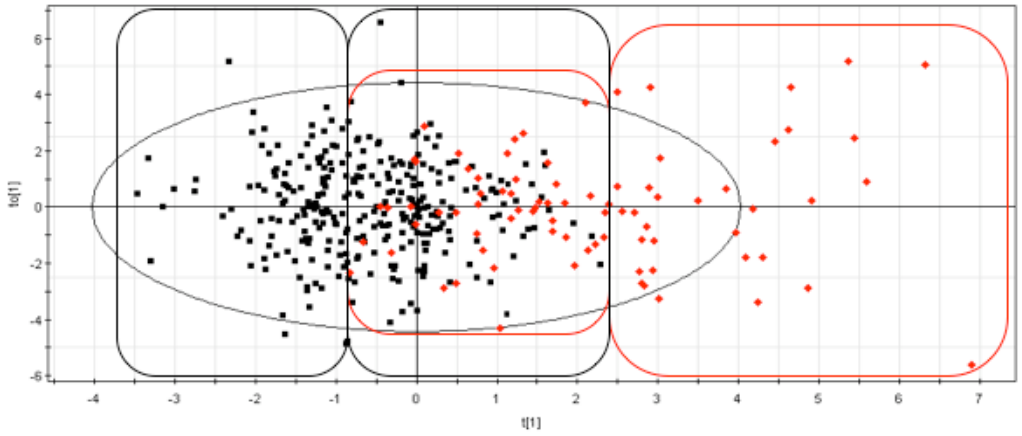
Figure 4.10: Coefficient plot for normal vs. CRC OPLS model



4.5.5 Model Analysis

As there is overlap between the normal group and the CRC group in the OPLS model generated, it was unclear what characteristics contributed to the overlap, that is, what clinical characteristics do some of the patients with cancer have that makes them behave like normal and vice versa. For this, the overlapping and non-overlapping groups in the model were studied separately. Four groups were generated: 1) Overlapping normal, 2) Non-overlapping normal, 3) Overlapping CRC, and 4) Non-overlapping CRC groups. Since we are working with the OPLS model, only one direction of separation had to be taken into consideration (left-right). To generate the subgroups, we took all the normal (black square) data points to the left of the left-most CRC (red diamond) data point and made this the *Non-overlapping normal group*; and the remainder normal data points the *Overlapping normal group*. The same process was done for the CRC data points. (Figure 4.11)

Figure 4.11: OPLS scatter plot of normal (black squares) vs. CRC (red diamonds) model with overlapping and non-overlapping groups defined



We looked at each combination of the different subgroups and examined the metabolites that contribute to the separation separately in an attempt to narrow down the list of metabolites that would give us a more powerful model. Next we looked at clinical characteristics that may be different between the subgroups to determine the best traits to stratify the groups to generate more predictive models.

4.5.5.1 Subdividing The Normal Group

4.5.5.1.1 Metabolites

- CRC vs. Non-overlapping normals

When the CRC group was plotted against the Non-overlapping normals group, that is, the group of normals that was the most different from the CRC group, the top metabolites that drove the separation were: dimethylamine, creatinine, hypoxanthine, 3-indoxylsulfate, and methanol (table 4.3).

- CRC vs. Overlapping normals

When the adenoma group was plotted against the Overlapping normal group, that is, the group of normals that somehow resembled the CRC group metabolomically, the top metabolites that drove the separation were: urea, hypoxanthine, adipate, dimethylamine, and 3-indoxylsulfate (table 4.3).

- Overlapping normals Vs. Non-overlapping normals

The Overlapping and the Non-overlapping groups of normal were plotted together on a scatter plot, without the CRC group, to see what the metabolomic difference was between these two groups of normals. The top 5 metabolites that contributed to the separation between these two groups of Normal were: methanol, creatinine, creatine, guanidoacetate and hypoxanthine (table 4.3).

Table 4.3: Summary of metabolites from sub-models of normal subjects

CRC vs. Non-overlapping Normal	CRC vs. Overlapping Normal	Overlapping vs. Non-overlapping Normal
Dimethylamine	Urea	Methanol*
Creatinine	Hypoxanthine	Creatinine*
Hypoxanthine	Adipate	Creatine*
3-Indoxylsulfate	Dimethylamine	Guanidoacetate*
Methanol	3-Indoxylsulfate	Hypoxanthine*
Guanidoacetate	Methanol	3-Indoxylsulfate*
Urea**	Creatinine	π-methylhistidine*
Creatine	Acetone	Dimethylamine*
Cis-Aconitate	3-hydroxybutyrate	Threonine
π-methylhistidine	Guanidoacetate	Cis-Aconitate*
β-alanine**	β-alanine	Glycine
Adipate**	Trigonelline	Methylguanidine
Pyruvate**	Isoleucine	Carnitine
Citrate**	Valine	Formate
Serine**	Lactate	2-Hydroxyisobutyrate

* Metabolites in Overlapping vs. Non-overlapping Normal model that are also part of CRC vs. Non-overlapping Normal model. ** Metabolites in CRC vs. Non-overlapping Normal model that are not in Overlapping vs. Non-overlapping Normal model.

It was hypothesized that if we eliminated those metabolites that separate Overlapping & Non-overlapping normal from the main model developed from the Normal vs. CRC training set, perhaps we could minimize the difference between the two normal groups and achieve greater separation between normal and CRC. However, this was not the case as the R^2 and Q^2 values are much lower than that for the main model (table 4.4).

Table 4.4: Summary of sub-model characteristics

Model	R^2Y	Q^2
Main Model (Normal vs. CRC Training)	0.478	0.355
Main model excluding the 9 metabolites in Overlapping vs. Non-overlapping Normal model that are also part of CRC vs. Non-overlapping Normal model (denoted by *)	0.373	0.226
Main model using only 6 metabolites in CRC vs. Non-overlapping Normal model that are not in Overlapping vs. Non-overlapping Normal model (denoted by **)	0.206	0.172

4.5.5.1.2 Clinical Parameters

Using logistic regression, clinical characteristics such as age, gender, family history, etc of the normal group were tested for their odds of predicting the dichotomous outcome of overlap with the CRC group or not. The odds ratios and p-values are summarized in the table below (table 4.5). This is an exploratory analysis to identify potential factors for stratification.

Table 4.5: Clinical characteristics of the normal group and the odds of overlapping with CRC group

Variable		Odds Ratio	p-value
Gender		2.491	<0.001*
Age		1.020	0.175
Smoking		1.574	0.293
Diabetes		1.045	0.937
Family History	CRC	1.266	0.369
	1st degree CRC/polyp	1.130	0.616
	Any cancer	1.482	0.389
Symptoms	GI bleed	0.189	0.138
	Change bowel habits	2.667	0.143

* $p \leq 0.05$

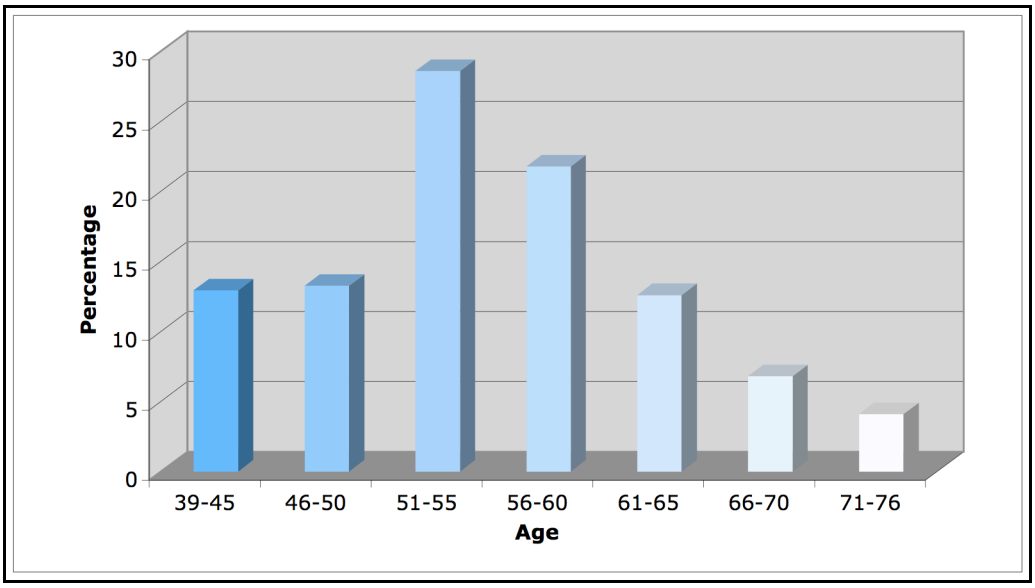
Demographics

In the normal group, the odds of resembling or overlapping with the CRC group for males was 2.49 (95% CI 1.52, 4.07) times that of females ($p<0.01$). That is, females were more different than the cancers compared to the males.

Within the normal group, age did not change the odds of overlapping with CRC, OR 1.02 (95% CI 0.99, 1.05). For every 10 years increase in age, the odds of overlapping with CRC increased about 22% (OR=1.22), however, the 95% CI for the OR is 0.915, 1.627, which included 1, therefore this was not statistically significant. One of the limitations of using logistic regression on continuous variables is that the model may have a threshold effect, a saturation effect, or a binary effect that cannot be shown, i.e. logistic regression models assume that the change from age 20 to 21 is the same as that from 60 to 61, which is not the case in terms of risks for colorectal cancer. Therefore, we tested the age by categories

of 5 years (figure 4.12) and found that none of the categories significantly predicted overlap with CRC. Although the odds of overlapping with CRC in those that were over the age of 70 was four times that of those that were younger than 45, but this was not statistically significant (p = 0.096) and there might not be enough numbers to see this effect. This was reflected by the wide 95% confidence interval (0.779, 21.019).

Figure 4.12: Age of normal subjects divided by category



Smoking History

Within the normal group, being a smoker or an ex-smoker did not increase the odds of overlapping with the CRC group. Since metabolism is a reflection of current status, ex-smokers could be classified as non-smokers and there still was no statistically significant difference (OR = 1.57; 95% CI = 0.68, 3.66; p = 0.293),

but since there were only 26 smokers, there might not be enough numbers to show a difference.

Diabetes

Having diabetes did not increase the odds of overlapping with the CRC group (OR = 1.04; 95% CI = 0.35, 3.09; p = 0.937), although there were only 14 out of 294 subjects who had diabetes, and again this might be too few to show a difference.

Family history

Within the normal group, having a family history of CRC did not increase the odds of overlapping with the CRC group (OR = 1.26; 95% CI = 0.76, 2.11; p = 0.369). Even when we divided the groups into those with a first-degree relative with CRC and those that don't, this highest risk group did not have increased odds of overlapping with the CRC group (OR = 1.13; 95% CI = 0.70, 1.82; p = 0.616). Having a family history of *any* cancer(s) did not increase the odds of overlapping with the cancer group either (OR = 1.48; 95% CI = 0.61, 3.63; p = 0.389).

Gastrointestinal Symptoms

Within the normal group, having symptoms of gastrointestinal bleeding (OR = 0.19; 95% CI = 0.02, 1.71; p=0.138) or changes in bowel habits (OR = 2.67; 95% CI = 0.72, 9.90; p=0.143) did not increase the odds of overlapping with the CRC group, although the number of subjects with GI symptoms in this screening population was small.

Overall

After adjusting for age, smoking status, diabetes, family history of CRC, family history of any cancer, GI bleeding, and bowel habits, gender was still statistical significant in predicting overlap with the CRC group.

4.5.5.2 Subdividing The CRC Group

4.5.5.2.1 Metabolites

- Normal vs. Non-overlapping CRC

When the Normal group was plotted against the Non-overlapping CRC group, that is, the group of CRCs that was the most different from the normal group, the top metabolites that drove the separation were: hypoxanthine, urea, adipate, acetone, and 3-hydroxybutyrate (table 4.6).

- Normal vs. Overlapping CRC

When the Normal group was plotted against the Overlapping CRC group, that is, the group of CRCs that somehow resembled the normal group metabolomically, the top metabolites that drove the separation were 3-Indoxylsulfate, creatinine, dimethylamine, methanol, and 4-hydroxyphenylacetate (table 4.6).

- Overlapping Cancers Vs. Non-overlapping CRC

The Overlapping and the Non-overlapping groups of CRC were plotted together on a scatter plot, without the normal group, to see what the metabolomic

difference was between these two groups of CRC. The top 5 metabolites that contributed to the separation between these two groups of CRC were urea, hypoxanthine, trigonelline, n, n-dimethylglycine, and guanidoacetate (table 4.6).

Table 4.6: Summary of metabolites from sub-models of CRC patients

Normal vs. Non-overlapping CRC	**Normal vs. Overlapping CRC**	**Overlapping vs. Non-overlapping CRC**
Hypoxanthine	3-Indoxylsulfate	Urea*
Urea	Creatinine	Hypoxanthine*
Adipate	Dimethylamine	Trigonelline*
Acetone	Butyrate	N,N-Dimethylglycine
3-Hydroxybutyrate**	β-Alanine	Guanidoacetate*
Dimethylamine	Methanol	Adipate*
Methanol	π-Methylhistidine	Tyrosine
Isoleucine	Asparagine	Lactate*
Lactate	Hypoxanthine	Betaine
Creatinine**	Creatine	Dimethylamine*
O-Acetylcarnitine**	Guanidoacetate	Acetone*
Guanidoacetate	4-Hydroxyphenylacetate	Methanol*
Trigonelline	Carnitine	Acetate
Citrate	Tyrosine	Isoleucine*
3-Aminoisobutyrate**	Trimethylamine	Citrate*

* Metabolites in Overlapping vs. Non-overlapping CRC model that are also in Normal vs. Non-overlapping CRC model. ** Metabolites in Normal vs. Non-overlapping CRC model that are not in Overlapping vs. Non-overlapping CRC model.

Again we hypothesized that if we eliminated those metabolites that separate Overlapping & Non-overlapping CRC from the main normal vs. CRC model developed from the training set, perhaps we could minimize the difference between the two CRC groups and achieve greater separation between normal and CRC. However, again this was not the case as the model characteristics for the new models are much lower than that for the original model, shown in table 4.7.

Table 4.7: Summary of sub-model characteristics

Model	R^2Y	Q^2
Main model (Normal vs. CRC Training)	0.478	0.355
Main model excluding 11 metabolites in Overlapping vs. Non-overlapping CRC model that are also in Normal vs. Non-overlapping CRC model (denoted by *)	0.313	0.208
Main model using only 4 metabolites in Normal vs. Non-overlapping CRC model that are not in Overlapping vs. Non-overlapping CRC model (denoted by **)	0.137	0.129

4.5.5.2.2 Clinical Parameters

Logistic regression analysis was used to test the significance of various clinical and pathological variables of CRC patients on the dichotomous outcome of overlapping or non-overlapping with normals. The odds ratios and p-values are summarized in table 4.8. This is an exploratory analysis to identify potential factors for stratification.

Table 4.8: Clinical characteristics of the CRC group and the odds of overlapping with normal group

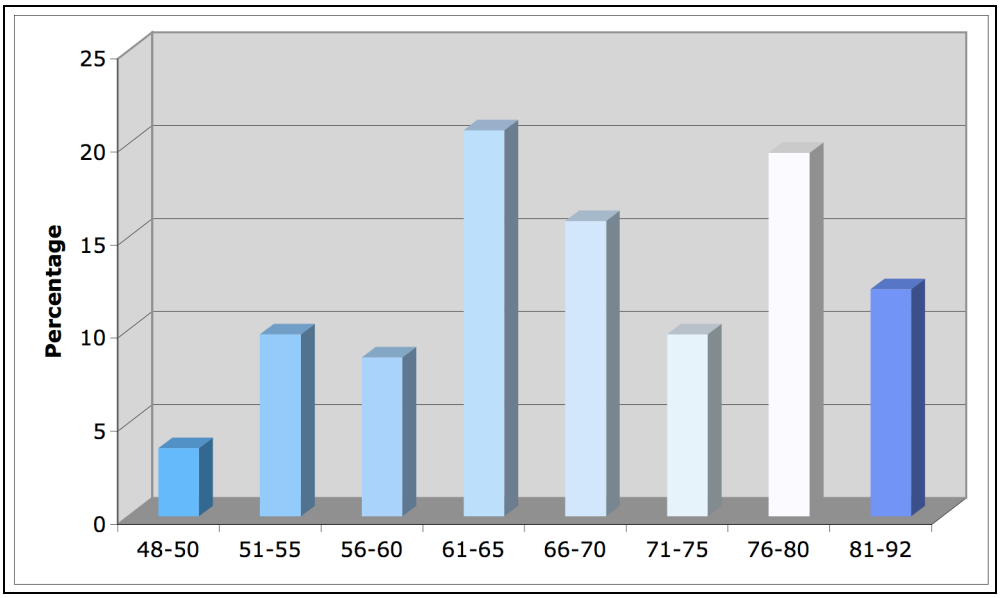
Variable		Odds Ratio	p-value
Gender		2.533	0.044*
Age		1.000	0.993
Smoking		0.583	0.353
Diabetes		0.731	0.552
Family history	CRC	1.327	0.599
	Any cancer	0.285	0.043*
Fasting		0.345	0.032*
Symptoms	GI bleed	1.316	0.548
	Change bowel habits	1.446	0.412
Location of cancer	Rectal vs. colon	0.545	0.222
	Left vs. right	1.503	0.390
Pathology of Cancer	Lymphatic	0.530	0.206
	Vascular	0.685	0.598
	Perineural	1.088	0.902
	Lymphocytic	0.873	0.795
	Grade	ns	ns
Cancer stage	Stage 2 vs. 1	0.500	0.396
	Stage 3 vs. 1	0.266	0.071
	Stage 4 vs. 1	0.194	0.046*
CEA (>5 μg/L vs. <=5 μg/L)		0.441	0.149

* $p \leq 0.05$

Demographics

In the CRC group, the odds of resembling or overlapping with the normal group for males was 2.53 (95% CI 1.02, 6.26) times that of females (p=0.04). That is, again, females were more different than the normals compared to the males. Age did not change the odds of overlapping with normals, OR 1.00 (95% CI 0.96, 1.04). Even when age was subdivided into 5-year categories (figure 4.13), there was no one category that statistically significantly predicts overlapping with normals.

Figure 4.13: Age of CRC patients divided by category



Smoking

Within the CRC group, being a smoker or an ex-smoker did not increase the odds of overlapping with the normal group. When ex-smokers were classified together with non-smokers, there was still no statistically significant difference (OR=0.58; 95% CI=0.19, 1.82; p=0.353), but again, since there were only 15 smokers, there might not be enough power to show this difference.

Diabetes

CRC patients who have diabetes did not have increased odds of overlapping with the normal group (OR=0.73; 95% CI=0.26, 2.05; p=0.552), although there were only 19 out of 82 subjects who had diabetes.

Family history

Within the CRC group, there were 20 patients who have a positive family history of CRC and this did not increase the odds of overlapping with the normal group. However, having a family history of *any* cancer(s) had 0.21 (p=0.02) the odds of overlapping with the normal group compared to those that did not have a family history. That is, the metabolomic fingerprint of those CRC patients that have a family history of any neoplastic process had statistically significant increased odds of being different than that of the normal group.

Fasting

Since some of the CRC patients were recruited soon after their colonoscopy, they were still in the fasting state. Within the CRC group, the odds of overlapping with the normal group in those that were fasting was 0.28 times (95% CI=0.11, 0.76; p=0.012) that of those that are not fasting. That is, the metabolomic fingerprint of those that were fasting were more different than the fingerprint of the normals (not fasting).

Gastrointestinal Symptoms

Within the CRC group, having symptoms of gastrointestinal bleeding (OR=1.23, 95% CI=0.50, 3.03; p=0.656) or changes in bowel habits (OR=1.45; 95% CI=0.60, 3.50; p=0.412) did not increase the odds of overlapping with the normal group.

Location of CRC

When colon and rectal cancer were separated and compared in their odds of overlapping with normal, the rectal cancers were more likely to be different than the normals, but this difference was not statistically significant (OR=0.55; 95% CI=0.21, 1.44; p=0.222). There was also no significant difference when left-sided (including transverse, descending colon, sigmoid, and rectal) cancers were compared to right-sided ones (OR=1.50; 95% CI=0.59, 3.81; p=0.390).

Cancer Pathology

Pathological characteristics of the CRC specimens such as lymphatic invasion, vascular invasion, perineural invasion, and lymphocytic response, grade, and stage, were examined to see if there was any that contributed to overlapping with normal. None of these characteristics significantly predicted overlap with the normal group. The TMN stage of the colorectal cancer was also analyzed and it was found that the higher the stage (*i.e.* more advanced the cancer), the more likely it was to be different than the normal group. Compared to stage 1 CRC, stage 4 or metastatic CRC was one-fifth times as likely to overlap with the normal group (OR=0.194, 95% CI =0.101, 2.477; p=0.046), that is, metastatic CRC was more than five times as likely to be different than the normals compared to stage 1 CRC.

CEA

Of the 82 CRC patients, 63 had CEA measured prior to their surgery. The highest CEA value was 4669.7 μg/L. As the distribution of CEA was quite skewed, the values were categorized into those within the normal range (less or equal to 5 μg/L) and those that are abnormal (above 5 μg/L). There were 45 people in the normal CEA category and 37 in the abnormal CEA category. This reiterates that CEA was not an accurate biomarker for CRC and should not be used for screening purposes. When the two categories of CEA were tested for their significance on the dichotomous outcome of overlapping or non-overlapping with normals, the abnormal CEA group had more than twice the odds of being different than the normals, but this was not statistically significant (OR=0.44; 95% CI=0.15, 1.34; p=0.149).

Summary

In summary, *gender*, *family history of any cancer*, and *fasting* were statistically significant differences between the overlapping and non-overlapping CRC subgroups. When these factors were tested with an overall logistic regression test, they were still all statistically different. In fact, the effect size is even bigger when all the variables were in the model, so ideally we should do 4-way or 6-way stratification. However this would reduce the number of samples even more, thus making it hard to use projection-based methods to analyze the data.

4.5.5.3 Stratification

The above method of model analysis allowed us to narrow down the specific clinical characteristics by which to stratify the models – these were gender and family history of any cancer. We also chose to stratify by family history of CRC as this is a known risk factor for CRC[6]. We could not stratify based on fasting as nobody in the normal group was fasting, but we did eliminate those in the CRC group that were fasting to see if this would change the model significantly.

4.5.5.3.1 Gender

Separate male and female OPLS models were built for normal vs. CRC and their model characteristics are shown in table 4.9. It was clear that the female model had a numerically better R^2Y, a comparable Q^2 value, and a numerically better AUC compared to the main normal vs. CRC model developed from the training set. Thus the urine metabolomics test for CRC should work better for females, although it was unclear whether the differences were statistically or clinically significant. The scatter plots for the male and female models are shown as figure 4.14 and 4.15, respectively.

Table 4.9: Characteristics of gender-stratified models compared to the main normal vs. CRC model

Model	R^2Y	Q^2	Sens	Spec	AUC	AUC 95% CI
Main Normal vs. CRC Training Set Model	0.478	0.355	93%	72%	0.9314	0.9017, 0.9611
Male Model	0.437	0.269	94%	71%	0.9394	0.9017, 0.9772
Female Model	0.564	0.346	94%	73%	0.9588	0.9243, 0.9934

Figure 4.14: OPLS scatter plot of the male model of normal (blue diamonds) vs. CRC (pink squares)

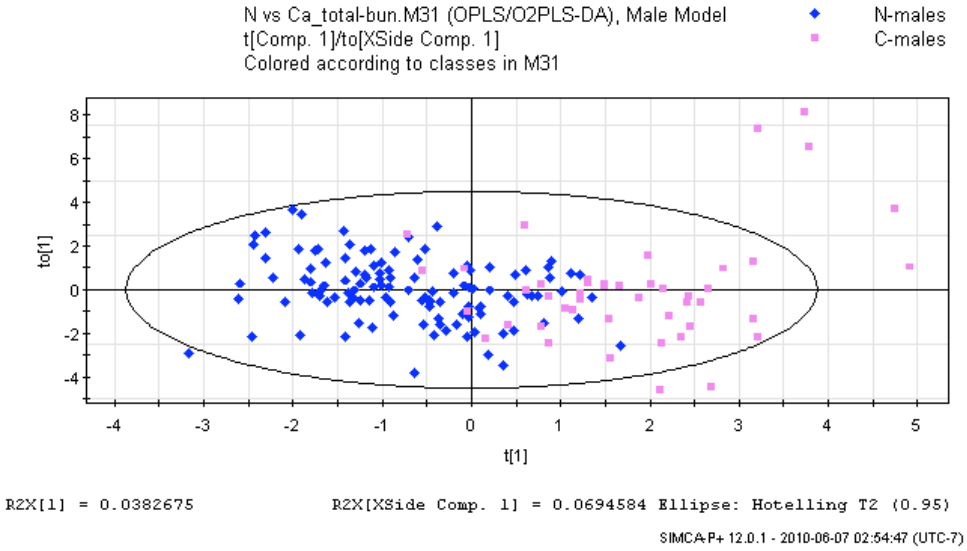
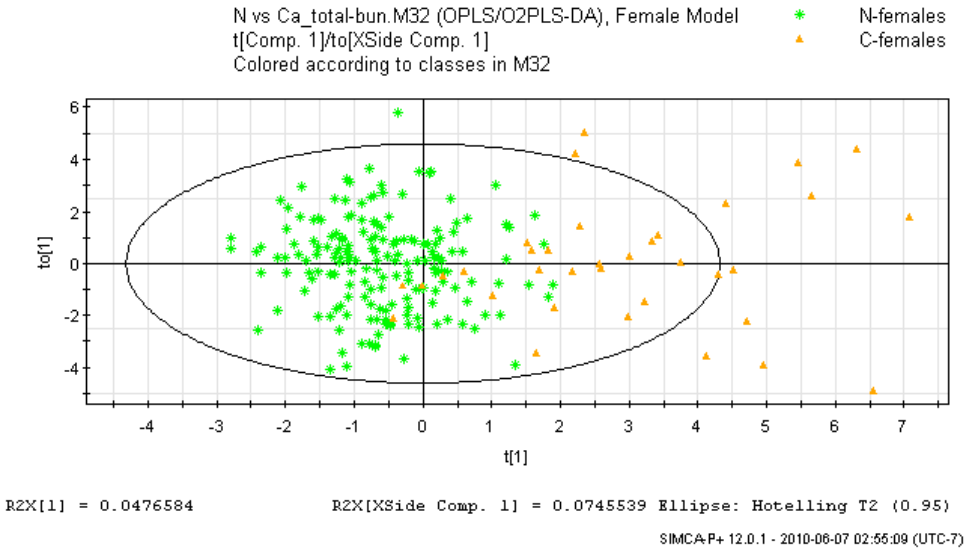


Figure 4.15: OPLS scatter plot of the female model of normal (green diamonds) vs. CRC (orange triangles)



4.5.5.3.2 Family History of Any Cancer

The normal and CRC patients were also stratified by family history of any cancer and one OPLS model was built for those with a positive family history of any cancer and one for those without a family history of any cancer. The model characteristics are listed in table 4.10. The positive family history model had numerically better R^2Y, Q^2, and AUC values compared to the main normal vs. CRC model developed from the training set and the no/unknown family history of cancer model had a numerically higher $R^2Y$and AUC but a much lower Q^2 compared to the main normal vs. CRC model. This is overall suggestive that stratifying by family history of cancer can increase the accuracy of this screening urine metabolomic test, although again, the statistical and clinical significance of the differences in the model characteristics are unclear. The scatter plots for the stratified models are shown as figure 4.16 and 4.17, respectively.

Table 4.10: Characteristics of models stratified by family history of any cancer compared to the main normal vs. CRC model

Model	R^2Y	Q^2	Sens	Spec	AUC	AUC 95% CI
Main Normal vs. CRC Training Set Model	0.478	0.355	93%	72%	0.9314	0.9017, 0.9611
No/Unknown family history of any cancer	0.638	0.146	96%	86%	0.9776	0.9532, 1.000
Family history of any cancer	0.533	0.416	93%	81%	0.9525	0.9246, 0.9805

Figure 4.16: Normal (orange triangles) vs. CRC (pink squares) OPLS scatter plot of the no/unknown family history of any cancer groups

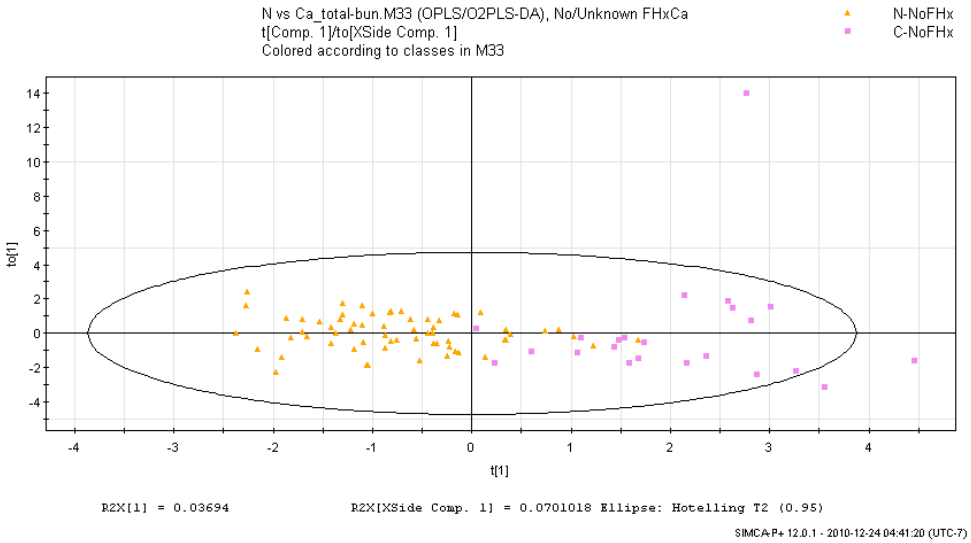
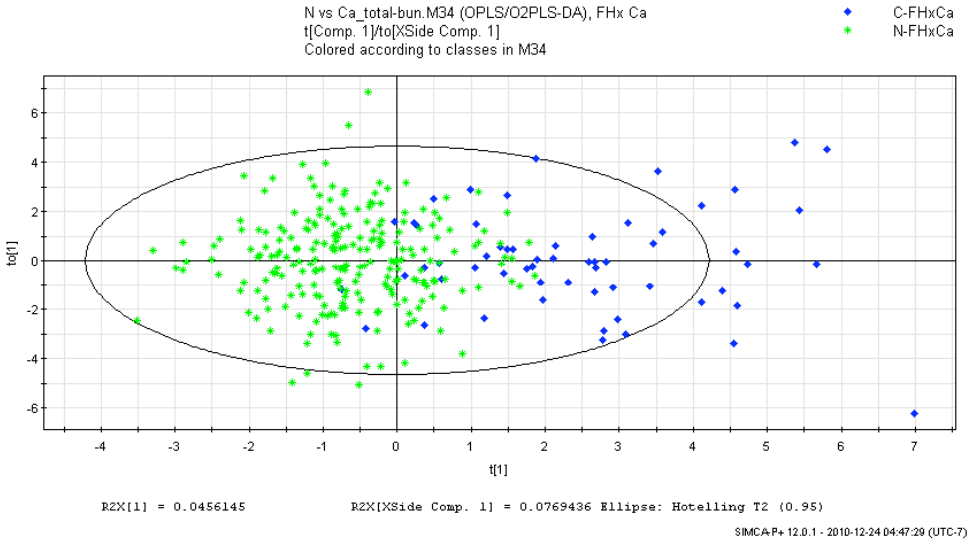


Figure 4.17: Normal (green stars) vs. CRC (blue diamonds) OPLS scatter plot of the positive family history of any cancer groups



4.5.5.3.3 Family History of Colorectal Cancer

The normal and CRC patients were also stratified by family history of CRC and one OPLS model was built for those with a positive family history of CRC and one for those without a family history of CRC. The model characteristics are listed in table 4.11. The no/unknown family history model had better R^2Y, comparable Q^2, and higher AUC values compared to the main normal vs. CRC model developed from the training set and the positive family history of CRC model had a comparable $R^2Y$and higher AUC but a lower Q^2 compared to the main normal vs. CRC model. Thus, stratifying by family history of CRC could increase the accuracy of this screening test, albeit by a small amount. Again, it was unclear whether the numerical differences in model characteristics had any statistical or clinical significance. The scatter plots for the stratified models are shown as figure 4.18 and 4.19, respectively.

Table 4.11: Characteristics of models stratified by family history of CRC compared to the main normal vs. CRC model

Model	R^2Y	Q^2	Sens	Spec	AUC	AUC 95% CI
Main Normal vs. CRC Training Set Model	0.478	0.355	93%	72%	0.9314	0.9017, 0.9611
No/Unknown family history CRC	0.587	0.350	95%	79%	0.9616	0.9377, 0.9856
Family history of CRC	0.479	0.267	95%	94%	0.9804	0.9641, 0.9967

Figure 4.18: Normal (black squares) vs. CRC (red diamonds) OPLS scatter plot of the no/unknown family history of CRC groups

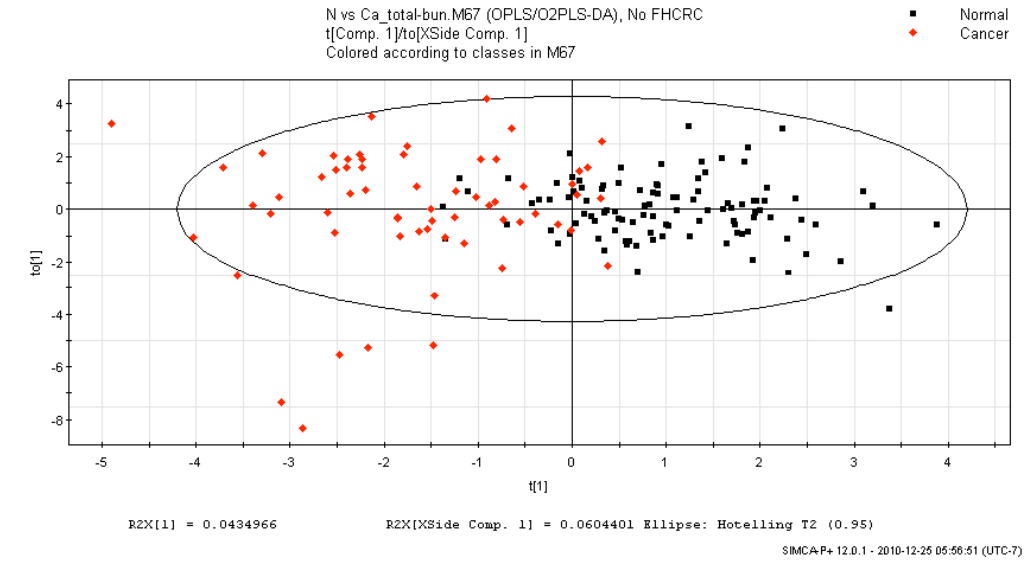
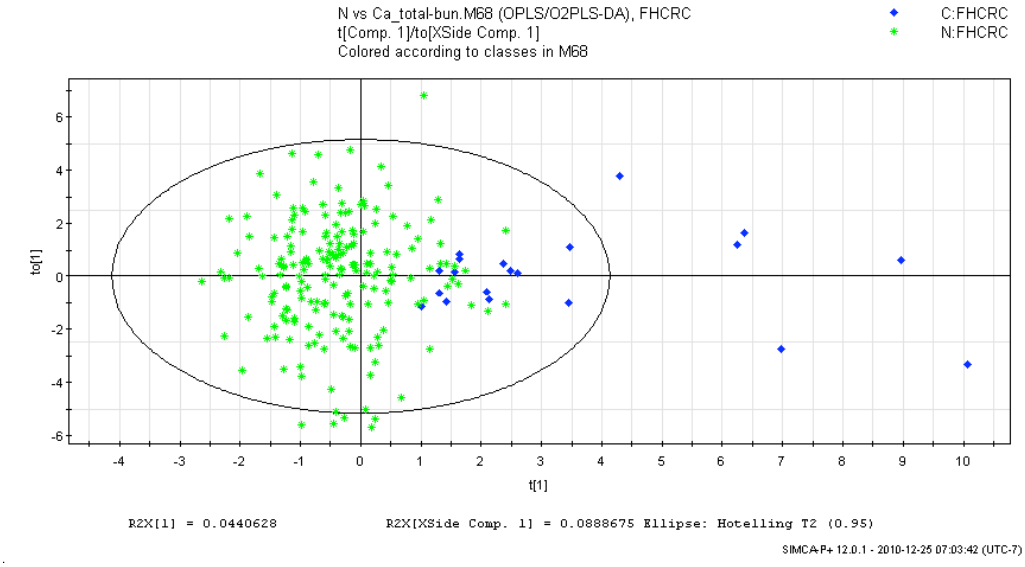


Figure 4.19: Normal (green stars) vs. CRC (blue diamonds) OPLS scatter plot of the positive family history of CRC groups



4.5.5.3.4 Fasting

Since fasting at the time of urine collection, compared to non-fasting, in CRC patients increased the metabolomic profile difference relative to the profiles of the normals (not fasting), some of the separation of normal and CRC may be overestimated. We excluded the 28 CRC patients that were fasting and the OPLS model was rebuilt with the normal group vs. the 54 non-fasting CRC patients, the model characteristics are summarized in table 4.12. The sensitivity and specificity values of the non-fasting model were not that different from the main normal vs. CRC model but as expected, the separation of the two groups was not as good as in the main normal vs. CRC model, i.e. the R^2Y and Q^2 values are numerically lower. It is unclear whether this is statistically or clinically significant. The differences in the model characteristics may be due to the fact that some of the previously seen separation between the two groups was actually from fasting rather than from the disease state of CRC, but this effect may also be due to a smaller number of CRC patients in this new model. To resolve this, a bigger model of non-fasting CRC patients needs to be built and analyzed.

Table 4.12: Main normal vs. CRC model compared to the normal vs. non-fasting CRC model

Model	R^2Y	Q^2	Sens	Spec	AUC	AUC 95% CI
Main Normal vs. CRC Training Set Model	0.478	0.355	93%	72%	0.9314	0.9017, 0.9611
Normal vs. Non-Fasting CRC	0.436	0.295	93%	76%	0.9365	0.9016, 0.9714

4.5.6 Validation with Testing Set

The robustness of the metabolomics model is reflected by how well it predicts unknowns. Following our development of the metabolomics model with the un-blinded training set (see above sections 4.5.2) we next used the blinded testing set of 184 urine samples (34 cancer, 150 normal) to validate the ability of the metabolomics model to distinguish normal from CRC patients.

To demonstrate that the testing set samples were matched and representative of the training set, the clinical characteristics of the training set normal subjects were compared to those of the testing set normal subjects (table 4.13). Other than family history of any cancer (92% vs. 77%, $p<0.001$) and changes in bowel habit (4% vs. 0%, $p=0.009$), there were no statistically significant differences between the training and the testing set. The CRC patients of the training set were also compared to the CRC patients of the testing set (table 4.14). Again, other than family history of any cancer (79% vs. 53%, $p=0.015$), the two groups were fairly similar to each other.

Table 4.13: Clinical characteristics of ***normal*** subjects in training vs. testing set

		Training Set [n=294] N (%)	**Testing Set [n=150] N (%)**	**p-value**
Male:Female		117:177	71:79	0.128
Average age (years ± SEM)		55.3±0.5	56.4±0.7	0.148
Smoking		26 (9)	15 (10)	0.802
Diabetes		14 (5)	10 (7)	0.401
Family History	CRC	191 (69)	91 (65)	0.386
	Any cancer	230 (92)	115 (77)	<0.001*
Symptoms	GI bleed	5 (2)	2 (1)	0.772
	Change bowel habits	13 (4)	0 (0)	0.009*

Note: Not all % are calculated with the denominator of the total in each group as some clinical information was missing or unknown. * $p \leq 0.05$

Table 4.14: Clinical characteristics of ***CRC*** patients in training vs. testing set

			Training Set (n=82) N (%)	Testing Set (n=34) N (%)	p-value
Male:Female			47:35	23:11	0.300
Average age (years ± SEM)			68.6±1.2	67.9±2.0	0.756
Smoking			15 (20)	4 (12)	0.322
Diabetes			19 (23)	8 (24)	0.967
Family history	CRC		20 (25)	7 (21)	0.668
	Any cancer		60 (79)	18 (53)	0.015*
Symptoms	GI bleed		49 (60)	22 (65)	0.619
	Change bowel habits		43 (52)	19 (56)	0.735
Location of cancer	Rectal vs. colon		23 (28)	7 (21)	0.404
	Left vs. right		55 (67)	21 (62)	0.584
Pathology of Cancer	Lymphatic		26 (35)	6 (21)	0.183
	Vascular		9 (12)	4 (14)	0.774
	Perineural		11 (15)	2 (7)	0.268
	Lymphocytic		36 (56)	11 (48)	0.487
	Grade	Well	62 (84)	24 (80)	0.644
		Moderate	3 (4)	3 (10)	0.239
		High	9 (12)	3 (10)	0.755
Cancer stage	Stage 1		15 (18)	6 (18)	0.989
	Stage 2		18 (22)	11 (33)	0.204
	Stage 3		33 (40)	8 (24)	0.105
	Stage 4		16 (20)	8 (24)	0.572
CEA (>5 vs. <=5)			18 (29)	10 (40)	0.299

Note: Not all % are calculated with the denominator of the total in each group as some clinical information was missing or unknown. * $p \leq 0.05$

Diagnostic accuracies were calculated using the same cutoff (0.212925) from the original model that resulted in a sensitivity of 92.7% and specificity of 71.8%. The sensitivity and specificity from the validation samples were 85.3% and 52.7%, respectively.

When only the testing *CRC* samples were introduced as the prediction set to the metabolomic normal vs. CRC model developed from the training samples (figure

4.20), these testing CRC samples showed up on cancer side of the model (figure 4.21). However, when the 150 testing set *normal* samples were introduced blindly to the original normal vs. CRC model developed from the training set, the samples scatter on both sides of the plot (figures 4.22), which was as expected with the calculated validation specificity of 52.7%.

Figure 4.20: Original normal (black squares) vs. CRC (red diamonds) OPLS scatter plot

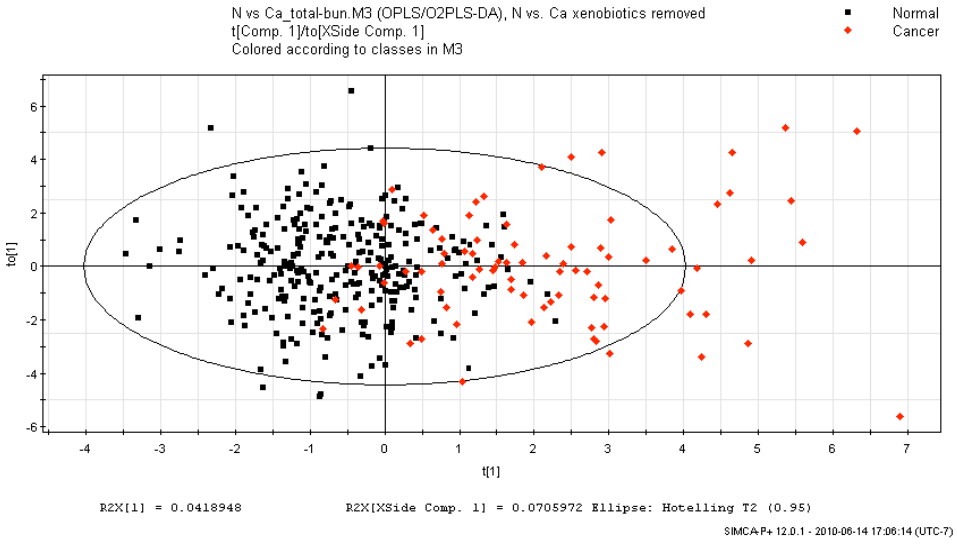


Figure 4.21: Testing set CRC samples (blue squares) superimposed on normal (black triangles) vs. CRC (red diamonds) OPLS scatter plot

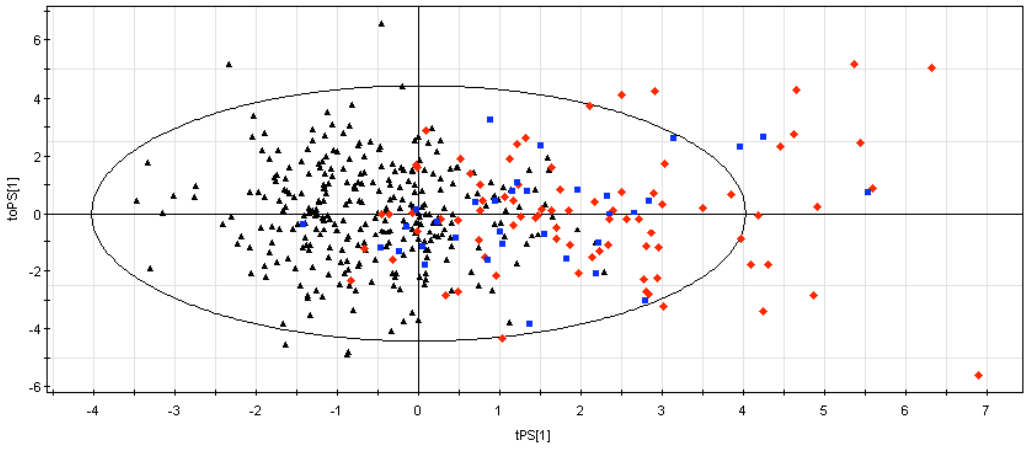
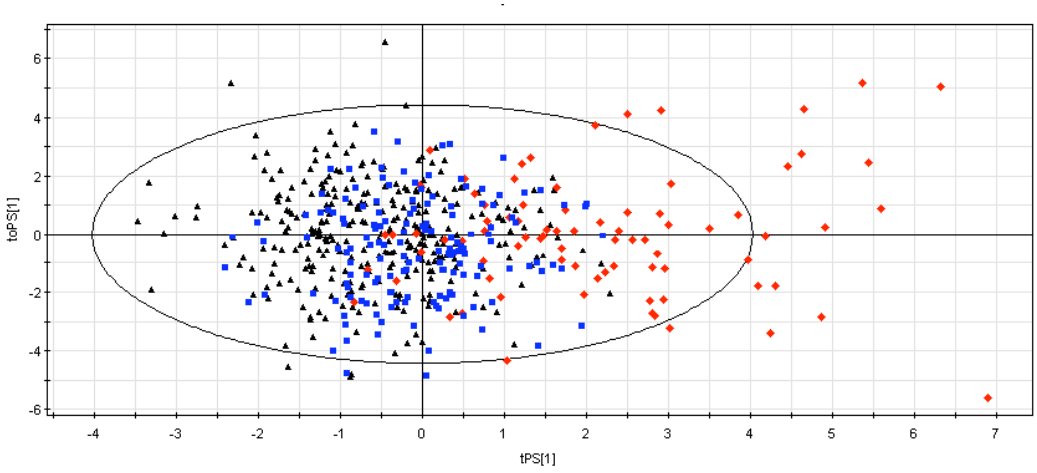


Figure 4.22: Testing set normal samples (blue squares) superimposed on the normal (black triangles) vs. CRC (red diamonds) OPLS scatter plot



4.5.7 Commercialization

To commercialize urine metabolomics as a screening test for CRC, accuracy of the test is very important, but for it to become a population-based test, the cost needs to be reasonable. The normal vs. CRC model is currently built using 69

metabolites, but if we can produce a model of acceptable diagnostic accuracies using fewer metabolites, then the cost of the test would be much lower. From the variable importance plot of the current model, we know, in order of importance, the metabolites that contribute the most in the separation of the two groups. We can capitalize on this by taking just the top metabolites and see what kind of models we can build with them. The results are summarized in table 4.15. Using the concentrations of only the top 10 metabolites, namely hypoxanthine, creatinine, dimethylamine, 3-indoxylsulfate, methanol, adipate, urea, guanidoacetate, 3-hydroxybutyrate, and acetone, a reasonable OPLS model can be built (R^2Y of 0.356, Q^2 of 0.343) with sensitivity and specificity of 86.6% and 75.5% respectively, and an AUC of 0.8723.

Table 4.15: Model characteristics and diagnostic accuracies of OPLS models built with top contributing metabolites

Model	R^2Y	Q^2	Sens	Spec	AUC
Main CRC Model	0.478	0.355	92.7%	71.8%	0.9314
Top 5 metabolites	0.27	0.256	85.4%	67.7%	0.8362
Top 10 metabolites	0.356	0.343	86.6%	75.5%	0.8723
Top 15 metabolites	0.351	0.337	86.6%	75.2%	0.8786
Top 20 metabolites	0.402	0.349	89.0%	74.5%	0.8988
Top 25 metabolites	0.437	0.381	89.0%	74.8%	0.9150
Top 30 metabolites	0.448	0.387	91.5%	64.3%	0.9174

To ensure that the top metabolites in the variable importance plot were indeed more important in establishing the metabolomic fingerprint of colorectal cancer, we validated the process above by attempting to build models using the bottom metabolites in the VIP list. It took 50 metabolites before we could even build an OPLS model, and it was a poorly predictive one (R^2Y of 0.224 and Q^2 of only 0.057). This validates the uniqueness of top metabolites as a diagnostic tool.

4.6 Discussion

4.6.1 Summary

Using projection-based methods, the role of urine metabolomics in distinguishing normal subjects vs. colorectal cancer patients was examined in this chapter. PCA was unable to separate the two groups but since each PCA model is generated based on the direction in the data demonstrating the highest variation, i.e. gender, age, diet, lifestyle, genes, unknown factors, etc. which might be distinctly different from the direction separating the classes.[7, 8], this is not surprising.

The training set of 294 normal subjects and 82 CRC patients was used to build the OPLS model ($R^2Y = 0.478$, $Q^2 = 0.355$), which was internally validated using permutation testing. Sensitivity and specificity were calculated to be 92.7% and 71.8% respectively. The area under the curve was 0.9314. Although the normal and the CRC patients were projected on different sides of the scatter plot, there was some overlap. Each group was then subdivided into overlapping and non-overlapping groups and analyzed to determine the best clinical characteristics by

which to stratify the groups to improve the model. When stratified by gender, family history of any cancer, and family history of CRC, the diagnostic accuracies improved. The main normal vs. CRC model developed from the training set was then externally validated with a blinded testing set of similar clinical characteristics and diagnostic accuracies of 85.3% and 52.7% were achieved for sensitivity and specificity, respectively. Lastly, models were built using only the top metabolites and diagnostic accuracies were calculated. Using only the top 10 metabolites, sensitivity of 86.6% and specificity of 75.5% could be achieved, suggesting commercialization potential for this test.

4.6.2 Patient and Disease Characteristics

It is obvious that the normal group was different than the CRC group in baseline characteristics. Since the normal subjects were recruited from a screening program and the CRC patients were recruited after they had a diagnosis of CRC, it is not surprising that the two groups were different in many ways, including gender, age, family history of CRC, family history of any cancer, smoking, diabetes, and GI symptoms. However, each of these factors was analyzed statistically to see if they falsely contribute to the separation of the two groups and they didn't; in fact, the diagnostic accuracies improved when the model was stratified by gender, family history of any cancer and family history of CRC.

Sixty percent of the CRC patients were diagnosed with stage 3 or 4 CRC, that is, the lymph nodes were already involved or there was evidence of distance

metastasis, but the screening pilot study (SCOPE) only identified 2 malignant polyps in 1200 asymptomatic subjects. This is partially explained by the fact that the CRC patients are a lot older than the screening group and most of them had not undergone regular screening in the past. About one-third of the CRC patients had right-sided cancer that would have been missed if flexible sigmoidoscopy was used as a screening test.

4.6.3 Model Characteristics

The goodness of an OPLS model is represented by R^2 and Q^2. As mentioned previously, R^2Y is the percent of variation of Y explained by the model and thus is a measure of how well the model fits the data. Q^2 is the percent of variation of Y predicted by the model according to 7-fold cross validation and thus is a measure of how well the model predicts new data. A large R^2Y (close to 1) is necessary condition for a good model and a large Q^2Y ($Q2Y > 0.5$) indicates good predictivity.[9] The R^2Y for this study was 0.478 and the Q^2Y was 0.355.

It can be difficult to produce high R^2Y and Q^2Y values in human studies due to the high variability between individuals. Although a couple of metabolomic studies in the literature were able to generate higher R^2Y and Q^2Y values than the current study, there were potential factors to suggest over fitting in those cases. For example, Chan et al. 2009[10] was able to produce an OPLS model of normal vs. with CRC with R^2Y and Q^2Y of 0.622 and 0.518, respectively. However, the number in each group was 22 and 25 only and yet there were 1101 variables

analyzed. It is always easy to find random multivariate correlations when the number of variables greatly exceeds the number of samples.[11] Qiu et al. 2010[12] was able to construct a normal vs. CRC OPLS model with R^2Y and Q^2Y of 0.763 and 0.467 respectively. Again their sample sizes for the groups were 60 and 63, but 187 metabolites were analyzed.

Table 4.16 illustrates the R^2Y and Q^2Y values of different models that can be built using our data and demonstrates that higher R^2Y and Q^2Y values can be obtained with smaller sample sized models. However, larger sample sizes are more representative of the mean and variance of the population and therefore the models built using our original sample size is a more accurate reflection of reality.

Table 4.16: Model characteristics and diagnostic accuracies of OPLS models built with top contributing metabolites

Model (total n)	# Normals	# CRC	R^2Y	Q^2
Main CRC Model (376)	294	82	0.478	0.355
20	10	10	0.975	0.688
40	20	20	0.870	0.652
60	30	30	0.651	0.463
90	60	30	0.761	0.516
120	90	30	0.705	0.470
150	90	60	0.734	0.579
164	82	82	0.728	0.603

4.6.4 Diagnostic Accuracies

Urine metabolomics is being investigated for its role as a screening test for CRC. It is not meant to replace the gold standard colonoscopy but rather to replace current non-invasive fecal tests. For screening tests in general, a high sensitivity is more important than a high specificity, since a falsely negative result (missed cancer) is of much more consequence than a falsely positive result (unnecessary colonoscopy) for the patient. Specificity is obviously important as well since too many unnecessary interventions will be costly to society, not to mention the patient anxiety that a false positive test can cause. For this reason, we have chosen the diagnostic accuracies to reflect the highest sensitivities but with a

specificity of at least 50%. Area under the receiver operating characteristics curve (AUC) of > 0.9 represents an excellent test, while a value over 0.8 is still likely to be good.[11] In our case, the main normal vs. CRC model had an AUC of 0.9314.

4.6.5 Metabolites and Metabolic Pathways

A more careful analysis of the top metabolites that drive the separation of normal and CRC and the underlying biochemical pathways involved in generating these metabolites may shine some light on the pathogenesis of CRC, or at least how human metabolism is affected by CRC.

The top 10 metabolites that separated the normal group from the CRC group are hypoxanthine, creatinine, dimethylamine, 3-indoxylsulfate, methanol, adipate, urea, guanidoacetate, 3-hydroxybutyrate, and acetone. Mapping the differential metabolites to their respective biochemical pathways as outlined in the Kyoto Encyclopedia of Genes and Genomes (KEGG, http://www.genome.jp/kegg/) revealed alterations mostly in nucleotide, amino acid, and microbial metabolisms.

Hypoxanthine (increased in CRC) is a central intermediate in purine nucleotide biosynthesis and is the extra cellular compound most directly related to intracellular ATP. Increases in hypoxanthine in biofluids have been associated with ATP depletion[13], and may underline the deteriorating state of the energy level in CRC patients.

Creatinine (increased in CRC), is a degradation product of creatine, a phosphorylated molecule specific to muscle energy metabolism and was found to be increased in the urine of patients with cancer and cachexia[14]. Creatinine is also involved in arginine and proline metabolism, along with urea (increased in normal), and guanidoacetate (increased in normal). In addition, urea is involved in purine and pyrimidine metabolism, and microbial metabolism in diverse environments[15]. Guanidoacetate is also a precursor of creatine, an essential substrate for muscle energy metabolism.

Dimethylamine (increased in CRC) is produced by degradation of dietary choline to trimethylamine, which is subsequently converted to dimethylamine by gut microflora.[16, 17] Methanol (increased in normal) is also a product of microbial metabolism.[15]

3-indoxylsulfate is increased in the CRC group. It is a dietary protein metabolite, and also the metabolite of the common amino acid tryptophan.[18] Up-regulation of tryptophan metabolism in patients with CRC has been suggested in the literature.[12] 3-indoxylsulfate also strongly decreases the levels of glutathione, one of the most active antioxidant systems of the cell[19], thus suggesting a possible mechanism of pathogenesis for CRC.

3-hydroxybutyrate is a ketone body that is raised in ketosis. It is involved in butanoate (carbohydrate) metabolism and synthesis and degradation of ketones. This metabolite is increased in CRC patients and this is not due to the fact that some CRC patients were fasting at the time of urine collection. In fact, when the fasting CRC patients were excluded and the model was rebuilt with just those that were not fasting, 3-hydroxybutyrate moved up in the ranks to number three in order of importance in separating normal and CRC. This is also seen for acetone. The increase in 3-hydroxybutyrate and acetone in CRC patients cannot be explained by diabetes either. Exploratory analysis revealed that the patients with increased 3-hydroxybutryrate and acetone are not those with diabetes and vice versa. Acetone is also typically derived from acetoacetate through the action of microbial acetoacetate decarboxylases found in gut microflora such as *Clostridium acetobutylicum*.[20]

Adipate is particularly interesting. It is a food additive and acidity regulator and was thought to be part of a probable carcinogenic DEHA (Di (2-ethylhexyl) Adipate) as DEHA caused liver tumors in mice.[21] In our subjects, adipate was present in 23 out of 82 (28.0%) CRC patients but only in 7 out of 294 (2.4%) normal subjects. The highest adipate level in one CRC patient was 13536 μM, while that in the normal group was 64 μM. According to KEGG[15], adipate is involved in microbial metabolism in diverse environments. It is unclear whether in CRC, bacteria in the body are producing more adipate or that CRC is in part caused by increased adipate in the body.

β-alanine, which is the 12th metabolite that separates normal from CRC, is another metabolite that has a different distribution in the normal group compared to the CRC group. It is found in 61 out of 294 or 20.7% of the normal subjects and only 2 out of 82 or 2.4% of CRC patients. β-alanine is an amino acid formed in vivo by the degradation of dihydrouracil and carnosine. It is also involved in pyrimidine metabolism, propanoate metabolism, and pantothenate (vitamin B5) and CoA biosynthesis.[15] β-alanine is a rate-limiting precursor of carnosine. Muscle carnosine is increased with β-alanine supplementation, which also results in decreased fatigue in athletes and increase in total muscular work done.[22, 23] It is unclear whether the lack of β-alanine in CRC patients is a cause or a result of CRC. If the lack of β-alanine is a result of CRC, then this could potentially represent a mechanism of cancer-related fatigue.

Overall, in the limited literature on urine metabolomics and CRC, suggestions of increased tryptophan metabolism and altered gut microflora metabolism in patients with CRC were also observed in this study. In fact, many of the metabolites that separate normal from CRC (dimethylamine, methanol, adipate, urea, and acetone) are products of microbial metabolism. There have been many studies in the literature demonstrating the role of gastrointestinal microbiota in colorectal cancer by production of toxic and genotoxic bacterial metabolites.[24-26] As there are many more bacteria cells than human cells in the colon, it is not

surprising that the differences in the metabolites of normal versus CRC patients are reflections of different microbiota rather than human metabolism.

The top contributing metabolites identified in this study, however, were different than those identified in the only other study in the literature on urine metabolomics and colorectal cancer.[12] This may be due to several differences between the two studies. First Qiu's study used GC-MS to quantify the metabolites and our study used NMR and consequently the library of compounds to identify the metabolite peaks were different as well. In addition, Qiu's study only identified and structurally validated 40% of the differential variables detected. Secondly, the populations of patients were different between the two studies. All of Qiu's samples were collected from people in Shanghai, whereas the samples for this study were collected from subjects in Northern Alberta (Edmonton and Grande Prairie). The differences in ethnicity, climate, and diet can significantly change the metabolite profile of individuals. Thirdly Qiu's controls were healthy volunteers, whereas the controls for this study are colonoscopy-negative individuals. Also, since there were more variables (187) than subjects (123 total; 63 control and 60 normal) in Qiu's study, there could be some random multivariate correlations.

4.6.6 Limitations

There are some limitations to this study. There were several occasions where information provided by the patient on the questionnaires was inconsistent with

what was identified in their medical records. In such cases and where appropriate, the information regarding patients' past medical history, family history and medications was taken from the patients' charts, as this was likely a more reliable source. However information such as family history was only as good a patients' reporting. Additional information not provided by the patients but found on their medical charts were supplemented to the database. There can be bias/misinterpretations in filling out the questionnaires.

The number of metabolites analyzed from the urine samples is limited by the number of metabolites contained in the Chenomx compound library. As Chenomx is continually expanding their library, some of the newer metabolites would have been missed in the older analyses.

Some of the CRC patients were fasting at the time of giving the urine sample. This was because these patients were identified and recruited either right after having a colonoscopy to confirm CRC or right before surgery. The state of fasting affects one's metabolic fingerprint, thus some of the separation between normal and CRC may be falsely due to separation between non-fasting and fasting. This was examined by building a separate model with the fasting CRC patients excluded. This new model did have lower R^2 and Q^2 values but it is unclear whether this is also due to a smaller sample size. This can be further investigated as more CRC patients are recruited and analyzed.

There were some differences in baseline characteristics between the training set and the testing set. This was because the testing set samples were not methodologically matched to the training set samples, but were simply subsequently collected samples. To improve the robustness of this model, the validation set should be comparable to the training set. This can be done by increasing the number of validation samples or by matching the baseline characteristics of the testing set to those of the training set.

4.6.7 Bias

4.6.7.1 Disease Progression Bias

Urine samples are collected at the education session for the normal group and the pre-admission clinic for the CRC group. Ideally the results of the index test and the reference standard are collected on the same patients at the same time to minimize misclassification due to spontaneous recovery or to progression to a more advanced stage of disease (*disease progression bias*). However, it is unknown at this time whether the bowel preparation and the perioperative medications used for colonoscopy (SCOPE subjects) and surgery (CRC patients) affect one's metabolomic profile, therefore the study was designed to collect the urine while the subjects are in a natural state of hydration and health. Also, practically speaking, the yield of urine collection is higher with this study design. Moreover, disease progression bias may not be a problem for cancer as it would be for infectious diseases as cancer does not spontaneously recover or progress over the course of days.[27]

4.6.7.2 Misclassification

Differential misclassification can occur if a CRC is missed and if the urine metabolomics test is positive, then there's bias towards null, i.e. the difference is harder to detect. The polyp and CRC miss rate for colonoscopy is 2-6%. *Nondifferential misclassification* can result when other unknown cancers such as breast, ovarian etc. gives a positive test for urine metabolomics. This can happen in both controls and cases.[27]

4.6.7.3 Spectrum Bias

This study population is representative of the patients who will receive the test in practice, that is, asymptomatic average and high-risk individuals undergoing CRC screening. This is an advantage over existing studies in that it does not merely consist of a group of healthy controls and a group known to have the target disorder (*spectrum bias*).[27]

4.6.7.4 Partial Verification & Incorporation Bias

Partial verification bias does not exist as all study patients went on to receive confirmation of the diagnosis by the reference standard. *Incorporation bias* does not exist as the index test (urine metabolomics) was not used in establishing the final diagnosis.

4.6.7.5 Review Bias

Blinding was achieved as the index test was interpreted without the knowledge of the result of the reference standard test, and vice versa (i.e. no *review bias*). The index test was analyzed and interpreted in an independent laboratory.

4.6.8 Strengths of Study

This is the largest study to demonstrate that urine metabolomics can separate normal subjects from CRC patients. In addition to the large sample size of this study, a significant strength is that the normals or controls in this study have all had a colonoscopy and found to be normal – no inflammation or polyps. The other strength of this study over existing ones in the literature is the amount of clinical information such as family history, fasting state, and pathological features of the CRC, gathered and analyzed. This allows for detailed analysis of the variations seen in the metabolomic model and stratification based on clinical characteristics such as gender and family history. Existing CRC screening guidelines dictate different ages to start screening for those with a positive family history of CRC compared to those without[28] and several studies have suggested to develop sex-specific recommendations for CRC screening[29, 30]. The urine metabolomics test may be sex or age-specific as well.

Whenever possible, all subjects' urine samples were collected in their normal state of diet, hydration and activities etc. as to avoid confounding factors. This is reflective of the situation that this test would be used in the future if it becomes a population-based screening test.

Some common types of errors in design and analysis of metabolomics experiments have been addressed in this study.[11] The sample size is sufficient as

the number of samples in this study exceeds the number of variables otherwise it is a lot easier to find random multivariate correlations. Over fitting is avoided by using an independent/blind testing set which is held back from model optimization and used only to test the robustness of prediction in the final phase of the study. Potential biases have been discussed and confounding variables such as gender, smoking, family history etc. have been addressed.

4.7 Conclusions

With 655,000 deaths worldwide per year, CRC is the third leading cause of cancer-related death in the Western World. Current population-based fecal occult blood testing has low compliance and sensitivity. Urine metabolomics has been shown to distinguish healthy subjects from CRC patients with high accuracy and can represent a novel, highly sensitive, patient-accepted screening test for CRC.

4.8 References

1. "Colon and Rectal Cancer" National Cancer Institute: http://www.cancer.gov/cancertopics/types/colon-and-rectal, Accessed June 23, 2010.
2. Canadian Cancer Society: http://www.cancer.ca/Alberta-NWT/About cancer/Cancer statistics/Stats at a glance/Colorectal cancer.aspx?sc_lang=en&r=1, Accessed Nov 25, 2010.
3. Burch JA, Soares-Weiser K, St John DJB, Duffy S, Smith S, Kleijnen J, Westwood M. Diagnostic accuracy of faecal occult blood tests used in screening for colorectal cancer: a systematic review. Journal of Medical Screening 2007;14:132-137.
4. Whitlock EP, Lin JS, Liles E, Beil TL, Fu RW. Screening for Colorectal Cancer: A Targeted, Updated Systematic Review for the US Preventive Services Task Force. Annals of Internal Medicine 2008;149:638-+.
5. Chang D, Weljie A, Newton J. Leveraging latent information in NMR spectra for robust predictive models. Pacific Symposium on Biocomputing 2007 2007:115-126.
6. Wilschut JA, Habbema JDF, Ramsey SD, Boer R, Looman CWN, van Ballegooijen M. Increased risk of adenomas in individuals with a family history of colorectal cancer: results of a meta-analysis. Cancer Causes & Control 2010;21:2287-2293.
7. Bylesjo M, Rantalainen M, Cloarec O, Nicholson JK, Holmes E, Trygg J. OPLS discriminant analysis: combining the strengths of PLS-DA and SIMCA classification. Journal of Chemometrics 2006;20:341-351.
8. Trygg J, Holmes E, Lundstedt T. Chemometrics in metabonomics. Journal of Proteome Research 2007;6:469-479.
9. SIMCA-P+ v12.0.1 Analysis Advisor (Umetrics, Umea, Sweden).
10. Chan ECY, Koh PK, Mal M, Cheah PY, Eu KW, Backshall A, Cavill R, Nicholson JK, Keun HC. Metabolic Profiling of Human Colorectal Cancer Using High-Resolution Magic Angle Spinning Nuclear Magnetic Resonance (HR-MAS NMR) Spectroscopy and Gas Chromatography Mass Spectrometry (GC/MS). Journal of Proteome Research 2009;8:352-361.
11. Broadhurst DI, Kell DB. Statistical strategies for avoiding false discoveries in metabolomics and related experiments. Metabolomics 2006;2:171-196.
12. Qiu YP, Cai GX, Su MM, Chen TL, Liu YM, Xu Y, Ni Y, Zhao AH, Cai SJ, Xu LX, Jia W. Urinary Metabonomic Study on Colorectal Cancer. Journal of Proteome Research 2010;9:1627-1634.
13. Harkness RA. HYPOXANTHINE, XANTHINE AND URIDINE IN BODY-FLUIDS, INDICATORS OF ATP DEPLETION. Journal of Chromatography-Biomedical Applications 1988;429:255-278.
14. Eisner R, Stretch C, Eastman T, Xia JG, Hau D, Damaraju S, Greiner R, Wishart DS, Baracos VE. Learning to predict cancer-associated skeletal muscle wasting from H-1-NMR profiles of urinary metabolites. Metabolomics 2011;7:25-34.
15. Kyoto Encyclopedia of Genes and Genomes (KEGG): http://www.genome.jp/kegg/, Accessed Nov 2010 to Jan 2011.
16. Salek RM, Maguire ML, Bentley E, Rubtsov DV, Hough T, Cheeseman M, Nunez D, Sweatman BC, Haselden JN, Cox RD, Connor SC, Griffin JL. A metabolomic comparison of urinary changes in type 2 diabetes in mouse, rat, and human. Physiological Genomics 2007;29:99-108.
17. Asatoor AM, Simenhof.Ml. ORIGIN OF URINARY DIMETHYLAMINE. Biochimica Et Biophysica Acta 1965;111:384-&.
18. Human Metabolomic Database: http://www.hmdb.ca/metabolites/HMDB00682, Accessed Dec. 2010.
19. Dou L, Jourde-Chiche N, Faure V, Cerini C, Berland Y, Dignat-George F, Brunet P. The uremic solute indoxyl sulfate induces oxidative stress in endothelial cells. Journal of Thrombosis and Haemostasis 2007;5:1302-1308.
20. Moat AG, Foster JW, Spector MP. Microbial physiology. Fourth edition. Microbial physiology. Fourth edition 2002:i-xx, 1-715.

21. Kluwe WM, McConnell EE, Huff JE, Haseman JK, Douglas JF, Hartwell WV. CARCINOGENICITY TESTING OF PHTHALATE-ESTERS AND RELATED-COMPOUNDS BY THE NATIONAL TOXICOLOGY PROGRAM AND THE NATIONAL-CANCER-INSTITUTE. Environmental Health Perspectives 1982;45:129-133.
22. Derave W, Oezdemir MS, Harris RC, Pottier A, Reyngoudt H, Koppo K, Wise JA, Achten E. beta-alanine supplementation augments muscle carnosine content and attenuates fatigue during repeated isokinetic contraction bouts in trained sprinters. Journal of Applied Physiology 2007;103:1736-1743.
23. Hill CA, Harris RC, Kim HJ, Harris BD, Sale C, Boobis LH, Kim CK, Wise JA. Influence of beta-alanine supplementation on skeletal muscle carnosine concentrations and high intensity cycling capacity. Amino Acids 2007;32:225-233.
24. Rowland IR. The Role of the Gastrointestinal Microbiota in Colorectal Cancer. Current Pharmaceutical Design 2009;15:1524-1527.
25. Hope ME, Hold GL, Kain R, El-Omar EM. Sporadic colorectal cancer - role of the commensal microbiota. Fems Microbiology Letters 2005;244:1-7.
26. Scanlan PD, Shanahan F, Clune Y, Collins JK, O'Sullivan GC, O'Riordan M, Holmes E, Wang YL, Marchesi JR. Culture-independent analysis of the gut microbiota in colorectal cancer and polyposis (vol 10, pg 789, 2008). Environmental Microbiology 2008;10:1382-1382.
27. Whiting P, Rutjes AWS, Reitsma JB, Bossuyt PMM, Kleijnen J. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. BMC Med Res Methodol 2003;3:25.
28. Rex DK, Johnson DA, Anderson JC, Schoenfeld PS, Burke CA, Inadomi JM. American College of Gastroenterology Guidelines for Colorectal Cancer Screening 2008. American Journal of Gastroenterology 2009;104:739-750.
29. Lieberman D. Race, gender, and colorectal cancer screening. American Journal of Gastroenterology 2005;100:2756-2758.
30. Brenner H, Hoffmeister M, Arndt V, Haug U. Gender differences in colorectal cancer: implications for age at initiation of screening. British Journal of Cancer 2007;96:828-831.

4.9 Appendix

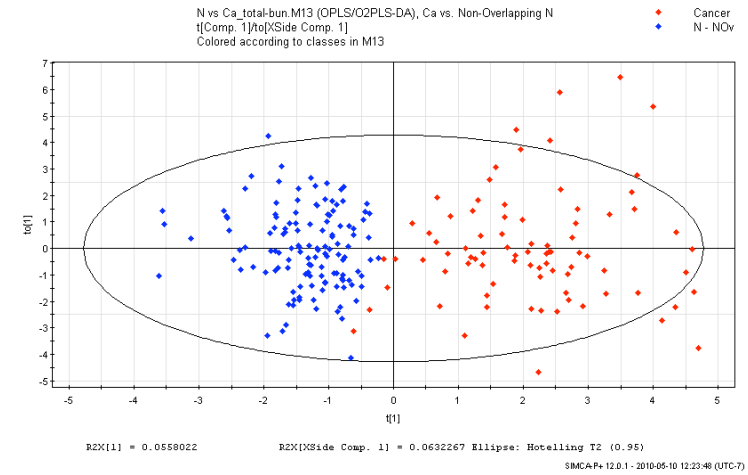
Training Set Normal vs. CRC Baseline Characteristics

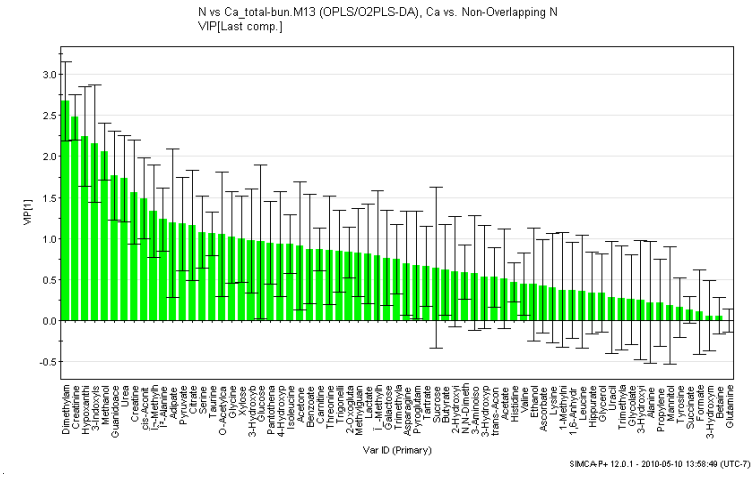
```
. cs sex nc if training ==1

                 | NC                    |
                 |   Exposed   Unexposed |      Total
-----------------+-----------------------+------------
           Cases |        47         117 |        164
        Noncases |        35         177 |        212
-----------------+-----------------------+------------
           Total |        82         294 |        376
                 |                       |
            Risk |  .5731707    .3979592 |   .4361702
                 |                       |
                 |      Point estimate   |    [95% Conf. Interval]
                 |-----------------------+------------------------
 Risk difference |         .1752115      |    .0544165    .2960066
      Risk ratio |         1.440275      |     1.14003    1.819595
 Attr. frac. ex. |         .3056882      |    .1228301    .4504271
 Attr. frac. pop |         .0876058      |
                 +-----------------------------------------------
                              chi2(1) =     8.00  Pr>chi2 = 0.0047


. cs famhx_cca nc if training ==1

                 | NC                    |
                 |   Exposed   Unexposed |      Total
-----------------+-----------------------+------------
           Cases |        20         191 |        211
        Noncases |        60          85 |        145
-----------------+-----------------------+------------
           Total |        80         276 |        356
                 |                       |
            Risk |       .25     .692029 |   .5926966
                 |                       |
                 |      Point estimate   |    [95% Conf. Interval]
                 |-----------------------+------------------------
 Risk difference |         -.442029      |   -.5514354   -.3326226
      Risk ratio |         .3612565      |    .2451742    .5323002
 Prev. frac. ex. |         .6387435      |    .4676998    .7548258
 Prev. frac. pop |         .1435379      |
                 +-----------------------------------------------
                              chi2(1) =    50.20  Pr>chi2 = 0.0000


. cs fh_any_ca nc if training ==1

                 | NC                    |
                 |   Exposed   Unexposed |      Total
-----------------+-----------------------+------------
           Cases |        60         230 |        290
        Noncases |        19          21 |         40
-----------------+-----------------------+------------
           Total |        79         251 |        330
                 |                       |
            Risk |  .7594937    .9163347 |   .8787879
                 |                       |
                 |      Point estimate   |    [95% Conf. Interval]
                 |-----------------------+------------------------
 Risk difference |         -.156841      |   -.2571182   -.0565638
      Risk ratio |         .8288387      |    .7280922    .9435257
 Prev. frac. ex. |         .1711613      |    .0564743    .2719078
 Prev. frac. pop |          .040975      |
                 +-----------------------------------------------
                              chi2(1) =    13.88  Pr>chi2 = 0.0002
```

```
. cs smoke nc if training ==1

                 | NC                    |
                 |   Exposed   Unexposed |      Total
-----------------+-----------------------+------------
           Cases |        15          26 |         41
        Noncases |        60         253 |        313
-----------------+-----------------------+------------
           Total |        75         279 |        354
                 |                       |
            Risk |        .2      .09319 |   .1158192
                 |                       |
                 |      Point estimate   |    [95% Conf. Interval]
                 |-----------------------+------------------------
 Risk difference |           .10681      |      .01007    .2035501
      Risk ratio |         2.146154      |     1.19909    3.841228
 Attr. frac. ex. |         .5340502      |     .166034    .7396666
 Attr. frac. pop |         .1953842      |
                 +-----------------------------------------------
                              chi2(1) =     6.59  Pr>chi2 = 0.0103


. cs dm nc if training ==1

                 | NC                    |
                 |   Exposed   Unexposed |      Total
-----------------+-----------------------+------------
           Cases |        19          14 |         33
        Noncases |        63         280 |        343
-----------------+-----------------------+------------
           Total |        82         294 |        376
                 |                       |
            Risk |  .2317073     .047619 |    .087766
                 |                       |
                 |      Point estimate   |    [95% Conf. Interval]
                 |-----------------------+------------------------
 Risk difference |         .1840883      |    .0895777    .2785988
      Risk ratio |         4.865854      |    2.551678    9.278809
 Attr. frac. ex. |         .7944862      |     .608101    .8922276
 Attr. frac. pop |         .4574315      |
                 +-----------------------------------------------
                              chi2(1) =    27.14  Pr>chi2 = 0.0000


. cs sympt_gibleed nc if training ==1

                 | NC                    |
                 |   Exposed   Unexposed |      Total
-----------------+-----------------------+------------
           Cases |        49           5 |         54
        Noncases |        33         288 |        321
-----------------+-----------------------+------------
           Total |        82         293 |        375
                 |                       |
            Risk |   .597561    .0170648 |       .144
                 |                       |
                 |      Point estimate   |    [95% Conf. Interval]
                 |-----------------------+------------------------
 Risk difference |         .5804961      |    .4733244    .6876679
      Risk ratio |         35.01707      |    14.42348    85.01386
 Attr. frac. ex. |         .9714425      |    .9306686    .9882372
 Attr. frac. pop |         .8814941      |
                 +-----------------------------------------------
                              chi2(1) =   175.15  Pr>chi2 = 0.0000
```

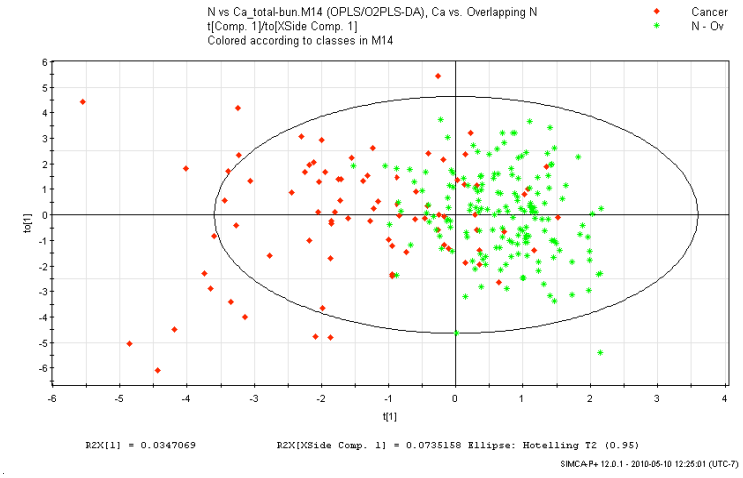
```
. cs sympt_bowelhabit nc if training ==1

                 | NC                     |
                 |   Exposed   Unexposed  |      Total
-----------------+------------------------+-----------
           Cases |        43          13  |         56
        Noncases |        39         279  |        318
-----------------+------------------------+-----------
           Total |        82         292  |        374
                 |                        |
            Risk |  .5243902    .0445205  |   .1497326
                 |                        |
                 |      Point estimate    |    [95% Conf. Interval]
                 |------------------------+------------------------
 Risk difference |         .4798697       |    .3692193    .5905201
      Risk ratio |         11.77861       |    6.661512    20.82646
 Attr. frac. ex. |         .9151004       |    .8498839    .9519842
 Attr. frac. pop |         .7026663       |
                 +-------------------------------------------------
                              chi2(1) =   115.80  Pr>chi2 = 0.0000


. ttest age, by(nc), if training==1

Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |     294     55.2619    .4731272    8.112442    54.33075    56.19306
       1 |      82    68.63415    1.164505    10.54504    66.31715    70.95115
---------+--------------------------------------------------------------------
combined |     376    58.17819    .5309904    10.29629     57.1341    59.22228
---------+--------------------------------------------------------------------
    diff |           -13.37224    1.086158               -15.50798   -11.2365
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t = -12.3115
Ho: diff = 0                                     degrees of freedom =      374

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 1.0000
```

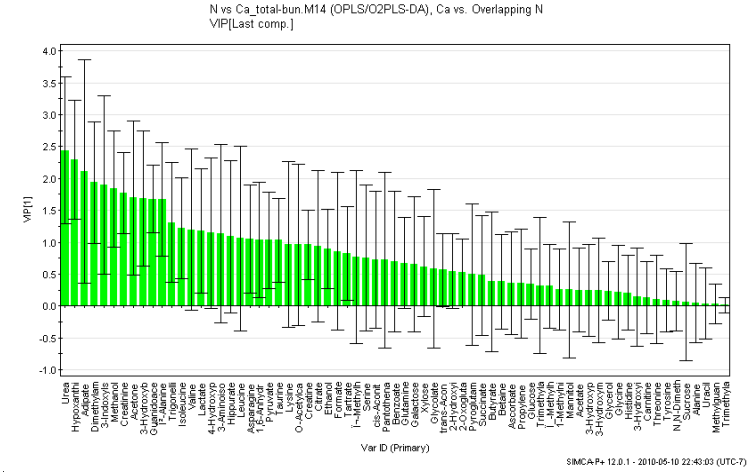
CRC vs. Non-Overlapping Normals



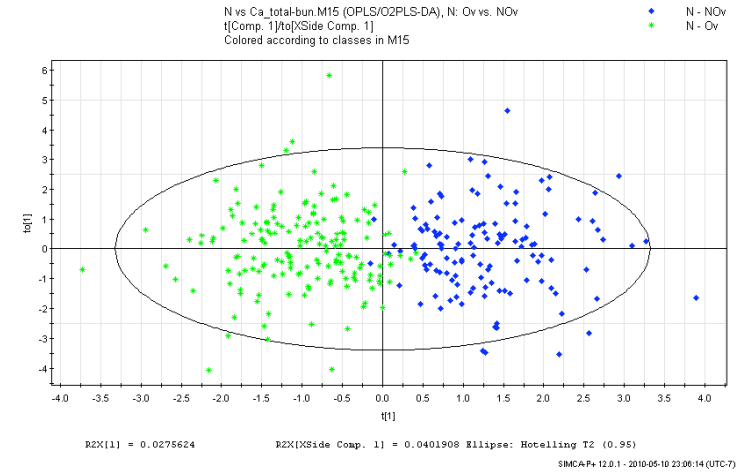


CRC vs. Overlapping Normals





Overlapping vs. Non-Overlapping Normals



Logistic Regression of Ov vs. NonOv Normals on predicting overlap with adenoma

Gender

```
. logistic overlapping_with_cancer S_sex

Logistic regression                                     Number of obs   =        294
                                                        LR chi2(1)      =      13.80
                                                        Prob > chi2     =     0.0002
Log likelihood = -194.67543                             Pseudo R2       =     0.0342

------------------------------------------------------------------------------
overlappin~r | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       S_sex |   2.491071   .6241658     3.64   0.000     1.524432    4.070655
------------------------------------------------------------------------------
```

In the normal group, the odds of resembling or overlapping with the cancer group for males is 2.49 (95% CI 1.52, 4.07) times that of females ($p<0.01$). That is, females are more different than the cancers compared to the males.

Age

```
. logistic overlapping_with_cancer age

Logistic regression                                     Number of obs   =        294
                                                        LR chi2(1)      =       1.86
                                                        Prob > chi2     =     0.1726
Log likelihood = -200.64549                             Pseudo R2       =     0.0046

------------------------------------------------------------------------------
overlappin~r | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   1.020107   .0149642     1.36   0.175     .9911952    1.049862
------------------------------------------------------------------------------
```

Within the normal group, age does not change the odds of overlapping with cancers, OR 1.02 (95% CI 0.99, 1.05).

```
. logit overlapping_with_cancer age

Iteration 0:   log likelihood = -201.57565
Iteration 1:   log likelihood = -200.64564
Iteration 2:   log likelihood = -200.64549

Logistic regression                                     Number of obs   =        294
                                                        LR chi2(1)      =       1.86
                                                        Prob > chi2     =     0.1726
Log likelihood = -200.64549                             Pseudo R2       =     0.0046

------------------------------------------------------------------------------
overlappin~r |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0199073   .0146692     1.36   0.175    -.0088438    .0486585
       _cons |  -.8523324   .8167304    -1.04   0.297    -2.453095    .7484298
------------------------------------------------------------------------------
```

$e^{0.0199 \times 10} = 1.220$

95% CI = $e^{10\,(0.0146) \pm 1.96\,(10)\,(0.0146)} = 0.915, 1.627$

For every 10 years increase in age, the odds of overlapping with cancer increases about 22%. The 95% CI for the OR is 0.915, 1.627. This includes 1, therefore there is no significance.

```
. generate age_cat = .
(294 missing values generated)

. replace age_cat=1 if age<=45 & age !=.
(38 real changes made)

. replace age_cat=2 if age>45 & age <=50 & age !=.
(39 real changes made)

. replace age_cat=3 if age>50 & age <=55 & age !=.
(84 real changes made)

. replace age_cat=4 if age>55& age <=60 & age !=.
(64 real changes made)

. replace age_cat=5 if age>60 & age <=65 & age !=.
(37 real changes made)

. replace age_cat=6 if age>65 & age <=70 & age !=.
```

```
(20 real changes made)

. replace age_cat=7 if age>70 & age !=.
(12 real changes made)

. tabulate age_cat

    age_cat |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |         38       12.93       12.93
          2 |         39       13.27       26.19
          3 |         84       28.57       54.76
          4 |         64       21.77       76.53
          5 |         37       12.59       89.12
          6 |         20        6.80       95.92
          7 |         12        4.08      100.00
------------+-----------------------------------
      Total |        294      100.00


. xi:logistic overlapping_with_cancer i.age_cat
i.age_cat         _Iage_cat_1-7       (naturally coded; _Iage_cat_1 omitted)

Logistic regression                               Number of obs   =        294
                                                  LR chi2(6)      =       7.13
                                                  Prob > chi2     =     0.3090
Log likelihood = -198.01048                       Pseudo R2       =     0.0177

------------------------------------------------------------------------------
overlappin~r | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 _Iage_cat_2 |   .8521303   .3896334    -0.35   0.726     .3477741    2.087924
 _Iage_cat_3 |   .8904762   .3496407    -0.30   0.768     .4124797    1.922393
 _Iage_cat_4 |   1.109347   .4580767     0.25   0.802     .4938393    2.492009
 _Iage_cat_5 |   1.494505   .7089571     0.85   0.397     .5898046    3.786926
 _Iage_cat_6 |   .6623377   .3678585    -0.74   0.458     .2230114    1.967125
 _Iage_cat_7 |   4.047619   3.401909     1.66   0.096     .7794512    21.01892
------------------------------------------------------------------------------
```

Smoking History

```
. xi:logistic overlapping_with_cancer i.s_smoke
i.s_smoke         _Is_smoke_0-2       (naturally coded; _Is_smoke_0 omitted)

Logistic regression                               Number of obs   =        279
                                                  LR chi2(2)      =       3.32
                                                  Prob > chi2     =     0.1900
Log likelihood = -190.00159                       Pseudo R2       =     0.0087

------------------------------------------------------------------------------
overlappin~r | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 _Is_smoke_1 |   1.629347   .7034889     1.13   0.258      .699033    3.797777
 _Is_smoke_2 |   3.019084   2.451486     1.36   0.174     .6147555    14.82682
------------------------------------------------------------------------------
```

Within the normal group, being a smoker or an ex-smoker does not increase the odds of overlapping with the cancer group.

Since metabolism is a reflection of current status, ex-smokers could be classified as non-smokers.

```
. logistic overlapping_with_cancer s_smoke_YN

Logistic regression                               Number of obs   =        279
                                                  LR chi2(1)      =       1.14
```

```
                                                Prob > chi2     =     0.2848
Log likelihood =  -191.0902                     Pseudo R2       =     0.0030

------------------------------------------------------------------------------
overlappin~r | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  s_smoke_YN |   1.574074    .678634     1.05   0.293     .6761527    3.664422
------------------------------------------------------------------------------
```

Still no statistically significant difference, but smokers only 26, so may not be enough numbers to show a difference.

Diabetes

```
. logistic overlapping_with_cancer diabetes

Logistic regression                             Number of obs   =        294
                                                LR chi2(1)      =       0.01
                                                Prob > chi2     =     0.9371
Log likelihood = -201.57254                     Pseudo R2       =     0.0000

------------------------------------------------------------------------------
overlappin~r | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    diabetes |   1.044586   .5779933     0.08   0.937     .3531465    3.089822
------------------------------------------------------------------------------
```

No statistically significant difference.

Family history

```
. logistic overlapping_with_cancer s_fhcca

Logistic regression                             Number of obs   =        276
                                                LR chi2(1)      =       0.81
                                                Prob > chi2     =     0.3688
Log likelihood = -189.04556                     Pseudo R2       =     0.0021

------------------------------------------------------------------------------
overlappin~r | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     s_fhcca |   1.265432    .331326     0.90   0.369     .7574782    2.114013
------------------------------------------------------------------------------
```

Within the normal group, having a family history of CRC does not increase the odds of overlapping with the cancer group.

```
. logistic overlapping_with_cancer st_degree_relative_with_crc

Logistic regression                             Number of obs   =        294
                                                LR chi2(1)      =       0.25
                                                Prob > chi2     =     0.6166
Log likelihood = -201.45029                     Pseudo R2       =     0.0006

------------------------------------------------------------------------------
overlappin~r | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
st_degree_~c |   1.129957   .2756153     0.50   0.616     .7005508     1.82257
------------------------------------------------------------------------------
```

1st degree relative – no difference.

```
. logistic overlapping_with_cancer s_fhca

Logistic regression                                     Number of obs   =        251
                                                        LR chi2(1)      =       0.74
                                                        Prob > chi2     =     0.3894
Log likelihood = -171.43385                             Pseudo R2       =     0.0022

------------------------------------------------------------------------------
overlappin~r | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      s_fhca |   1.481633   .6768466     0.86   0.389     .6051883    3.627359
------------------------------------------------------------------------------
```

The data on the family history of various cancers were combined and it was found that within the normal group, those having a family history of *any* cancer(s) does not increase the odds of overlapping with the cancer group.

Symptoms

```
. logistic overlapping_with_cancer s_GIbleed

Logistic regression                                     Number of obs   =        293
                                                        LR chi2(1)      =       2.82
                                                        Prob > chi2     =     0.0933
Log likelihood = -199.34162                             Pseudo R2       =     0.0070

------------------------------------------------------------------------------
overlappin~r | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   s_GIbleed |   .1890244   .2125285    -1.48   0.138     .0208677    1.712227
------------------------------------------------------------------------------

. logistic overlapping_with_cancer s_GIhabit

Logistic regression                                     Number of obs   =        292
                                                        LR chi2(1)      =       2.47
                                                        Prob > chi2     =     0.1161
Log likelihood = -198.68493                             Pseudo R2       =     0.0062

------------------------------------------------------------------------------
overlappin~r | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   s_GIhabit |   2.666667   1.784574     1.47   0.143     .7183429    9.899327
------------------------------------------------------------------------------
```

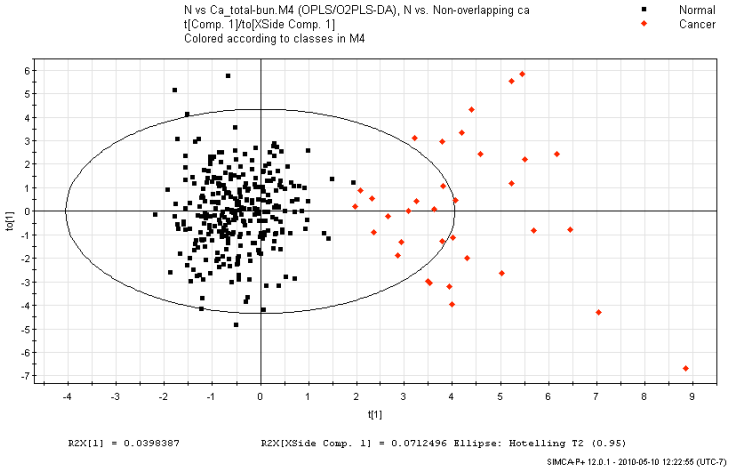
Within the normal group, having symptoms of gastrointestinal bleeding or changes in bowel habits does not increase the odds of overlapping with the cancer group.

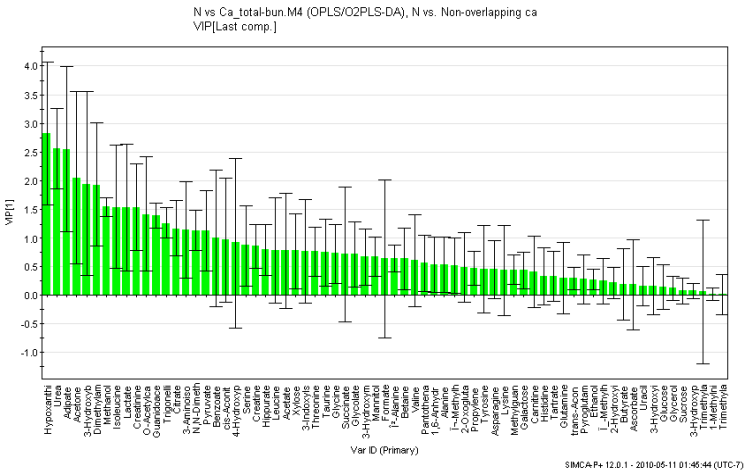
Overall test

```
. logistic overlapping_with_cancer sex age smoke dm famhx_cca fh_any_ca
sympt_gibleed symp
> t_bowelhabit if nc ==0

Logistic regression                                     Number of obs   =        231
                                                        LR chi2(8)      =      22.13
                                                        Prob > chi2     =     0.0047
Log likelihood =   -147.229                             Pseudo R2       =     0.0699
```

overlappin~r	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
sex	2.919098	.8853455	3.53	0.000	1.610958	5.289482
age	1.034737	.0192592	1.83	0.067	.9976696	1.073181
smoke	1.694832	.9341804	0.96	0.338	.5753734	4.992332
dm	1.659519	1.175265	0.72	0.474	.4141564	6.649668
famhx_cca	1.283903	.5316059	0.60	0.546	.5702801	2.890521
fh_any_ca	1.443366	.8532534	0.62	0.535	.4530899	4.597998
sympt_gibl~d	.1435988	.1706661	-1.63	0.102	.0139799	1.47502
sympt_bowe~t	3.797069	3.104427	1.63	0.103	.764749	18.8529

Normal vs. Non-overlapping CRC

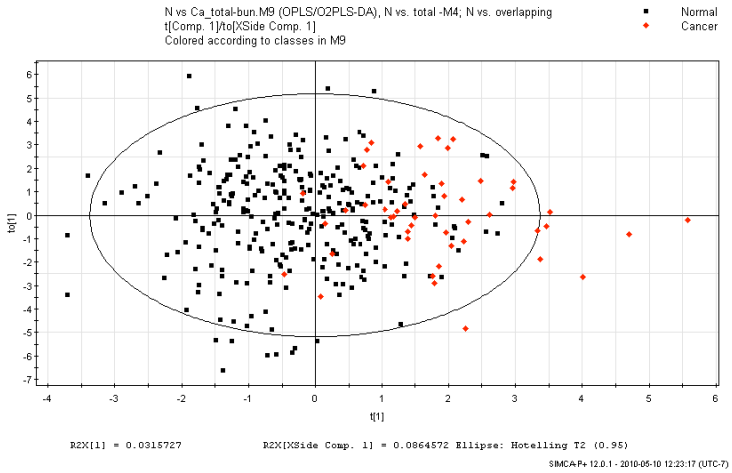


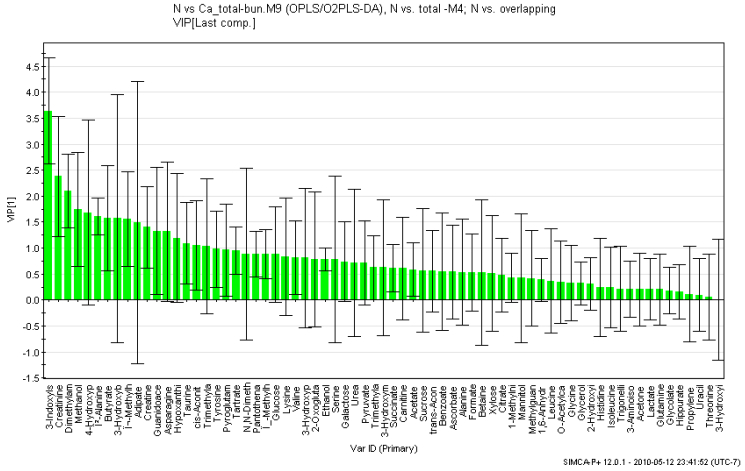


The top five metabolites for normals vs. non-overlapping cancers are:

- Hypoxanthine
- Urea
- Adipate
- Acetone
- 3-Hydroxybutyrate

Normal vs. Overlapping CRC

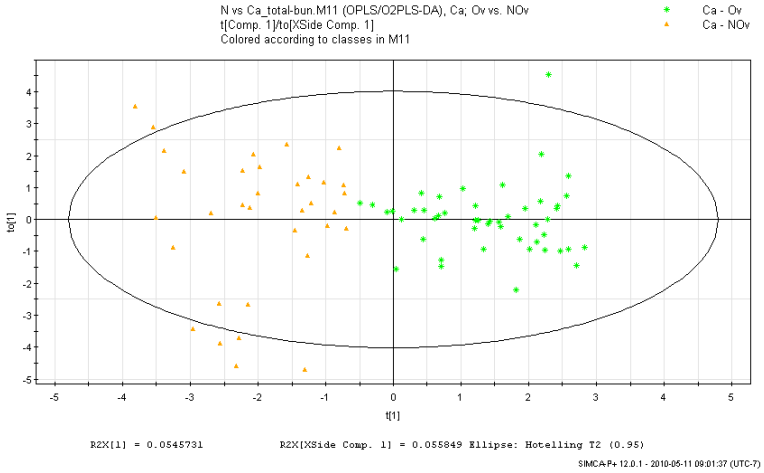


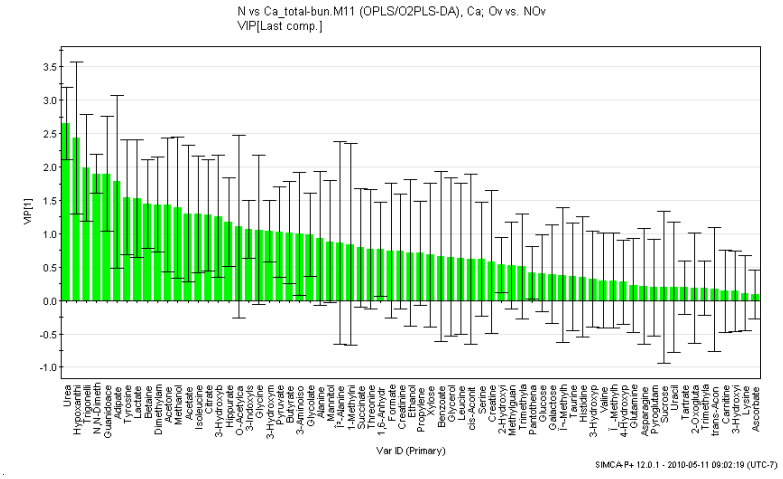


The top five metabolites are for normals vs. overlapping cancers:

- 3-Indoxylsulfate
- Creatinine
- Dimethylamine
- Methanol
- 4-Hydroxyphenylacetate

Overlapping vs. Non-Overlapping CRC





The top five metabolites for these two groups of cancers are:

- Urea
- Hypoxanthine
- Trigonelline
- N, N-Dimethylglycine
- Guanidoacetate

Logistic regression of overlapping and non-overlapping CRC on predicting overlapping with normal

```
. logistic overla_with_normal S_sex

Logistic regression                                     Number of obs   =         82
                                                        LR chi2(1)      =       4.15
                                                        Prob > chi2     =     0.0417
Log likelihood = -53.564057                             Pseudo R2       =     0.0373

------------------------------------------------------------------------------
overla_wit~l | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       S_sex |   2.533333   1.169308     2.01   0.044      1.02519    6.260088
------------------------------------------------------------------------------
```

Age

```
. logistic overla_with_normal age

Logistic regression                                     Number of obs   =         82
                                                        LR chi2(1)      =       0.00
                                                        Prob > chi2     =     0.9925
Log likelihood = -55.637028                             Pseudo R2       =     0.0000

------------------------------------------------------------------------------
overla_wit~l | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .9997992   .0213833    -0.01   0.993      .958755    1.042601
------------------------------------------------------------------------------
```

Within the cancer group, age does not change the odds of overlapping with normals, OR 1.00 (95% CI 0.96, 1.04).

```
. generate age_cat=.
(82 missing values generated)

. replace age_cat=1 if age<=50 & age !=.
(3 real changes made)

. replace age_cat=2 if age>50 & age<=55 & age !=.
(8 real changes made)

. replace age_cat=3 if age>55 & age<=60 & age !=.
(7 real changes made)

. replace age_cat=4 if age>60 & age<=65 & age !=.
(17 real changes made)

. replace age_cat=5 if age>65 & age<=70 & age !=.
(13 real changes made)

. replace age_cat=6 if age>70 & age<=75 & age !=.
(8 real changes made)

. replace age_cat=7 if age>75 & age<=80 & age !=.
(16 real changes made)

. replace age_cat=8 if age>80 & age !=.
(10 real changes made)

. tabulate age_cat

    age_cat |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |          3        3.66        3.66
          2 |          8        9.76       13.41
          3 |          7        8.54       21.95
          4 |         17       20.73       42.68
          5 |         13       15.85       58.54
          6 |          8        9.76       68.29
          7 |         16       19.51       87.80
          8 |         10       12.20      100.00
------------+-----------------------------------
      Total |         82      100.00


. xi:logistic overla_with_normal i.age_cat
i.age_cat         _Iage_cat_1-8       (naturally coded; _Iage_cat_1 omitted)

Logistic regression                               Number of obs   =         82
                                                  LR chi2(7)      =       8.72
                                                  Prob > chi2     =     0.2730
Log likelihood = -51.274834                       Pseudo R2       =     0.0784

------------------------------------------------------------------------------
overla_wit~l | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 _Iage_cat_2 |   .8333333   1.188292    -0.13   0.898     .0509387    13.63294
 _Iage_cat_3 |   .6666667   .9622504    -0.28   0.779     .0393827    11.28528
 _Iage_cat_4 |        .35   .4620606    -0.80   0.426      .026323    4.653733
 _Iage_cat_5 |       2.75   3.976493     0.70   0.484     .1616208    46.79162
 _Iage_cat_6 |        1.5    2.20794     0.28   0.783     .0837831    26.85505
 _Iage_cat_7 |   .3888889   .5150346    -0.71   0.476     .0290079    5.213571
 _Iage_cat_8 |        .75   1.038328    -0.21   0.835     .0497309    11.31088
------------------------------------------------------------------------------
```

Smoking History

```
. xi:logistic overla_with_normal i.s_smoke
i.s_smoke         _Is_smoke_0-2       (naturally coded; _Is_smoke_0 omitted)

Logistic regression                               Number of obs   =         75
                                                  LR chi2(2)      =       2.47
                                                  Prob > chi2     =     0.2906
Log likelihood = -49.940654                       Pseudo R2       =     0.0241

------------------------------------------------------------------------------
overla_wit~l | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 _Is_smoke_1 |   .8166667   .5203342    -0.32   0.751     .2342647    2.846969
 _Is_smoke_2 |       1.96   1.047664     1.26   0.208     .6874953     5.58782
------------------------------------------------------------------------------
```

Within the cancer group, being a smoker or an ex-smoker does not increase the odds of overlapping with the normal group.

Since metabolism is a reflection of current status, ex-smokers could be classified as non-smokers.

```
. logistic overla_with_normal s_smoke_YN

Logistic regression                               Number of obs   =         75
                                                  LR chi2(1)      =       0.86
                                                  Prob > chi2     =     0.3527
Log likelihood =  -50.74455                       Pseudo R2       =     0.0084

------------------------------------------------------------------------------
overla_wit~l | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  s_smoke_YN |   .5833333   .3387864    -0.93   0.353     .1868783    1.820853
------------------------------------------------------------------------------
```

Still no statistically significant difference, but smokers only 15, so may not be enough numbers to show a difference.

Diabetes

```
. logistic overla_with_normal diabetes

Logistic regression                               Number of obs   =         82
                                                  LR chi2(1)      =       0.35
                                                  Prob > chi2     =     0.5527
Log likelihood = -55.460787                       Pseudo R2       =     0.0032

------------------------------------------------------------------------------
overla_wit~l | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    diabetes |   .7309942   .3850243    -0.59   0.552     .2603603    2.052358
------------------------------------------------------------------------------
```

No statistically significant difference.

Family history

```
. logistic overla_with_normal s_fhcca
```

```
Logistic regression                                     Number of obs   =         80
                                                        LR chi2(1)      =       0.01
                                                        Prob > chi2     =     0.9129
Log likelihood = -53.834946                             Pseudo R2       =     0.0001

------------------------------------------------------------------------------
overla_wit~l | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     s_fhcca |   1.061776   .5823271     0.11   0.913     .3624054    3.110794
------------------------------------------------------------------------------
```

Within the cancer group, having a family history of CRC does not increase the odds of overlapping with the normal group.

```
. logistic overla_with_normal s_fhoca

Logistic regression                                     Number of obs   =         79
                                                        LR chi2(1)      =       4.93
                                                        Prob > chi2     =     0.0264
Log likelihood = -50.450127                             Pseudo R2       =     0.0466

------------------------------------------------------------------------------
overla_wit~l | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     s_fhoca |   .3461538   .1697502    -2.16   0.031     .1323881     .905085
------------------------------------------------------------------------------
```

However, within the cancer group, those having a family history of other cancers (other than colon, rectal, uterine, breast, ovarian cancer) have 0.35 (p=0.03) the odds of overlapping with the normal group compared to those that do not have a family history. That is, the metabolomic profile of those cancer patients that have a family history of 'other' cancers have statistically significant increased odds of being different than the normal group.

```
. logistic overla_with_normal s_fhca

Logistic regression                                     Number of obs   =         78
                                                        LR chi2(1)      =       6.61
                                                        Prob > chi2     =     0.0101
Log likelihood = -49.106464                             Pseudo R2       =     0.0631

------------------------------------------------------------------------------
overla_wit~l | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      s_fhca |   .2075893   .1413695    -2.31   0.021     .0546426     .788639
------------------------------------------------------------------------------
```

The data on the family history of various cancers were combined and it was found that within the cancer group, those having a family history of *any* cancer(s) have 0.21 (p=0.02) the odds of overlapping with the normal group compared to those that do not have a family history. That is, the metabolomic profiles of those cancer patients that have a family history of any neoplastic process have statistically significant increased odds of being different than that of the normal group.

Fasting

```
. logistic overlap_with_normal fasting if nc==1

Logistic regression                                     Number of obs   =         74
                                                        LR chi2(1)      =       4.76
```

```
                                                Prob > chi2     =     0.0292
Log likelihood = -48.480415                     Pseudo R2       =     0.0468

------------------------------------------------------------------------------
overlap_wi~l | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     fasting |    .345098   .1710117    -2.15   0.032     .1306573    .9114888
------------------------------------------------------------------------------
```

Within the cancer group, the odds of overlapping with the normal group in those that are fasting is 0.35 times (p=0.032) that of those that are not fasting. That is, the metabolomic profiles of those that are fasting are more different than the profiles of the normals.

Symptoms

```
. logistic overlap_with_normal sympt_gibleed if nc==1

Logistic regression                             Number of obs   =         82
                                                LR chi2(1)      =       0.36
                                                Prob > chi2     =     0.5476
Log likelihood = -55.456237                     Pseudo R2       =     0.0033

------------------------------------------------------------------------------
overlap_wi~l | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
sympt_gibl~d |   1.315789   .6003622     0.60   0.548     .5380271    3.217871
------------------------------------------------------------------------------


. logistic overla_with_normal s_GIhabit

Logistic regression                             Number of obs   =         82
                                                LR chi2(1)      =       0.67
                                                Prob > chi2     =     0.4115
Log likelihood = -55.299831                     Pseudo R2       =     0.0061

------------------------------------------------------------------------------
overla_wit~l | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   s_GIhabit |   1.446429   .6512329     0.82   0.412     .5984891    3.495729
------------------------------------------------------------------------------
```

Within the cancer group, having symptoms of gastrointestinal bleeding or changes in bowel habits does not increase the odds of overlapping with the normal group.

Location of Cancer

Location of Cancer (colon vs. rectal)
Colon = 0; rectal = 1

```
. logistic overla_with_normal s_location1

Logistic regression                             Number of obs   =         82
                                                LR chi2(1)      =       1.50
                                                Prob > chi2     =     0.2211
Log likelihood = -54.888452                     Pseudo R2       =     0.0135

------------------------------------------------------------------------------
overla_wit~l | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 s_location1 |    .545045    .270731    -1.22   0.222     .2058872    1.442897
```

```
------------------------------------------------------------------------------
```

No significant difference.

Location of Cancer (left vs. right sided)
Right = 0, everything else = 1

```
. logistic overla_with_normal s_location2

Logistic regression                               Number of obs   =         82
                                                  LR chi2(1)      =       0.74
                                                  Prob > chi2     =     0.3906
Log likelihood = -55.268547                       Pseudo R2       =     0.0066

------------------------------------------------------------------------------
overla_wit~l | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 s_location2 |   1.503401   .7137317     0.86   0.390     .5928865    3.812223
------------------------------------------------------------------------------
```

No difference.

CEA

```
. logistic overla_with_normal cea

Logistic regression                               Number of obs   =         63
                                                  LR chi2(1)      =       7.73
                                                  Prob > chi2     =     0.0054
Log likelihood =  -38.83666                       Pseudo R2       =     0.0905

------------------------------------------------------------------------------
overla_wit~l | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         cea |   .9506619   .0292049    -1.65   0.100     .8951106    1.009661
------------------------------------------------------------------------------
Note: 2 failures and 0 successes completely determined.
```

CEA not statistically significant, but CEA distribution very skewed, therefore catergorized.

```
. generate cea_cat = .
(82 missing values generated)

. replace cea_cat = 1 if cea > 0 & cea <= 5
(45 real changes made)

. replace cea_cat = 2 if cea > 5
(37 real changes made)


. logistic overla_with_normal cea_cat

Logistic regression                               Number of obs   =         63
                                                  LR chi2(1)      =       2.10
                                                  Prob > chi2     =     0.1471
Log likelihood = -41.652122                       Pseudo R2       =     0.0246

------------------------------------------------------------------------------
overla_wit~l | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     cea_cat |   .4413793    .250454    -1.44   0.149     .1451478    1.342188
------------------------------------------------------------------------------
```

```
. logistic overlap_with_normal li if nc ==1

Logistic regression                                     Number of obs   =         74
                                                        LR chi2(1)      =       1.60
                                                        Prob > chi2     =     0.2061
Log likelihood = -47.756974                             Pseudo R2       =     0.0165

------------------------------------------------------------------------------
overlap_wi~l | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          li |    .530303   .2660679    -1.26   0.206     .1983592    1.417737
------------------------------------------------------------------------------

. logistic overlap_with_normal vi if nc ==1

Logistic regression                                     Number of obs   =         74
                                                        LR chi2(1)      =       0.27
                                                        Prob > chi2     =     0.6007
Log likelihood = -48.419336                             Pseudo R2       =     0.0028

------------------------------------------------------------------------------
overlap_wi~l | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          vi |   .6845238    .492328    -0.53   0.598     .1671789    2.802823
------------------------------------------------------------------------------

. logistic overlap_with_normal pni if nc ==1

Logistic regression                                     Number of obs   =         71
                                                        LR chi2(1)      =       0.02
                                                        Prob > chi2     =     0.9013
Log likelihood = -47.150644                             Pseudo R2       =     0.0002

------------------------------------------------------------------------------
overlap_wi~l | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         pni |   1.087838   .7404992     0.12   0.902     .2865135    4.130316
------------------------------------------------------------------------------

. logistic overlap_with_normal lymphocyctic_resp if nc ==1

Logistic regression                                     Number of obs   =         64
                                                        LR chi2(1)      =       0.07
                                                        Prob > chi2     =     0.7945
Log likelihood = -42.306129                             Pseudo R2       =     0.0008

------------------------------------------------------------------------------
overlap_wi~l | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
lymphocyct~p |   .8730159   .4556764    -0.26   0.795     .3138571    2.428356
------------------------------------------------------------------------------



. xi:logistic overlap_with_normal i.gradepath
i.gradepath       _Igradepath_1-3     (naturally coded; _Igradepath_1 omitted)

note: _Igradepath_2 != 0 predicts success perfectly
      _Igradepath_2 dropped and 3 obs not used


Logistic regression                                     Number of obs   =         71
                                                        LR chi2(1)      =       0.16
                                                        Prob > chi2     =     0.6858
Log likelihood = -47.535101                             Pseudo R2       =     0.0017
```

```
------------------------------------------------------------------------------
overlap_wi~l | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
_Igradepat~3 |   1.351351   1.017582     0.40   0.689     .3088905    5.911968
------------------------------------------------------------------------------

. tab overlap_with_normal gradepath gradepath
too many variables specified
r(103);

. tab overlap_with_normal gradepath

Overlap_wi |            GradePath
 th_normal |         1          2          3 |     Total
-----------+---------------------------------+----------
         0 |        25          0          3 |        28
         1 |        37          3          6 |        46
-----------+---------------------------------+----------
     Total |        62          3          9 |        74
```

No statistically significant differences based on lymphovascular invasion, perineural invasion, lymphocytic response or grade. Although all moderately differentiated tumors are in the overlapping group…

Cancer Stage

```
. xi:logistic overla_with_normal i.stage_of_cancer
i.stage_of_ca~r   _Istage_of__1-4     (naturally coded; _Istage_of__1 omitted)

Logistic regression                               Number of obs   =         82
                                                  LR chi2(3)      =       5.70
                                                  Prob > chi2     =     0.1271
Log likelihood = -52.787022                       Pseudo R2       =     0.0512

------------------------------------------------------------------------------
overla_wit~l | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
_Istage_of~2 |         .5   .4082472    -0.85   0.396      .100918    2.477259
_Istage_of~3 |    .265625   .1948295    -1.81   0.071      .063086    1.118419
_Istage_of~4 |   .1944444   .1592346    -2.00   0.046     .0390597    .9679709
------------------------------------------------------------------------------

. tab overlap_with_normal stage_of_cancer

Overlap_wi |               STAGE_of_Cancer
 th_normal |         1          2          3          4 |     Total
-----------+---------------------------------------------+----------
         0 |         3          6         16          9 |        34
         1 |        12         12         17          7 |        48
-----------+---------------------------------------------+----------
     Total |        15         18         33         16 |        82
```

Stage 4 cancers more likely not to overlap compared to stage 1 cancers, i.e. more different than normals, OR 0.19 (p=0.05), but stage 2 and 3 cancers are not statistically different than stage 1 cancers.

In summary, gender, family history of any cancer, and fasting are statistically significant differences between the overlapping and non-overlapping cancer subgroups. When these factors are tested with an Overall test, they are still all statistically different.

```
. logistic overla_with_normal S_sex s_fasting s_fhca

Logistic regression                                     Number of obs   =         71
                                                        LR chi2(3)      =      18.04
                                                        Prob > chi2     =     0.0004
Log likelihood = -38.997903                             Pseudo R2       =     0.1878

------------------------------------------------------------------------------
overla_wit~l | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       S_sex |   3.500294   2.080689     2.11   0.035     1.091748    11.22242
   s_fasting |   .1812088   .1102056    -2.81   0.005      .055018    .5968346
      s_fhca |   .1793053   .1326809    -2.32   0.020     .0420463    .7646427
------------------------------------------------------------------------------
```

Training Set vs. Testing Set

. insheet using training_and_testing_set_normal_cancer.txt, clear
(94 vars, 560 obs)

Normal

```
. ttest age, by(training), if nc==0

Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |     150       56.44    .6635597    8.126913     55.1288     57.7512
       1 |     294     55.2619    .4731272    8.112442    54.33075    56.19306
---------+--------------------------------------------------------------------
combined |     444    55.65991    .3857057    8.127323    54.90187    56.41795
---------+--------------------------------------------------------------------
    diff |            1.178095    .8144886               -.4226564    2.778847
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =   1.4464
Ho: diff = 0                                     degrees of freedom =      442

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.9256         Pr(|T| > |t|) = 0.1488          Pr(T > t) = 0.0744
```

```
. cs sex training if nc ==0

                 | TRAINING              |
                 |   Exposed   Unexposed |      Total
-----------------+-----------------------+-----------
           Cases |       117          71 |        188
        Noncases |       177          79 |        256
-----------------+-----------------------+-----------
           Total |       294         150 |        444
                 |                       |
            Risk |  .3979592    .4733333 |   .4234234
                 |                       |
                 |      Point estimate   |    [95% Conf. Interval]
                 |-----------------------+------------------------
 Risk difference |        -.0753741      |   -.1729176    .0221693
      Risk ratio |         .8407588      |    .6749364    1.047322
 Prev. frac. ex. |         .1592412      |   -.0473216    .3250636
 Prev. frac. pop |         .1054435      |
                 +-------------------------------------------------
                              chi2(1) =     2.31  Pr>chi2 = 0.1284
```

```
. cs famhx_cca training if nc ==0

                | TRAINING               |
                |   Exposed   Unexposed  |      Total
----------------+------------------------+------------
          Cases |       191          91  |        282
       Noncases |        85          49  |        134
----------------+------------------------+------------
          Total |       276         140  |        416
                |                        |
           Risk |   .692029         .65  |   .6778846
                |                        |
                |      Point estimate    |    [95% Conf. Interval]
                |------------------------+------------------------
 Risk difference |          .042029       |   -.0539331     .137991
      Risk ratio |          1.06466       |     .921133    1.230551
 Attr. frac. ex. |          .060733       |   -.0856195    .1873557
 Attr. frac. pop |         .0411348       |
                +-------------------------------------------------
                              chi2(1) =     0.75  Pr>chi2 = 0.3860


. cs fh_any_ca training if nc ==0

                | TRAINING               |
                |   Exposed   Unexposed  |      Total
----------------+------------------------+------------
          Cases |       230         115  |        345
       Noncases |        21          35  |         56
----------------+------------------------+------------
          Total |       251         150  |        401
                |                        |
           Risk |  .9163347    .7666667  |   .8603491
                |                        |
                |      Point estimate    |    [95% Conf. Interval]
                |------------------------+------------------------
 Risk difference |          .149668       |    .0738087    .2255273
      Risk ratio |          1.195219      |    1.085951    1.315481
 Attr. frac. ex. |          .1633333      |    .0791485    .2398219
 Attr. frac. pop |          .1088889      |
                +-------------------------------------------------
                              chi2(1) =    17.50  Pr>chi2 = 0.0000


. cs smoke training if nc ==0

                | TRAINING               |
                |   Exposed   Unexposed  |      Total
----------------+------------------------+------------
          Cases |        26          15  |         41
       Noncases |       253         134  |        387
----------------+------------------------+------------
          Total |       279         149  |        428
                |                        |
           Risk |    .09319    .1006711  |   .0957944
                |                        |
                |      Point estimate    |    [95% Conf. Interval]
                |------------------------+------------------------
 Risk difference |        -.0074812       |   -.0666226    .0516602
      Risk ratio |          .925687       |    .5062179    1.692742
 Prev. frac. ex. |          .074313       |   -.6927421    .4937821
 Prev. frac. pop |         .0484424       |
                +-------------------------------------------------
                              chi2(1) =     0.06  Pr>chi2 = 0.8022

. cs dm training if nc ==0

                | TRAINING               |
                |   Exposed   Unexposed  |      Total
----------------+------------------------+------------
          Cases |        14          10  |         24
```

```
       Noncases |       280         140  |        420
----------------+------------------------+-----------
          Total |       294         150  |        444
                |                        |
           Risk |   .047619    .0666667  |   .0540541
                |                        |
                |      Point estimate    |    [95% Conf. Interval]
                |------------------------+------------------------
 Risk difference |        -.0190476       |    -.065803    .0277078
      Risk ratio |         .7142857       |    .3250475    1.569629
 Prev. frac. ex. |         .2857143       |   -.5696295    .6749525
 Prev. frac. pop |         .1891892       |
                +-------------------------------------------------
                             chi2(1) =     0.70  Pr>chi2 = 0.4012
```

```
. cs sympt_gibleed training if nc ==0

                | TRAINING               |
                |   Exposed   Unexposed  |      Total
----------------+------------------------+-----------
          Cases |         5           2  |          7
       Noncases |       288         147  |        435
----------------+------------------------+-----------
          Total |       293         149  |        442
                |                        |
           Risk |  .0170648    .0134228  |   .0158371
                |                        |
                |      Point estimate    |    [95% Conf. Interval]
                |------------------------+------------------------
 Risk difference |          .003642       |   -.0200504    .0273345
      Risk ratio |         1.271331       |    .2496096    6.475242
 Attr. frac. ex. |         .2134228       |   -3.006256    .8455656
 Attr. frac. pop |         .1524449       |
                +-------------------------------------------------
                             chi2(1) =     0.08  Pr>chi2 = 0.7719
```

```
. cs sympt_bowelhabit training if nc ==0

                | TRAINING               |
                |   Exposed   Unexposed  |      Total
----------------+------------------------+-----------
          Cases |        13           0  |         13
       Noncases |       279         149  |        428
----------------+------------------------+-----------
          Total |       292         149  |        441
                |                        |
           Risk |  .0445205           0  |   .0294785
                |                        |
                |      Point estimate    |    [95% Conf. Interval]
                |------------------------+------------------------
 Risk difference |         .0445205       |    .0208642    .0681769
      Risk ratio |                .       |           .           .
 Attr. frac. ex. |                1       |           .           .
 Attr. frac. pop |                1       |
                +-------------------------------------------------
                             chi2(1) =     6.84  Pr>chi2 = 0.0089
```

CRC group

```
. ttest age, by(training), if nc==1

Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
```

```
       0 |      34    67.94118    2.019487    11.77553     63.8325    72.04985
       1 |      82    68.63415    1.164505    10.54504    66.31715    70.95115
---------+--------------------------------------------------------------------
combined |     116    68.43103    1.009492    10.87256    66.43143    70.43064
---------+--------------------------------------------------------------------
    diff |           -.6929699     2.22652               -5.103688    3.717749
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =  -0.3112
Ho: diff = 0                                     degrees of freedom =      114

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.3781         Pr(|T| > |t|) = 0.7562          Pr(T > t) = 0.6219
```

```
. cs sex training if nc ==1
```

	TRAINING		
	Exposed	Unexposed	Total
Cases	47	23	70
Noncases	35	11	46
Total	82	34	116
Risk	.5731707	.6764706	.6034483
	Point estimate		[95% Conf. Interval]
Risk difference	-.1032999		-.2935324 .0869327
Risk ratio	.8472959		.6288244 1.14167
Prev. frac. ex.	.1527041		-.1416705 .3711756
Prev. frac. pop	.107946		

chi2(1) = 1.07 Pr>chi2 = 0.3006

```
. cs famhx_cca training if nc ==1
```

	TRAINING		
	Exposed	Unexposed	Total
Cases	20	7	27
Noncases	60	26	86
Total	80	33	113
Risk	.25	.2121212	.2389381
	Point estimate		[95% Conf. Interval]
Risk difference	.0378788		-.1308167 .2065743
Risk ratio	1.178571		.5516041 2.518166
Attr. frac. ex.	.1515152		-.8128943 .6028855
Attr. frac. pop	.1122334		

chi2(1) = 0.18 Pr>chi2 = 0.6677

```
. cs fh_any_ca training if nc ==1
```

	TRAINING		
	Exposed	Unexposed	Total
Cases	60	18	78
Noncases	19	16	35
Total	79	34	113

```
                |                       |
           Risk |  .7594937    .5294118 |   .6902655
                |                       |
                |      Point estimate   |    [95% Conf. Interval]
                |-----------------------+------------------------
 Risk difference |         .2300819      |    .0376489    .4225149
      Risk ratio |         1.434599      |    1.020762    2.016214
 Attr. frac. ex. |         .3029412      |    .0203398    .5040209
 Attr. frac. pop |         .2330317      |
                +-------------------------------------------------
                           chi2(1) =     5.89  Pr>chi2 = 0.0153
```

```
. cs smoke training if nc ==1
```

```
                | TRAINING              |
                |   Exposed   Unexposed |      Total
----------------+-----------------------+------------
          Cases |        15           4 |         19
       Noncases |        60          29 |         89
----------------+-----------------------+------------
          Total |        75          33 |        108
                |                       |
           Risk |        .2    .1212121 |   .1759259
                |                       |
                |      Point estimate   |    [95% Conf. Interval]
                |-----------------------+------------------------
 Risk difference |         .0787879      |   -.0647212    .2222969
      Risk ratio |             1.65      |    .5925316    4.594691
 Attr. frac. ex. |         .3939394      |   -.6876736    .7823575
 Attr. frac. pop |         .3110048      |
                +-------------------------------------------------
                           chi2(1) =     0.98  Pr>chi2 = 0.3219
```

```
. cs dm training if nc ==1
```

```
                | TRAINING              |
                |   Exposed   Unexposed |      Total
----------------+-----------------------+------------
          Cases |        19           8 |         27
       Noncases |        63          26 |         89
----------------+-----------------------+------------
          Total |        82          34 |        116
                |                       |
           Risk |  .2317073    .2352941 |   .2327586
                |                       |
                |      Point estimate   |    [95% Conf. Interval]
                |-----------------------+------------------------
 Risk difference |        -.0035868      |    -.172906    .1657324
      Risk ratio |         .9847561      |    .4779611    2.028919
 Prev. frac. ex. |         .0152439      |   -1.028919    .5220389
 Prev. frac. pop |         .0107759      |
                +-------------------------------------------------
                           chi2(1) =     0.00  Pr>chi2 = 0.9668
```

```
. cs sympt_gibleed training if nc ==1
```

```
                | TRAINING              |
                |   Exposed   Unexposed |      Total
----------------+-----------------------+------------
          Cases |        49          22 |         71
       Noncases |        33          12 |         45
----------------+-----------------------+------------
          Total |        82          34 |        116
                |                       |
           Risk |   .597561    .6470588 |    .612069
                |                       |
                |      Point estimate   |    [95% Conf. Interval]
```

```
                |-----------------------+-------------------------
 Risk difference |        -.0494978      |   -.2420296    .1430339
      Risk ratio |         .9235033      |    .6805655    1.253161
 Prev. frac. ex. |         .0764967      |   -.2531613    .3194345
 Prev. frac. pop |         .0540752      |
                +-------------------------------------------------
                              chi2(1) =     0.25  Pr>chi2 = 0.6185


. cs sympt_bowelhabit training if nc ==1

                | TRAINING              |
                |   Exposed   Unexposed |      Total
----------------+-----------------------+------------
          Cases |        43          19 |         62
       Noncases |        39          15 |         54
----------------+-----------------------+------------
          Total |        82          34 |        116
                |                       |
           Risk |  .5243902    .5588235 |   .5344828
                |                       |
                |      Point estimate   |    [95% Conf. Interval]
                |-----------------------+-------------------------
 Risk difference |        -.0344333      |   -.2332774    .1644108
      Risk ratio |         .9383825      |    .6527998      1.3489
 Prev. frac. ex. |         .0616175      |   -.3489003    .3472002
 Prev. frac. pop |         .0435572      |
                +-------------------------------------------------
                              chi2(1) =     0.11  Pr>chi2 = 0.7350



. generate or_tumorlocation_rc =.
(560 missing values generated)

. tab or_tumorlocation

OR_tumorlocation |      Freq.     Percent        Cum.
-----------------+-----------------------------------
          Lt COL |         11        9.48        9.48
      Rect Above |          7        6.03       15.52
         Rect At |          6        5.17       20.69
      Rect Below |         17       14.66       35.34
          Rt COL |         40       34.48       69.83
Rt COL & Sigmoid |          1        0.86       70.69
         Sigmoid |         29       25.00       95.69
      Transv COL |          5        4.31      100.00
-----------------+-----------------------------------
           Total |        116      100.00

. replace or_tumorlocation_rc = 0 if or_tumorlocation == "Lt COL"
(11 real changes made)

. replace or_tumorlocation_rc = 0 if or_tumorlocation == "Rt COL"
(40 real changes made)

. replace or_tumorlocation_rc = 0 if or_tumorlocation == "Rt COL & Sigmoid"
(1 real change made)

. replace or_tumorlocation_rc = 0 if or_tumorlocation == "Sigmoid"
(29 real changes made)

. replace or_tumorlocation_rc = 0 if or_tumorlocation == "Transv COL"
(5 real changes made)

. replace or_tumorlocation_rc = 1 if or_tumorlocation == "Rect Above"
(7 real changes made)

. replace or_tumorlocation_rc = 1 if or_tumorlocation == "Rect At"
(6 real changes made)
```

```
. replace or_tumorlocation_rc = 1 if or_tumorlocation == "Rect Below"
(17 real changes made)

. tab or_tumorlocation_rc or_tumorlocation_rc

or_tumorlo |  or_tumorlocation_rc
 cation_rc |         0          1 |     Total
-----------+----------------------+----------
         0 |        86          0 |        86
         1 |         0         30 |        30
-----------+----------------------+----------
     Total |        86         30 |       116


. cs or_tumorlocation_rc training if nc == 1

                 | TRAINING               |
                 |   Exposed   Unexposed  |      Total
-----------------+------------------------+------------
           Cases |        23           7  |         30
        Noncases |        59          27  |         86
-----------------+------------------------+------------
           Total |        82          34  |        116
                 |                        |
            Risk |  .2804878    .2058824  |   .2586207
                 |                        |
                 |      Point estimate    |    [95% Conf. Interval]
                 |------------------------+------------------------
 Risk difference |         .0746055       |   -.0925075    .2417184
      Risk ratio |         1.362369       |    .6463542    2.871568
 Attr. frac. ex. |         .2659847       |   -.5471394    .6517583
 Attr. frac. pop |         .2039216       |
                 +-------------------------------------------------
                              chi2(1) =     0.70  Pr>chi2 = 0.4036



. generate or_tumorlocation_lr = .
(560 missing values generated)

. replace or_tumorlocation_lr = 0 if or_tumorlocation == "Rt COL"
(40 real changes made)

. replace or_tumorlocation_lr = 1 if or_tumorlocation != "Rt COL"
(520 real changes made)

. cs or_tumorlocation_lr training if nc == 1

                 | TRAINING               |
                 |   Exposed   Unexposed  |      Total
-----------------+------------------------+------------
           Cases |        55          21  |         76
        Noncases |        27          13  |         40
-----------------+------------------------+------------
           Total |        82          34  |        116
                 |                        |
            Risk |  .6707317    .6176471  |   .6551724
                 |                        |
                 |      Point estimate    |    [95% Conf. Interval]
                 |------------------------+------------------------
 Risk difference |         .0530846       |   -.1393431    .2455124
      Risk ratio |         1.085947       |    .8005879    1.473017
 Attr. frac. ex. |         .0791444       |    -.249082    .3211214
 Attr. frac. pop |         .0572755       |
                 +-------------------------------------------------
                              chi2(1) =     0.30  Pr>chi2 = 0.5840
```

```
. cs li training if nc == 1

                 | TRAINING               |
                 |   Exposed   Unexposed  |      Total
-----------------+------------------------+------------
           Cases |        26           6  |         32
        Noncases |        48          22  |         70
-----------------+------------------------+------------
           Total |        74          28  |        102
                 |                        |
            Risk |  .3513514    .2142857  |   .3137255
                 |                        |
                 |      Point estimate    |    [95% Conf. Interval]
                 |------------------------+------------------------
 Risk difference |         .1370656       |     -.04983    .3239613
      Risk ratio |          1.63964       |    .7562373    3.554993
 Attr. frac. ex. |         .3901099       |   -.3223363    .7187055
 Attr. frac. pop |         .3169643       |
                 +-------------------------------------------------
                              chi2(1) =     1.77  Pr>chi2 = 0.1831

. cs vi training if nc == 1

                 | TRAINING               |
                 |   Exposed   Unexposed  |      Total
-----------------+------------------------+------------
           Cases |         9           4  |         13
        Noncases |        65          24  |         89
-----------------+------------------------+------------
           Total |        74          28  |        102
                 |                        |
            Risk |  .1216216    .1428571  |    .127451
                 |                        |
                 |      Point estimate    |    [95% Conf. Interval]
                 |------------------------+------------------------
 Risk difference |        -.0212355       |   -.1707183    .1282472
      Risk ratio |         .8513514       |    .2849328    2.543755
 Prev. frac. ex. |         .1486486       |   -1.543755    .7150672
 Prev. frac. pop |         .1078431       |
                 +-------------------------------------------------
                              chi2(1) =     0.08  Pr>chi2 = 0.7741

. cs pni training if nc == 1

                 | TRAINING               |
                 |   Exposed   Unexposed  |      Total
-----------------+------------------------+------------
           Cases |        11           2  |         13
        Noncases |        60          26  |         86
-----------------+------------------------+------------
           Total |        71          28  |         99
                 |                        |
            Risk |  .1549296    .0714286  |   .1313131
                 |                        |
                 |      Point estimate    |    [95% Conf. Interval]
                 |------------------------+------------------------
 Risk difference |          .083501       |   -.0437132    .2107152
      Risk ratio |         2.169014       |    .5129994    9.170815
 Attr. frac. ex. |          .538961       |   -.9493202    .8909584
 Attr. frac. pop |          .456044       |
                 +-------------------------------------------------
                              chi2(1) =     1.23  Pr>chi2 = 0.2679

. cs lymphocyctic_resp training if nc == 1

                 | TRAINING               |
                 |   Exposed   Unexposed  |      Total
-----------------+------------------------+------------
           Cases |        36          11  |         47
        Noncases |        28          12  |         40
-----------------+------------------------+------------
```

```
           Total |        64          23  |         87
                 |                        |
            Risk |     .5625    .4782609  |   .5402299
                 |                        |
                 |      Point estimate    |    [95% Conf. Interval]
                 |------------------------+------------------------
 Risk difference |         .0842391       |   -.1533473    .3218255
      Risk ratio |         1.176136       |    .7289227    1.897728
 Attr. frac. ex. |         .1497585       |   -.3718876     .473054
 Attr. frac. pop |         .1147086       |
                 +-------------------------------------------------
                           chi2(1) =     0.48  Pr>chi2 = 0.4869


. cs igradepath_1 training if nc == 1

                 | TRAINING               |
                 |   Exposed   Unexposed  |      Total
-----------------+------------------------+------------
           Cases |        62          24  |         86
        Noncases |        12           6  |         18
-----------------+------------------------+------------
           Total |        74          30  |        104
                 |                        |
            Risk |  .8378378          .8  |   .8269231
                 |                        |
                 |      Point estimate    |    [95% Conf. Interval]
                 |------------------------+------------------------
 Risk difference |         .0378378       |   -.1281163     .203792
      Risk ratio |         1.047297       |     .853106    1.285692
 Attr. frac. ex. |         .0451613       |   -.1721872    .2222088
 Attr. frac. pop |         .0325581       |
                 +-------------------------------------------------
                           chi2(1) =     0.21  Pr>chi2 = 0.6440

. cs _Igradepath_2 training if nc == 1

                 | TRAINING               |
                 |   Exposed   Unexposed  |      Total
-----------------+------------------------+------------
           Cases |         3           3  |          6
        Noncases |        71          27  |         98
-----------------+------------------------+------------
           Total |        74          30  |        104
                 |                        |
            Risk |  .0405405          .1  |   .0576923
                 |                        |
                 |      Point estimate    |    [95% Conf. Interval]
                 |------------------------+------------------------
 Risk difference |        -.0594595       |   -.1758364    .0569174
      Risk ratio |         .4054054       |    .0866463    1.896832
 Prev. frac. ex. |         .5945946       |   -.8968322    .9133537
 Prev. frac. pop |         .4230769       |
                 +-------------------------------------------------
                           chi2(1) =     1.39  Pr>chi2 = 0.2387

. cs _Igradepath_3 training if nc == 1

                 | TRAINING               |
                 |   Exposed   Unexposed  |      Total
-----------------+------------------------+------------
           Cases |         9           3  |         12
        Noncases |        65          27  |         92
-----------------+------------------------+------------
           Total |        74          30  |        104
                 |                        |
            Risk |  .1216216          .1  |   .1153846
                 |                        |
                 |      Point estimate    |    [95% Conf. Interval]
                 |------------------------+------------------------
 Risk difference |         .0216216       |   -.1090309    .1522741
```

```
      Risk ratio |         1.216216       |    .3534132    4.185418
 Attr. frac. ex. |         .1777778       |   -1.829549    .7610752
 Attr. frac. pop |         .1333333       |
                 +-------------------------------------------------
                              chi2(1) =     0.10  Pr>chi2 = 0.7545


. generate istage_of_cancer_1 = .
(560 missing values generated)


. replace istage_of_cancer_1 = 1 if stage_of_cancer == 1 & stage_of_cancer !=.
(21 real changes made)

. replace istage_of_cancer_1 = 0 if stage_of_cancer != 1 & stage_of_cancer !=.
(94 real changes made)

. cs istage_of_cancer_1 training if nc == 1

                 | TRAINING               |
                 |   Exposed   Unexposed  |      Total
-----------------+------------------------+------------
           Cases |        15           6  |         21
        Noncases |        67          27  |         94
-----------------+------------------------+------------
           Total |        82          33  |        115
                 |                        |
            Risk |  .1829268    .1818182  |   .1826087
                 |                        |
                 |      Point estimate    |    [95% Conf. Interval]
                 |------------------------+------------------------
 Risk difference |         .0011086       |   -.1548364    .1570537
      Risk ratio |         1.006098       |    .4273615    2.368562
 Attr. frac. ex. |         .0060606       |   -1.339939     .577803
 Attr. frac. pop |          .004329       |
                 +-------------------------------------------------
                              chi2(1) =     0.00  Pr>chi2 = 0.9889


. cs _Istage_of__2 training if nc == 1

                 | TRAINING               |
                 |   Exposed   Unexposed  |      Total
-----------------+------------------------+------------
           Cases |        18          11  |         29
        Noncases |        64          22  |         86
-----------------+------------------------+------------
           Total |        82          33  |        115
                 |                        |
            Risk |  .2195122    .3333333  |   .2521739
                 |                        |
                 |      Point estimate    |    [95% Conf. Interval]
                 |------------------------+------------------------
 Risk difference |        -.1138211       |   -.2979259    .0702836
      Risk ratio |         .6585366       |    .3500418    1.238911
 Prev. frac. ex. |         .3414634       |   -.2389105    .6499582
 Prev. frac. pop |         .2434783       |
                 +-------------------------------------------------
                              chi2(1) =     1.62  Pr>chi2 = 0.2036

. cs _Istage_of__3 training if nc == 1

                 | TRAINING               |
                 |   Exposed   Unexposed  |      Total
-----------------+------------------------+------------
           Cases |        33           8  |         41
        Noncases |        49          25  |         74
-----------------+------------------------+------------
           Total |        82          33  |        115
                 |                        |
```

```
            Risk |   .402439    .2424242  |   .3565217
                 |                        |
                 |      Point estimate    |    [95% Conf. Interval]
                 |------------------------+------------------------
 Risk difference |         .1600148       |   -.0206637    .3406933
      Risk ratio |         1.660061       |      .85948     3.20636
 Attr. frac. ex. |         .3976125       |   -.1634942    .6881199
 Attr. frac. pop |         .3200296       |
                 +-------------------------------------------------
                              chi2(1) =     2.63  Pr>chi2 = 0.1051

. cs _Istage_of__4 training if nc == 1

                 | TRAINING               |
                 |   Exposed   Unexposed  |      Total
-----------------+------------------------+------------
           Cases |        16           8  |         24
        Noncases |        66          25  |         91
-----------------+------------------------+------------
           Total |        82          33  |        115
                 |                        |
            Risk |   .195122    .2424242  |   .2086957
                 |                        |
                 |      Point estimate    |    [95% Conf. Interval]
                 |------------------------+------------------------
 Risk difference |        -.0473023       |   -.2168197    .1222151
      Risk ratio |          .804878       |    .3815928    1.697696
 Prev. frac. ex. |          .195122       |   -.6976964    .6184072
 Prev. frac. pop |         .1391304       |
                 +-------------------------------------------------
                              chi2(1) =     0.32  Pr>chi2 = 0.5723


. generate cea_cat = .
(560 missing values generated)

. replace cea_cat=1 if cea > 5 & cea !=.
(28 real changes made)

. replace cea_cat=0 if cea <= 5 & cea !=.
(60 real changes made)

. cs cea_cat training if nc == 1

                 | TRAINING               |
                 |   Exposed   Unexposed  |      Total
-----------------+------------------------+------------
           Cases |        18          10  |         28
        Noncases |        45          15  |         60
-----------------+------------------------+------------
           Total |        63          25  |         88
                 |                        |
            Risk |  .2857143          .4  |   .3181818
                 |                        |
                 |      Point estimate    |    [95% Conf. Interval]
                 |------------------------+------------------------
 Risk difference |        -.1142857       |   -.3363713    .1077998
      Risk ratio |         .7142857       |    .3847034    1.326227
 Prev. frac. ex. |         .2857143       |   -.3262272    .6152966
 Prev. frac. pop |         .2045455       |
                 +-------------------------------------------------
                              chi2(1) =     1.08  Pr>chi2 = 0.2992
```

5.0 NMR Spot Urine Metabolomics as a New and Highly Sensitive Screening Test for Colorectal Adenomatous Polyps

5.1 Abstract

Background: Colorectal cancer (CRC) is a major public health concern. Adenomatous polyps are precursors of CRC and their identification is the basis for population-based colon cancer screening programs. Current non-invasive, fecal-based screening methods have poor diagnostic sensitivities (range 10-30%) for adenomatous polyps and limited patient uptake due to their fecal nature. Novel, patient-acceptable, highly sensitive CRC screening modalities are urgently required. Metabolomics is a new science that identifies patterns of small molecule metabolites and has been shown to predict health and disease states.

Aim: The aim of this study was to use metabolomics from a spot urine sample to develop a diagnostic test that would distinguish healthy subjects from patients with colonic polyps. We achieved this aim by building and refining a metabolomics model that estimated the sensitivity and specificity of adenomatous polyps relative to the gold standard of colonoscopy.

Methods: Through a prospective controlled study, urine samples were collected from 354 subjects with normal colonoscopies, 243 subjects with colonic adenomatous polyps (215 tubular, 28 villous) and 110 subjects with hyperplastic polyps. One-dimensional nuclear magnetic resonance (NMR) spectra were acquired using an Oxford 600Hz NMR spectrometer with a Varian VNMRS two-channel console. The 1H NMR spectrum of each urine sample was analyzed

using Chenomx NMRSuite v7.0 (Chenomx, Inc. Edmonton, Canada). The first 294 of the normal and 200 of the adenoma urine specimens were used as a training set to establish the diagnostic metabolomic model using SIMCA-P+ v12.0.1 (Umetrics, Umea, Sweden). The model built was then validated with the remaining 60 normal and 43 adenoma samples, as well as the hyperplastic samples (the testing set).

Results: A two-component orthogonal partial least squares (OPLS) model for normal vs. adenoma was built; $R^2Y = 0.396$ (model's fit of data), $Q^2 = 0.25$ (model's predictability of data in 7-fold cross-validation). The model had a sensitivity and specificity of 89.5% and 71.8%, respectively. A receiver operating characteristics (ROC) curve was generated and area under the curve (AUC) was calculated to be 0.891 (95% CI 0.864, 0.919). Validation of the model with 103 blinded samples resulted in sensitivity and specificity values of 72.1% and 40.0%, respectively. When the hyperplastic samples were introduced blindly into the adenoma model, exploratory analysis showed that they were more similar to the adenomatous polyps than the normals.

Conclusions: This is the first study to demonstrate that NMR urine metabolomics, as a diagnostic test, has the ability to distinguish normal healthy subjects from patients with adenomatous polyps with far superior accuracy than that of current fecal-based screening tests. Urine metabolomics has the potential to become an accurate, non-invasive, and inexpensive screening tool for CRC.

5.2 Introduction

Colorectal cancer can be curable if detected early and even preventable if identified in the adenomatous polyp stage. The development of colorectal carcinoma is a multi-step process that typically develops over decades and requires mutational activation of many oncogenes coupled with the mutational inactivation of tumor suppressor genes. Approximately 80% of the colorectal carcinomas develop through the loss of heterozygosity (LOH) pathway involving inactivation of APC gene. Other genes involved in the LOH pathway include K-*ras*, *DCC* and *p53*.[1] The cascade of events described by Fearon and Vogelstein begins with a series of genetic mutations that transforms normal colonic mucosa through adenoma to carcinoma.[2] Adenomatous polyps are precursors of CRC and their identification is the basis for population-based CRC screening programs. Once identified, adenomatous polyps can be removed endoscopically. Contrary to adenomatous polyps, hyperplastic polyps are benign growths in the colon that have no malignant potential.

Current world-wide population-based screening uses guaiac-based fecal occult blood tests which only have a sensitivity 10-30% in detecting adenomatous polyps.[3, 4] Newer fecal immunochemical tests (FITs) use antibodies directed against human globin and are thus more specific for colorectal bleeding. Initial clinical trials have demonstrated that the diagnostic performance of several FITs

is superior to standard guaiac-based tests in detecting both colonic adenomas (20-67%) and cancers (61-91%).[5]

It was shown in chapter 4 that urine metabolomics can distinguish healthy normals from CRC, but detecting adenomatous polyps is the key in preventing CRC. This chapter examines urine metabolomic fingerprint for colonic adenomatous polyps.

5.3 Objectives

- To identify the most influential metabolites that contribute to the separation, in the spot urine metabolomic fingerprint, between patients with colorectal adenomatous polyps and those without colorectal adenomatous polyps
- In patients with colorectal adenomatous polyps and those without colorectal adenomatous polyps, using an un-blinded training set of the metabolomic fingerprint data defined above, to build and refine a model that would estimate the sensitivity and specificity of colorectal adenomatous polyps relative to the gold standard of colonoscopy
- Using a blinded testing set of the metabolomic fingerprint data, to further confirm the statistical models and the sensitivity and specificity of the spot urine metabolomic fingerprint as a diagnostic test that would distinguish patients with colorectal adenomatous polyps from those without.

- To compare the sensitivity and specificity of the spot urine metabolomics test with those of fecal occult blood test (Hemoccult II) and fecal immune tests, namely Hemoccult II, Hemoccult ICT, and MagStream HemSp/HT.

5.4 Material & Methods

5.4.1 Recruitment and Sample Collection

Urine samples were collected from 707 subjects aged 39-76 through a regional population-based screening program called SCOPE (Stop COlorectal cancer through Prevention and Education) in Edmonton, Alberta between April 2008 and October 2009. The screening population consisted of average and high-risk individuals (personal or family history of CRC). The midstream urine samples were collected from the screening subjects at the end of the education session with subjects in their normal states, *i.e.* no diet modifications. The urine sample containers were pre-coated with sodium azide drops (27.3mg/mL) to prevent bacterial growth. The urine samples were frozen at -80°C within 24 hours of collection and if the samples could not be frozen immediately, they were stored at 4°C within 4 hours of collection. All urine samples were collected prior to colonoscopy, which was the gold standard for diagnosis. All polyps identified via colonoscopy were removed using standard endoscopic techniques (polypectomy snare with electrocautery or polypectomy forceps). All removed colonic tissue was sent for histological analysis. Pathologists were blinded to the urine metabolomics results. Each case was classified according to the most severe lesion found on colonoscopy based on polyp size, number and histology. 354 out

of 707 had normal colonoscopies, 243 out of 707 had adenomatous polyps, and 110 out of 707 subjects had hyperplastic polyps. A screen relevant neoplasm (SRN) was defined as any adenoma 1.0 cm or greater in size, any adenoma with villous components or high-grade dysplasia on histology or carcinoma of any size. Subjects were excluded if they were anuric, oliguric, have end-stage renal failure, or on hemo or peritoneal dialysis. Ethics approval for this study was obtained from the Health Research Ethics Board at the University of Alberta.

5.4.2 Sample Analysis

All urine samples were stored at -80°C until they were ready to be analyzed. The day prior to NMR acquisition, each sample was thawed to room temperature and was diluted (1:10) with internal standard consisting of 5 mM sodium 2,2-dimethyl-2-silapentane-5-sulfonate (DSS), 100 mM imidazole, 0.2% sodium azide in 99% D2O. The samples were stored at 4°C overnight. On the day of NMR acquisition, each sampled was adjusted to a pH between 6.7 and 6.8 and aliquoted into 5mm NMR tubes. One-dimensional nuclear magnetic resonance spectra was acquired using an Oxford 600Hz NMR spectrometer with a Varian VNMRS two channel console and running VNMRJ software version 2.2C on a RHEL 4 host computer in the Canadian National High Field NMR Centre (NANUC), Edmonton, Alberta. All samples were run at a sweep width (sw) of 7225.43 Hz. The saturation frequency (sfrq), transmitter offset (tof) and pulse width (pw) were all individually calibrated at the start of each day. The tof typically ranged from (-213 to -215 Hz) and the pw ranged from 6 to 8

microseconds. Shims were optimized until an acceptable line width value was obtained at relative peak heights of: 50% (< 1.0 Hz), 0.55% (< 12.0 Hz), and 0.11% (< 20.0 Hz) were achieved. Water suppression was performed. Spectra were collected at 25°C through a total of 32 scans over a period of 3.5 minutes; a total recycle delay of 5 seconds was also used (*i.e.* 1 second recovery delay/saturation and a 4 second acquisition). The 1H NMR spectrum of each urine sample was analyzed and quantitated using the targeted profiling technique[6] as implemented in Chenomx NMRSuite v7.0 (Chenomx, Inc. Edmonton, Canada). The quantification process was done independently by two individuals and verified by a third individual to optimize accuracy. 294 metabolites were considered and 72 were found to be significant.

The spectral acquisition and quantification process were performed without the knowledge of the pathology results.

5.4.3 Data Analysis

The first 294 normal and 200 adenoma samples were used as a training set to establish the diagnostic metabolomic model of normal vs. adenoma using projection-based methods and logistic regression with the aid of SIMCA-P+ v12.0.1 (Umetrics, Umea, Sweden) and STATA/SE 10.1 (TX, USA). The metabolite concentrations were normalized (to total metabolite concentration except urea) to account for the dilutional differences in the urine samples. Log transformation was done to account for the non-normal distributive nature of the

concentrations. Finally, those metabolites that are not products of normal human metabolism, i.e. xenobiotics, such as ibuprofen and salicylurate, were excluded. Diagnostic accuracies (such as sensitivity, specificity, and AUC for ROC curve) were calculated from the statistical model and the model was then validated with 103 blinded urine samples (60 normal, 43 adenoma), the testing set. The clinical characteristics of the two groups were compared and stratified models were built. Chi squared test was used to compare proportional outcomes and student's t-test was used to compare continuous outcomes. Lastly the urine metabolomic results were compared to the fecal test results.

5.4.4 Fecal Test Collection and Analysis

As part of the SCOPE pilot study, all screening subjects were required to obtain stool samples for three fecal tests, namely the Hemoccult II® (Beckman Coulter Canada Inc.), Hemoccult ICT® (Beckman Coulter Inc. USA) and the MagStream HemSp/HT® (Fujirebio Inc, Japan and Fujirebio Diagnostics Inc, USA) occult blood tests. Each participant was instructed on the proper use of each test kit during the educational sessions. All study participants completed the three occult blood kits from the same bowel movement on each of two consecutive days at home 10 days prior to the booked date of the colonoscopy. No medication or dietary restrictions were required prior to or during the stool collection except that vitamin C supplements were to be discontinued three days prior. Stool samples were not to be collected three days prior, during or three days after a menstrual period, if they had bleeding hemorrhoids or if there was blood in their urine. The

Hemoccult II and Hemoccult ICT stool specimens were to be obtained from 2 portions of 1 bowel movement and applied to the test card windows. MagStream HemSp samples were to be obtained from multiple areas of the same bowel movements using the collection probes (Hemetubes) provided. Once all stool collections were complete, samples from day 1 and day 2 were placed in separate plastic biohazard bag and delivered in person to an outpatient collection site. Samples were stored at 4°C and analyzed within 4 days of receipt. All occult blood stool samples were analyzed at a single laboratory according to standard manufacturer's instructions. Each occult blood test was analyzed independently by trained laboratory personnel who were blinded to patient history and colonoscopy results.[7]

5.5 Results

5.5.1 Training Set Subject Characteristics

Table 5.1 lists the demographics and clinical characteristics comparing the normal and the adenoma groups. It is a little surprising that the two groups are quite different from each other. There are more females in the normal group (60%) compared to the adenoma group (41%) ($p<0.001$). The average age for the adenoma group (59.4 ± 0.6 years) is approximately 4 years older than that of the normal group (55.3 ± 0.5 years) ($p<0.001$). There are more people in the normal group with positive family history of CRC (69%) or any cancer (92%) compared to the adenoma group (61%, 71% respectively) ($p=0.055$, $p<0.001$, respectively), and surprisingly more people in the normal group who have altered bowel habits

compared to the adenoma group (4% vs. 0.5%, p=0.010). Smoking and diabetes are equally prevalent for the two groups.

Table 5.1: Patient characteristics

	NORMAL [N=294] N (%)	ADENOMA [N= 200] N (%)	p-VALUE
Male:Female	117:177	118:82	<0.001*
Average age (years ± SEM)	55.3 ± 0.5	59.4 ± 0.6	<0.001*
FHx of Colon or Rectal cancer	191 (69)	112 (61)	0.055
FHx of Any Cancer	230 (92)	141 (71)	<0.001*
Smoking	26 (9)	29 (15)	0.058
Diabetes	14 (5)	8 (4)	0.687
GI Bleeding	5 (2)	6 (3)	0.340
Change in Bowel Habit	13 (4)	1 (0.5)	0.010*

Note: Not all % are calculated with the denominator of the total in each group as some clinical information was missing or unknown.
* $p \leq 0.05$

Adenomatous polyps were classified according to pathology – 87% of the adenomas were tubular and 13% were tubulovillous or villous (figure 5.1). Where multiple adenomas were found in the same patient, they were classified by the largest or most histologically advanced lesion found. The presence of high-grade dysplasia was also noted. A screen relevant neoplasm (SRN) was defined as any adenoma 1.0 cm or greater in size, any adenoma with villous components or high-grade dysplasia on histology or carcinoma of any size. There were 52 out of 200 or 26% adenoma subjects with screening relevant neoplasms (figure 5.2). The adenomas were fairly evenly distributed along the colon and rectum (figure 5.3) although the locations of the polyps were endoscopically determined by the gastroenterologists/surgeons and there can be a lot of inter-observer variability.

Figure 5.1: Adenoma by type

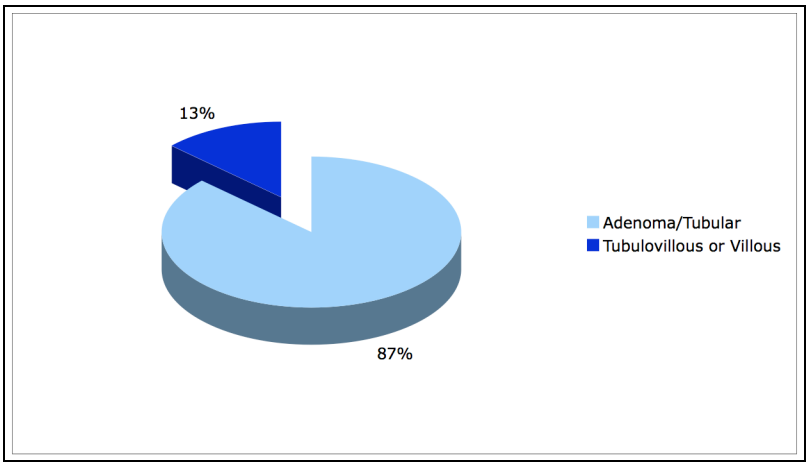


Figure 5.2: Screen relevant neoplasm

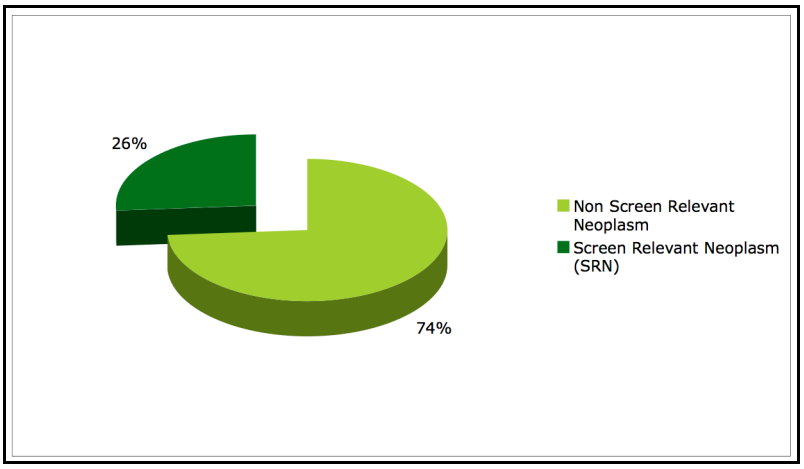
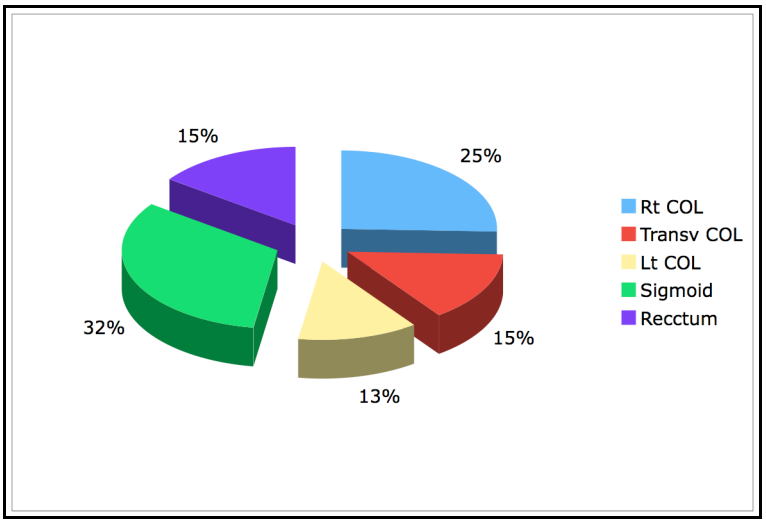


Figure 5.3: Adenoma by location



5.5.2 Building the Models

Unsupervised principal component analysis (PCA) was unable to generate a statistically significant model to separate out the two groups. However, using two-component separation, a supervised orthogonal partial least squares (OPLS) model was built with R^2Y of 0.396, and Q^2 of 0.250. The OPLS scatter plot shown below (figure 5.4) illustrates the normal group in black squares and the cancer group in red diamonds. A crude exploratory data analysis shows that although there's some degree of overlap, it is clear that the two groups are showing up in different areas of the plot. A partial least squared discriminate analysis (PLS-DA) model was also built and extra components were added so the data can be represented using a three-dimensional scatter plot (figure 5.5).

Figure 5.4: OPLS Scatter plot of normal (black squares) vs. adenoma (red diamonds)

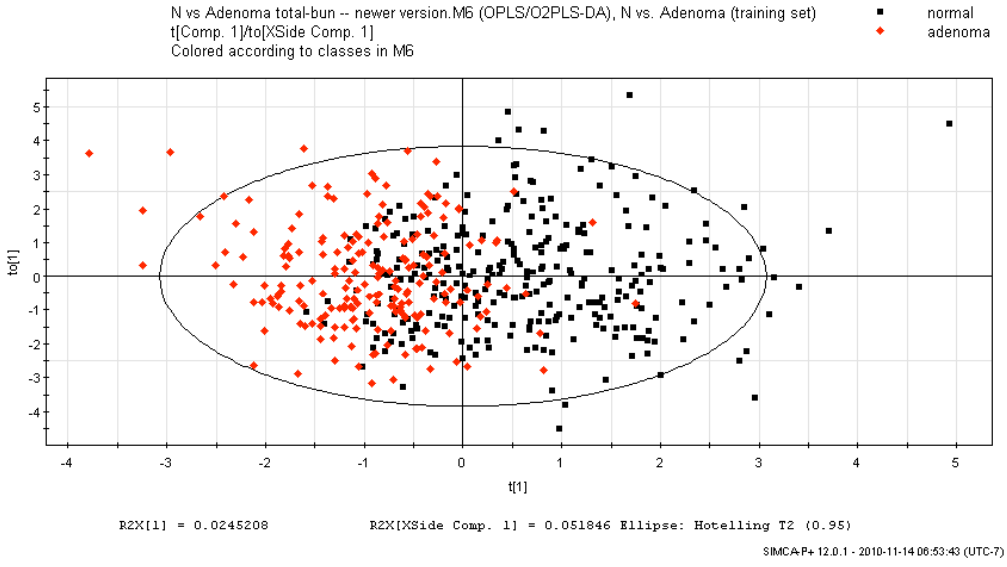
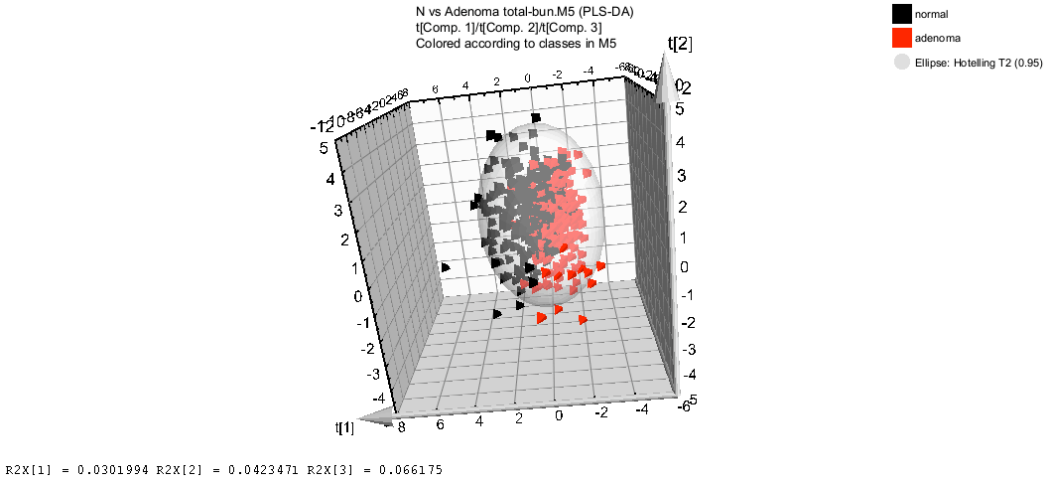
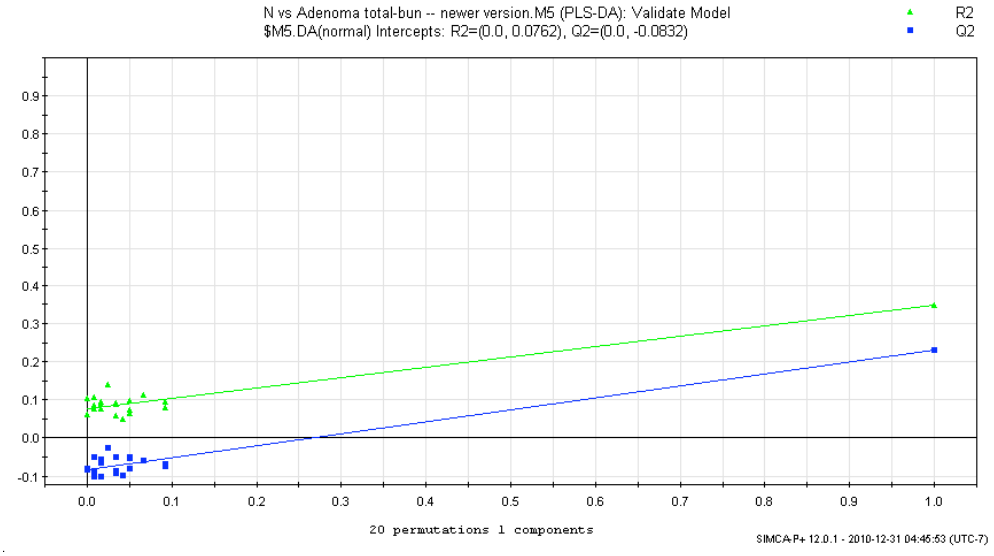


Figure 5.5: PLS-DA 3-D scatter plot of normal (black pyramids) vs. adenoma (red pyramids)



The PLS-DA model was again internally validated using permutation tests and the validation plot is shown as figure 5.6. Twenty models were generated and the R^2 and Q^2 values for these models are shown on the left side of the validation plot. The R^2 and Q^2 values of the original normal vs. adenoma model are shown far to the right. This validation plot shows that all the R^2 and Q^2 values for the randomly generated models are lower than the original points to the right, that is, the model is not over-fit.

Figure 5.6: Validation plot



5.5.3 Diagnostic Accuracies

A spectrum of sensitivity and specificity were again calculated using the Observed vs. Predicted plot (figure 5.7). Three representative pairs of diagnostic accuracies are listed in table 5.2. Receiver operating characteristics (ROC) curve is generated (figure 5.8) from the range of sensitivity and specificity values and AUC is calculated to be 0.8913 (95% CI 0.8639, 0.9187).

Figure 5.7: Observed vs. predicted plot of normal (black squares) vs. adenoma (red diamonds) model

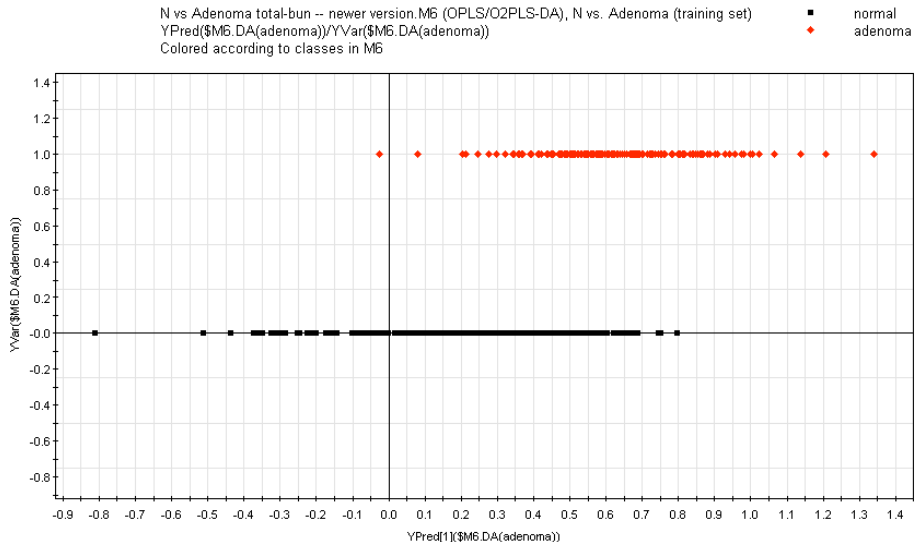
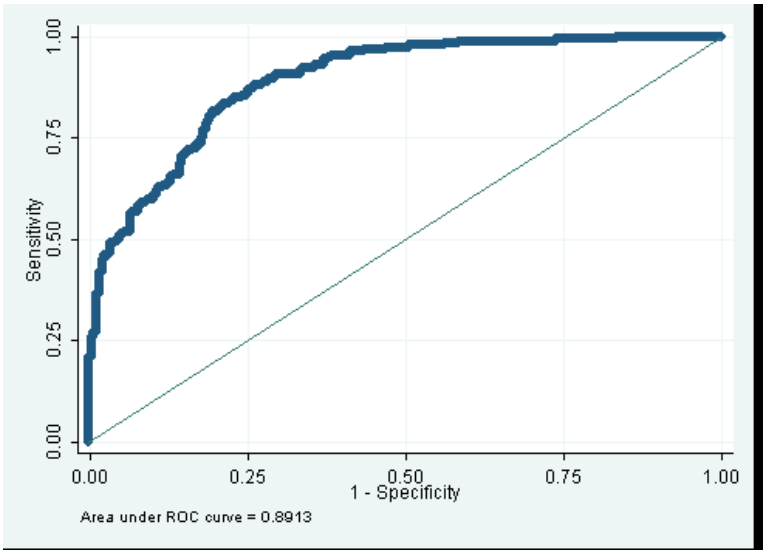


Table 5.2: Representative diagnostics for normal vs. adenoma OPLS model

Cut off level	Sensitivity	Specificity	R^2Y	Q^2	AUC
0.276579	97.50%	53.06%	0.396	0.25	0.8913
0.422109	89.50%	71.77%			
0.491614	81.00%	80.61%			

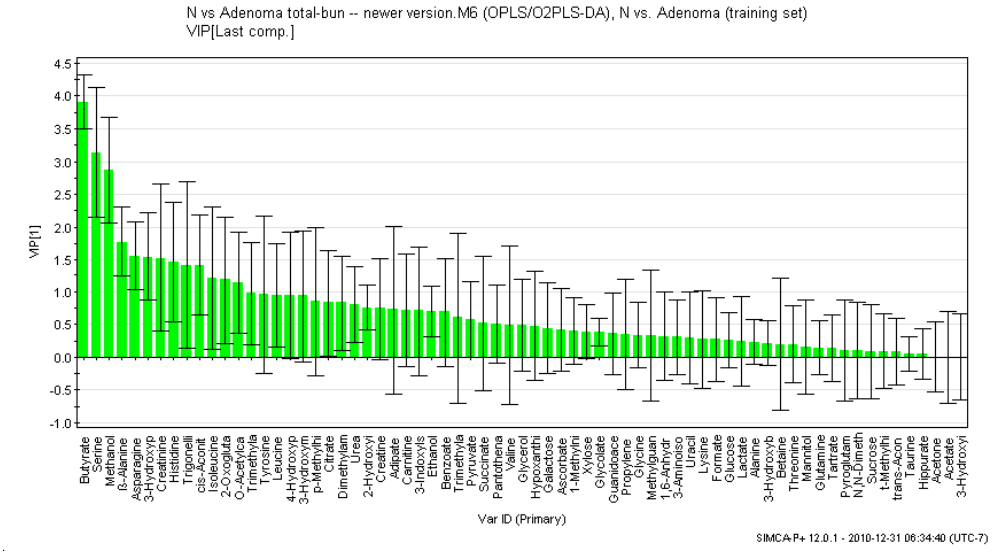
Figure 5.8: Receiver operating characteristics (ROC) curve for normal vs. adenoma OPLS model



5.5.4 Metabolites

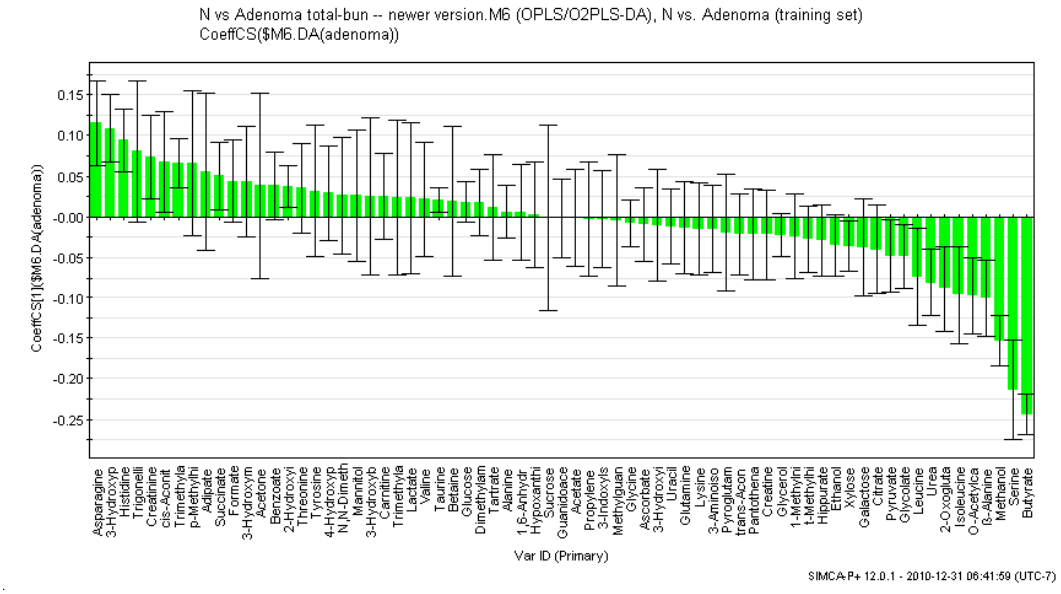
The most contributing metabolites to the separation between normal and adenoma are shown in figure 5.9.

Figure 5.9: Variable importance plot of normal vs. adenoma OPLS model



The top 10 metabolites that contribute to the separation of normal and adenoma (in order of importance) are: butyrate, serine, methanol, β-alanine, asparagine, 3-hydroxyphenylacetate, creatinine, histidine, trigonelline, and cis-aconitate. Specifically, the top 5 metabolites that are higher in concentration in the adenoma samples are asparagine, 3-hydroxyphenylacetate, histidine, trigonelline, and creatinine; and those that are higher in normal samples are butyrate, serine, methanol, β-alanine, and O-acetylcarnitine. This is shown by the coefficient plot (figure 5.10).

Figure 5.10: Coefficient plot for normal vs. adenoma OPLS model

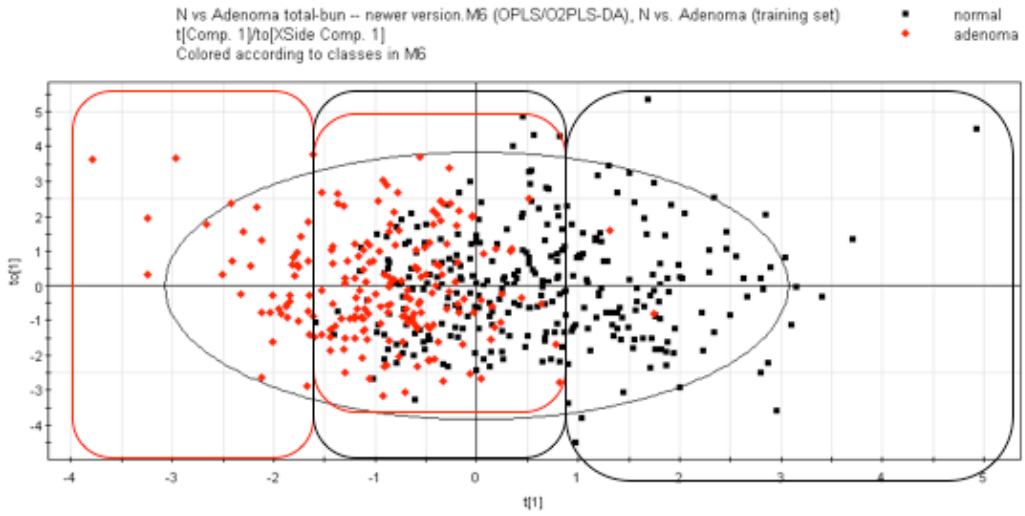


5.5.5 Model Analysis

As there is overlap between the normal group and the adenoma group in the OPLS model generated, it was unclear what characteristics contributed to the overlap, that is, what clinical characteristics do some of the patients with adenoma have that makes them behave like normal and vice versa. For this, the overlapping and non-overlapping groups in the model were studied separately. Four groups were generated: 1) Normal Overlapping, 2) Normal Non-overlapping, 3) Adenoma Overlapping, and 4) Adenoma non-overlapping groups. Since we are working with the OPLS model, only one direction of separation had to be taken into consideration (left-right). To generate the subgroups, we took all the normal (black square) data points to the right of the right-most adenoma (red diamond) data point (excluding the two obvious outliers) and made this the *Non-overlapping Normal group*; and the remainder normal data points the *Overlapping*

Normal group. The same process was done for the adenoma data points. (Figure 5.11)

Figure 5.11: OPLS scatter plot of normal (black squares) vs. adenoma (red diamonds) model with overlapping and non-overlapping groups defined



We looked at each combination of the different subgroups and examined the metabolites that contribute to the separation separately in an attempt to narrow down the list of metabolites that would give us a more powerful model. Next we looked at clinical characteristics that may be different between the subgroups to determine the best traits to stratify the groups to generate more predictive models.

5.5.5.2 Subdividing The Normal Group

5.5.5.2.1 Metabolites

- Adenoma vs. Non-overlapping Normal

When the adenoma group was plotted against the non-overlapping normal group, that is, the group of normal that is the most different from the adenoma group, the

top metabolites that drove the separation were: butyrate, serine, methanol, β-alanine, and isoleucine (table 5.3).

- Adenoma vs. Overlapping Normal

When the adenoma group was plotted against the overlapping normal group, that is, the group of normal that somehow resembles the adenoma group metabolomically, the top metabolites that drove the separation were: butyrate, methanol, cis-aconitate, asparagine, and serine (table 5.3).

- Overlapping Normal Vs. Non-overlapping Normal

The overlapping group and the non-overlapping group of normal are plotted together on a scatter plot, without the adenoma group, to see what the metabolomic difference is between these two groups of normal. The top 5 metabolites that contributed to the separation between these two groups of Normal were: butyrate, serine, leucine, methanol, and β-alanine (table 5.3).

Table 5.3: Summary of metabolites from sub models of normal subjects

Adenoma vs. Non-overlapping Normal	Adenoma vs. Overlapping Normal	Overlapping vs. Non-overlapping Normal
Butyrate	Butyrate	Butyrate*
Serine	Methanol	Serine*
Methanol	Cis-Aconitate	Leucine*
β-alanine	Asparagine	Methanol*
Isoleucine	Serine	β-alanine*
Leucine	Tyrosine	Isoleucine*
Trigonelline**	Histidine	Uracil*
2-oxoglutarate	Urea	2-oxoglutarate*
O-Acetylcarnitine	Creatinine	Valine
Creatinine**	3-Hydroxyphenylacetate	Pyroglutamate
3-Hydroxyphenylacetate**	Valine	O-Acetylcarnitine*
Asparagine**	π-methylhistidine	Threonine
Citrate**	2-Hydroxyisobutyrate	Glutamine
Uracil	Trimethylamine	Urea
3-Hydroxymandelate**	Trigonelline	Methylguanidine

* Metabolites in Overlapping vs. Non-overlapping Normal model that are also part of Adenoma vs. Non-overlapping Normal model. ** Metabolites in Adenoma vs. Non-overlapping Normal model not in Overlapping vs. Non-overlapping Normal model.

It was hypothesized that if we eliminated those top metabolites that separate overlapping & non-overlapping normal from the main normal vs. adenoma model, perhaps we could minimize the difference between the two normal groups and achieve greater separation between normal and CRC. However, this was not the case as the R^2 and Q^2 values are not as good as that for the main model (table 5.4).

Table 5.4: Summary of sub-model characteristics

Model	R^2Y	Q^2
Main model (Normal vs. Adenoma)	0.396	0.250
Main model excluding the 9 metabolites in Overlapping vs. Non-overlapping Normal model that are also part of Adenoma vs. Non-overlapping Normal model (denoted by *)	0.122	0.009
Main model using only 6 metabolites in Adenoma vs. Non-overlapping Normal model not in Overlapping vs. Non-overlapping Normal model (denoted by **)	0.089	0.059

5.5.5.2.2 Clinical Parameters

Using logistic regression, clinical characteristics such as age, gender, family history, etc of the normal group were tested for their odds of predicting the dichotomous outcome of overlap with the adenoma group or not. The odds ratios and p-values are summarized in the table below (table 5.5). This is an exploratory analysis to identify potential factors for stratification.

Table 5.5: Clinical characteristics of the normal group and the odds of overlapping with adenoma group

Variable		Odds Ratio	p-value
Gender		1.232	0.391
Age		1.015	0.314
Smoking		2.055	0.118
Diabetes		0.943	0.916
Family History	CRC	1.122	0.662
	Any cancer	0.723	0.502
Symptoms	GI bleed	1.071	0.940
	Change bowel habits	0.601	0.371

Demographics

In the normal group, gender and age did not change the odds of overlapping with the adenoma group, OR 1.23 (95% CI 0.76, 1.98), p = 0.391 and OR 1.01 (0.99, 1.04), p = 0.314, respectively. Even when age was tested by categories of 5 years, none of the categories significantly predicts overlap with adenoma. Although the odds of overlapping with adenoma in those that were over the age of 70 was almost three times that of those that were younger than 45, but this was not statistically significant (p = 0.205) and there might not be enough numbers to see this effect. This was reflected by the wide 95% confidence interval (0.557, 15.264) of the odds ratio.

Smoking History

Within the normal group, being a smoker did not increase the odds of overlapping with the adenoma group (OR = 2.05, 95% CI = 0.83, 5.06; p = 0.118), but again since there were only 26 smokers in the normal group, there might not be enough numbers to show a difference.

Diabetes

Having diabetes did not increase the odds of overlapping with the CRC group (OR = 0.94, 95% CI = 0.32, 2.79; p=0.916), although there were only 14 out of 294 subjects who had diabetes, and this might be too few to show a difference.

Family history

Within the normal group, having a family history of CRC did not increase the odds of overlapping with the CRC group (OR = 1.12; 95% CI = 0.67, 1.88; p=0.662). Having a family history of *any* cancer(s) did not increase the odds of overlapping with the cancer group either (OR = 0.72; 95% CI = 0.28, 1.86; p=0.502).

Gastrointestinal Symptoms

Within the normal group, having symptoms of gastrointestinal bleeding (OR = 1.07; 95% CI = 0.18, 6.51; p=0.940) or changes in bowel habits (OR = 0.60; 95% CI = 0.20, 1.84; p=0.371) did not increase the odds of overlapping with the CRC group, although the number of subjects with GI symptoms in this screening population was small.

5.5.5.3 Subdividing The Adenoma Group

5.5.5.3.1 Metabolites

- Normal vs. Non-overlapping Adenoma

When the normal group was plotted against the non-overlapping adenoma group, that is, the group of adenoma that was the most different from the normal group, the top metabolites that drove the separation were: serine, trigonelline, trimethylamine, butyrate, and asparagine (table 5.6).

- Normal vs. Overlapping Adenoma

When the normal group was plotted against the overlapping adenoma group, that is, the group of adenomas that somehow resembled the normal group metabolomically, the top metabolites that drove the separation were butyrate, serine, methanol, β-alanine, and creatinine (table 5.6).

- Overlapping Cancers Vs. Non-overlapping Adenoma

The overlapping group and the non-overlapping group of adenoma were plotted together on a scatter plot, without the normal group, to see what the metabolomic difference was between these two groups of adenoma. The top 5 metabolites that contributed to the separation between these two groups of normal were trimethylamine, trigonelline, asparagine, acetate, and histidine (table 5.6).

Table 5.6: Summary of metabolites from sub models of Adenoma

Normal vs. Non-overlapping Adenoma	Normal vs. Overlapping Adenoma	Overlapping vs. Non-overlapping Adenoma
Serine	Butyrate	Trimethylamine*
Trigonelline	Serine	Trigonelline*
Trimethylamine	Methanol	Asparagine*
Butyrate**	β-alanine	Acetate
Asparagine	Creatinine	Histidine*
3-Hydroxyphenylacetate	3-Hydroxyphenylacetate	Isoleucine*
Histidine	Cis-Aconitate	Pyruvate*
Methanol**	Histidine	4-hydroxyphenylacetate*
2-Oxoglutarate	Asparagine	Benzoate*
Isoleucine	Trimethylamine N-oxide	2-Oxoglutarate*
Benzoate	3-Hydroxymandelate	3-Hydroxyphenylacetate*
4-Hydroxyphenylacetate	Adipate	Serine*
Pyruvate	Carnitine	Citrate*
Cis-Aconitate**	O-Acetylcarnitine	3-Hydroxyisovalerate
Citrate	3-Indoxylsulfate	Trimethylamine N-oxide

* Metabolites in Overlapping vs. Non-overlapping adenoma model that are also in Normal vs. Non-overlapping adenoma model. ** Metabolites in Normal vs. Non-overlapping adenoma model not in Overlapping vs. Non-overlapping adenoma model.

Again we hypothesized that if we eliminated those metabolites that separate Overlapping & Non-overlapping adenoma from the main model, perhaps we could minimize the difference between the two adenoma groups and achieve greater separation between normal and adenoma. However, his was not the case as shown in table 5.7.

Table 5.7: Summary of sub-model characteristics

Model	R^2Y	Q^2
Main model (Normal vs. Adenoma)	0.396	0.250
Main model excluding 12 metabolites in Overlapping vs. Non-overlapping adenoma model that are also in Normal vs. Non-overlapping adenoma model (denoted by*)	0.237	0.127
Main model using only 3 metabolites in Normal vs. Non-overlapping adenoma model not in Overlapping vs. Non-overlapping adenoma model (denoted by**)	0.186	0.181

5.5.5.3.2 Clinical Parameters

Logistic regression analysis was used to test the significance of various clinical and pathological variables of adenoma patients on the dichotomous outcome of overlapping or non-overlapping with normals. The odds ratios and p-values are summarized in table 5.8. This is an exploratory analysis to identify potential factors for stratification.

Table 5.8: Clinical characteristics of the adenoma group and the odds of overlapping with normal group

Variable		Odds Ratio	p-value
Gender		0.434	0.031*
Age		0.989	0.632
Smoking		0.304	0.006*
Diabetes		0.425	0.255
Family history	CRC	1.113	0.766
	Any cancer	0.710	0.393
Symptoms	GI bleed	-	-
	Change bowel habits	-	-
Location of Adenoma	Rectal vs. colon	0.6875	0.410
	Left vs. right	0.753	0.497
Pathology of Adenoma	Villous vs. Tubular	1.134	0.812
	Screening Relevant Neoplasm vs. not	0.633	0.225

Note: There are too few people with symptoms to make any meaningful conclusions.
* $p \leq 0.05$

Demographics

In the adenoma group, the odds of resembling or overlapping with the normal group for males was 0.43 (95% CI 0.20, 0.93) times that of females (p=0.03). That is, males were more different than the normal compared to the females. Age did not change the odds of overlapping with normal, OR 0.99 (95% CI 0.95, 1.03); p=0.632. Even when age was subdivided into 5-year categories, there was not one category that statistically significantly predicts overlapping with normal.

Smoking

Within the adenoma group, being a smoker did decrease the odds of overlapping with the normal group (OR = 0.30; 95% CI = 0.13, 0.71; p = 0.006). That is, smokers were more likely to be different than normal.

Diabetes

Adenoma patients who have diabetes had decreased odds of overlapping with the normal group (OR=0.42; 95% CI=0.10, 1.85; p=0.255), but this was not statistically significant, although there were only 8 out of 200 subjects who have diabetes.

Family history

Within the adenoma group, a positive family history of CRC (OR = 1.11; 95% CI = 0.55, 2.25; p=0.766) and a positive family history of *any* cancer(s) (OR = 0.71; 95% CI = 0.32, 1.56; p=0.393) did not increase the odds of overlapping with the normal group.

Gastrointestinal Symptoms

Since there were only six people with GI bleeding in the adenoma group and one person with altered bowel habits, there was not enough numbers to make meaningful conclusions regarding the effects of gastrointestinal symptoms on the effects of overlapping with normal.

Location of Adenoma

When colon and rectal adenomas were separated and compared in their odds of overlapping with normal, the rectal adenomas were more likely to be different than the normals, but this difference was not statistically significant (OR=0.69; 95% CI=0.28, 1.68; p=0.410). There was also no statistically significant

difference when left-sided (including transverse, descending colon, sigmoid, and rectal) adenomas were compared to right-sided ones (OR=0.75; 95% CI=0.33, 1.71; p=0.497).

Adenoma Pathology

In the adenoma group, there were 174 subjects with tubular adenomas and 26 with tubulovillous or villous adenomas. Villous adenomas were more likely to become malignant than tubular ones[1], but logistic regression revealed that the villous adenomas were not more likely to be different than normal compared to the tubular ones (OR = 1.13; 95% CI = 0.40, 3.21; p=0.812).

A screen relevant neoplasm (SRN) was defined as any adenoma 1.0 cm or greater in size, any adenoma with villous components or high-grade dysplasia on histology or carcinoma of any size. There were 52 out of 200 or 26% adenoma subjects with screening relevant neoplasms. The SRNs were more likely to be different than the normals compared to the non-SRNs (OR = 0.63; 95% CI = 0.30, 1.32; p=0.225), but this did not reach statistical significance.

Summary

In summary, *gender* and *smoking* were statistically significant differences between the overlapping and non-overlapping adenoma subgroups. When these factors were tested with an overall logistic regression test, they were still all statistically different. In fact, the effect size was even bigger when both variables

were in the model, so ideally we should do 4-way stratification. However, since there were only 55 smokers in total, four-way stratification would reduce the number of samples further thus making it hard to use projection-based methods to analyze the data.

5.5.5.4 Stratification

The above method of model analysis allowed us to narrow down the specific clinical characteristics by which to stratify the models – these were gender and smoking. We also chose to stratify by family history of colorectal cancer as this is a known risk factor for CRC[8], as well as family history of any cancer since this stratification was performed for the normal vs. CRC analysis.

5.5.5.4.1 Gender

Separate male and female OPLS models were built for normal vs. adenoma and their model characteristics are shown in table 5.9. Contrary to what was seen for the normal vs. CRC analysis, the *male* model had a numerically better R^2Y, a comparable Q^2 value, and a numerically better AUC compared to the main model, suggesting that this urine metabolomics test for adenomas would work better for males, but the statistical and clinical significance for this was unclear. The scatter plots for the male and female models are shown as figure 5.12 and 5.13, respectively. Note when the male model was autofit, only one component was generated, hence an extra component was added to generate a proper scatter plot

shown here, but the R^2Y and Q^2 values in table 5.9 are those from the autofit model.

Table 5.9: Characteristics of gender-stratified models compared to the normal vs. adenoma training set model

Model	R^2Y	Q^2	Sens	Spec	AUC	AUC 95% CI
Main Normal vs. Adenoma Training Set Model	0.396	0.250	90%	72%	0.8913	0.8639, 0.9187
Male Model	0.436	0.276	90%	74%	0.9027	0.8655, 0.9398
Female Model	0.369	0.164	89%	63%	0.8836	0.8421, 0.9250

Figure 5.12: OPLS scatter plot of the male model of normal (orange triangle) vs. adenoma (blue diamonds)

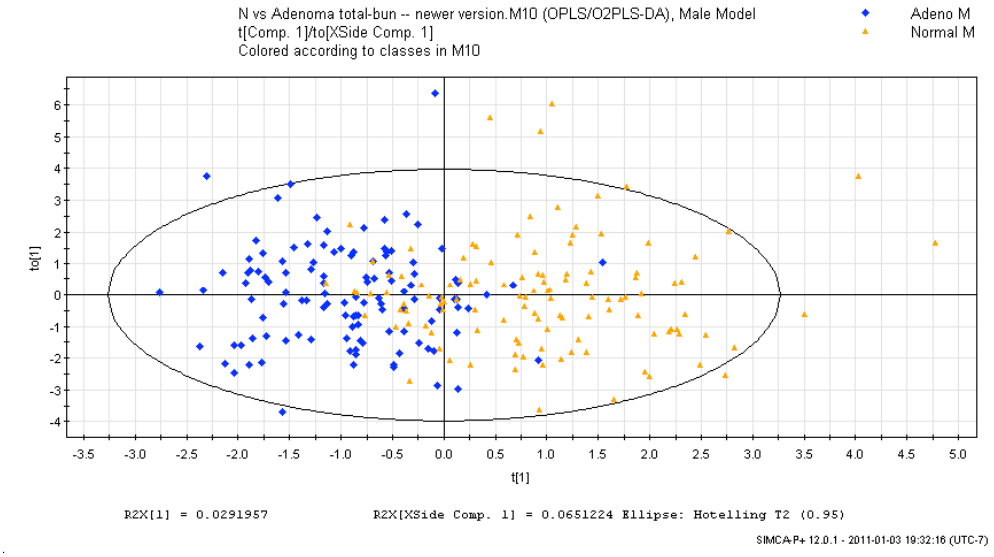
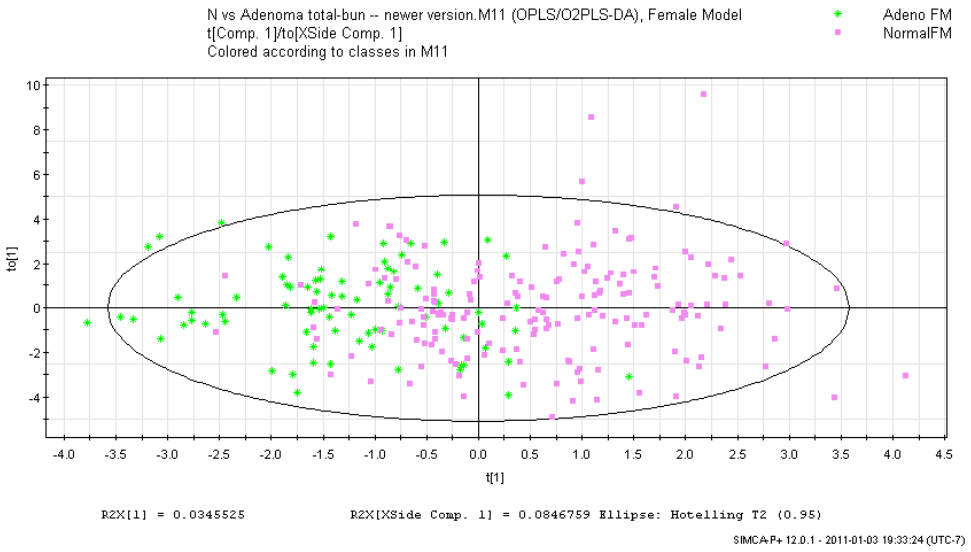


Figure 5.13: OPLS scatter plot of the female model of normal (pink squares) vs. adenoma (green stars)



5.5.5.4.2 Smoking

When stratified models for smoking were compared to the main adenoma model, the diagnostic accuracies were not improved at all (table 5.10). Although the smoking model had a better AUC, but the Q^2 for the model was only 0.04. This was largely due to the fact that there were only 55 people in both the normal and adenoma groups who smoke. The scatter plots for the smoking and non-smoking/ex-smoking/unknown models are shown as figure 5.14 and 5.15, respectively. Again, an extra component had to be generated to create a scatter plot for representation of the data for the smoking model. The scatter plot for the smoking model (figure 5.14) gives a false impression that the separation between the two groups is very good, but in fact, the Q^2 for the model is only 0.04, re-emphasizing that the model characteristics need to be interpreted together with the scatter plots to get an accurate idea of what the data shows. It is also much easier

to find random multivariate correlations when the number of variables exceeds the number of samples, as is the case for the smoking model.

Table 5.10: Characteristics of smoking-stratified models compared to the main normal vs. adenoma training set model

Model	R^2Y	Q^2	Sens	Spec	AUC	AUC 95% CI
Main Normal vs. Adenoma Training Set Model	0.396	0.250	90%	72%	0.8913	0.8639, 0.9187
Smoking Model	0.547	0.040	90%	88%	0.9430	0.8836, 1.000
Non/Ex Smoking/Unknown Model	0.378	0.223	90%	66%	0.8819	0.8514, 0.9124

Figure 5.14: OPLS scatter plot of the smoking model of normal (black squares) vs. adenoma (red diamonds)

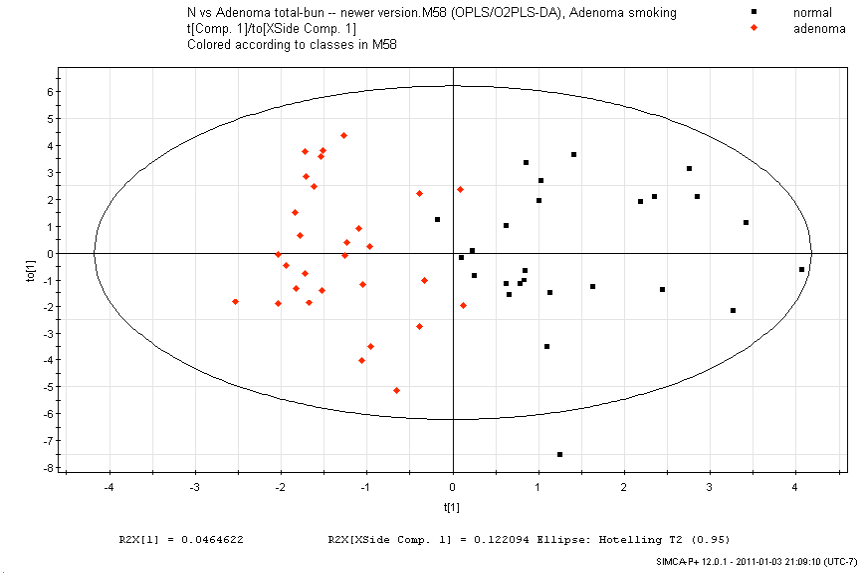
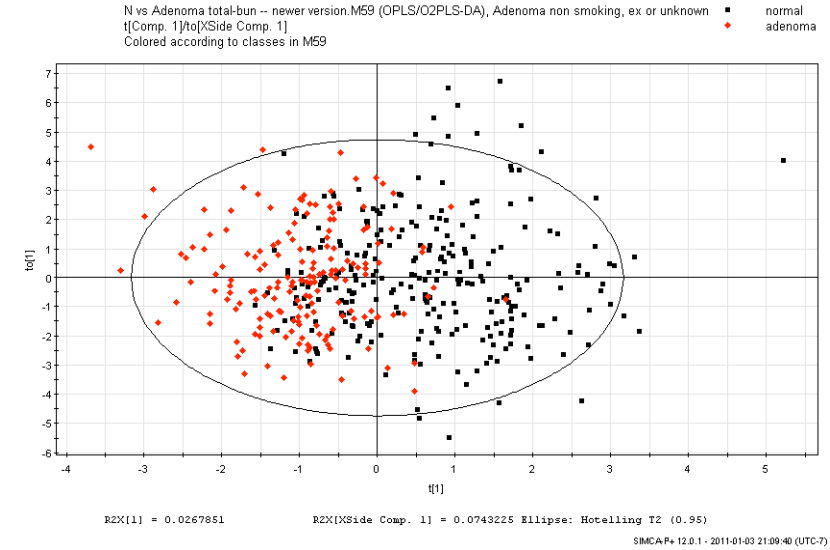


Figure 5.15: OPLS scatter plot of the non/ex-smoking/unknown model of normal (black squares) vs. adenoma (red diamonds)



5.5.5.4.3 Family History of Any Cancer

The normal and adenoma patients were also stratified by family history of any cancer and one OPLS model was built for those with a positive family history of any cancer and one for those without. The model characteristics are listed in table 5.11. The no/unknown family history model had a much better numerical R^2Y, Q^2, and AUC values compared to the main model and the positive family history of any cancer model had comparable model characteristic and diagnostic accuracies to the main model. This is overall suggestive that stratifying by family history of cancer can increase the accuracy of this screening urine metabolomic test, but the statistical and clinical significance of this is unclear. The scatter plots for the stratified models are shown as figure 5.16 and 5.17, respectively.

Table 5.11: Characteristics of adenoma models stratified by family history of any cancer compared to the main normal vs. adenoma training set model

Model	R^2Y	Q^2	Sens	Spec	AUC	AUC 95% CI
Main Normal vs. Adenoma Training Set Model	0.396	0.250	90%	72%	0.8913	0.8639, 0.9187
No/Unknown family history of any cancer	0.573	0.306	95%	88%	0.9502	0.9097, 0.9907
Family history of any cancer	0.407	0.234	90%	73%	0.8957	0.8641, 0.9272

Figure 5.16: Normal (black squares) vs. adenoma (red diamonds) OPLS scatter plot of the no/unknown family history of any cancer groups

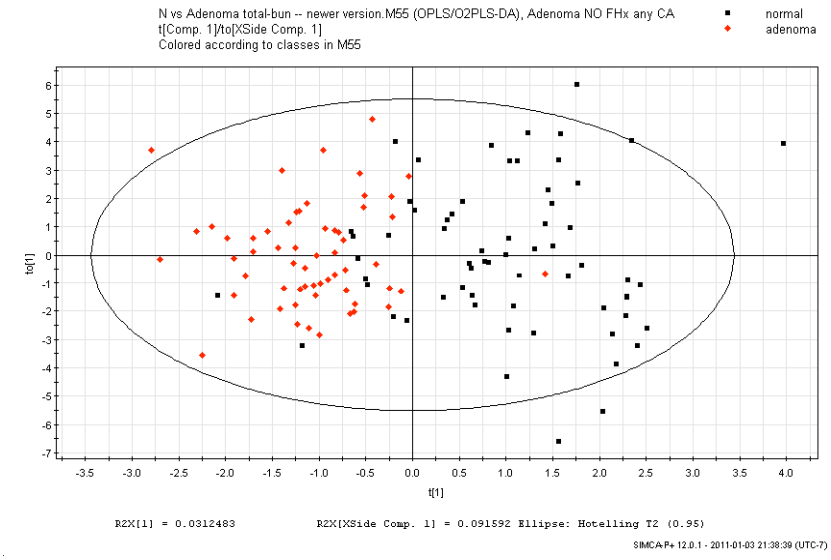
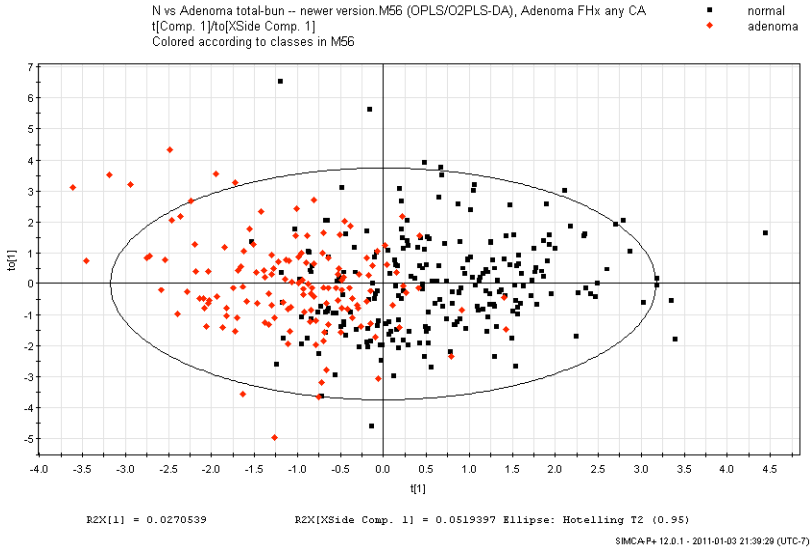


Figure 5.17: Normal (black squares) vs. adenoma (red diamonds) OPLS scatter plot of the positive family history of any cancer groups



5.5.5.4.4 Family History of Colorectal Cancer

The normal and CRC patients were also stratified by family history of CRC and one OPLS model was built for those with a positive family history of CRC and one for those without. The model characteristics are listed in table 5.12. The no/unknown family history model had numerically better R^2Y, comparable Q^2, and higher AUC values compared to the main model and the positive family history of CRC model had a comparable $R^2Y$and higher AUC but a lower Q^2 compared to the main model. This is overall suggestive that stratifying by family history of CRC can increase the accuracy of this screening test, but it is unclear whether this is statistically or clinically significant. The scatter plots for the stratified models are shown as figure 5.18 and 5.19, respectively.

Table 5.12: Characteristics of models stratified by family history of CRC compared to the main model

Model	R^2Y	Q^2	Sens	Spec	AUC	AUC 95% CI
Main Normal vs. Adenoma Training Set Model	0.396	0.250	90%	72%	0.8913	0.8639, 0.9187
No/Unknown family history CRC	0.502	0.267	91%	80%	0.9296	0.8934, 0.9659
Family history of CRC	0.414	0.213	90%	74%	0.8990	0.8645, 0.9336

Figure 5.18: Normal (black squares) vs. adenoma (red diamonds) OPLS scatter plot of the no/unknown family history of CRC groups

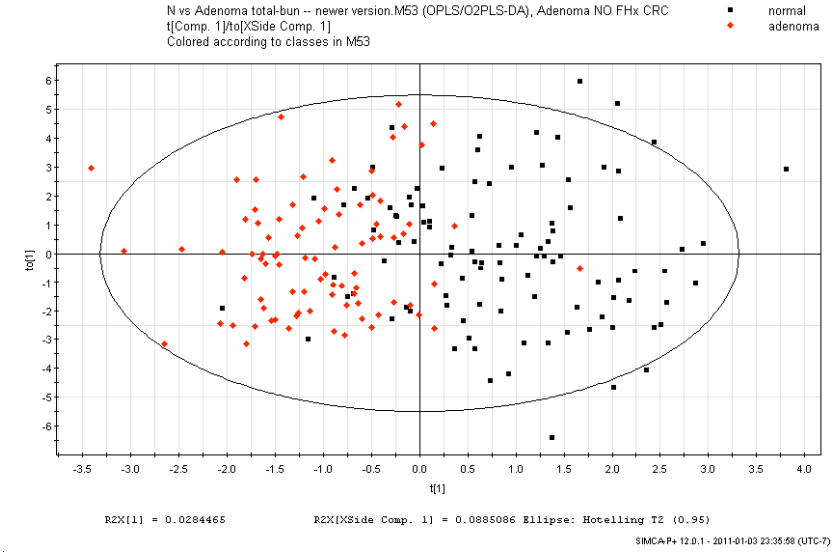
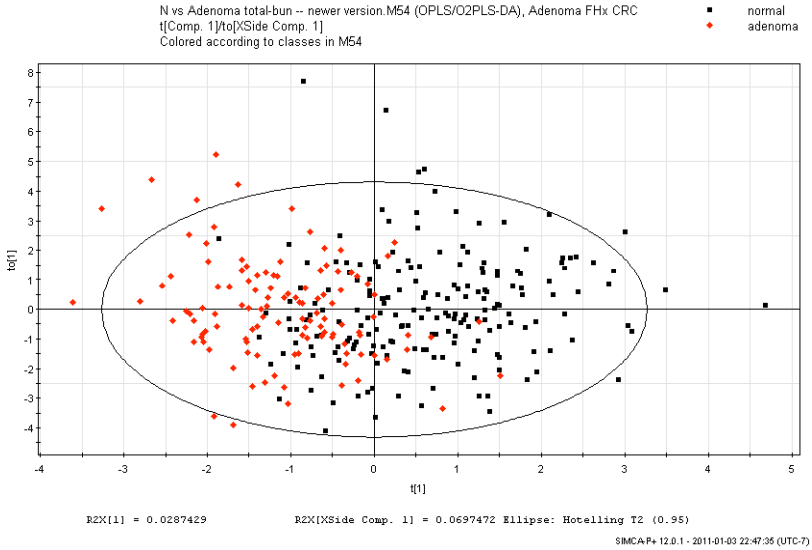


Figure 5.19: Normal (black squares) vs. adenoma (red diamonds) OPLS Scatter plot of the positive family history of CRC groups



5.5.6 Validation with Testing Set

The robustness the metabolomics model is reflected by how well it predicts unknowns. Following our development of the metabolomics model with the un-blinded training set (see above sections 5.5.2), we next used the blinded testing set of 103 urine samples (43 adenoma, 60 normal) to validate the ability of the metabolomics model to distinguish normal from patients with colorectal adenomas.

Ideally the validation samples should be matched and completely representative of the training set. The clinical characteristics of the training set normal subjects were compared to those of the testing set normal subjects (table 5.13), and the same was done for the adenoma subjects (table 5.14). As these validation samples are simply subsequently collected normal and adenoma samples and not

methodologically matched, there are some differences between the training and the testing set. In the normal group, the number of subjects with a positive family history of CRC (69% vs. 54%, p=0.031) and family history of any cancer (92% vs. 65%, p<0.001) were significantly more in the training set compared to the testing set. The gender distribution and number of people with changes in bowel habits were also approaching statistical significance. In the adenoma group, the testing set subjects were older than the training set subjects (62.2±1.1 vs. 59.4±0.6; p = 0.034), and again the number of people with a positive family history of CRC was significantly more in the training set than the testing set (61% vs. 38%, p=0.008).

Table 5.13: Clinical characteristics of ***normal*** subjects in the training set vs. testing set

		Training Set [n=294] N (%)	**Testing Set [n=60] N (%)**	**p-value**
Male:Female		117:177	31:29	0.089
Average age (years ± SEM)		55.3±0.5	55.7±1.1	0.684
Smoking		26 (9)	6 (10)	0.839
Diabetes		14 (5)	4 (7)	0.541
Family History	CRC	191 (69)	31 (54)	0.031*
	Any cancer	230 (92)	39 (65)	<0.001*
Symptoms	GI bleed	5 (2)	2 (3)	0.410
	Change bowel habits	13 (4)	0 (0)	0.096

Note: Not all % are calculated with the denominator of the total in each group as some clinical information was missing or unknown.
* p≤0.05

Table 5.14: Clinical characteristics of ***adenoma*** patients in the training set vs. testing set

		Training Set (n=200) N (%)	**Testing Set (n=43) N (%)**	**p-value**
Male:Female		118:82	27:16	0.646
Average age (years ± SEM)		59.4±0.6	62.2±1.1	0.034*
Smoking		29 (15)	10 (24)	0.166
Diabetes		8(4)	4 (9)	0.145
Family History	CRC	112 (61)	15(38)	0.008*
	Any cancer	141 (71)	27 (63)	0.298
Symptoms	GI bleed	6 (3)	1 (2)	0.810
	Change bowel habits	1 (0.5)	0 (0)	0.642
Location of Adenoma	Rectal vs. colon	30 (15)	8 (19)	0.555
	Left vs. right	149 (75)	27 (63)	0.119
Pathology of Adenoma	Villous vs. Tubular	26 (13)	2 (5)	0.120
	Screening Relevant Neoplasm vs. not	52 (26)	13 (30)	0.570

Note: Not all % are calculated with the denominator of the total in each group as some clinical information was missing or unknown.
* $p \leq 0.05$

Diagnostic accuracies were calculated using the same cutoff (0.491614) from the original model that resulted in a sensitivity of 81.0% and specificity of 80.6%. The sensitivity and specificity from the validation samples are 72.1% and 40.0%, respectively.

When only the testing *adenoma* samples were introduced to the main normal vs. adenoma model (figure 5.20) as the prediction set, these samples showed up on adenoma side of the model (figure 5.21). However, when the 60 *normal* samples were introduced to the original model blindly, the samples still tend to be more on the adenoma side of the plot (figure 5.22), which is as expected with the calculated validation specificity of 40.0%.

Figure 5.20: Original normal (black squares) vs. adenoma (red diamonds) OPLS scatter plot

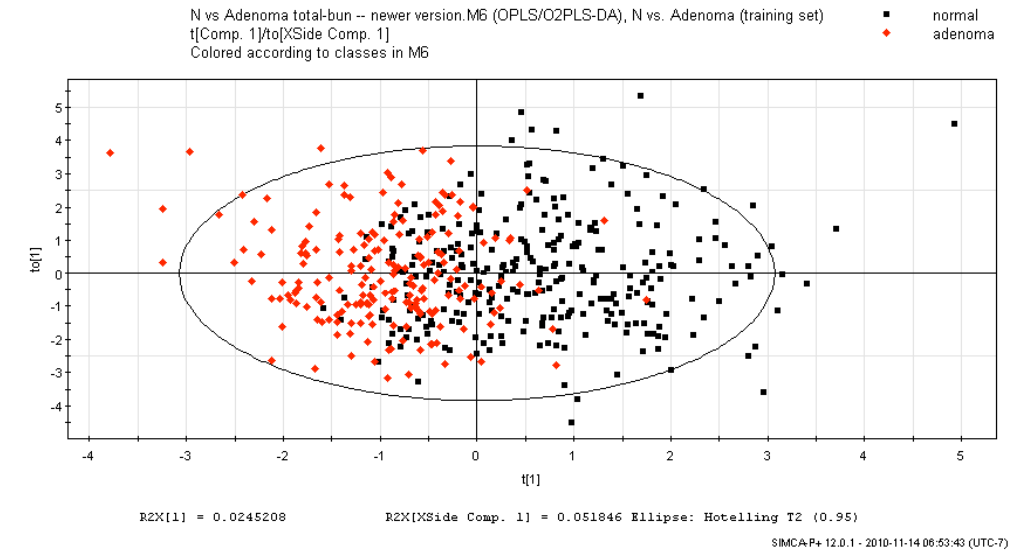


Figure 5.21: Testing set adenoma samples (blue squares) superimposed on normal (black triangles) vs. adenoma (red diamonds) OPLS scatter plot

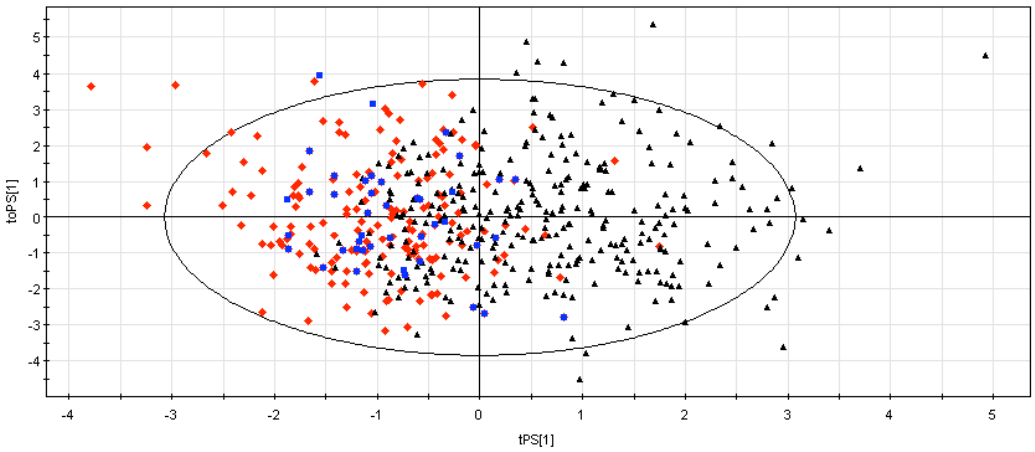
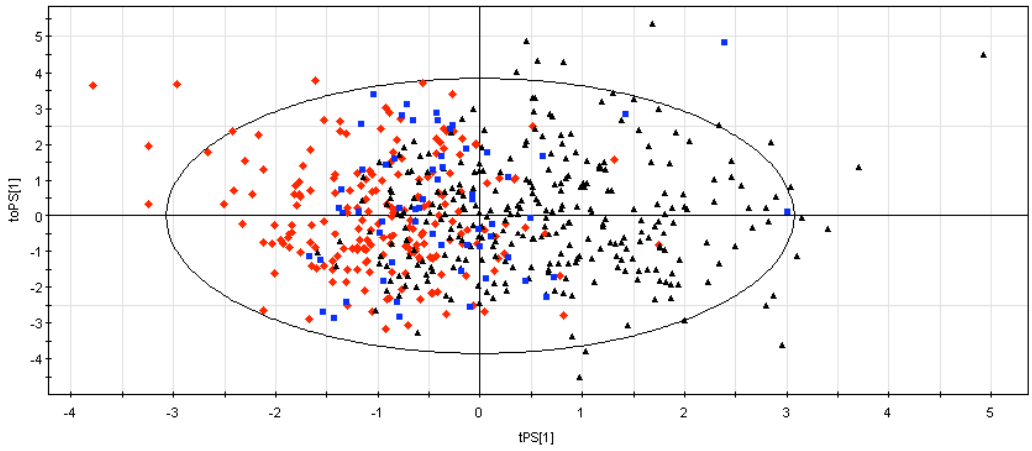


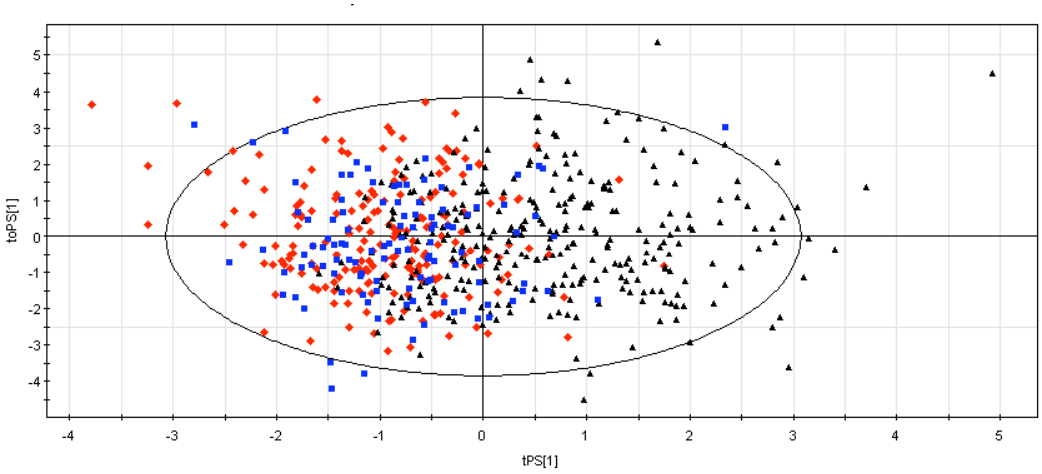
Figure 5.22: Testing set normal samples (blue squares) superimposed on normal (black triangles) vs. adenoma (red diamonds) OPLS scatter plot



5.5.7 Validation With Hyperplastic Polyps

Hyperplastic polyps are benign growth of the colon that have no malignant potential. When 110 urine samples from patients with hyperplastic polyps were introduced blindly to the Normal vs. Adenoma model, exploratory analysis shows that the hyperplastic polyps were more alike with the adenomatous polyps than the normals (figure 5.23). This is further confirmed when we attempted to establish an OPLS model between hyperplastic polyps and adenomatous polyps. A meaningful model to separate the two groups could not be constructed; R^2Y = 0.126, Q^2 = -0.0771. Since hyperplastic polyps are not pre-cancerous, we expected them to behave more like normals than adenomas, however it seems the model was more predictive of a growth and not powerful enough to distinguish the type of tumor.

Figure 5.23: Hyperplastic samples (blue squares) superimposed on normal (black triangles) vs. adenoma (red diamonds) OPLS scatter plot



5.5.8 Adenoma Model vs. CRC Model

The normal vs. CRC model (chapter 4) is very different than the normal vs. adenoma model. The metabolites that drive the separation between the normal and CRC groups are completely different than those that drive the separation between normal and adenoma; the top ten in each model are summarized in table 5.15. Creatinine and methanol are the only two metabolites in common.

Table 5.15: Comparison of top ten contributing metabolites from the normal vs. CRC model and the normal vs. adenoma model

N vs. CRC Model	N vs. Adenoma Model
Hypoxanthine	Butyrate
Dimethylamine	Serine
Creatinine	Methanol
Urea	β-alanine
3-Indoxylsulfate	Asparagine
Adipate	3- Hydroxyphenylacetate
Methanol	Creatinine
Guanidoacetate	Histidine
3-Hydroxybutyrate	Trigonelline
Acetone	Cis-Aconitate

While histologically it is believed that normal colonic mucosa transforms to carcinoma through adenomatous polyps[2], metabolomically this spectrum of events is not so clear. We attempted to investigate this matter further by superimposing the CRC training set onto the normal vs. adenoma model (figure 5.24) as well as superimposing the adenoma training set onto the normal vs. CRC model (figure 5.25). When the CRC samples are tested in the adenoma model, exploratory data analysis revealed that the CRC samples were distributed on both sides of the plot, although there's slightly more on the adenoma side and particularly a few outliers on the adenoma side. When the adenoma samples are tested in the CRC model, they are evenly distributed on both sides of the plot, suggesting that there is no resemblance of adenomatous polyps to CRC metabolomically.

Figure 5.24: CRC training samples (blue squares) superimposed on normal (black triangles) vs. adenoma (red diamonds) OPLS scatter plot

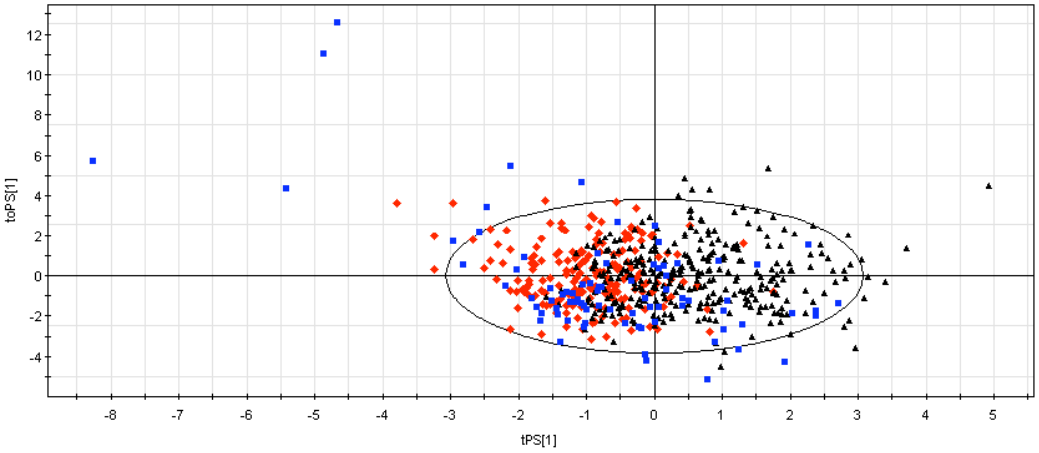
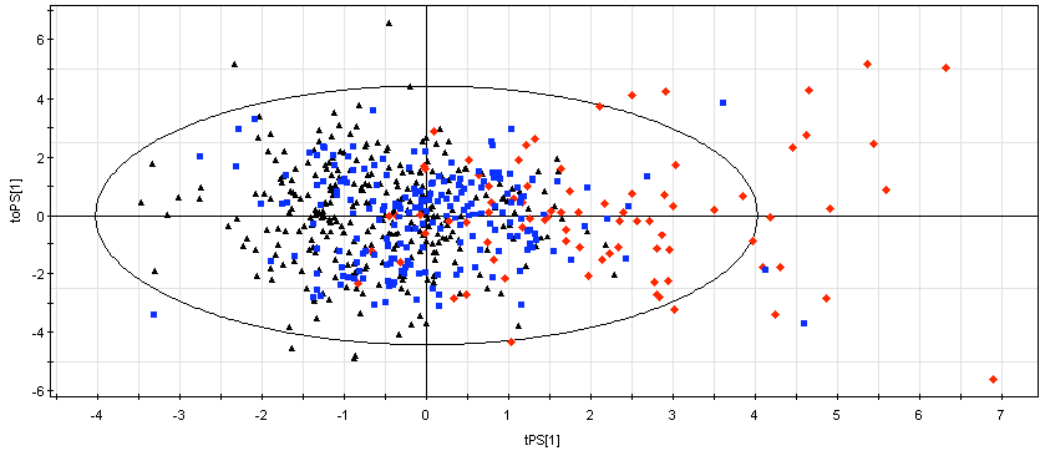


Figure 5.25: Adenoma training samples (blue squares) in normal (black triangles) vs. CRC (red diamonds) OPLS scatter plot



5.5.9 Comparison of Urine Metabolomic Test to Fecal Tests

The diagnostic accuracies of urine metabolomic test for adenoma was compared to the three fecal tests as part of the SCOPE pilot study and the sensitivity and specificity for each test are summarized in table 5.16. The diagnostics of the fecal tests were calculated from the raw SCOPE trial data. Urine metabolomics far outperformed the currently used Hemoccult II FOBT in sensitivity (89.5% vs.

3.0%). The newer fecal immune tests had slightly higher sensitivities (13.8% for Hemoccult ICT and 18.8% for MagStream HemSp/HT) but still far inferior to the urine metabolomics test. The specificity of the urine metabolomics test was not as high as that of the fecal tests (71.8% vs. 99.0%), but as mentioned previously, it is more important for a screening test to have a higher sensitivity than specificity. Even when the adenomas were divided into villous and tubular subgroups, then sensitivity of the MagStream HemSp/HT test merely reached 50% for villous adenomas.

Table 5.16: Diagnostic accuracies of the urine metabolomics test for adenomas compared to fecal tests

Test	Sensitivity (%)	Specificity (%)
Urine Metabolomics	89.5	71.8
Hemoccult II	3.0	99.0
Hemoccult ICT	13.8	94.2
MagStream HemSp/HT	18.8	92.8

5.5.10 Commercialization

To commercialize urine metabolomics as a screening test for adenoma, accuracy of the test is very important, but for it to become a population-based test, the cost needs to be reasonable. The normal vs. adenoma model is currently built using 69 metabolites, but if we can produce a model of acceptable diagnostic accuracies using fewer metabolites, then the cost of the test would be much lower. From the variable importance plot of the current model, we know, in order of importance, the metabolites that contribute most in the separation of the two groups. We can capitalize on this by taking the top metabolites and see what kind of models we

can build with them. The results are summarized in table 5.17. Using the concentrations of the top 10 metabolites, namely butyrate, serine, methanol, β-alanine, creatinine, asparagines, 3-hydroxyphenylacetate, histidine, trigonelline, and cis-aconitate, a reasonable OPLS model could be built (R^2Y of 0.301, Q^2 of 0.284) with sensitivity and specificity of 87.5% and 60.9% respectively, and an AUC of 0.8474.

Table 5.17: Model characteristics and diagnostic accuracies of OPLS models built with top contributing metabolites

Model	R^2Y	Q^2	Sens	Spec	AUC
Main Adenoma Model	0.396	0.250	89.5%	71.8%	0.8913
Top 5 metabolites	0.256	0.250	87.0%	61.2%	0.8314
Top 10 metabolites	0.301	0.284	87.5%	60.9%	0.8474
Top 15 metabolites	0.333	0.304	87.5%	62.9%	0.8639
Top 20 metabolites	0.336	0.298	88.5%	62.2%	0.8615
Top 25 metabolites	0.361	0.313	87.5%	69.7%	0.8771
Top 30 metabolites	0.362	0.303	87.5%	68.7%	0.8768

To ensure that the top metabolites in the variable importance plot were indeed more important in establishing the metabolomic fingerprint of colorectal adenoma, we validated the process above by attempting to build models using the bottom metabolites in the VIP list. It took 65 metabolites before we could even

build an OPLS model, and it was a poorly predictive one (R^2Y of 0.195 and Q^2 of only 0.030). This validates the uniqueness of top metabolites as a diagnostic tool.

5.6 Discussion

5.6.1 Summary

The role of urine metabolomics in distinguishing subjects with a normal colonoscopy from subjects found to have colorectal adenoma(s) was examined in this chapter. The testing set of 294 normal subjects and 200 adenoma patients was used to build the OPLS model ($R^2Y = 0.396$, $Q^2 = 0.250$), which was internally validated using permutation testing. A spectrum of diagnostic accuracies, namely sensitivity and specificity were calculated and the area under the curve was found to be 0.8913. A representative pair of sensitivity and specificity was 89.5% and 71.8% respectively. When the study subjects were stratified by gender, family history of any cancer, and family history of CRC, the diagnostic accuracies improved. The main model was then externally validated with a blinded testing set of 103 urine samples and sensitivity and specificity of 72.1% and 40.0% were achieved.

To our surprise, the subjects with hyperplastic polyps resembled the subjects with adenomatous polyps rather than the normal subjects. And interestingly, the metabolites that drive the separation between normal and adenoma are completely different than those that drive the separation between normal and CRC.

Models were then built using only the top metabolites and diagnostic accuracies were calculated. Using only the top 10 metabolites, sensitivity of 87.5% and specificity of 60.9% could be achieved, suggesting commercialization potential for this test. The diagnostic accuracies of the urine metabolomics test for adenoma are far superior than those of the fecal occult blood test and newer fecal immunochemical tests.

5.6.2 Patient and Disease Characteristics

The normal and the adenoma group are quite different from each other in terms of age, gender, family history of CRC and family history of any cancer, and gastrointestinal symptoms, but each of these factors was analyzed statistically to see if they falsely contribute to the separation of the two groups and they didn't; in fact, the diagnostic accuracies improved when the model was stratified by gender, family history of any cancer and family history of CRC.

It was anticipated that villous adenomas or screen relevant neoplasms would be more different than tubular ones, but this was not the case.

5.6.3 Metabolites and Metabolic Pathways

The top 10 metabolites that separated the normal group from the adenoma group were butyrate, serine, methanol, β-alanine, asparagine, 3-hydroxyphenylacetate, creatinine, histidine, trigonelline, and cis-aconitate.

Butyrate was only found to be present in 3 out of 200 (1.5%) adenoma patients while it was found in 109 out of 294 (37%) of normal patients. Butyrate is a short-chain fatty acid generated by microbial fermentation of dietary fibre.[9] Short-chain fatty acids in general are one of the proposed health-promoting effects of prebiotics. Butyrate has been shown to increase apoptosis in both colon adenoma and cancer cell lines in a p53-independent way, thus contributing to the protection against CRC.[10] It also influences a wide array of cellular functions affecting colonic health, as such that besides being anti-carcinogenic[11], it may have anti-inflammatory potential[12], affect the intestinal barrier[13] and play a role in satiety[14] and oxidative stress.[15] Epidemiological studies have been inconclusive and direct evidence for a protective effect of butyrate on colorectal carcinogenesis in humans is lacking.[16]

Serine was another metabolite that was present more in the normal group than the adenoma group – 35 out of 200 (17.5%) adenoma vs. 157 out of 294 (53%) normals. Serine is an amino acid derived from glycine that plays a central role in cellular proliferation and altered levels of serine and glycine have been noted in patients with psychiatric disorders[17] and neurological abnormalities[18]. Serine is an active component of serine protease, which is a group of enzymes that cleaves peptides. Certain serine proteases have been shown to act as tumor suppressors.[19] Furthermore, certain serine protease *inhibitors* have been reported to promote angiogenesis, induce tumor cell migration, and enhance the invasive potential of pancreatic, breast and lung cancer cells[20-22 23-25]. It unknown whether

higher urine levels of serine reflect a higher systemic level and whether higher levels of serine in the normal group allows for more serine proteases to form and thus offers a protective effect for colonic adenomas. As serine is also involved in microbial metabolism[26], this difference in the presence of serine between the normal group and the adenoma group could also represent a difference in the microbiota of the two groups.

Since methanol is mainly a product of microbial metabolism, the differences in its presence in the two groups may reflect the differences in the microbiota of the two groups.[26]

β-alanine was only present in 8 out of 200 or 4% of adenoma patients and 61 out of 294 or 20.7% of normal subjects. As mentioned in chapter 4, it was only present in 2 out of 82 or 2.4% of CRC patients as well. This metabolite was discussed in detail in Chapter 4. Histidine is a metabolite that is present more in the adenoma subjects compared to the normal group. In the β-alanine metabolism pathway [KEGG] [26], carnosine either metabolizes to β-alanine or histidine, hence in adenoma patients, carnosine may be preferentially metabolizing to histidine rather than β-alanine.

Asparagine was found in more adenoma subjects than normal subjects. It is a non-essential amino acid involved in alanine, aspartate, and glutamate metabolism, cyanoamino acid metabolism, and nitrogen metabolism. It is present

in abnormal concentrations in neurological[18] and psychological disorders[27] as well as leukemia[28]. However, its role in colorectal and colonic adenoma is unclear.

3- Hydroxyphenylacetate, present in more adenoma patients than normal subjects, is a product of phenylalanine metabolism and a substrate of tyrosine metabolism. It is also a metabolite of microbial metabolism.

Creatinine and cis-aconitate are both increased in adenoma patients. Creatinine is involved in arginine and proline metabolism and is increased in the urine of patients with cancer and cachexia[29]. Cis-aconitate is a TCA cycle intermediate but is also a metabolite of microbial metabolism.

Trigonelline is an alkaloid originating from dietary sources, particularly coffee[30], therefore the difference in this metabolite between the two groups may simply be a reflection of dietary differences.

Overall, 5 of the top 10 metabolites could be products of microbial metabolism (butyrate, serine, methanol, 3-hydroxyphenylacetate, and cis-aconitate), emphasizing the importance of microbiota in the development of adenoma and CRC. The human colonic microbiota consists of approximately 10^{14} bacterial cells and more than 1000 different bacterial species, and it plays a pivotal role for the maintenance of human health[31] and several studies have indicated the importance of the intestinal microbiota in the development of various conditions

including inflammatory bowel disease[32], cancer[33], and even obesity[34].

5.6.4 Colonic Adenoma vs. CRC

Metabolomically speaking, colonic adenomas are different than CRCs. Since adenoma is an intermediate step in the pathway of normal colonic epithelium's progression to CRC, it was anticipated that the metabolites that drive the separation between normal and adenoma would be quite similar to those that drive the separation between normal and CRC, but the concentrations would not be as high. However, completely different metabolomic fingerprints were seen in the two models. This may be because the adenoma metabolomic fingerprint simply reflects intermediate genetic changes in the multi-step process of the colorectal cancer pathway.

5.6.5 Limitations

As mentioned in the previous chapter, there may have been potential misinterpretation on the patients' part in filling out the questionnaire and moreover, the analysis was limited by the number of metabolites contained in the Chenomx compound library.

There were also some differences in baseline characteristics between the training set and the testing populations. This is because the testing set samples were not methodologically matched to the training set samples, but were simply subsequently collected samples. To improve the robustness of this model, the

validation set should be completely comparable to the training set. This can be done by increasing the number of validation samples or by selecting only those subjects that have matching baseline characteristics to the training set to validate the model.

5.6.6 Bias

The concepts of disease progression bias, misclassification, spectrum bias, partial verification bias, incorporation bias, and review bias addressed in chapter 4 also apply to this study.

5.6.7 Strengths of Study

This is the largest study to demonstrate that urine metabolomics can separate subjects with normal colons from patients with colonic adenoma. The controls in this study are not merely healthy volunteers but rather colonoscopy-negative controls. This eliminates metabolomic fingerprints associated with other colonic disease and disorders. The robustness of the model was internally tested with permutation testing and also externally validated with a blinded testing set. The sample size was large enough to avoid random correlations when using multivariate analysis. Potential biases and confounders have been addressed.

5.7 Conclusions

CRC is a preventable disease if identified at the adenomatous polyp stage. Since the development of CRC from normal colonic epithelium takes years and multiple

genetic mutations need to occur, regular screening can detect CRC in its early stages or in the pre-cancerous adenomatous polyp stage. In this chapter, urine metabolomics has been demonstrated to distinguish normal healthy subjects from patients with adenomatous polyps with far-superior accuracy than that of current guaiac-based and immunochemical fecal tests. Spot urine metabolomics test has the potential to become a new and highly sensitive screening tool for CRC and colonic adenomas.

5.8 References

1. Gordon PH, Nivatvongs S. Principles and Practice of Surgery for the Colon, Rectum, and Anus. Informa Healthcare, 2006.
2. Fearon ER, Vogelstein B. A GENETIC MODEL FOR COLORECTAL TUMORIGENESIS. Cell 1990;61:759-767.
3. Allison JE, Tekawa IS, Ransom LJ, Adrain AL. A comparison of fecal occult-blood tests for colorectal-cancer screening. New England Journal of Medicine 1996;334:155-159.
4. Imperiale TF, Ransohoff DF, Itzkowitz SH, Turnbull BA, Ross ME, Colorectal Canc Study G. Fecal DNA versus fecal occult blood for colorectal-cancer screening in an average-risk population. New England Journal of Medicine 2004;351:2704-2714.
5. Whitlock EP, Lin JS, Liles E, Beil TL, Fu RW. Screening for Colorectal Cancer: A Targeted, Updated Systematic Review for the US Preventive Services Task Force. Annals of Internal Medicine 2008;149:638-+.
6. Chang D, Weljie A, Newton J. Leveraging latent information in NMR spectra for robust predictive models. Pacific Symposium on Biocomputing 2007 2007:115-126.
7. Wong CK, Fedorak RN, Prosser C, Stewart ME, Van Zanten S, Sadowski DC. A comparison of guaiac and immunochemical fecal occult blood tests to colonoscopy for the detection of colonic polyps and cancer: *submitted* 2010.
8. Wilschut JA, Habbema JDF, Ramsey SD, Boer R, Looman CWN, van Ballegooijen M. Increased risk of adenomas in individuals with a family history of colorectal cancer: results of a meta-analysis. Cancer Causes & Control 2010;21:2287-2293.
9. McMillan L, Butcher SK, Pongracz J, Lord JM. Opposing effects of butyrate and bile acids on apoptosis of human colon adenoma cells: differential activation of PKC and MAP kinases. British Journal of Cancer 2003;88:748-753.
10. Hague A, Manning AM, Hanlon KA, Huschtscha LI, Hart D, Paraskeva C. SODIUM-BUTYRATE INDUCES APOPTOSIS IN HUMAN COLONIC TUMOR-CELL LINES IN A P53-INDEPENDENT PATHWAY - IMPLICATIONS FOR THE POSSIBLE ROLE OF DIETARY FIBER IN THE PREVENTION OF LARGE-BOWEL CANCER. International Journal of Cancer 1993;55:498-505.
11. Scheppach W, Bartram HP, Richter F. ROLE OF SHORT-CHAIN FATTY-ACIDS IN THE PREVENTION OF COLORECTAL-CANCER. European Journal of Cancer 1995;31A:1077-1080.
12. Segain JP, de la Bletiere DR, Bourreille A, Leray V, Gervois N, Rosales C, Ferrier L, Bonnet C, Blottiere HM, Galmiche JP. Butyrate inhibits inflammatory responses through NF kappa B inhibition: implications for Crohn's disease. Gut 2000;47:397-403.
13. Finnie IA, Dwarakanath AD, Taylor BA, Rhodes JM. COLONIC MUCIN SYNTHESIS IS INCREASED BY SODIUM-BUTYRATE. Gut 1995;36:93-99.
14. Cherbut C. Motor effects of short-chain fatty acids and lactate in the gastrointestinal tract. Proceedings of the Nutrition Society 2003;62:95-99.
15. Rosignoli P, Fabiani R, De Bartolomeo A, Spinozzi F, Agea E, Pelli MA, Morozzi G. Protective activity of butyrate on hydrogen peroxide-induced DNA damage in isolated human colonocytes and HT29 tumour cells. Carcinogenesis 2001;22:1675-1680.
16. Hamer HM, Jonkers D, Venema K, Vanhoutvin S, Troost FJ, Brummer RJ. Review article: the role of butyrate on colonic function. Alimentary Pharmacology & Therapeutics 2008;27:104-119.
17. Hashimoto K, Fukushima T, Shimizu E, Komatsu N, Watanabe H, Shinoda N, Nakazato M, Kumakiri C, Okada S, Hasegawa H, Imai K, Iyo M. Decreased serum levels of D-serine in patients with schizophrenia - Evidence in support of the N-methyl-D-aspartate receptor hypofunction hypothesis of schizophrenia. Archives of General Psychiatry 2003;60:572-576.
18. Fonteh AN, Harrington RJ, Tsai A, Liao P, Harrington MG. Free amino acid and dipeptide changes in the body fluids from Alzheimer's disease subjects. Amino Acids 2007;32:213-224.

19. Selzer-Plon J, Bornholdt J, Friis S, Bisgaard HC, Lothe IMB, Tveit KM, Kure EH, Vogel U, Vogel LK. Expression of prostasin and its inhibitors during colorectal cancer carcinogenesis. Bmc Cancer 2009;9.
20. Bajou K, Noel A, Gerard RD, Masson V, Brunner N, Holst-Hansen C, Skobe M, Fusenig NE, Carmeliet P, Collen D, Foidart JM. Absence of host plasminogen activator inhibitor 1 prevents cancer invasion and vascularization. Nature Medicine 1998;4:923-928.
21. Bajou K, Peng H, Laug WE, Maillard C, Noel A, Foidart JM, Martial JA, DeClerck YA. Plasminogen activator inhibitor-1 protects endothelial cells from FasL-mediated apoptosis. Cancer Cell 2008;14:324-334.
22. Liu GH, Shuman MA, Cohen RL. COEXPRESSION OF UROKINASE, UROKINASE RECEPTOR AND PAI-1 IS NECESSARY FOR OPTIMUM INVASIVENESS OF CULTURED LUNG-CANCER CELLS. International Journal of Cancer 1995;60:501-506.
23. Buchholz M, Biebl A, Neessbe A, Wagner M, Iwamura T, Leder G, Adler G, Gress TM. SERPINE2 (protease nexin I) promotes extracellular matrix production and local invasion of pancreatic tumors in vivo. Cancer Research 2003;63:4945-4951.
24. Candia BJ, Hines WC, Heaphy CM, Griffith JK, Orlando RA. Protease nexin-1 expression is altered in human breast cancer. Cancer Cell Int 2006;6:16.
25. Yang SF, Dong QG, Yao M, Shi MP, Ye JD, Zhao LX, Su JZ, Gu WY, Xie WH, Wang K, Du YZ, Li Y, Huang Y. Establishment of an experimental human lung adenocarcinoma cell line SPC-A-1BM with high bone metastases potency by Tc-99m-MDP bone scintigraphy. Nuclear Medicine and Biology 2009;36:313-321.
26. Kyoto Encyclopedia of Genes and Genomes (KEGG): http://www.genome.jp/kegg/, Accessed Nov 2010 to Jan 2011.
27. Do KQ, Lauer CJ, Schreiber W, Zollinger M, Gutteckamsler U, Cuenod M, Holsboer F. GAMMA-GLUTAMYLGLUTAMINE AND TAURINE CONCENTRATIONS ARE DECREASED IN THE CEREBROSPINAL-FLUID OF DRUG-NAIVE PATIENTS WITH SCHIZOPHRENIC DISORDERS. Journal of Neurochemistry 1995;65:2652-2662.
28. Peng CT, Wu KH, Lan SJ, Tsai JJP, Tsai FJ, Tsai CH. Amino acid concentrations in cerebrospinal fluid in children with acute lymphoblastic leukemia undergoing chemotherapy. European Journal of Cancer 2005;41:1158-1163.
29. Eisner R, Stretch C, Eastman T, Xia JG, Hau D, Damaraju S, Greiner R, Wishart DS, Baracos VE. Learning to predict cancer-associated skeletal muscle wasting from H-1-NMR profiles of urinary metabolites. Metabolomics 2011;7:25-34.
30. Casal S, Oliveira M, Alves MR, Ferreira MA. Discriminate analysis of roasted coffee varieties for trigonelline, nicotinic acid, and caffeine content. Journal of Agricultural and Food Chemistry 2000;48:3420-3424.
31. O'Keefe SJD. Nutrition and colonic health: the critical role of the microbiota. Current Opinion in Gastroenterology 2008;24:51-58.
32. Takaishi H, Matsuki T, Nakazawa A, Takada T, Kado S, Asahara T, Kamada N, Sakuraba A, Yajima T, Higuchi H, Nagamu I, Ogata H, Iwao Y, Nomoto K, Tanaka R, Hibi T. Imbalance in intestinal microflora constitution could be involved in the pathogenesis of inflammatory bowel disease. International Journal of Medical Microbiology 2008;298:463-472.
33. Lupton JR. Microbial degradation products influence colon cancer risk: the butyrate controversy. Journal of Nutrition 2004;134:479-482.
34. Cani PD, Delzenne NM, Amar J, Burcelin R. Role of gut microflora in the development of obesity and insulin resistance following high-fat diet feeding. Pathologie Biologie 2008;56:305-309.

5.9 Appendix

Normal vs. Adenoma Group

```
. cs sex na if training ==1

                 | NA                    |
                 |   Exposed   Unexposed |      Total
-----------------+-----------------------+------------
           Cases |       118         117 |        235
        Noncases |        82         177 |        259
-----------------+-----------------------+------------
           Total |       200         294 |        494
                 |                       |
            Risk |       .59    .3979592 |   .4757085
                 |                       |
                 |      Point estimate   |    [95% Conf. Interval]
                 |-----------------------+------------------------
 Risk difference |         .1920408      |     .103855    .2802267
      Risk ratio |         1.482564      |    1.235901    1.778457
 Attr. frac. ex. |         .3254929      |    .1908734    .4377149
 Attr. frac. pop |          .163439      |
                 +-------------------------------------------------
                             chi2(1) =    17.60  Pr>chi2 = 0.0000


. cs famhx_cca na if training ==1

                 | NA                    |
                 |   Exposed   Unexposed |      Total
-----------------+-----------------------+------------
           Cases |       112         191 |        303
        Noncases |        73          85 |        158
-----------------+-----------------------+------------
           Total |       185         276 |        461
                 |                       |
            Risk |  .6054054     .692029 |   .6572668
                 |                       |
                 |      Point estimate   |    [95% Conf. Interval]
                 |-----------------------+------------------------
 Risk difference |        -.0866236      |   -.1756562     .002409
      Risk ratio |         .8748267      |    .7601904     1.00675
 Prev. frac. ex. |         .1251733      |     -.00675    .2398096
 Prev. frac. pop |         .0502323      |
                 +-------------------------------------------------
                             chi2(1) =     3.69  Pr>chi2 = 0.0548


. cs fh_any_ca na if training ==1

                 | NA                    |
                 |   Exposed   Unexposed |      Total
-----------------+-----------------------+------------
           Cases |       141         230 |        371
        Noncases |        58          21 |         79
-----------------+-----------------------+------------
           Total |       199         251 |        450
                 |                       |
            Risk |  .7085427    .9163347 |   .8244444
                 |                       |
                 |      Point estimate   |    [95% Conf. Interval]
                 |-----------------------+------------------------
 Risk difference |        -.2077919      |   -.2796234   -.1359604
      Risk ratio |         .7732357      |    .7020123    .8516853
 Prev. frac. ex. |         .2267643      |    .1483147    .2979877
 Prev. frac. pop |         .1002802      |
                 +-------------------------------------------------
```

```
                           chi2(1) =    33.11  Pr>chi2 = 0.0000
```

```
. cs smoke na if training ==1
```

```
                 | NA                    |
                 |   Exposed   Unexposed |      Total
-----------------+-----------------------+------------
           Cases |        29          26 |         55
        Noncases |       164         253 |        417
-----------------+-----------------------+------------
           Total |       193         279 |        472
                 |                       |
            Risk |  .1502591      .09319 |   .1165254
                 |                       |
                 |      Point estimate   |    [95% Conf. Interval]
                 |-----------------------+------------------------
 Risk difference |         .0570691      |   -.0037987    .1179369
      Risk ratio |         1.612395      |    .9813688    2.649176
 Attr. frac. ex. |         .3798047      |   -.0189849    .6225242
 Attr. frac. pop |         .2002607      |
                 +-----------------------------------------------
                           chi2(1) =     3.61  Pr>chi2 = 0.0575
```

```
. cs dm na if training ==1
```

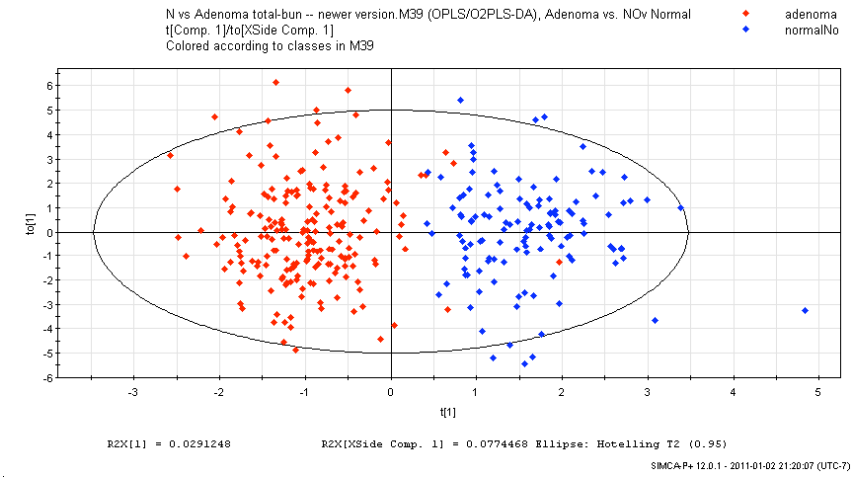
```
                 | NA                    |
                 |   Exposed   Unexposed |      Total
-----------------+-----------------------+------------
           Cases |         8          14 |         22
        Noncases |       192         280 |        472
-----------------+-----------------------+------------
           Total |       200         294 |        494
                 |                       |
            Risk |       .04    .047619  |   .0445344
                 |                       |
                 |      Point estimate   |    [95% Conf. Interval]
                 |-----------------------+------------------------
 Risk difference |         -.007619      |     -.04409    .0288519
      Risk ratio |              .84      |     .359071    1.965071
 Prev. frac. ex. |              .16      |   -.9650707     .640929
 Prev. frac. pop |         .0647773      |
                 +-----------------------------------------------
                           chi2(1) =     0.16  Pr>chi2 = 0.6870
```

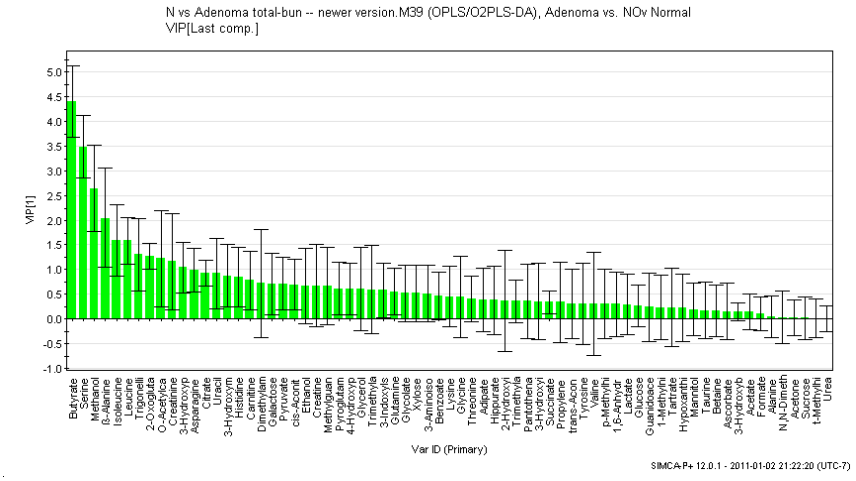
```
. cs sympt_gibleed na if training ==1
```

```
                 | NA                    |
                 |   Exposed   Unexposed |      Total
-----------------+-----------------------+------------
           Cases |         6           5 |         11
        Noncases |       194         288 |        482
-----------------+-----------------------+------------
           Total |       200         293 |        493
                 |                       |
            Risk |       .03   .0170648  |   .0223124
                 |                       |
                 |      Point estimate   |    [95% Conf. Interval]
                 |-----------------------+------------------------
 Risk difference |         .0129352      |   -.0149727     .040843
      Risk ratio |            1.758      |    .5439239    5.681978
 Attr. frac. ex. |         .4311718      |   -.8384924     .824005
 Attr. frac. pop |         .2351846      |
                 +-----------------------------------------------
                           chi2(1) =     0.91  Pr>chi2 = 0.3397
```

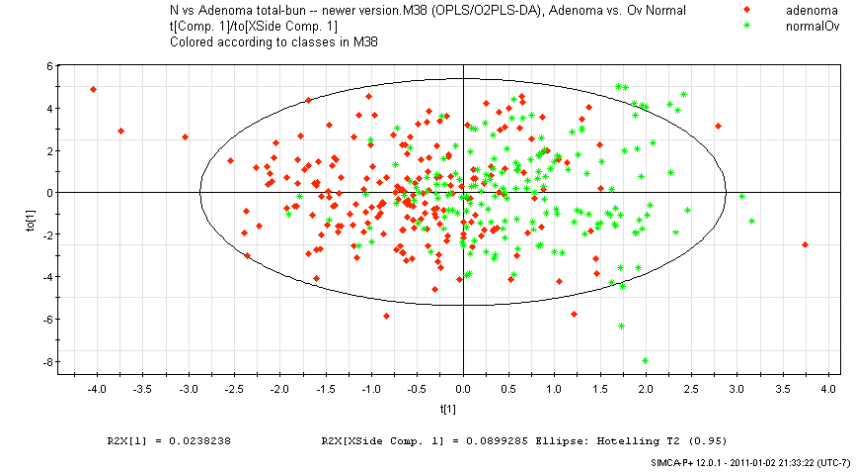
```
. cs sympt_bowelhabit na if training ==1

                 | NA                    |
                 |   Exposed   Unexposed |      Total
-----------------+-----------------------+-----------
           Cases |         1          13 |         14
        Noncases |       199         279 |        478
-----------------+-----------------------+-----------
           Total |       200         292 |        492
                 |                       |
            Risk |      .005    .0445205 |   .0284553
                 |                       |
                 |      Point estimate   |    [95% Conf. Interval]
                 |-----------------------+------------------------
 Risk difference |        -.0395205      |    -.065117   -.0139241
      Risk ratio |         .1123077      |    .0148094    .8516893
 Prev. frac. ex. |         .8876923      |    .1483107    .9851906
 Prev. frac. pop |         .3608505      |
                 +-----------------------------------------------
                              chi2(1) =     6.71  Pr>chi2 = 0.0096


. ttest age, by(na), if training==1

Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |     294     55.2619    .4731272    8.112442    54.33075    56.19306
       1 |     200       59.44    .5528655    7.818699    58.34977    60.53023
---------+--------------------------------------------------------------------
combined |     494     56.95344    .3710253    8.246449    56.22446    57.68243
---------+--------------------------------------------------------------------
    diff |           -4.178095    .7328069               -5.617912   -2.738278
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =  -5.7015
Ho: diff = 0                                     degrees of freedom =      492

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 1.0000
```

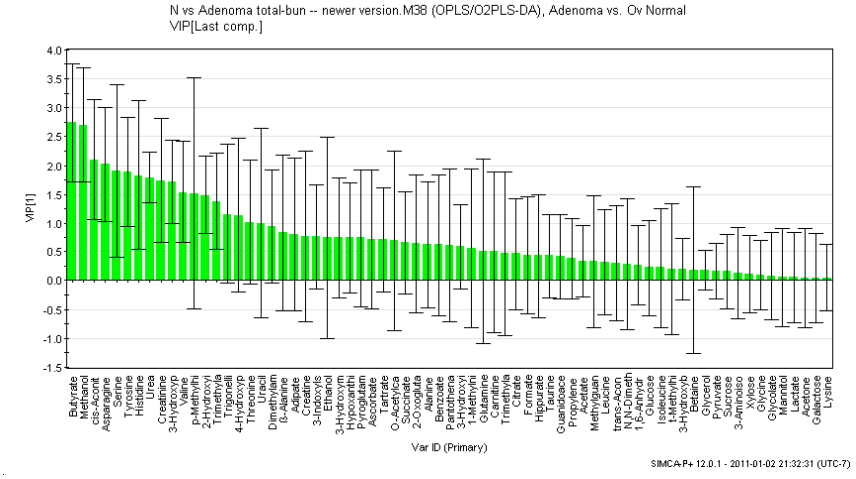
Adenoma vs. Non-overlapping Normal



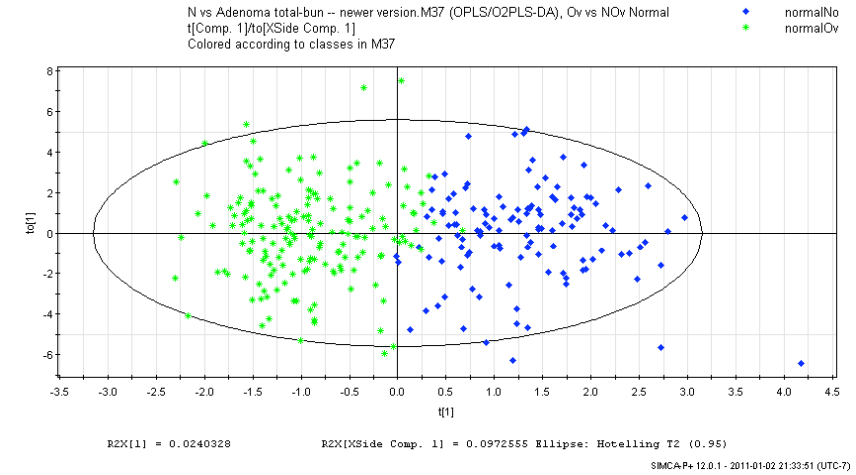


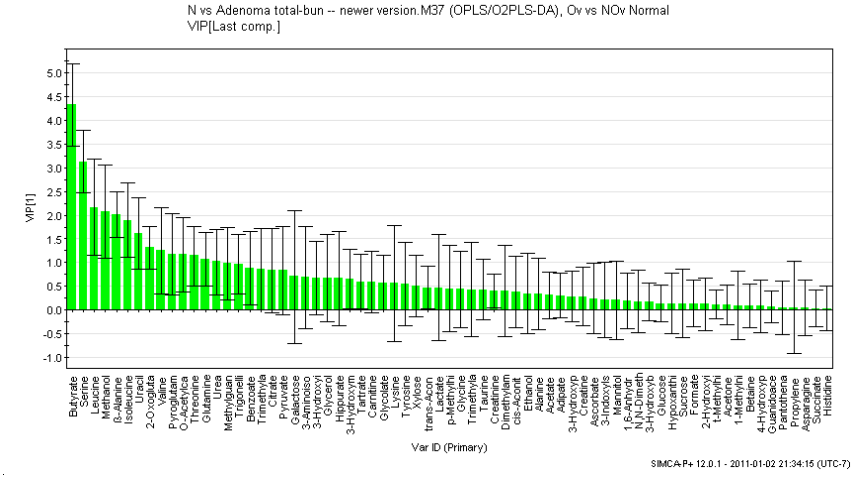
Adenoma vs. Overlapping Normal





Overlapping vs. Non-Overlapping Normal





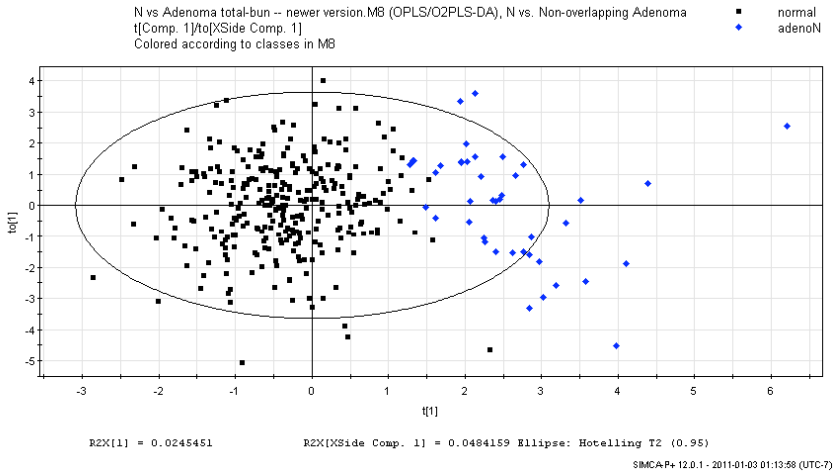
Logistic regression of overlapping vs. non-overlapping normals on predicting overlap with adenoma

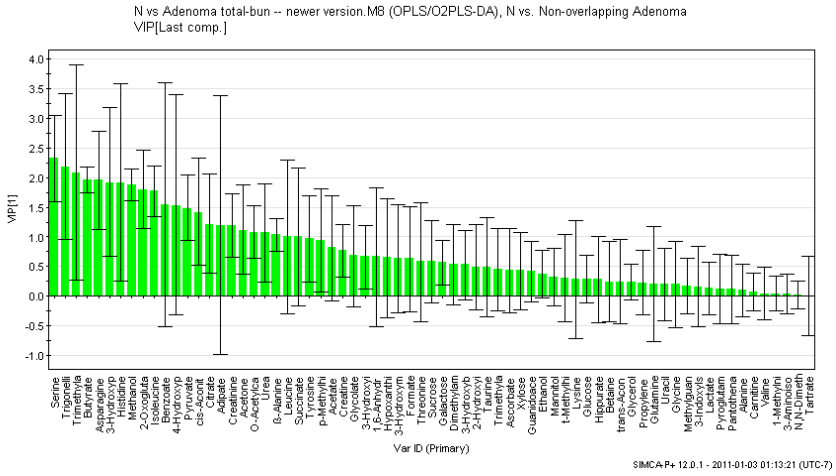
```
. logistic overlap_w_adenoma sex

Logistic regression                               Number of obs   =        294
                                                  LR chi2(1)      =       0.74
                                                  Prob > chi2     =     0.3898
Log likelihood = -199.14304                       Pseudo R2       =     0.0019

------------------------------------------------------------------------------
overlap_w_~a | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |      1.232   .2995005     0.86   0.391     .7650375    1.983986
------------------------------------------------------------------------------


. logistic overlap_w_adenoma age

Logistic regression                               Number of obs   =        294
                                                  LR chi2(1)      =       1.02
```

```
                                                Prob > chi2     =     0.3128
Log likelihood = -199.00348                     Pseudo R2       =     0.0026

------------------------------------------------------------------------------
overlap_w_~a | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   1.014918   .0149368     1.01   0.314     .9860611     1.04462
------------------------------------------------------------------------------


. logistic overlap_w_adenoma smoke

Logistic regression                             Number of obs   =        279
                                                LR chi2(1)      =       2.66
                                                Prob > chi2     =     0.1032
Log likelihood =  -188.0823                     Pseudo R2       =     0.0070

------------------------------------------------------------------------------
overlap_w_~a | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       smoke |   2.054563   .9451159     1.57   0.118     .8339883    5.061499
------------------------------------------------------------------------------


. logistic overlap_w_adenoma dm

Logistic regression                             Number of obs   =        294
                                                LR chi2(1)      =       0.01
                                                Prob > chi2     =     0.9158
Log likelihood = -199.50725                     Pseudo R2       =     0.0000

------------------------------------------------------------------------------
overlap_w_~a | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          dm |   .9430894   .5220193    -0.11   0.916     .3187098    2.790682
------------------------------------------------------------------------------


. logistic overlap_w_adenoma famhx_cca

Logistic regression                             Number of obs   =        276
                                                LR chi2(1)      =       0.19
                                                Prob > chi2     =     0.6622
Log likelihood = -188.30439                     Pseudo R2       =     0.0005

------------------------------------------------------------------------------
overlap_w_~a | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   famhx_cca |   1.121809   .2948914     0.44   0.662     .6701348    1.877912
------------------------------------------------------------------------------


. logistic overlap_w_adenoma fh_any_ca

Logistic regression                             Number of obs   =        251
                                                LR chi2(1)      =       0.46
                                                Prob > chi2     =     0.4958
Log likelihood = -168.93426                     Pseudo R2       =     0.0014

------------------------------------------------------------------------------
overlap_w_~a | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   fh_any_ca |   .7234043   .3486446    -0.67   0.502     .2812837    1.860448
------------------------------------------------------------------------------


. logistic overlap_w_adenoma sympt_gibleed

Logistic regression                             Number of obs   =        293
```

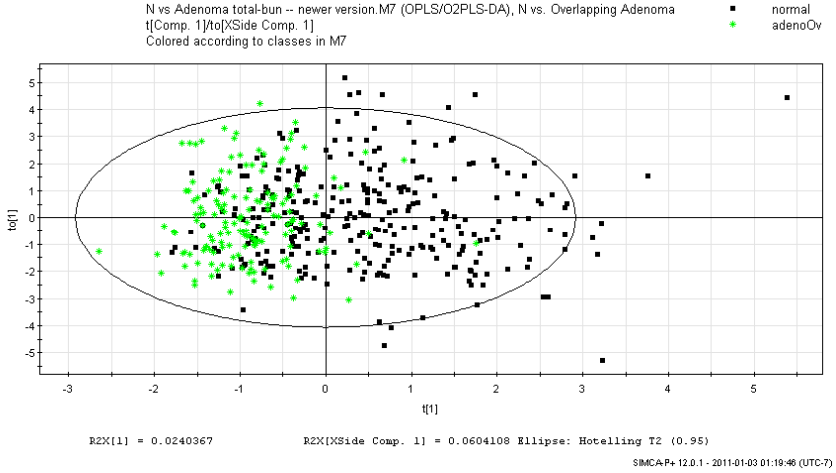
```
                                                LR chi2(1)      =       0.01
                                                Prob > chi2     =     0.9401
Log likelihood = -198.97272                     Pseudo R2       =     0.0000

------------------------------------------------------------------------------
overlap_w_~a | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
sympt_gibl~d |   1.071429   .9864239     0.07   0.940      .176317     6.51077
------------------------------------------------------------------------------



. logistic overlap_w_adenoma sympt_bowelhabit

Logistic regression                             Number of obs   =        292
                                                LR chi2(1)      =       0.80
                                                Prob > chi2     =     0.3707
Log likelihood = -198.03521                     Pseudo R2       =     0.0020

------------------------------------------------------------------------------
overlap_w_~a | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
sympt_bowe~t |   .6010453   .3422883    -0.89   0.371     .1968602    1.835086
------------------------------------------------------------------------------


. generate age_cat_n = .
(597 missing values generated)

. replace age_cat_n = 1 if age <=45 & training == 1 & na==0
(38 real changes made)

. replace age_cat_n = 2 if age >45 & age <=50 & training == 1 & na==0
(39 real changes made)

. replace age_cat_n = 3 if age >50 & age <=55 & training == 1 & na==0
(84 real changes made)

. replace age_cat_n = 4 if age >55 & age <=60 & training == 1 & na==0
(64 real changes made)

. replace age_cat_n = 5 if age >60 & age <=65 & training == 1 & na==0
(37 real changes made)

. replace age_cat_n = 6 if age >65 & age <=70 & training == 1 & na==0
(20 real changes made)

. replace age_cat_n = 7 if age >70 & age <=75 & training == 1 & na==0
(10 real changes made)

. replace age_cat_n = 7 if age >75 & training == 1 & na==0
(2 real changes made)


. xi:logistic overlap_w_adenoma i.age_cat_n
i.age_cat_n       _Iage_cat_n_1-7     (naturally coded; _Iage_cat_n_1 omitted)

Logistic regression                             Number of obs   =        294
                                                LR chi2(6)      =       4.85
                                                Prob > chi2     =     0.5632
Log likelihood =  -197.0877                     Pseudo R2       =     0.0122

------------------------------------------------------------------------------
overlap_w_~a | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
_Iage_cat_~2 |    .754902   .3519603    -0.60   0.546     .3027138     1.88256
_Iage_cat_~3 |    .740991   .2976893    -0.75   0.456     .3371686    1.628466
_Iage_cat_~4 |        .75   .3151672    -0.68   0.494     .3291295    1.709054
_Iage_cat_~5 |   .6862745   .3232919    -0.80   0.424     .2725911    1.727763
_Iage_cat_~6 |   1.083333   .6250331     0.14   0.890     .3496699     3.35634
_Iage_cat_~7 |   2.916667   2.462909     1.27   0.205      .557327    15.26383
```

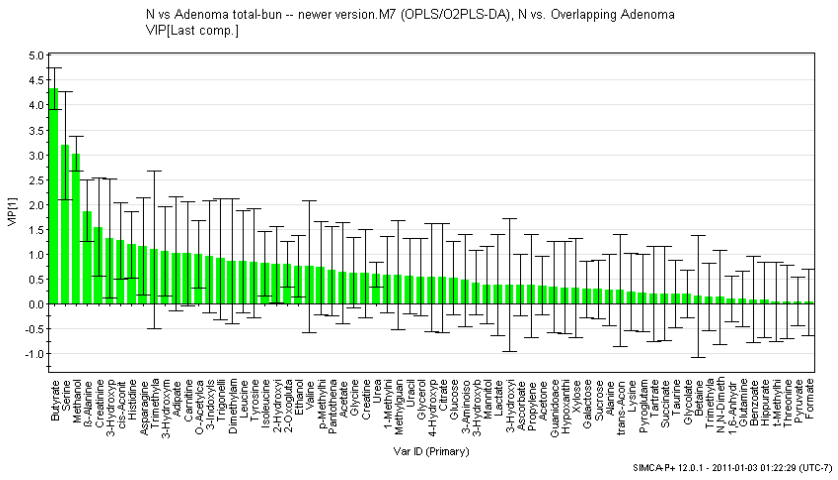
Normal vs. Non-overlapping Adenoma



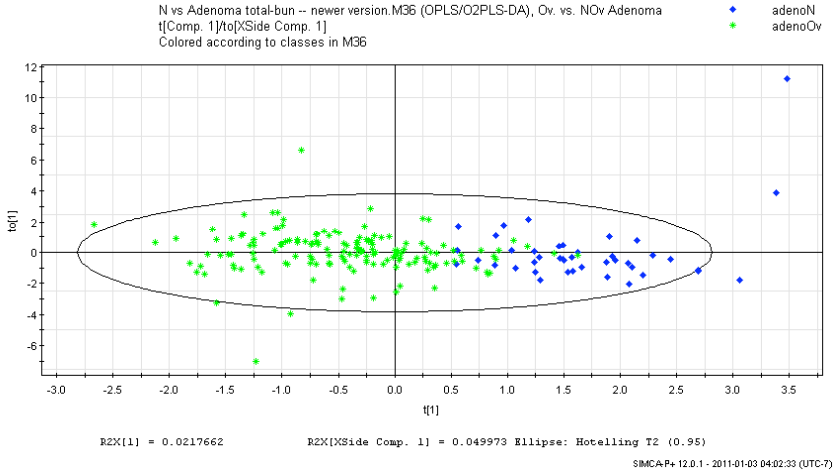


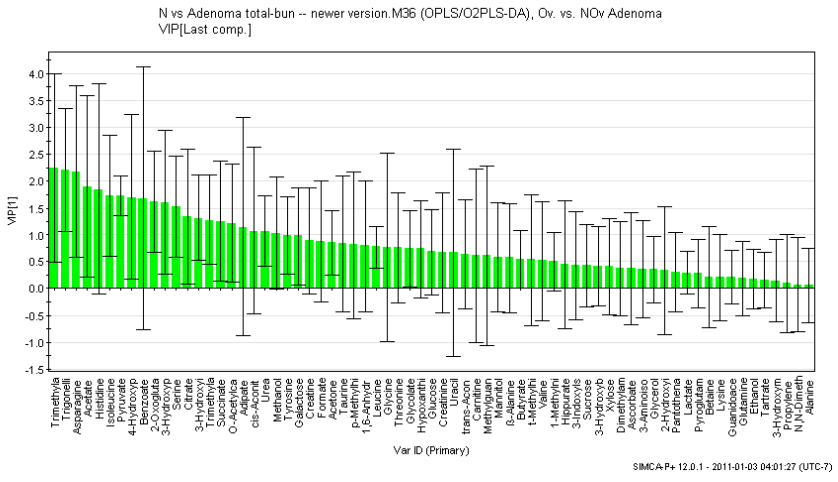
Normal vs. Overlapping Adenoma





Overlapping vs. Non-Overlapping Adenoma





Logistic regression of overlapping vs. non-overlapping adenomas on predicting overlap with normal

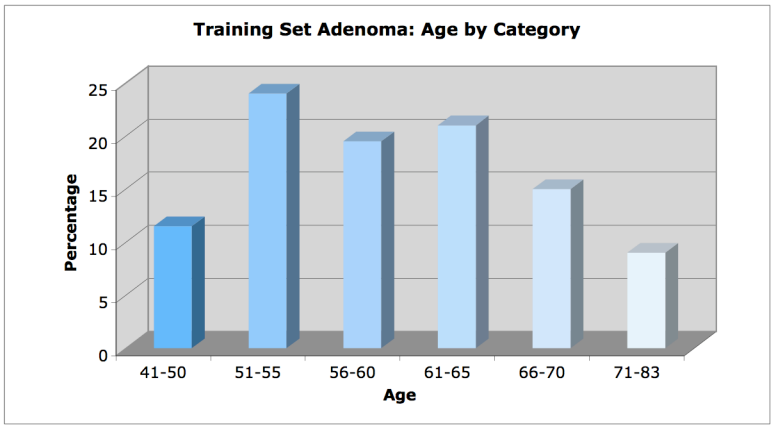
```
. logistic overlap_w_normal sex

Logistic regression                               Number of obs   =        200
                                                  LR chi2(1)      =       5.03
                                                  Prob > chi2     =     0.0249
Log likelihood = -100.27704                       Pseudo R2       =     0.0245

------------------------------------------------------------------------------
overlap_w_~l | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |   .4348024   .1676927    -2.16   0.031     .2041757    .9259332
------------------------------------------------------------------------------



. logistic overlap_w_normal age

Logistic regression                               Number of obs   =        200
                                                  LR chi2(1)      =       0.23
                                                  Prob > chi2     =     0.6323
Log likelihood = -102.67686                       Pseudo R2       =     0.0011

------------------------------------------------------------------------------
overlap_w_~l | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .9894264    .021969    -0.48   0.632     .9472914    1.033435
------------------------------------------------------------------------------



. logistic overlap_w_normal smoke

Logistic regression                               Number of obs   =        193
                                                  LR chi2(1)      =       7.26
                                                  Prob > chi2     =     0.0070
Log likelihood = -96.182492                       Pseudo R2       =     0.0364

------------------------------------------------------------------------------
overlap_w_~l | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       smoke |    .304321   .1305558    -2.77   0.006     .1312685    .7055103
------------------------------------------------------------------------------



. logistic overlap_w_normal dm

Logistic regression                               Number of obs   =        200
                                                  LR chi2(1)      =       1.19
                                                  Prob > chi2     =     0.2751
Log likelihood = -102.19571                       Pseudo R2       =     0.0058

------------------------------------------------------------------------------
overlap_w_~l | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          dm |   .4248366   .3194789    -1.14   0.255     .0973005    1.854936
------------------------------------------------------------------------------



. logistic overlap_w_normal famhx_cca

Logistic regression                               Number of obs   =        185
                                                  LR chi2(1)      =       0.09
                                                  Prob > chi2     =     0.7665
Log likelihood =  -97.81216                       Pseudo R2       =     0.0005
```

```
------------------------------------------------------------------------------
overlap_w_~l | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   famhx_cca |   1.113095   .4008858     0.30   0.766     .5495012    2.254738
------------------------------------------------------------------------------


. logistic overlap_w_normal fh_any_ca

Logistic regression                               Number of obs   =        199
                                                  LR chi2(1)      =       0.76
                                                  Prob > chi2     =     0.3845
Log likelihood = -102.17687                       Pseudo R2       =     0.0037

------------------------------------------------------------------------------
overlap_w_~l | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   fh_any_ca |   .7096354   .2849676    -0.85   0.393     .3230124    1.559019
------------------------------------------------------------------------------


. logistic overlap_w_normal sympt_gibleed

note: sympt_gibleed != 0 predicts success perfectly
      sympt_gibleed dropped and 6 obs not used


Logistic regression                               Number of obs   =        194
                                                  LR chi2(0)      =       0.00
                                                  Prob > chi2     =          .
Log likelihood = -101.35252                       Pseudo R2       =     0.0000

------------------------------------------------------------------------------
overlap_w_~l | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
------------------------------------------------------------------------------


. logistic overlap_w_normal sympt_bowelhabit

note: sympt_bowelhabit != 0 predicts success perfectly
      sympt_bowelhabit dropped and 1 obs not used


Logistic regression                               Number of obs   =        199
                                                  LR chi2(0)      =      -0.00
                                                  Prob > chi2     =          .
Log likelihood = -102.55494                       Pseudo R2       =    -0.0000

------------------------------------------------------------------------------
overlap_w_~l | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
------------------------------------------------------------------------------


. generate polyp_location_rc = .
(597 missing values generated)

. replace polyp_location_rc = 1 if polyp_location == "Rt COL"
(67 real changes made)

. replace polyp_location_rc = 1 if polyp_location == "Transv COL"
(32 real changes made)

. replace polyp_location_rc = 1 if polyp_location == "Lt COL"
(30 real changes made)

. replace polyp_location_rc = 1 if polyp_location == "Sigmoid"
```

```
(76 real changes made)

. replace polyp_location_rc = 2 if polyp_location == "Rectum"
(38 real changes made)


. tab polyp_location_rc

polyp_locat |
     ion_rc |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |        205       84.36       84.36
          2 |         38       15.64      100.00
------------+-----------------------------------
      Total |        243      100.00

. logistic overlap_w_normal polyp_location_rc

Logistic regression                               Number of obs   =        200
                                                  LR chi2(1)      =       0.65
                                                  Prob > chi2     =     0.4198
Log likelihood = -102.46587                       Pseudo R2       =     0.0032

------------------------------------------------------------------------------
overlap_w_~l | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
polyp_loca~c |      .6875   .3129592    -0.82   0.410     .2817049    1.677842
------------------------------------------------------------------------------


. generate polyp_location_lr = .
(597 missing values generated)

. replace polyp_location_lr = 1 if polyp_location == "Rt COL"
(67 real changes made)

. replace polyp_location_lr = 1 if polyp_location == "Lt COL"
(30 real changes made)

. replace polyp_location_lr = 2 if polyp_location == "Transv COL"
(32 real changes made)

. replace polyp_location_lr = 2 if polyp_location == "Lt COL"
(30 real changes made)

. replace polyp_location_lr = 2 if polyp_location == "Sigmoid"
(76 real changes made)

. replace polyp_location_lr = 2 if polyp_location == "Rectum"
(38 real changes made)

. tab polyp_location_lr

polyp_locat |
     ion_lr |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |         67       27.57       27.57
          2 |        176       72.43      100.00
------------+-----------------------------------
      Total |        243      100.00

. logistic overlap_w_normal polyp_location_lr

Logistic regression                               Number of obs   =        200
                                                  LR chi2(1)      =       0.48
                                                  Prob > chi2     =     0.4897
Log likelihood = -102.55275                       Pseudo R2       =     0.0023

------------------------------------------------------------------------------
overlap_w_~l | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
```

```
polyp_loca~r |   .7532468   .3140636    -0.68   0.497     .3326839    1.705465
------------------------------------------------------------------------------

. browse

. generate path_vt = .
(597 missing values generated)

. replace path_vt = 1 if path == "T"
(170 real changes made)

. replace path_vt = 1 if path == "A"
(41 real changes made)

. replace path_vt = 2 if path == "V"
(27 real changes made)

. replace path_vt = 1 if path == "T-"
(4 real changes made)

. replace path_vt = 2 if path == "V-"
(1 real change made)

. tab path_vt

    path_vt |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |        215       88.48       88.48
          2 |         28       11.52      100.00
------------+-----------------------------------
      Total |        243      100.00

. logistic overlap_w_normal path_vt

Logistic regression                               Number of obs   =        200
                                                  LR chi2(1)      =       0.06
                                                  Prob > chi2     =     0.8106
Log likelihood = -102.76262                       Pseudo R2       =     0.0003

------------------------------------------------------------------------------
overlap_w_~l | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     path_vt |   1.134307   .6022994     0.24   0.812      .400641    3.211482
------------------------------------------------------------------------------
```

Age of adenoma patients divided by category



```
. generate age_cat_a = .
(597 missing values generated)

. replace age_cat_a = 1 if age <=50 & training == 1 & na==1
(20 real changes made)

. replace age_cat_a = 2 if age >50 & age <=55 & training == 1 & na==1
(48 real changes made)

. replace age_cat_a = 3 if age >55 & age <=60 & training == 1 & na==1
(39 real changes made)

. replace age_cat_a = 4 if age >60 & age <=65 & training == 1 & na==1
(42 real changes made)

. replace age_cat_a = 5 if age >65 & age <=70 & training == 1 & na==1
(30 real changes made)

. replace age_cat_a = 6 if age >70 & age <=75 & training == 1 & na==1
(17 real changes made)

. replace age_cat_a = 6 if age >75 & training == 1 & na==1
(1 real change made)

. browse

. tab age_cat_a

  age_cat_a |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |         23       11.50       11.50
          2 |         48       24.00       35.50
          3 |         39       19.50       55.00
          4 |         42       21.00       76.00
          5 |         30       15.00       91.00
          6 |         18        9.00      100.00
------------+-----------------------------------
      Total |        200      100.00

. xi:logistic overlap_w_normal i.age_cat_a
i.age_cat_a       _Iage_cat_a_1-6     (naturally coded; _Iage_cat_a_1 omitted)

Logistic regression                               Number of obs   =        200
                                                  LR chi2(5)      =       5.30
                                                  Prob > chi2     =     0.3799
Log likelihood = -100.13927                       Pseudo R2       =     0.0258

------------------------------------------------------------------------------
overlap_w_~l | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
_Iage_cat_~2 |   1.052632   .7081884     0.08   0.939     .2815826     3.93502
_Iage_cat_~3 |   1.157895   .8184287     0.21   0.836     .2897469    4.627212
_Iage_cat_~4 |   .4210526   .2695311    -1.35   0.177     .1200747    1.476459
_Iage_cat_~5 |   .6917293   .4836999    -0.53   0.598     .1756805    2.723635
_Iage_cat_~6 |   1.052632   .8823487     0.06   0.951     .2035976     5.44227
------------------------------------------------------------------------------


. logistic overlap_w_normal srn if training == 1

Logistic regression                               Number of obs   =        200
                                                  LR chi2(1)      =       1.43
                                                  Prob > chi2     =     0.2317
Log likelihood = -102.07628                       Pseudo R2       =     0.0070

------------------------------------------------------------------------------
overlap_w_~l | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         srn |   .6333333   .2384834    -1.21   0.225     .3027681    1.324813
```

```
------------------------------------------------------------------------------


. logistic overlap_w_normal sex smoke

Logistic regression                               Number of obs   =        193
                                                  LR chi2(2)      =      13.79
                                                  Prob > chi2     =     0.0010
Log likelihood = -92.915551                       Pseudo R2       =     0.0691

------------------------------------------------------------------------------
overlap_w_~l | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |   .3695259   .1510271    -2.44   0.015     .1658649    .8232567
       smoke |    .287499   .1269998    -2.82   0.005     .1209557    .6833546
------------------------------------------------------------------------------

.
```

Training vs. Testing Set

Normals

```
. tab training na

           |          NA
  Training |         0          1 |     Total
-----------+----------------------+----------
         0 |        60         43 |       103
         1 |       294        200 |       494
-----------+----------------------+----------
     Total |       354        243 |       597



. ttest age, by(training), if na==0

Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |      60    55.73333    1.080612    8.370381    53.57103    57.89563
       1 |     294     55.2619    .4731272    8.112442    54.33075    56.19306
---------+--------------------------------------------------------------------
combined |     354    55.34181    .4329874     8.14661    54.49025    56.19337
---------+--------------------------------------------------------------------
    diff |            .4714286    1.155427               -1.800981    2.743838
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =   0.4080
Ho: diff = 0                                     degrees of freedom =      352

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.6582         Pr(|T| > |t|) = 0.6835          Pr(T > t) = 0.3418



. cs sex training if na==0

                 | Training               |
                 |   Exposed   Unexposed  |      Total
-----------------+------------------------+------------
           Cases |       117          31  |        148
        Noncases |       177          29  |        206
-----------------+------------------------+------------
           Total |       294          60  |        354
                 |                        |
            Risk |  .3979592    .5166667  |   .4180791
                 |                        |
```

```
                |      Point estimate    |    [95% Conf. Interval]
                |------------------------+------------------------
 Risk difference |        -.1187075       |   -.2569782    .0195632
      Risk ratio |         .7702436       |    .5808339     1.02142
 Prev. frac. ex. |         .2297564       |   -.0214196    .4191661
 Prev. frac. pop |         .1908147       |
                +-------------------------------------------------
                              chi2(1) =     2.89  Pr>chi2 = 0.0893
```

```
. cs famhx_cca training if na==0

                | Training               |
                |   Exposed   Unexposed  |      Total
----------------+------------------------+------------
          Cases |       191          31  |        222
       Noncases |        85          26  |        111
----------------+------------------------+------------
          Total |       276          57  |        333
                |                        |
           Risk |   .692029    .5438596  |   .6666667
                |                        |
                |      Point estimate    |    [95% Conf. Interval]
                |------------------------+------------------------
 Risk difference |         .1481693       |    .0078654    .2884733
      Risk ratio |          1.27244       |    .9905459    1.634558
 Attr. frac. ex. |         .2141086       |   -.0095443    .3882137
 Attr. frac. pop |         .1842105       |
                +-------------------------------------------------
                              chi2(1) =     4.67  Pr>chi2 = 0.0307
```

```
. cs fh_any_ca training if na==0

                | Training               |
                |   Exposed   Unexposed  |      Total
----------------+------------------------+------------
          Cases |       230          39  |        269
       Noncases |        21          21  |         42
----------------+------------------------+------------
          Total |       251          60  |        311
                |                        |
           Risk |  .9163347         .65  |   .8649518
                |                        |
                |      Point estimate    |    [95% Conf. Interval]
                |------------------------+------------------------
 Risk difference |         .2663347       |      .14088    .3917893
      Risk ratio |         1.409746       |    1.166503     1.70371
 Attr. frac. ex. |         .2906522       |    .1427367    .4130457
 Attr. frac. pop |          .248513       |
                +-------------------------------------------------
                              chi2(1) =    29.41  Pr>chi2 = 0.0000
```

```
. cs smoke training if na==0

                | Training               |
                |   Exposed   Unexposed  |      Total
----------------+------------------------+------------
          Cases |        26           6  |         32
       Noncases |       253          53  |        306
----------------+------------------------+------------
          Total |       279          59  |        338
                |                        |
           Risk |    .09319    .1016949  |   .0946746
                |                        |
                |      Point estimate    |    [95% Conf. Interval]
                |------------------------+------------------------
```

```
 Risk difference |         -.008505       |   -.0928345    .0758246
      Risk ratio |          .916368       |    .3947803    2.127083
 Prev. frac. ex. |          .083632       |   -1.127083    .6052197
 Prev. frac. pop |         .0690335       |
                 +-------------------------------------------------
                              chi2(1) =     0.04  Pr>chi2 = 0.8393
```

```
. cs dm training if na==0
```

```
                 | Training               |
                 |   Exposed   Unexposed  |      Total
-----------------+------------------------+------------
           Cases |        14           4  |         18
        Noncases |       280          56  |        336
-----------------+------------------------+------------
           Total |       294          60  |        354
                 |                        |
            Risk |   .047619    .0666667  |   .0508475
                 |                        |
                 |      Point estimate    |    [95% Conf. Interval]
                 |------------------------+------------------------
 Risk difference |        -.0190476       |    -.086696    .0486008
      Risk ratio |         .7142857       |    .2435531    2.094837
 Prev. frac. ex. |         .2857143       |   -1.094837    .7564469
 Prev. frac. pop |         .2372881       |
                 +-------------------------------------------------
                              chi2(1) =     0.37  Pr>chi2 = 0.5405
```

```
. cs sympt_gibleed training if na==0
```

```
                 | Training               |
                 |   Exposed   Unexposed  |      Total
-----------------+------------------------+------------
           Cases |         5           2  |          7
        Noncases |       288          58  |        346
-----------------+------------------------+------------
           Total |       293          60  |        353
                 |                        |
            Risk |  .0170648    .0333333  |     .01983
                 |                        |
                 |      Point estimate    |    [95% Conf. Interval]
                 |------------------------+------------------------
 Risk difference |        -.0162685       |   -.0640484    .0315114
      Risk ratio |         .5119454       |    .1017059    2.576922
 Prev. frac. ex. |         .4880546       |   -1.576922    .8982941
 Prev. frac. pop |         .4050992       |
                 +-------------------------------------------------
                              chi2(1) =     0.68  Pr>chi2 = 0.4102
```

```
. cs sympt_bowelhabit training if na==0
```

```
                 | Training               |
                 |   Exposed   Unexposed  |      Total
-----------------+------------------------+------------
           Cases |        13           0  |         13
        Noncases |       279          60  |        339
-----------------+------------------------+------------
           Total |       292          60  |        352
                 |                        |
            Risk |  .0445205           0  |   .0369318
                 |                        |
                 |      Point estimate    |    [95% Conf. Interval]
                 |------------------------+------------------------
 Risk difference |         .0445205       |    .0208642    .0681769
      Risk ratio |                .       |           .           .
```

```
 Attr. frac. ex. |               1       |           .           .
 Attr. frac. pop |               1       |
                 +-------------------------------------------------
                              chi2(1) =     2.77  Pr>chi2 = 0.0958
```

Adenoma

```
. ttest age, by(training), if na==1

Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |      43    62.18605    1.053001    6.904987    60.06101    64.31109
       1 |     200       59.44    .5528655    7.818699    58.34977    60.53023
---------+--------------------------------------------------------------------
combined |     243    59.92593    .4954417    7.723172       58.95    60.90185
---------+--------------------------------------------------------------------
    diff |            2.746047    1.288832                .2072319    5.284861
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =   2.1306
Ho: diff = 0                                     degrees of freedom =      241

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.9829         Pr(|T| > |t|) = 0.0341          Pr(T > t) = 0.0171
```

```
. cs sex training if na==1

                 | Training                |
                 |   Exposed   Unexposed   |      Total
-----------------+-------------------------+------------
           Cases |       118          27   |        145
        Noncases |        82          16   |         98
-----------------+-------------------------+------------
           Total |       200          43   |        243
                 |                         |
            Risk |        .59     .627907   |   .5967078
                 |                         |
                 |      Point estimate     |    [95% Conf. Interval]
                 |-------------------------+------------------------
 Risk difference |         -.037907        |   -.1976529    .1218389
      Risk ratio |         .9396296        |     .726343    1.215547
 Prev. frac. ex. |         .0603704        |   -.2155468     .273657
 Prev. frac. pop |         .0496875        |
                 +-------------------------------------------------
                              chi2(1) =     0.21  Pr>chi2 = 0.6457
```

```
. cs famhx_cca training if na==1

                 | Training                |
                 |   Exposed   Unexposed   |      Total
-----------------+-------------------------+------------
           Cases |       112          15   |        127
        Noncases |        73          25   |         98
-----------------+-------------------------+------------
           Total |       185          40   |        225
                 |                         |
            Risk |  .6054054        .375   |   .5644444
                 |                         |
                 |      Point estimate     |    [95% Conf. Interval]
                 |-------------------------+------------------------
 Risk difference |         .2304054        |    .0646676    .3961432
      Risk ratio |         1.614414        |    1.064308    2.448852
 Attr. frac. ex. |         .3805804        |    .0604226    .5916455
 Attr. frac. pop |         .3356299        |
                 +-------------------------------------------------
```

```
                           chi2(1) =     7.10  Pr>chi2 = 0.0077
```

```
. cs fh_any_ca training if na==1
```

```
                | Training               |
                |   Exposed   Unexposed  |      Total
----------------+------------------------+------------
          Cases |       141          27  |        168
       Noncases |        58          16  |         74
----------------+------------------------+------------
          Total |       199          43  |        242
                |                        |
           Risk |  .7085427    .627907   |   .6942149
                |                        |
                |      Point estimate    |    [95% Conf. Interval]
                |------------------------+------------------------
 Risk difference |         .0806357       |   -.0770314    .2383029
      Risk ratio |          1.12842       |     .881684    1.444204
 Attr. frac. ex. |          .113805       |   -.1341932     .307577
 Attr. frac. pop |          .095515       |
                +-------------------------------------------------
                           chi2(1) =     1.08  Pr>chi2 = 0.2980
```

```
. cs smoke training if na==1
```

```
                | Training               |
                |   Exposed   Unexposed  |      Total
----------------+------------------------+------------
          Cases |        29          10  |         39
       Noncases |       164          32  |        196
----------------+------------------------+------------
          Total |       193          42  |        235
                |                        |
           Risk |  .1502591    .2380952  |   .1659574
                |                        |
                |      Point estimate    |    [95% Conf. Interval]
                |------------------------+------------------------
 Risk difference |        -.0878362       |   -.2261595    .0504872
      Risk ratio |         .6310881       |    .3339057    1.192768
 Prev. frac. ex. |         .3689119       |   -.1927684    .6660943
 Prev. frac. pop |         .3029787       |
                +-------------------------------------------------
                           chi2(1) =     1.92  Pr>chi2 = 0.1656
```

```
. cs dm training if na==1
```

```
                | Training               |
                |   Exposed   Unexposed  |      Total
----------------+------------------------+------------
          Cases |         8           4  |         12
       Noncases |       192          39  |        231
----------------+------------------------+------------
          Total |       200          43  |        243
                |                        |
           Risk |       .04    .0930233  |   .0493827
                |                        |
                |      Point estimate    |    [95% Conf. Interval]
                |------------------------+------------------------
 Risk difference |        -.0530233       |   -.1439895     .037943
      Risk ratio |              .43       |    .1355934    1.363636
 Prev. frac. ex. |              .57       |   -.3636358    .8644066
 Prev. frac. pop |         .4691358       |
                +-------------------------------------------------
                           chi2(1) =     2.12  Pr>chi2 = 0.1454
```

```
. cs sympt_gibleed training if na==1

                | Training               |
                |   Exposed   Unexposed  |      Total
----------------+------------------------+------------
          Cases |         6           1  |          7
       Noncases |       194          42  |        236
----------------+------------------------+------------
          Total |       200          43  |        243
                |                        |
           Risk |       .03    .0232558  |   .0288066
                |                        |
                |      Point estimate    |    [95% Conf. Interval]
                |------------------------+------------------------
 Risk difference |         .0067442       |   -.0441302    .0576186
      Risk ratio |             1.29       |    .1593635    10.44217
 Attr. frac. ex. |         .2248062       |   -5.274963    .9042344
 Attr. frac. pop |          .192691       |
                +-------------------------------------------------
                              chi2(1) =     0.06  Pr>chi2 = 0.8104


. cs sympt_bowelhabit training if na==1

                | Training               |
                |   Exposed   Unexposed  |      Total
----------------+------------------------+------------
          Cases |         1           0  |          1
       Noncases |       199          43  |        242
----------------+------------------------+------------
          Total |       200          43  |        243
                |                        |
           Risk |      .005           0  |   .0041152
                |                        |
                |      Point estimate    |    [95% Conf. Interval]
                |------------------------+------------------------
 Risk difference |              .005      |   -.0047753    .0147753
      Risk ratio |                .       |           .           .
 Attr. frac. ex. |                1       |           .           .
 Attr. frac. pop |                1       |
                +-------------------------------------------------
                              chi2(1) =     0.22  Pr>chi2 = 0.6422



. generate path_vt_01 = .
(597 missing values generated)

. replace path_vt_01 = 0 if path_vt == 1
(215 real changes made)

. replace path_vt_01 = 1 if path_vt == 2
(28 real changes made)

. cs path_vt_01 training if na == 1

                | Training               |
                |   Exposed   Unexposed  |      Total
----------------+------------------------+------------
          Cases |        26           2  |         28
       Noncases |       174          41  |        215
----------------+------------------------+------------
          Total |       200          43  |        243
                |                        |
           Risk |       .13    .0465116  |   .1152263
                |                        |
                |      Point estimate    |    [95% Conf. Interval]
```

```
                |-----------------------+-------------------------
 Risk difference |         .0834884      |    .0051669    .1618099
      Risk ratio |           2.795       |    .6892547    11.33402
 Attr. frac. ex. |         .6422182      |   -.4508425      .91177
 Attr. frac. pop |         .5963455      |
                +-------------------------------------------------
                              chi2(1) =     2.42  Pr>chi2 = 0.1198


. cs srn training if na == 1

                | Training              |
                |   Exposed   Unexposed |      Total
----------------+-----------------------+------------
          Cases |        52          13 |         65
       Noncases |       148          30 |        178
----------------+-----------------------+------------
          Total |       200          43 |        243
                |                       |
           Risk |       .26    .3023256 |   .2674897
                |                       |
                |      Point estimate   |    [95% Conf. Interval]
                |-----------------------+-------------------------
 Risk difference |        -.0423256      |   -.1924546    .1078034
      Risk ratio |              .86      |    .5160583    1.433172
 Prev. frac. ex. |              .14      |   -.4331715    .4839417
 Prev. frac. pop |         .1152263      |
                +-------------------------------------------------
                              chi2(1) =     0.32  Pr>chi2 = 0.5695



. generate polyp_location_lr_01 = .
(597 missing values generated)

. replace polyp_location_lr_01 = 0 if polyp_location_lr == 1
(67 real changes made)

. replace polyp_location_lr_01 = 1 if polyp_location_lr == 2
(176 real changes made)


. cs polyp_location_lr_01 training if na == 1

                | Training              |
                |   Exposed   Unexposed |      Total
----------------+-----------------------+------------
          Cases |       149          27 |        176
       Noncases |        51          16 |         67
----------------+-----------------------+------------
          Total |       200          43 |        243
                |                       |
           Risk |      .745     .627907 |   .7242798
                |                       |
                |      Point estimate   |    [95% Conf. Interval]
                |-----------------------+-------------------------
 Risk difference |          .117093      |   -.0395001    .2736862
      Risk ratio |         1.186481      |    .9296349    1.514292
 Attr. frac. ex. |         .1571718      |   -.0756911    .3396252
 Attr. frac. pop |         .1330603      |
                +-------------------------------------------------
                              chi2(1) =     2.43  Pr>chi2 = 0.1190

. generate polyp_location_rc_01 = .
(597 missing values generated)


. replace polyp_location_rc_01 = 0 if polyp_location_rc == 1
```

```
(205 real changes made)

. replace polyp_location_rc_01 = 1 if polyp_location_rc == 2
(38 real changes made)

. cs polyp_location_rc_01 training if na == 1

                 | Training               |
                 |   Exposed   Unexposed  |      Total
-----------------+------------------------+------------
           Cases |        30           8  |         38
        Noncases |       170          35  |        205
-----------------+------------------------+------------
           Total |       200          43  |        243
                 |                        |
            Risk |       .15    .1860465  |   .1563786
                 |                        |
                 |      Point estimate    |    [95% Conf. Interval]
                 |------------------------+------------------------
 Risk difference |        -.0360465       |   -.1624483    .0903552
      Risk ratio |           .80625       |    .3976245    1.634806
 Prev. frac. ex. |           .19375       |   -.6348063    .6023755
 Prev. frac. pop |          .159465       |
                 +-------------------------------------------------
                              chi2(1) =     0.35  Pr>chi2 = 0.5549
```

6.0 The Metabolomic Fingerprint of Colorectal Cancer Remains After Curative Treatment

6.1 Abstract

Background: The urine metabolomic fingerprint of CRC could represent an early detection method for this common disease, but it s unclear whether this urine fingerprint persists in patients after curative treatment of CRC.

Aim: The aim of this chapter was to use projection-based methods to assess whether there are any differences in the urine metabolomic fingerprint of pre and post curative treatment CRC patients. Whether the CRC-predictive metabolites changed after curative resection therapy was also studied.

Methods: Urine samples were collected from 23 CRC patients at 3 months to 1 year after curative treatment of the CRC. The urine samples were analyzed using an Oxford 600Hz nuclear magnetic resonance (NMR) spectrometer with a Varian VNMRS two-channel console. The 1H NMR spectrum of each urine sample was analyzed using Chenomx NMRSuite v7.0 (Chenomx, Inc. Edmonton, Canada) and the metabolite concentrations were subsequently compared to the pre-treatment ones of the same patients. Projection-based models were used to separate the pre and post-treatment samples.

Results: When analyzed as a group, the pre-treatment CRC urine metabolomic fingerprint was not different from the post-treatment urine metabolomic

fingerprint. Six out of 23 CRC patients showed a recovery tendency towards normal. Only 2 of 10 CRC-predictive metabolites, hypoxanthine and 3-hydroxybutyrate, returned towards normal following CRC curative resection and treatment.

Conclusions: This study was not able to demonstrate a difference in the metabolomic fingerprint of CRC after curative treatment.

6.2 Introduction

Colorectal cancer is among the leading causes of death in North America, but it can be curable with surgical or a combination of surgical and medical treatments if identified early. However, about half of those that are thought to be curatively treated will develop recurrent or metastatic disease within 3 to 5 years of treatment, despite the absence of clinical, histological, and biochemical evidence of remaining overt disease after resection. The availability of validated biological markers for detection of complete resolution of disease after treatment and for early detection of recurrent disease can be one way to increase survival in these colorectal cancer patients.[1] Several studies have been published with distinguishing metabolites for CRC[2-9], but few have addressed what happens to these metabolites after the CRC has been cured and whether any of the metabolites could be used to detect CRC recurrence. The urine metabolomic fingerprint for post-treatment CRC is being investigated in this chapter.

Two recently published studies showed a clear and significant separation between the urines of pre-op, post-op colorectal cancer patients and healthy controls using advanced statistical methods.[7, 9] Ma et al. used UPLC/MS (ultra high performance liquid chromatography / mass spectroscopy) to examine the urine samples of 24 colorectal cancer patients both before and after their cancer operations, and that of 9 controls. They noted that when compared to the healthy controls, pre-op colorectal cancer patients had significantly increased levels of low-molecular weight compounds 283 and 234, and these compounds decreased significantly after the operation.[9] Qiu et al. examined the urine metabolite profile of 60 CRC patients using GC/MS and showed metabolic alterations between the preoperative and postoperative states.[7]

Given the results of the recent publications, we wanted to test this phenomenon in our population of colorectal cancer patients -- that is, to examine what happens to the metabolomic urinalysis of our CRC patients after they are cured by their surgical +/- medical treatments. Curative treatment was defined as not having any residual macroscopic cancer after surgery. We hypothesized that the metabolomic fingerprint of colorectal cancer would change post treatment.

6.3 Objectives

- To assess, using projection-based modeling, differences in the urine metabolomics of CRC patients before and after curative resection therapy

- To determine if the CRC-predictive metabolites changed after curative resection therapy

6.4 Material & Methods

6.4.1 Recruitment and Sample Collection

All patients in the training and testing CRC groups who met the inclusion criteria were contacted for the post op study via telephone between 3 months to 1 year after their CRC treatment when their diet, activities, medications etc. would have returned to baseline. Those who were treated with adjuvant chemotherapy were recruited 3 months to 1 year after they had completed their treatment to remove the effects that chemotherapy would have on metabolism. Patients were not recruited for the post-op study if they refused to participate, were unreachable by telephone, lived out of town, did not have curative treatment(s), still undergoing adjuvant treatment, or were deceased. Upon enrolling into the study, subjects were asked whether there were any changes in their medical conditions, medications, and family history since they were enrolled into the pre-op study. Clinical information such as the stage of cancer, adjuvant therapy, and CEA levels was collected from the patients' medical records. Specifically the CEA level around the time of post op urine collection was noted. The urine samples were collected from the patients in their normal state of diet and activity and in containers coated with sodium azide. Patients were contacted by telephone and urine containers were couriered to their place of residence.

From January 2009 to August 2010, 116 CRC patients were screened for this study and urine samples were collected from 23 patients (18 from the training set and 5 from the testing set). At the time of screening, 33 patients were either out of town, could not be reached via telephone or refused to participate in the study; 15 patients had either unresectable or metastatic CRC; 11 patients were deceased; 17 were still undergoing adjuvant treatment; and the rest were either past the 1 year post op time point or were still within 3 months of surgery.

6.4.2 Sample Analysis

All urine samples were stored at -80°C until they were ready to be analyzed. The day prior to NMR acquisition, each sample was thawed to room temperature and was diluted (1:10) with internal standard consisting of 5 mM sodium 2,2-dimethyl-2-silapentane-5-sulfonate (DSS), 100 mM imidazole, 0.2% sodium azide in 99% D2O. The samples were stored at 4°C overnight. On the day of NMR acquisition, each sampled was adjusted to a pH between 6.7 and 6.8 and aliquoted into 5mm NMR tubes. One-dimensional nuclear magnetic resonance spectra was acquired using an Oxford 600Hz NMR spectrometer with a Varian VNMRS two channel console and running VNMRJ software version 2.2C on a RHEL 4 host computer in the Canadian National High Field NMR Centre (NANUC), Edmonton, Alberta. All samples were run at a sweep width (sw) of 7225.43 Hz. The saturation frequency (sfrq), transmitter offset (tof) and pulse width (pw) were all individually calibrated at the start of each day. The tof typically ranged from (-213 to -215 Hz) and the pw ranged from 6 to 8

microseconds. Shims were optimized until an acceptable line width value was obtained at relative peak heights of: 50% (< 1.0 Hz), 0.55% (< 12.0 Hz), and 0.11% (< 20.0 Hz) were achieved. Water suppression was performed. Spectra were collected at 25°C through a total of 32 scans over a period of 3.5 minutes; a total recycle delay of 5 seconds was also used (*i.e.* 1 second recovery delay/saturation and a 4 second acquisition). The 1H NMR spectrum of each urine sample was analyzed and quantitated using the targeted profiling technique[10] as implemented in Chenomx NMRSuite v7.0 (Chenomx, Inc. Edmonton, Canada). The quantification process was done independently by two individuals and verified by a third individual to optimize accuracy. 294 metabolites were considered and 72 were significant.

The spectral acquisition and quantification process were performed without the knowledge of the pathology results.

6.4.3 Data Analysis

The twenty-three pairs of samples were analyzed using projection-based methods with the aid of SIMCA-P+ v12.0.1 (Umetrics, Umea, Sweden). The metabolite concentrations were normalized (to total metabolite concentration except urea) to account for the dilutional differences in the urine samples. Log transformation was done to account for the non-normal distributive nature of the concentrations. Finally, those metabolites that are not products of normal human metabolism, i.e. xenobiotics, such as ibuprofen and salicylurate, were excluded. The pre-

treatment and post-treatment groups were compared to each other but also the post-treatment group was checked against the normal vs. CRC model. In addition, the concentrations of the top contributing metabolites in the CRC model were examined in the pre and post-treatment groups to see if there were any differences between the two states. This was statistically analyzed using paired student's t-test (STATA/SE 10.1 (TX, USA)) as the two groups of data were from the same population of patients.

6.5 Results

6.5.1 Patient Characteristics

The demographics and characteristics of the patients enrolled in the post-treatment study are listed in table 6.1. The average time of the post-treatment urine collection was 8.8 months from surgery.

Table 6.1: Post curative treatment patient characteristics

			Post op Patients (n=23) N (%)
Male:Female			17:6
Age at Diagnosis (years ± SEM)			69.4 ± 2.0
Smoking			5 (23)
Diabetes			5 (22)
Family history	CRC		4 (17)
	Any cancer		16 (70)
Symptoms	GI bleed		11 (48)
	Change bowel habits		12 (52)
Location of cancer	Rectal vs. colon		8 (35)
	Left vs. right		17 (74)
Pathology of Cancer	Lymphatic		4 (17)
	Vascular		2 (8)
	Perineural		0 (0)
	Lymphocytic		6 (33)
	Grade	Well	22 (96)
		Moderate	0
		High	1(4)
Cancer stage	Stage 1		13 (57)
	Stage 2		5 (22)
	Stage 3		5 (22)
	Stage 4		0 (0)
Pre-op CEA (>5 vs. <=5)			4 (22)
Post-op CEA (at time of urine collection) (>5 vs. <=5)			1 (5)
Adjuvant Chemotherapy			3 (13)

Note: Not all % are calculated with the denominator of the total in each group as some clinical information was missing or unknown.

6.5.2 Building and Analyzing the Models

An unsupervised principal component analysis (PCA) model was built using the pre and the post-treatment CRC sample concentrations. Auto-fitting by SIMCA

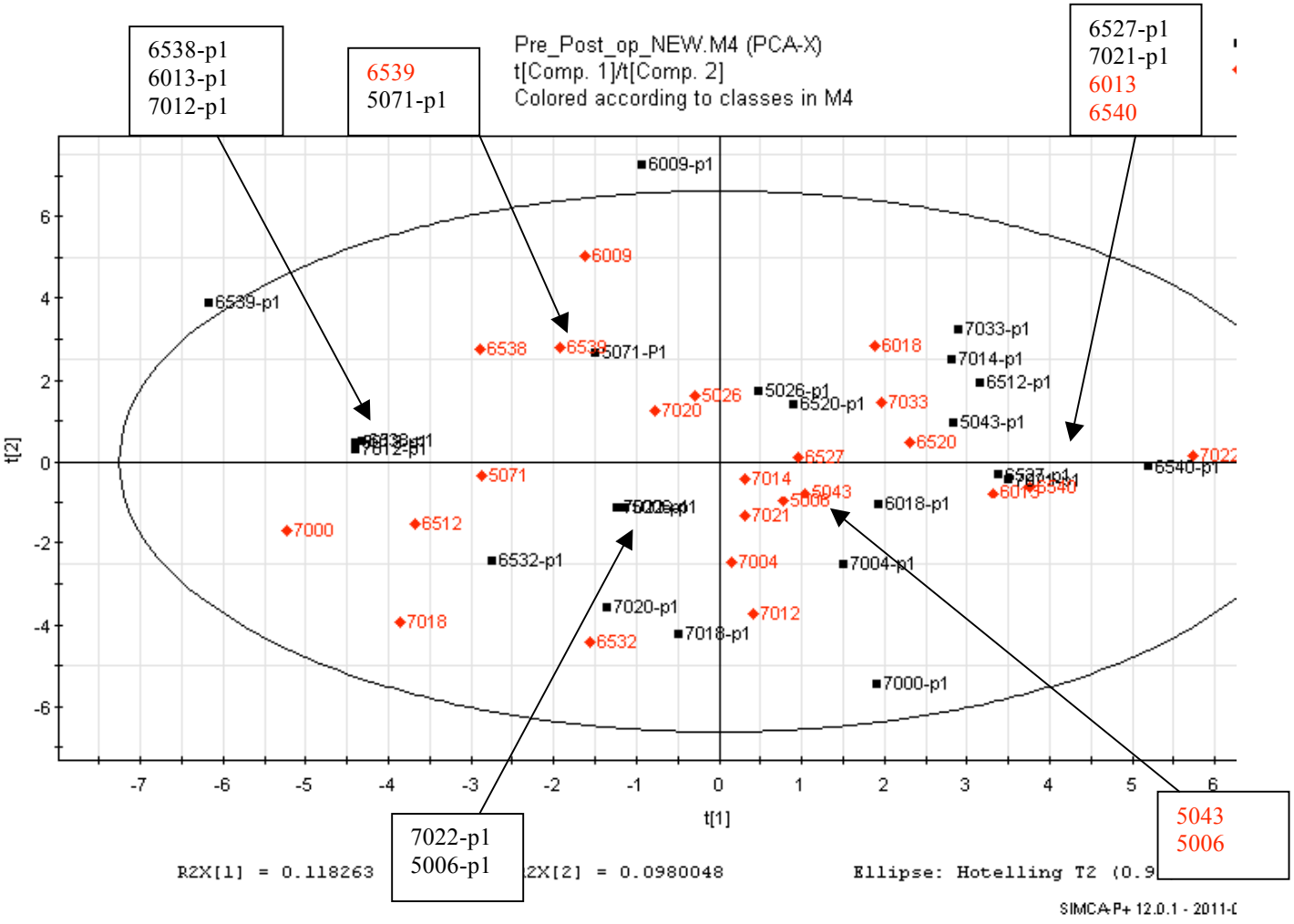
resulted in a two component model and a scatter plot was generated (figure 6.1). The sample labels were shown on the plot to compare the individual pairs of pre and post-treatment samples. The overlapped sample labels are clarified in the textboxes shown in figure 6.1. Careful analysis of the scatter plot revealed that there were three categories of samples. Ten pairs of pre and post-treatment samples were within the same quadrant and fairly close to each other on the scatter plot (6009, 6539, 6538, 6532, 7018, 7004, 7021, 6540, 6520, 7033). The other 13 pairs were all on different quadrants on the scatter plot, nine across one quadrant (5071, 7000, 7020, 5006, 7014, 5043, 6527, 5026, 6018) and four across two (7012, 6013, 6512, 7022).

The clinical characteristics were analyzed against these categories to see if there were any other correlating factors and there were not. For example, the three patients who received adjuvant chemotherapy (7021, 5043, and 6013) were evenly distributed with one in each category. There was also no specific pattern for the three out the four subjects with elevated preoperative CEA that returned to normal in the post-treatment state (6539, 7018, and 6013).

It is unclear whether these categories identified above are significant or even meaningful since they are generated from a PCA plot and as previously stated, PCA is unsupervised and can be separating based on many other factors and not necessarily based on the pre and post-treatment groups. Supervised PLS or OPLS

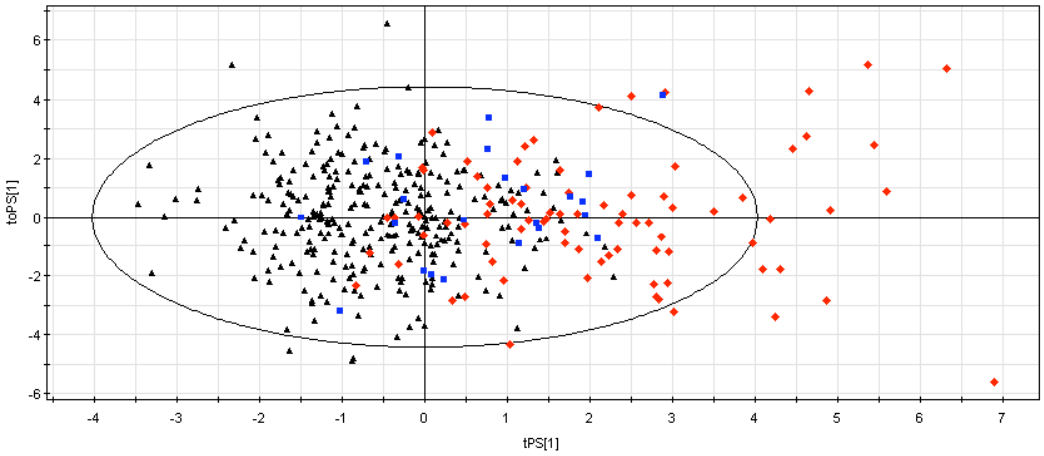
models were attempted to separate the pre and post-treatment groups, but could not be generated as the Q^2Y values were negative.

Figure 6.1: Scatter plot of pre-treatment (red diamonds) vs. post-treatment (black squares) CRC patients PCA



The post-treatment samples were superimposed onto the original normal versus CRC model OPLS scatter plot to assess whether the post-treatment samples were more alike the CRC samples or the normal samples (figure 6.2). It is fairly clear from exploratory data analysis that the post-treatment samples are distributed mostly on the CRC side of the plot. More objectively, when Y-predicted values were generated for the post-treatment samples, 17 out of 23 were higher than the cutoff of 0.212925, that is, on the cancer side of the plot.

Figure 6.2: Post treatment CRC samples (blue squares) superimposed on normal (black triangles) vs. CRC (red diamonds) OPLS scatter plot



To further illustrate any changes in post-treatment CRC samples compared to the pre-treatment ones, a new Normal (n=294) vs. Pre-treatment CRC (n=23) OPLS model was built (scatter plot shown in figure 6.3). The post-treatment CRC samples (n=23) were superimposed onto this scatter plot and it can be seen that three of the samples actually migrated to the normal side of the OPLS scatter plot (6009, 6512, 7033) shown in figure 6.4. Three other samples also migrated towards the normal side of the scatter plot and have Y-predicted values lower than

the cancer cutoff of 0.212925, that is, they would have been interpreted as normal (6538, 6539, and 7014) (figure 6.5). This means that 6 of the 23 CRC patients showed a recovering tendency towards normal state after they have had their curative treatment(s). Interestingly, all these patients had early stage CRC (5/6 stage 1, 1/6 stage 2) and 5 out of these 6 patients did not have a family history of CRC.

Figure 6.3: Normal (black squares) vs. pre-treatment CRC (red diamonds) OPLS scatter plot

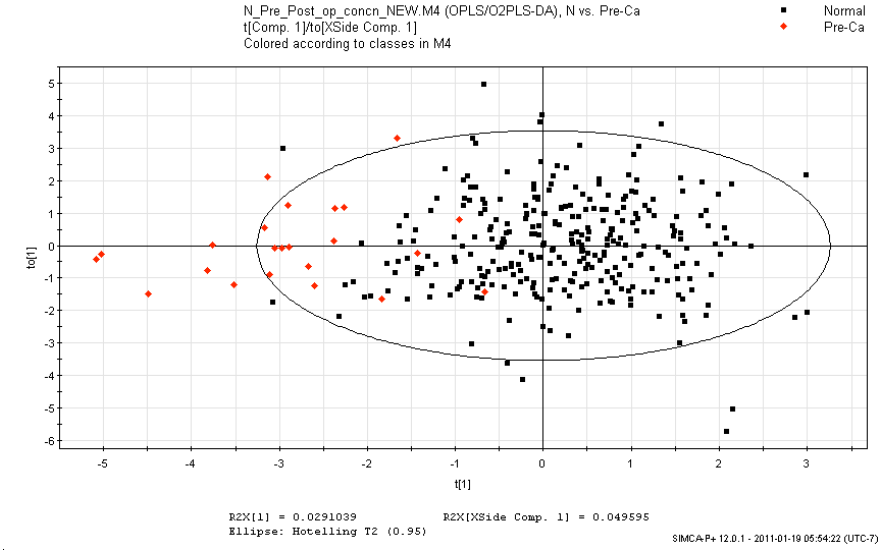


Figure 6.4: Post-treatment CRC samples (blue squares) superimposed on the normal (black triangles) vs. pre-treatment CRC samples (red diamonds) model specifically showing the three CRC samples that migrated to the normal side of the scatter plot (7033, 6512, 6009)

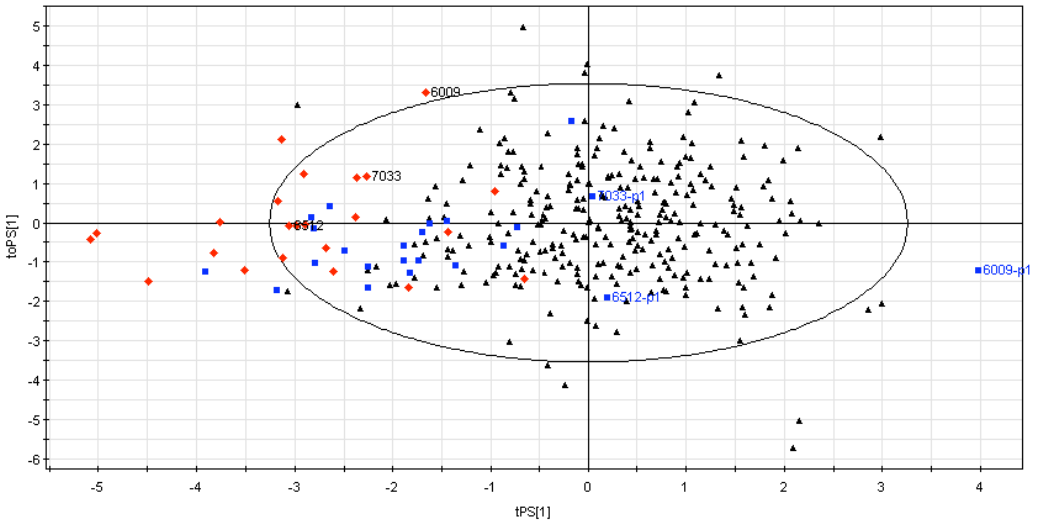
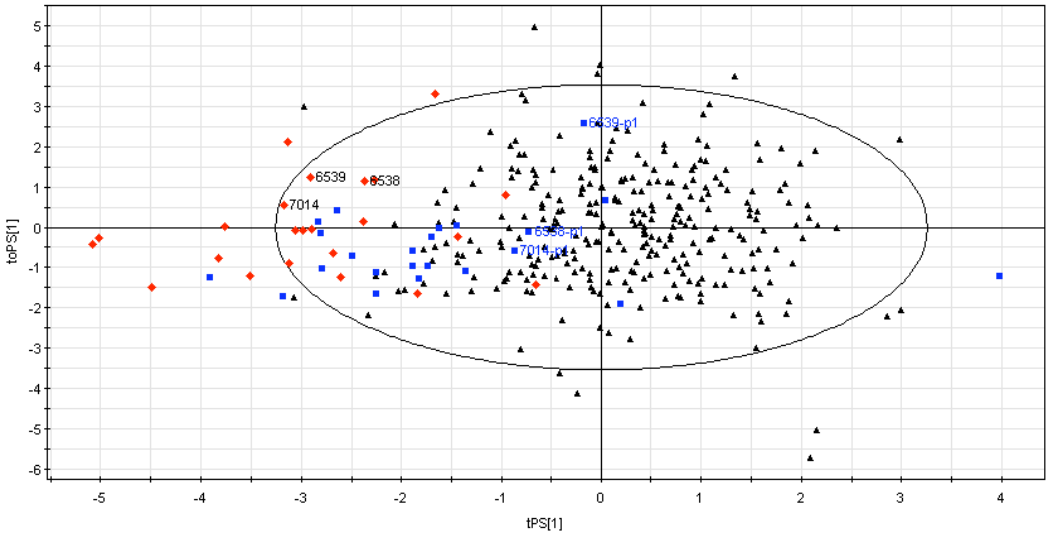
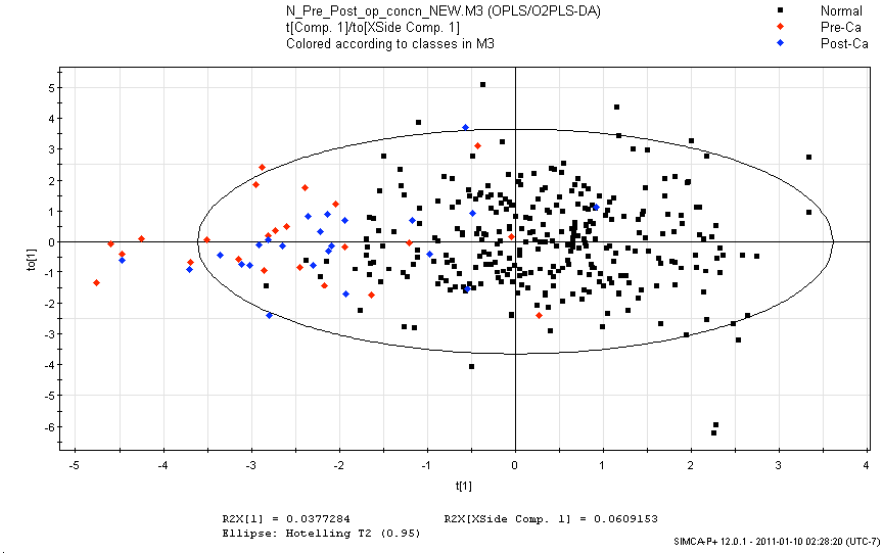


Figure 6.5: Post-treatment CRC samples (blue squares) superimposed on the normal (black triangles) vs. pre-treatment CRC samples (red diamonds) model specifically showing the other three CRC samples that migrated towards the normal side of the scatter plot (6538, 6539, 7014)



When the three groups (normal, pre-treatment, and post-treatment CRC patients) were plotted on the same OPLS scatter plot, there was not a good separation between the pre (red diamonds) and post-treatment (blue diamonds) groups. (Figure 6.6)

Figure 6.6: OPLS scatter plot of normal (black squares) vs. pre-treatment CRC (red diamonds) and post treatment CRC patients (blue diamonds)



6.5.3 Metabolites

The top ten contributing metabolites in the normal versus CRC model (hypoxanthine, creatinine, dimethylamine, 3-indoxylsulfate, methanol, adipate, urea, guanidoacetate, 3-hydroxybutyrate, and acetone) were analyzed to see if their levels significantly changed after the curative treatment of CRC. Paired student's t-test was used to compare the concentrations of the metabolite concentrations and the results are shown in table 6.2. Also shown in table 6.2 is

the average concentration of the metabolites in the normal control group from the previous normal vs. CRC analysis, to determine whether the post-treatment changes in the metabolite concentration were towards that of the normal or not.

Table 6.2: Comparison of the concentrations of the top 10 contributing metabolites generated from the normal vs. CRC model in the pre-treatment and post-treatment samples using paired student's t-test; also shown is the average metabolite concentration in the normal control group

Metabolites	Pre-treatment CRC Average Metabolite Concentration (μM)	Post-treatment CRC Average Metabolite Concentration (μM)	p-value	Normal Control Average Metabolite Concentration (μM)
Hypoxanthine	55.5	79.9	0.398	18.2
Creatinine	10,162.4	20,973.6	0.064	5584.6
Dimethylamine	374.1	743.2	0.082	185.4
3-indoxylsulfate	259.8	526.0	0.056	126.1
Methanol	20.0	82.3	0.029	76.8
Adipate	7.0	0	0.328	0.7
Urea	150,026.4	343,737.4	0.074	141,544.5
Guanidoacetate	107.6	289.2	0.052	180.6
3-Hydroxybutyrate	19.7	3.7	0.058	13.3
Acetone	14.1	53.3	0.139	9.5

Several metabolites were quite different between the pre and post-treatment groups; in fact, seven out of ten metabolites had p-values <0.1, namely creatinine, dimethylamine, 3-indoxylsulfate, methanol, urea, guanidoacetate, and 3-hydroxybutyrate. When these seven metabolites were more carefully analyzed, it was determined that for three of them (methanol, guanidoacetate, and 3-hydroxybutyrate), the direction of change from the pre to the post-treatment states

was in the same direction as that towards the normal concentration. For example, the average concentration of methanol was 76.8μM in the normal group and 20.0μM in the pre-treatment CRC group and this increased to 82.3μM in the post-treatment CRC group. This showed a recovering tendency towards healthy state in the post-treatment samples. However, for the other four metabolites (creatinine, dimethylamine, 3-indoxylsulfate, and urea), the direction of change was the opposite.

Examining the raw concentrations of the urine samples overlooks the effects of different hydration states, thus the normalized concentrations of the metabolites would likely give a more realistic representation of any differences between the pre and post-treatment groups (table 6.3). Normalization was to total metabolite concentration minus urea concentration, *i.e.* ([metabolite]/([total metabolite]-[urea]). When the normalized data is examined, only hypoxanthine and 3-hydroxybutyrate levels are different in the post-treatment state compared to the pre-treatment state, both showing a recovering tendency towards healthy state.

Table 6.3: Comparison of the *normalized concentrations* of the top 10 contributing metabolites generated from the normal vs. CRC model in the pre-treatment and post-treatment samples using paired student's t-test; also shown is the *normalized* average metabolite concentration in the normal control group

Metabolites	Pre-treatment CRC Average Normalized Metabolite Concentration	Post-treatment CRC Average Normalized Metabolite Concentration	p-value	Normal Control Average Normalized Metabolite Concentration
Hypoxanthine	0.00193	0.00111	0.031	0.00092
Creatinine	0.41833	0.40870	0.734	0.33153
Dimethylamine	0.01562	0.01435	0.259	0.01137
3-indoxylsulfate	0.01020	0.00992	0.858	0.00667
Methanol	0.00130	0.00188	0.209	0.00484
Adipate	0.00014	0	0.328	0.00005
Urea	7.00514	7.24626	0.784	10.42251
Guanidoacetate	0.00473	0.00576	0.344	0.01132
3-Hydroxybutyrate	0.00059	0.00013	0.033	0.00058
Acetone	0.00094	0.00080	0.550	0.00082

6.6 Discussion

6.6.1 Summary

The urine metabolomic fingerprint of post-treatment CRC patients was studied in this chapter. Twenty-three urine samples were collected from CRC patients who underwent curative surgical resection with or without adjuvant chemotherapy approximately 3 months to 1 year post-treatment. Using projection-based methods, a model to separate the pre and the post-treatment groups was attempted but could not be built due to the lack of difference between the two groups. When the post-treatment group samples were validated against the normal vs. CRC

model (from chapter 4), they were more like the CRC patients, and when all three groups were represented on in an OPLS scatter plot, the pre and post-treatment samples could not be separated. However, a more targeted analysis of the pre and post-treatment samples using a new normal vs. pre-treatment CRC OPLS model demonstrated that 6 out of the 23 CRC patients' metabolomic fingerprints showed a recovery tendency towards normal. All of these six patients had early stage CRC and 5 out of 6 had no family history of CRC. This could suggest that early stage CRC metabolomic fingerprints are more likely to revert to normal after treatment or that patients without a positive family history of CRC are not 'genetically pre-dispositioned to have CRC' and thus their metabolomic fingerprint of CRC disappears after treatment, however, the number of patients in this study is too small to draw any definitive conclusions. This will be explored further as more post-treatment patients are recruited.

When the top CRC-predictive metabolites were analyzed, there were some significant differences in seven out of the ten metabolites between the pre and post-treatment groups, however, less than half of these showed a recovering tendency towards normal. When the normalized concentrations were examined, only 2 of 10 CRC-predictive metabolites, namely hypoxanthine and 3-hydroxybutyrate, returned towards normal following CRC curative resection and treatment. This explains why the post-treatment group, for the most part, did not resemble the normal group when analyzed using the projection-based methods.

6.6.2 Comparison to Literature

Unlike what is shown in the literature[7, 9], the post-treatment patient samples in this study could not be easily separated from the pre-treatment ones using projection-based methods. There are several differences between this study and the two in the literature. In Qiu's study, the post-operative urine sample was collected on the seventh day after surgery and in Ma's study, the post operative urine sample was also collected in the immediate post op period, although the exact timing was unspecified. In contrast, for this study, the urine samples were collected 3 months to 1 year after the definitive treatments are finished. There are many metabolic changes in the immediate post-operative stages, such as recent bowel preps, fluid and electrolyte derangements, perioperative medications, and altered activity levels, therefore we elected to collect the post-treatment urine sample when the subjects' metabolisms would have presumably returned to their baselines. Another advantage that this study has over the others is that only those patients with curative intent were included, i.e. those with metastatic cancer were excluded. In Qiu's study, 9 of the CRC patients had stage IV or metastatic CRC, so after the colon resection surgery there would have been residual CRC cells in the body and hence metabolism would still be affected by these CRC cells. Therefore, the separation of the pre and post-operative patients demonstrated in the literature may not be due to the CRC being removed but rather a host of other potential factors. Moreover, both Ma & Qiu's study subjects are from China whereas the patients for this study are from Northern Alberta (Edmonton and

Grande Prairie). The differences in patient ethnicity, climate, and diet can also significantly change the subjects' metabolite profiles.

6.6.3 Limitations

The major limitation in this study is the small sample size, which makes it difficult to use projection-based multivariate methods to analyze the samples. After screening all 116 CRC patients included in the normal vs. CRC analysis (chapter 4), only 23 subjects fit the inclusion criteria so far. About a third of the patients were either out of town, cannot be reached via telephone or refused to participate in the study. Due to the geographical distributions of Alberta, many patients live in the small towns around the periphery of Edmonton and Grande Prairie and traveled to the city to have their surgeries, but many find it quite difficult or troublesome to travel to the city to participate in this follow up study. At the cut-off time for recruitment into this study (Aug 2010) there were still many CRC patients who were still undergoing adjuvant treatment or were within three months of their surgery date, therefore in a few months time there will be more patients that could qualify for this study.

As mentioned in the previous chapters, the number of metabolites analyzed for this post-treatment group is limited by the Chenomx compound library.

6.6.4 Strengths of Study

The advantages of this study, as stated above, are the timing of collection of the post-treatment samples and the exclusion of the metastatic CRC patients. These two modifications from the existing studies in the literature theoretically result in a more homogenous patient population and also remove the potential confounding perioperative changes in metabolism from the analysis.

6.7 Conclusions

This post-treatment study showed that when analyzed as a group, the pre-treatment CRC urine metabolomic fingerprint was not different from the post-treatment urine metabolomic fingerprint. However a quarter of the patients did show a recovery tendency towards normal after curative treatment of their CRC. Two of 10 CRC-predictive metabolites trend towards normal following CRC curative resection and treatment.

6.8 References

1. Sorensen NM, Sorensen IV, Wurtz SO, Schrohl AS, Dowell B, Davis G, Christensen IJ, Nielsen HJ, Brunner N. Biology and potential clinical implications of tissue inhibitor of metalloproteinases-1 in colorectal cancer treatment. Scandinavian Journal of Gastroenterology 2008;43:774-786.
2. Chan ECY, Koh PK, Mal M, Cheah PY, Eu KW, Backshall A, Cavill R, Nicholson JK, Keun HC. Metabolic Profiling of Human Colorectal Cancer Using High-Resolution Magic Angle Spinning Nuclear Magnetic Resonance (HR-MAS NMR) Spectroscopy and Gas Chromatography Mass Spectrometry (GC/MS). Journal of Proteome Research 2009;8:352-361.
3. Piotto M, Moussallieh FM, Dillmann B, Imperiale A, Neuville A, Brigand C, Bellocq JP, Elbayed K, Namer IJ. Metabolic characterization of primary human colorectal cancers using high resolution magic angle spinning H-1 magnetic resonance spectroscopy. Metabolomics 2009;5:292-301.
4. Denkert C, Budczies J, Weichert W, Wohlgemuth G, Scholz M, Kind T, Niesporek S, Noske A, Buckendahl A, Dietel M, Fiehn O. Metabolite profiling of human colon carcinoma - deregulation of TCA cycle and amino acid turnover. Molecular Cancer 2008;7.
5. Lean CL, Newland RC, Ende DA, Bokey EL, Smith ICP, Mountford CE. ASSESSMENT OF HUMAN COLORECTAL BIOPSIES BY H-1 MRS - CORRELATION WITH HISTOPATHOLOGY. Magnetic Resonance in Medicine 1993;30:525-533.
6. Qiu YP, Cai GX, Su MM, Chen TL, Zheng XJ, Xu Y, Ni Y, Zhao AH, Xu LX, Cai SJ, Jia W. Serum Metabolite Profiling of Human Colorectal Cancer Using GC-TOFMS and UPLC-QTOFMS. Journal of Proteome Research 2009;8:4844-4850.
7. Qiu YP, Cai GX, Su MM, Chen TL, Liu YM, Xu Y, Ni Y, Zhao AH, Cai SJ, Xu LX, Jia W. Urinary Metabonomic Study on Colorectal Cancer. Journal of Proteome Research 2010;9:1627-1634.
8. Ritchie SA, Ahiahonu PWK, Jayasinghe D, Heath D, Liu J, Lu YS, Jin W, Kavianpour A, Yamazaki Y, Khan AM, Hossain M, Su-Myat KK, Wood PL, Krenitsky K, Takemasa I, Miyake M, Sekimoto M, Monden M, Matsubara H, Nomura F, Goodenowe DB. Reduced levels of hydroxylated, polyunsaturated ultra long-chain fatty acids in the serum of colorectal cancer patients: implications for early screening and detection. Bmc Medicine 2010;8.
9. Ma Y-L, Qin H-L, Liu W-J, Peng J-Y, Huang L, Zhao X-P, Cheng Y-Y. Ultra-High Performance Liquid Chromatography-Mass Spectrometry for the Metabolomic Analysis of Urine in Colorectal Cancer. Digestive Diseases and Sciences 2009;54:2655-2662.
10. Chang D, Weljie A, Newton J. Leveraging latent information in NMR spectra for robust predictive models. Pacific Symposium on Biocomputing 2007 2007:115-126.

6.9 Appendix

```
. tab sex if postop == 0

        Sex |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |          6       26.09       26.09
          1 |         17       73.91      100.00
------------+-----------------------------------
      Total |         23      100.00

. sum age if postop==0

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
         age |        23     69.3913    9.731575         51         86

. sum age if
invalid syntax
r(198);

. sum age

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
         age |        23     69.3913    9.731575         51         86

. tab famhx_cca

  Famhx_cca |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |         19       82.61       82.61
          1 |          4       17.39      100.00
------------+-----------------------------------
      Total |         23      100.00

. tab fh_any_ca

  FH_ANY_CA |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |          7       30.43       30.43
          1 |         16       69.57      100.00
------------+-----------------------------------
      Total |         23      100.00

. tab smoke

      Smoke |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |         17       77.27       77.27
          1 |          5       22.73      100.00
------------+-----------------------------------
      Total |         22      100.00

. tab dm

         DM |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |         18       78.26       78.26
          1 |          5       21.74      100.00
------------+-----------------------------------
      Total |         23      100.00

. tab sympt_gibleed

Sympt_GIble |
         ed |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |         12       52.17       52.17
```

```
          1 |         11       47.83      100.00
------------+-----------------------------------
      Total |         23      100.00

. tab sympt_bowelhabit

Sympt_bowel |
      habit |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |         11       47.83       47.83
          1 |         12       52.17      100.00
------------+-----------------------------------
      Total |         23      100.00


. generate or_tumorlocation_rc =.

(46 missing values generated)

. replace or_tumorlocation_rc = 0 if or_tumorlocation == "Rt COL"

(6 real changes made)

. replace or_tumorlocation_rc = 0 if or_tumorlocation == "Lt COL"

(2 real changes made

. replace or_tumorlocation_rc = 0 if or_tumorlocation == "Sigmoid"

(6 real changes made)

. replace or_tumorlocation_rc = 0 if or_tumorlocation == "Rt COL & Sigmoid"

(1 real change made)

. replace or_tumorlocation_rc = 1 if or_tumorlocation == "Rect Below"

(5 real changes made)


. replace or_tumorlocation_rc = 1 if or_tumorlocation == "Rect At"

(1 real change made)

. replace or_tumorlocation_rc = 1 if or_tumorlocation == "Rect Above"

(2 real changes made)

. tab or_tumorlocation_rc


or_tumorloc |
   ation_rc |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |         15       65.22       65.22
          1 |          8       34.78      100.00
------------+-----------------------------------
      Total |         23      100.00

. generate or_tumorlocation_lr =.

(46 missing values generated)

. replace or_tumorlocation_lr = 0 if or_tumorlocation == "Rt COL"

(6 real changes made)

. replace or_tumorlocation_lr = 1 if or_tumorlocation == "Lt COL"
```

```
(2 real changes made)

. replace or_tumorlocation_lr = 1 if or_tumorlocation == "Sigmoid"

(6 real changes made)

. replace or_tumorlocation_lr = 1 if or_tumorlocation == "Rt COL & Sigmoid"

(1 real change made)

. replace or_tumorlocation_lr = 1 if or_tumorlocation == "Rect Below"

(5 real changes made)

. replace or_tumorlocation_lr = 1 if or_tumorlocation == "Rect At"

(1 real change made)

. replace or_tumorlocation_lr = 1 if or_tumorlocation == "Rect Above"

(2 real changes made)


. tab or_tumorlocation_lr
or_tumorloc |
   ation_lr |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |          6       26.09       26.09
          1 |         17       73.91      100.00
------------+-----------------------------------
      Total |         23      100.00

. tab li

         LI |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |         19       82.61       82.61
          1 |          4       17.39      100.00
------------+-----------------------------------
      Total |         23      100.00


. tab vi

         VI |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |         21       91.30       91.30
          1 |          2        8.70      100.00
------------+-----------------------------------
      Total |         23      100.00

. tab pni

        PNI |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |         21      100.00      100.00
------------+-----------------------------------
      Total |         21      100.00

. tab lymphocyctic_resp

Lymphocycti |
     c_resp |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |         12       66.67       66.67
          1 |          6       33.33      100.00
------------+-----------------------------------
      Total |         18      100.00
```

```
. tab gradepath

  GradePath |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |         22       95.65       95.65
          3 |          1        4.35      100.00
------------+-----------------------------------
      Total |         23      100.00


. tab stage

STAGE_of_Ca |
       ncer |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |         13       56.52       56.52
          2 |          5       21.74       78.26
          3 |          5       21.74      100.00
------------+-----------------------------------
      Total |         23      100.00

. generate CEA_cat
=exp required
r(100);

. generate CEA_cat =.
(47 missing values generated)

. replace CEA_cat = 0 if cea <=5 & cea !=.
(14 real changes made)

. replace CEA_cat = 1 if cea > 5 & cea !=.
(4 real changes made)

. tab CEA_cat

    CEA_cat |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |         14       77.78       77.78
          1 |          4       22.22      100.00
------------+-----------------------------------
      Total |         18      100.00

. generate postcea_cat = .
(47 missing values generated)

. replace postcea_cat = 0 if cea_at_post_op_collection <=5 &
cea_at_post_op_collection !=.
(19 real changes made)

. replace postcea_cat = 1 if cea_at_post_op_collection > 5 &
cea_at_post_op_collection !=.
(1 real change made)

. tab postcea_cat

postcea_cat |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |         19       95.00       95.00
          1 |          1        5.00      100.00
------------+-----------------------------------
      Total |         20      100.00

. tab post_op_chemoyn

Post_op_Che |
       moyn |      Freq.     Percent        Cum.
------------+-----------------------------------
```

```
          0 |         20       86.96       86.96
          1 |          3       13.04      100.00
------------+-----------------------------------
      Total |         23      100.00

. sum time_from_or

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
time_from_or |        23    8.782609    3.204493          4         14


. ttest hypoxanthine = hypoxanthine_post

Paired t test
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
hypoxa~e |      23    55.45652    12.37113    59.32983    29.80038    81.11267
hypoxa~t |      23    79.92609    29.67827     142.332    18.37711    141.4751
---------+--------------------------------------------------------------------
    diff |      23   -24.46956    28.36762    136.0463   -83.30041    34.36128
------------------------------------------------------------------------------
     mean(diff) = mean(hypoxanthine - hypoxanthine_p~t)           t =  -0.8626
 Ho: mean(diff) = 0                               degrees of freedom =       22

 Ha: mean(diff) < 0           Ha: mean(diff) != 0           Ha: mean(diff) > 0
 Pr(T < t) = 0.1988         Pr(|T| > |t|) = 0.3977          Pr(T > t) = 0.8012

. ttest creatinine = creatinine_post

Paired t test
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
creati~e |      23    10162.43    1432.794    6871.437    7190.998    13133.86
creati~t |      23    20973.58    5865.074    28127.91    8810.164       33137
---------+--------------------------------------------------------------------
    diff |      23   -10811.15    5553.426    26633.29   -22328.25    705.9481
------------------------------------------------------------------------------
     mean(diff) = mean(creatinine - creatinine_post)              t =  -1.9468
 Ho: mean(diff) = 0                               degrees of freedom =       22

 Ha: mean(diff) < 0           Ha: mean(diff) != 0           Ha: mean(diff) > 0
 Pr(T < t) = 0.0322         Pr(|T| > |t|) = 0.0644          Pr(T > t) = 0.9678

. ttest dimethylamine = dimethylamine_post

Paired t test
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
dimeth~e |      23    374.1261    57.36578    275.1166    255.1567    493.0954
dimeth~t |      23    743.1565    212.9026    1021.045    301.6236    1184.689
---------+--------------------------------------------------------------------
    diff |      23   -369.0304    202.5175    971.2398    -789.026    50.96515
------------------------------------------------------------------------------
     mean(diff) = mean(dimethylamine - dimethylamine_~t)          t =  -1.8222
 Ho: mean(diff) = 0                               degrees of freedom =       22

 Ha: mean(diff) < 0           Ha: mean(diff) != 0           Ha: mean(diff) > 0
 Pr(T < t) = 0.0410         Pr(|T| > |t|) = 0.0820          Pr(T > t) = 0.9590

. ttest indoxylsulfate= indoxylsulfate_post

Paired t test
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
indoxy~e |      23    259.8217    48.43363    232.2795    159.3765    360.2669
```

```
indoxy~t |      23    525.9696    146.1294    700.8122    222.9157    829.0235
---------+--------------------------------------------------------------------
    diff |      23   -266.1478    131.6651    631.4437   -539.2046    6.908911
------------------------------------------------------------------------------
     mean(diff) = mean(indoxylsulfate - indoxylsulfate~t)         t =  -2.0214
 Ho: mean(diff) = 0                               degrees of freedom =       22

 Ha: mean(diff) < 0           Ha: mean(diff) != 0           Ha: mean(diff) > 0
 Pr(T < t) = 0.0278         Pr(|T| > |t|) = 0.0556          Pr(T > t) = 0.9722

. ttest methanol= methanol_post

Paired t test
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
methanol |      23    19.96087    3.251067    15.59157    13.21857    26.70317
methan~t |      23    82.35652    27.70549    132.8709    24.89885    139.8142
---------+--------------------------------------------------------------------
    diff |      23   -62.39565    26.78432    128.4531   -117.9429   -6.848374
------------------------------------------------------------------------------
     mean(diff) = mean(methanol - methanol_post)                 t =  -2.3296
 Ho: mean(diff) = 0                               degrees of freedom =       22

 Ha: mean(diff) < 0           Ha: mean(diff) != 0           Ha: mean(diff) > 0
 Pr(T < t) = 0.0147         Pr(|T| > |t|) = 0.0294          Pr(T > t) = 0.9853

. ttest adipate= adipate_post

Paired t test
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
 adipate |      23    7.026087    7.026087    33.69593   -7.545126     21.5973
adipat~t |      23           0           0           0           0           0
---------+--------------------------------------------------------------------
    diff |      23    7.026087    7.026087    33.69593   -7.545126     21.5973
------------------------------------------------------------------------------
     mean(diff) = mean(adipate - adipate_post)                   t =   1.0000
 Ho: mean(diff) = 0                               degrees of freedom =       22

 Ha: mean(diff) < 0           Ha: mean(diff) != 0           Ha: mean(diff) > 0
 Pr(T < t) = 0.8359         Pr(|T| > |t|) = 0.3282          Pr(T > t) = 0.1641

. ttest urea= urea_post

Paired t test
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
    urea |      23    150026.4    19722.01    94583.46    109125.4    190927.3
urea_p~t |      23    343737.4    101332.3    485972.7    133587.1    553887.8
---------+--------------------------------------------------------------------
    diff |      23   -193711.1      103395    495864.9   -408139.2    20716.98
------------------------------------------------------------------------------
     mean(diff) = mean(urea - urea_post)                         t =  -1.8735
 Ho: mean(diff) = 0                               degrees of freedom =       22

 Ha: mean(diff) < 0           Ha: mean(diff) != 0           Ha: mean(diff) > 0
 Pr(T < t) = 0.0372         Pr(|T| > |t|) = 0.0743          Pr(T > t) = 0.9628

. ttest guanidoacetate= guandidoacetate_post

Paired t test
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
guanid~e |      23    107.5609    21.88041    104.9348    62.18367    152.9381
guandi~t |      23    289.1783    90.95651    436.2121     100.546    477.8105
---------+--------------------------------------------------------------------
    diff |      23   -181.6174    88.51358    424.4962   -365.1833    1.948542
```

```
------------------------------------------------------------------------------
     mean(diff) = mean(guanidoacetate - guandidoacetat~t)         t =  -2.0519
 Ho: mean(diff) = 0                                degrees of freedom =       22

 Ha: mean(diff) < 0           Ha: mean(diff) != 0           Ha: mean(diff) > 0
 Pr(T < t) = 0.0261         Pr(|T| > |t|) = 0.0523          Pr(T > t) = 0.9739

. ttest hydroxybutyrate= hydroxybutyrate_post

Paired t test
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
hydrox~e |      23    19.66087    7.455353    35.75462    4.199415    35.12233
hydrox~t |      23    3.747826    2.238558    10.73575   -.8946587    8.390311
---------+--------------------------------------------------------------------
    diff |      23    15.91304    7.952478    38.13875   -.5793869    32.40547
------------------------------------------------------------------------------
     mean(diff) = mean(hydroxybutyrate - hydroxybutyrat~t)        t =   2.0010
 Ho: mean(diff) = 0                                degrees of freedom =       22

 Ha: mean(diff) < 0           Ha: mean(diff) != 0           Ha: mean(diff) > 0
 Pr(T < t) = 0.9711         Pr(|T| > |t|) = 0.0579          Pr(T > t) = 0.0289

. ttest acetone= acetone_post

Paired t test
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
 acetone |      23        14.1    3.251755    15.59487    7.356272    20.84373
aceton~t |      23    53.26957    26.61296    127.6313    -1.92234    108.4615
---------+--------------------------------------------------------------------
    diff |      23   -39.16957    25.52557    122.4164   -92.10637    13.76724
------------------------------------------------------------------------------
     mean(diff) = mean(acetone - acetone_post)                    t =  -1.5345
 Ho: mean(diff) = 0                                degrees of freedom =       22

 Ha: mean(diff) < 0           Ha: mean(diff) != 0           Ha: mean(diff) > 0
 Pr(T < t) = 0.0696         Pr(|T| > |t|) = 0.1392          Pr(T > t) = 0.9304
```

7.0 General Discussion and Conclusion

The role of urine metabolomics in the detection of colorectal cancer and polyps was explored in this thesis project.

Chapter one highlighted the public health concerns of colorectal cancer screening in Canada by summarizing the current screening modalities, effectiveness of screening, current compliance rates for CRC screening and barriers to effective screening. Some suggestions on overcoming the barriers were given and Alberta's approach to CRC screening was elaborated.

Chapter two provided an up-to-date systematic review of the existing literature on the field of metabolomics and CRC in humans. Eight studies were included and the distinguishing metabolites from each study were summarized. Tissue and serum metabolomics were discussed in addition to urine metabolomics.

In chapter three, the methodology of recruitment, specimen processing and data analysis were provided in detail. As well, the technical concepts of NMR were discussed.

Chapter four demonstrated that using urine metabolomics and advanced statistical analysis, a robust OPLS model could be built to distinguish colonoscopy-negative controls from CRC patients with high sensitivity (92.7%). Using only the top ten

metabolites, sensitivity and specificity of 86.6% and 75.5% respectively were achieved, suggesting commercialization potential for this test.

Chapter five demonstrated that the urine metabolomic test diagnostics for precancerous adenomatous polyps (sensitivity of 89.5% and specificity of 71.8%) were far superior to the existing fecal tests. While the adenomatous polyp stage is believed to be an intermediate step between normal colonic epithelium and colorectal malignancy, the metabolomic fingerprint for colorectal adenomatous polyps was found to be completely different than that of CRC.

In chapter six, advanced statistical models could not show any differences in the metabolomic fingerprint of the pre and post-treatment CRC groups, however, when analyzed individually, 6 out of 23 CRC patients showed a recovering tendency towards normal and 2 of the top 10 CRC-predictive metabolites trended towards normal.

7.1 Future Directions

This study is a work in progress. Work is ongoing to collect urine samples from new CRC patients and post-treatment CRC patients to improve the normal vs. CRC model and further investigate the changes to the metabolomic fingerprint of CRC after curative treatment.

The existing data is also being re-analyzed with the samples randomly distributed first prior to being assigned to training and testing groups for both the CRC group and the adenoma group. This method will minimize any clinical differences between the groups and make the training and testing groups more comparable.

We plan to contact the adenoma patients from this study to obtain another urine sample in order to study whether the adenoma fingerprint changes after the polyps are removed. This has implications should the urine metabolomics test proves useful as a commercial screening test. That is, if the urine metabolomic signature remains after the polyps are removed, then this would represent an once-in-a-lifetime test rather than an annual test.

In this thesis, the hyperplastic polyps were visually shown to resemble adenomas rather than normals. This will be further investigated by building normal vs. hyperplastic models to see if the metabolite fingerprint of hyperplastic polyps also resembles that of adenomatous polyps. The hyperplastic polyps will also be validated against the adenoma model to establish whether the urine metabolomic test predicts growths or more specifically adenomatous growth.

Finally the CRC urine metabolomic fingerprint will be validated with the urine samples of patients with other types of common adenocarcinomas such as breast and prostate to determine whether it is specific enough for CRC or if it is a general adenocarcinoma fingerprint .