# USE OF FAMILY HISTORY TO IMPROVE RISK PREDICTION IN CLINICAL CARE: COLORECTAL CANCER AS AN EXAMPLE

by

David P. Taylor

A dissertation submitted to the faculty of
The University of Utah
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Biomedical Informatics

The University of Utah

May 2011

**STATEMENT OF DISSERTATION APPROVAL**

The dissertation of **David P. Taylor**

has been approved by the following supervisory committee members:

| | | |
|---|---|---|
| **Peter J. Haug** | , Chair | **12/3/2010** Date Approved |
| **Lisa A. Cannon-Albright** | , Member | **12/3/2010** Date Approved |
| **Marc S. Williams** | , Member | **12/3/2010** Date Approved |
| **Randall W. Burt** | , Member | **12/3/2010** Date Approved |
| **Joyce A. Mitchell** | , Member | **12/3/2010** Date Approved |

and by **Joyce A. Mitchell** , Chair of

the Department of **Biomedical Informatics**

and by Charles A. Wight, Dean of The Graduate School.

## ABSTRACT

Family history has been called the "cornerstone of individualized disease prevention" but it is underutilized in clinical practice. In order to use it more effectively, its role in assessing risk for disease needs to be better quantified and understood. Family history has been identified as an important risk factor for colorectal cancer (CRC) and risk prediction in CRC is potentially worthwhile because of the possibility of preventing the disease through application of individualized screening programs tailored to risk. The overall project objective was to explore how family history can be better utilized to predict who will develop CRC. First, we used the Utah Population Database (UPDB) to define familial risk for CRC in more detail than has previously been reported. Second, we explored whether individuals at increased familial risk for CRC or at increased risk based on other risk factors such as a personal history of CRC or adenomatous polyps, are more compliant with screening and surveillance recommendations using colonoscopy than those who are at normal risk. Third, we measured how well family history can predict who will develop CRC over a period of 20 years, using family history by itself as a risk factor, and also in combination with the risk factor, age. We found that increased numbers of affected first-degree relatives influence risk much more than affected relatives from the second or third degrees. However, when combined with a positive first-degree family history, a positive second- and third-degree family history can

significantly increase risk. Next, we found that colonoscopy rates were higher in those with risk factors, according to risk-specific guidelines, but improvements in compliance are still warranted. Lastly, it was determined that family history by itself is not a strong predictor of exactly who will acquire colorectal cancer within 20 years. However, stratification of risk using absolute risk probabilities may be more helpful in focusing screening on individuals who are more likely to develop the disease. Future work includes using these findings as a basis for a cost/benefit analysis to determine optimal screening recommendations and building tools to better capture and utilize family history data in an electronic health record system.

To Anissa, Benjamin, and Leah.

## TABLE OF CONTENTS

## LIST OF TABLES

## ABBREVIATIONS

ACS–MSTF–ACR - American College of Gastroenterology, the American Cancer Society, the Multi-Society Task Force on CRC, and the American College of Radiology

BRFSS - Behavioral Risk Factor Surveillance System

CRC – colorectal cancer

EMR – electronic medical record

FAP – familial adenomatous polyposis

FDR – first-degree relative

FOBT – fecal occult blood test

FRR – familial relative risk

FURTHeR - Federated Utah Research and Translational Health e-Repository

HIPAA - Health Insurance Portability and Accountability Act

HNPCC – hereditary nonpolyposis colorectal cancer

IBD – inflammatory bowel disease

ICD-9 – International Classification of Disease, version 9

IRB – institutional review board

NCCN – National Comprehensive Cancer Network

NHIS – National Health Information Survey

RGE – Resource for Genetic Epidemiology

ROC – Receiver Operator Characteristic

RR – relative risk

SDR – second-degree relative

SEER – Surveillance, Epidemiology, and End Results program

TDR – third-degree relative

UCR – Utah Cancer Registry

UPDB – Utah Population Database

USPSTF – United States Preventive Services Task Force

## ACKNOWLEDGMENTS

I would especially like to thank Peter Haug and Lisa Cannon-Albright for their guidance and support in so many ways, which was critical to the success of this project. Marc Williams, Randy Burt, and Joyce Mitchell also provided key contributions and I am grateful for their assistance. Carol Sweeney and Greg Stoddard provided much need epidemiological and statistical guidance. Jim Farnham and Kristina Allen-Brady also gave informal statistical advice that was invaluable. I would also like to acknowledge Steve Backus' programmatic and database assistance, which provided the foundation for these analyses. Lastly, support for this work was provided by a fellowship through the Homer Warner Center for Informatics Research at Intermountain Healthcare.

# CHAPTER 1

## INTRODUCTION

Risk prediction is an area of growing interest and attention in medicine and biomedical informatics, particularly through the development and application of statistical and machine learning models. In 1976, the first risk prediction model for a chronic disease was published: the Framingham Coronary Risk Prediction Model.[1] This model predicted an individual's risk for developing heart disease based on factors such as serum cholesterol level, blood pressure, smoking history, electrocardiogram results, and glucose intolerance. Work on the original model has been modified and expanded over the years to include a number of risk score profile tools that are used by physicians to make decisions about prevention and treatment.[2-7]

Predicting who is at risk for cancer has been a particular focus of effort, based on the prevalence and burden of this group of diseases. Risk prediction models for breast cancer were developed in the late 1980s and early 1990s. These models used risk factors such as age, age at menarche, and family history of breast cancer to predict the absolute risk that an individual would develop the disease over a particular period of time.[8, 9] Since then, breast cancer models have been updated and refined with new components addressing genetic susceptibility for the disease.[10-

[13] Risk models for other cancers such as colorectal, lung, prostate, ovarian, and melanoma have been developed as well.[14-18] The major applications of models for predicting risk and susceptibility in cancer are: (1) Identifying individuals at high risk (for appropriate screening), (2) Improving clinical decision making (e.g., genetic counseling), (3) Designing population prevention strategies, (4) Estimating the cost of the population burden of disease, (5) Enabling the creation of benefit-risk indices, and (6) Planning intervention trials (e.g., chemoprevention).[19]

The development and application of risk models are enabled by information technology and would be more difficult to implement without the assistance of the modern computer. The field of biomedical informatics may be considered an important contributor to risk modeling through the integration of medicine, computer science, information science and data management, and statistical analysis. However, a particularly topical area where biomedical informatics plays a key role is the application of risk prediction models for use in decision support in electronic medical record (EMR) systems. One may envision an EMR system with comprehensive, integrated decision support that uses data from a patient's record and other sources to assess the risk for developing a particular disease and provides the clinician and patient with actionable recommendations for mitigating the potential risk.

This application of a risk model may also be considered within the domain of personalized medicine. This is a medical model of growing interest that leverages information about an individual patient, particularly genomic and proteomic, in order to provide preventive and therapeutic options that are optimized to that

individual patient.[20] This movement is being facilitated by advances in genetic testing and molecular profiling technologies and aims to address the deficiencies of therapies that are primarily based on epidemiological studies of large cohorts. Currently, known factors such as family history, environment, social circumstances, and behavior are considered by clinicians in creating personalized treatment plans. While these risk factors will continue to be useful in the future, it is likely that prediction will further be enhanced by molecular profiling, providing the capability to address more individual differences and tailor specific and efficacious prevention and treatment.

## 1.1 Family history

Even as specific risk factors for disease are discovered through molecular profiling or other research methods, there is clearly benefit to be gained from the exploration of risk factors such as family history. When more sophisticated molecular profiling techniques are ready for clinical implementation, experts still believe family history will be play an important role in stratifying a patient's risk.[21] It has been called the "cornerstone of individualized disease prevention" and is a "free, well-proven, personalized genomic tool" that captures many of the interactions between multiple genes and environmental factors.[21]

Despite the promise of the role of family history, there is evidence that it is currently underutilized in clinical practice. There are several barriers present. First, patients are sometimes unaware of family members' health histories, even in first-degree relatives. Even if patients do have the information, patients and

clinicians may fail to recognize the potential value as it can be difficult to quantify and understand risk. Even those clinicians who understand the importance of family history in disease, are under extraordinary time pressure during an office visit that impacts their ability to collect, document and discuss the information and risk with the patient. Although family history may get documented in the patient's record, and sometimes electronically in an EMR, there is a lack of tools to help analyze and interpret the resulting risk. Improving or finding new methods to capture and utilize family history information is a worthwhile challenge with a clear role for informatics. In order to more effectively use family history in clinical practice, its role in assessing risk for disease needs to be better quantified and understood, particularly in the context of other risk factors.

Studying the role of family history in disease can be challenging because of the reasons mentioned above, most notably because it is not captured with consistency in clinical practice, let alone in a standardized way. Epidemiological studies have captured family history information at times, most often for first-degree relatives. Data are almost always self-reported so recall bias and inaccuracy are potential issues. In the state of Utah there is an unusual opportunity to investigate the influence of family history in disease, and particularly cancer, through the Utah Population Database (UPDB). The UPDB is a population-based genealogical resource that includes statewide cancer registry records from the Utah Cancer Registry (UCR), and other records such as birth and death certificates, driver's license data, and inpatient and outpatient medical and billing records from the University of Utah Health Sciences Center.[22] It was created in the early 1970s

based on genealogical records from the Utah Family History Library and contains genealogies for the original Utah pioneers (members of the Church of Jesus Christ of Latter-day Saints) and their modern-day descendants.[23, 24] This resource will be described in more detail in subsequent chapters, but it is worthwhile to note that this resource does much to address the limitations previously mentioned concerning family history availability and quality. The original Utah Genealogy included records of 1.6 million persons who were part of 6- to 7-generation pedigrees.[22] Although not all have linked genealogic data, the UPDB includes information for approximately 7 million persons today with some pedigrees >11 generations deep. The link between the UCR and the Utah genealogy is a major strength of the UPDB. The UCR was established in 1966 and includes records dating back to 1952 and since 1973 has been part of the Surveillance, Epidemiology, and End Results network of National Cancer Institute registries. Among those with cancer records in the UCR, 94% link to ≥1 records in the UPDB and 64.2% have family information. Cancer records are coded by disease site according to the International Classification of Diseases of Oncology and include information on site, stage, grade, age at diagnosis, histology, and patient survival.[25] In cases in which a person has multiple cancers the UCR is careful to report only independent primary sites of cancer.

While the UPDB is an extraordinary resource on its own, in recent years a link has been created between the UPDB and electronic records from Intermountain Healthcare. Intermountain Healthcare provides medical care for a large percentage of the Utah population and also has a long history of electronic documentation of clinical care and a rich enterprise data warehouse. Among available data are coded

data for visits, hospital stays, diagnoses, procedures, laboratory orders and results, medications, problems, and clinical findings, as well as various unstructured narrative notes and reports. Many of the individuals with genealogy data in the UPDB also have health care records at Intermountain. This has provided an opportunity to link high-quality family history data from the UPDB with clinical and billing data from Intermountain, enabling research previously not possible.

The UPDB has strict rules concerning appropriate uses of the data, and they cannot be used for clinical care purposes. For example, even if a group of individuals was identified at high risk for a particular disease based on family history in the UPDB, they could not be contacted for the purpose of clinical care or intervention due to the restrictions imposed by the legislation that created the resource. However, despite this limitation, the results of research based on UPDB data may help to confirm and/or quantify the value of utilizing self-reported family history data in clinical practice. It also may help to set a standard against which to benchmark other approaches for collecting family history.

## 1.2 Colorectal cancer

CRC is the second leading cause of death among cancers. In 2009, it was estimated that 147,000 cases would be newly diagnosed and that 50,000 deaths would be caused by the disease.[26] The lifetime risk of acquiring CRC is 5.5% for men and 5.1% for women. Although rates have slowly been declining since 1985, the age-adjusted incidences per 100,000 in the population are 61.2 and 44.8, for men and women, respectively.[26] Family history is a known risk factor for several cancers, but

colorectal cancer (CRC) is a particular area of interest in risk prediction because screening through a method like colonoscopy can detect and remove precancerous polyps before they can progress into cancer.[27, 28] Despite the declining incidence, screening rates are still not at recommended levels. It has been estimated that over half of deaths from CRC could be prevented through early detection.[29] Increased surveillance in those at increased risk may lead to the detection of more cases, and therefore a potentially greater mortality reduction, than uniform surveillance of the entire population.[30]

Family history has been well established as a risk factor for CRC.[31-34] As mentioned previously, risk models have been developed for CRC, although with a few exceptions,[14, 35, 36] they most often address highly-genetic forms of the disease such as Lynch syndrome (LS, also known as hereditary nonpolyposis colorectal cancer (HNPCC)) or familial adenomatous polyposis (FAP), which represent a relatively small percentage of CRC cases.[37-39] Although it is clear that rare forms of the disease such as LS and FAP are highly genetic, it has been established that genetics still play a role in a fair number of other cases of the disease. However, the majority of CRC cases are still considered sporadic without a known genetic contribution.

CRC is an example of a disease where risk modeling is potentially very useful because screening and treatment for the disease can both be financially costly. Screening through a technique such as colonoscopy, while effective, also has medical risks. A comprehensive decision model that takes into account the risk of acquiring the disease based on family history, age, and other factors; colonoscopy risk; and

costs of screening and treatment would be particularly useful in determining the "best care at the best cost" based on an individual's unique set of risk factors.

## 1.3 Study objectives

The overall goal of the project was to explore how family history can be better utilized to improve clinical care; specifically better predicting who will develop CRC. The strength of family history and cancer data resources available through the UPDB, electronic medical records from institutions such as Intermountain Healthcare, and local interest and expertise in clinical genetics, genetic epidemiology, and clinical decision support, created an ideal environment for exploring this topic. The project was divided into three phases that addressed different aspects of predicting risk for CRC. The objective of the first phase was to use the UPDB to define familial risk for CRC in more depth than has previously been reported. Typically in studies of family history and disease only first-degree relatives (e.g., parents, siblings, children) are considered. A principal reason may be that studies often use self-reported family history data and it may be difficult for probands to obtain reliable data on extended relatives from the second (e.g., aunts and uncles, grandparents) or third degrees (e.g., great-grandparents, cousins). Our hypothesis was that the influence of affected relatives from the second and third degrees may have a significant impact on relative risk for CRC. Another objective of this phase was to produce familial relative risk estimates for a large number of different "constellations" or combinations of affected first-, second-, and third-degree relatives to allow more individualized risk estimation based on an *exact* family

history. The sheer size of the UPDB, with respect to numbers of individuals with high-quality family history and with and without CRC, allowed investigation of a large number of different constellations of affected relatives. This first phase of research provided a foundation for subsequent phases that included subsets of the same UPDB population and also capitalized on the familial relative risk estimates for various constellations.

The second phase of the project used the link between the UPDB and Intermountain Healthcare medical data. This phase involved an exploration of whether individuals at increased familial risk for CRC or at increased risk based on other risk factors such as a personal history of CRC or adenomatous polyps are more compliant with screening and surveillance recommendations using the colonoscopy modality than those who are at normal risk. A group of living individuals who had both UPDB records and Intermountain data, and who had been seen as an inpatient or outpatient within the last 5 years at Intermountain Healthcare, were included in this study. Intermountain records provided coded and/or structured data on risk factors such as a previous history of CRC or adenoma and dates of procedures such as colonoscopy. Well-known screening and surveillance guidelines for those at increased risk based on these risk factors were adapted in order to measure compliance. The hypothesis was that individuals with a positive family history or other risk factors would be screened more frequently than those at normal risk, according to accepted guidelines.

The objective of the third and final phase of the project was to determine the clinical implications of risk prediction for CRC based on family history. Specifically,

the objective was to measure how well family history can predict who will develop CRC over a period of 20 years, using family history by itself as a risk factor, and also in combination with the risk factor of age. Although family history has been well-established as a risk factor for CRC, relative risk does not reflect that the incidence of CRC may be considered low overall in the population. Considering this fact, the absolute risk for the disease may or may not be strongly influenced by family history. However, the hypothesis was that family history does have predictive value that is clinically worthwhile, particularly for those at the highest levels of familial risk.

## 1.4 Summary

Risk prediction is an area of growing interest in biomedical informatics and medicine and the increasing availability of data has fueled this interest. Among diseases where risk prediction has been applied, cancer has been a particular area of focus considering its scope and impact. Among cancers, CRC is particularly interesting to researchers in risk prediction because of the possibility of preventing the disease through early screening. Family history has been identified as an important risk factor for CRC and plays a potentially important role in risk prediction. Even when more specific molecular risk factors are discovered for CRC and other cancers, it is likely that family history will still be worthwhile to consider. However, family history is currently underutilized in clinical practice for a variety of reasons including lack of patient awareness of family member medical histories, lack of recognition of value by patients and clinicians, not enough time for clinicians to

collect and make use of the data in clinical practice, and lack of tools such as risk models and clinical decision support to help estimate and interpret risk and provide recommendations. Although family history is considered one of the most important risk factors for CRC, it has not been clear whether (1) a very large study using population-based family history and cancer registry records would find differences in relative risk from published estimates, (2) extended relatives affected with CRC make a significant contribution to risk, (3) those with a positive family history and other clinical risk factors are more likely to have been screened according to common guidelines, and (4) family history is able to predict accurately who will get CRC over a 20-year period, particularly in certain risk categories. The following chapters present three papers that address the objectives and research questions that have been outlined above.

## 1.5 References

1. Kannel WB, McGee D, Gordon T. A general cardiovascular risk profile: the Framingham Study. Am J Cardiol 1976;38:46-51.

2. Wilson PW, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB. Prediction of coronary heart disease using risk factor categories. Circulation 1998;97:1837-47.

3. D'Agostino RB, Wolf PA, Belanger AJ, Kannel WB. Stroke risk profile: adjustment for antihypertensive medication. The Framingham Study. Stroke 1994;25:40-3.

4. Kannel WB, D'Agostino RB, Silbershatz H, Belanger AJ, Wilson PW, Levy D. Profile for estimating risk of heart failure. Arch Intern Med 1999;159:1197-204.

5. Wang TJ, Massaro JM, Levy D, et al. A risk score for predicting stroke or death in individuals with new-onset atrial fibrillation in the community: the Framingham Heart Study. Jama 2003;290:1049-56.

6. D'Agostino RB, Sr., Vasan RS, Pencina MJ, et al. General cardiovascular risk profile for use in primary care: the Framingham Heart Study. Circulation 2008;117:743-53.

7. Schnabel RB, Sullivan LM, Levy D, et al. Development of a risk score for atrial fibrillation (Framingham Heart Study): a community-based cohort study. Lancet 2009;373.

8. Gail MH, Brinton LA, Byar DP, et al. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. J Natl Cancer Inst 1989;81:1879-86.

9. Rosner B, Colditz GA. Nurses' health study: log-incidence mathematical model of breast cancer incidence. J Natl Cancer Inst 1996;88:359-64.

10. Shattuck-Eidens D, Oliphant A, McClure M, et al. BRCA1 sequence analysis in women at high risk for susceptibility mutations. Risk factor analysis and implications for genetic testing. Jama 1997;278:1242-50.

11. Parmigiani G, Berry D, Aguilar O. Determining carrier probabilities for breast cancer-susceptibility genes BRCA1 and BRCA2. Am J Hum Genet 1998;62:145-58.

12. Antoniou AC, Pharoah PD, McMullan G, et al. A comprehensive model for familial breast cancer incorporating BRCA1, BRCA2 and other genes. Br J Cancer 2002;86:76-83.

13. Antoniou AC, Pharoah PP, Smith P, Easton DF. The BOADICEA model of genetic susceptibility to breast and ovarian cancer. Br J Cancer 2004;91:1580-90.

14. Freedman AN, Slattery ML, Ballard-Barbash R, et al. Colorectal cancer risk prediction tool for white men and women without known susceptibility. J Clin Oncol 2009;27:686-93.

15. Spitz MR, Hong WK, Amos CI, et al. A risk model for prediction of lung cancer. J Natl Cancer Inst 2007;99:715-26.

16. Optenberg SA, Clark JY, Brawer MK, Thompson IM, Stein CR, Friedrichs P. Development of a decision-making tool to predict risk of prostate cancer: the Cancer of the Prostate Risk Index (CAPRI) test. Urology 1997;50:665-72.

17. Rosner BA, Colditz GA, Webb PM, Hankinson SE. Mathematical models of ovarian cancer incidence. Epidemiology 2005;16:508-15.

18. Fears TR, Guerry Dt, Pfeiffer RM, et al. Identifying individuals at high risk of melanoma: a practical predictor of absolute risk. J Clin Oncol 2006;24:3590-6.

19. Freedman AN, Seminara D, Gail MH, et al. Cancer risk prediction models: a workshop on development, evaluation, and application. J Natl Cancer Inst 2005;97:715-23.

20. Hamburg MA, Collins FS. The path to personalized medicine. N Engl J Med;363.

21. Guttmacher AE, Collins FS, Carmona RH. The family history--more important than ever. N Engl J Med 2004;351:2333-6.

22. Cannon Albright LA. Utah family-based analysis: past, present and future. Hum Hered 2008;65:209-20.

23. Skolnick MH. The Utah genealogical data base: a resource for genetic epidemiology. In: Cairns J LJ, Skolnick MH, eds. Banbury Report No 4; Cancer Incidence in defined populations. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory; 1980:285-97.

24. Skolnick MH. Prospects for population oncogenetics. In: Mulvihill JJ, Miller RW, Fraumeni JF, eds. Genetics of human cancer. New York: Raven Press; 1977:19-25.

25. World Health Organization. International classification of diseases for oncology. 3rd ed. Geneva, Switzerland: World Health Organization; 2000.

26. Jemal A, Siegel R, Ward E, Hao Y, Xu J, Thun MJ. Cancer statistics, 2009. CA Cancer J Clin 2009;59:225-49.

27. Levin B, Lieberman DA, McFarland B, et al. 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;134:1570-95.

28. Whitlock EP, Lin JS, Liles E, Beil TL, Fu R. Screening for colorectal cancer: a targeted, updated systematic review for the U.S. Preventive Services Task Force. Ann Intern Med 2008;149:638-58.

29. Colditz GA, Atwood KA, Emmons K, et al. Harvard report on cancer prevention volume 4: Harvard Cancer Risk Index. Risk Index Working Group, Harvard Center for Cancer Prevention. Cancer Causes Control 2000;11.

30. Hunt LM, Rooney PS, Hardcastle JD, Armitage NC. Endoscopic screening of relatives of patients with colorectal cancer. Gut 1998;42.

31. Johns LE, Houlston RS. A systematic review and meta-analysis of familial colorectal cancer risk. Am J Gastroenterol 2001;96.

32. Goldgar DE, Easton DF, Cannon-Albright LA, Skolnick MH. Systematic population-based assessment of cancer risk in first-degree relatives of cancer probands. J Natl Cancer Inst 1994;86.

33. Hemminki K, Li X. Familial colorectal adenocarcinoma from the Swedish Family-Cancer Database. Int J Cancer 2001;94.

34. Butterworth AS, Higgins JPT, Pharoah P. Relative and absolute risk of colorectal cancer for individuals with a family history: a meta-analysis. Eur J Cancer 2006;42.

35. Driver JA, Gaziano JM, Gelber RP, Lee IM, Buring JE, Kurth T. Development of a risk score for colorectal cancer in men. Am J Med 2007;120:257-63.

36. Selvachandran SN, Hodder RJ, Ballal MS, Jones P, Cade D. Prediction of colorectal cancer by a patient consultation questionnaire and scoring system: a prospective study. Lancet 2002;360:278-83.

37. Balmana J, Stockwell DH, Steyerberg EW, et al. Prediction of MLH1 and MSH2 mutations in Lynch syndrome. Jama 2006;296:1469-78.

38. Chen S, Wang W, Lee S, et al. Prediction of germline mutations and cancer risk in the Lynch syndrome. Jama 2006;296:1479-87.

39. Wijnen JT, Vasen HF, Khan PM, et al. Clinical findings with implications for genetic testing in families with clustering of colorectal cancer. N Engl J Med 1998;339:511-8.

## CHAPTER 2

# POPULATION-BASED FAMILY HISTORY-SPECIFIC RISKS FOR COLORECTAL CANCER: A CONSTELLATION APPROACH

David P Taylor, Randall W Burt, Marc S Williams, Peter J Haug, and Lisa A Cannon-Albright

This article was published in *Gastroenterology*; Volume 138; Authors: Taylor DP, Burt RW, Williams MS, Haug PJ, Cannon-Albright LA; Pages 877-885; 

GASTROENTEROLOGY 2010;138:877–885

# Population-Based Family History–Specific Risks for Colorectal Cancer: A Constellation Approach

DAVID P. TAYLOR,[*] RANDALL W. BURT,[‡,¶] MARC S. WILLIAMS,[*,§] PETER J. HAUG,[*,‖] and LISA A. CANNON–ALBRIGHT[¶]

*[*]Department of Biomedical Informatics, [‡]Huntsman Cancer Institute, University of Utah, Salt Lake City, Utah; [§]Clinical Genetics Institute, [‖]Homer Warner Center for Informatics Research, Intermountain Healthcare; and [¶]Department of Internal Medicine, University of Utah, Salt Lake City, Utah*



**BACKGROUND & AIMS:** Colorectal cancer (CRC) risk estimates based on family history typically include only close relatives. We report familial relative risk (FRR) in probands with various combinations, or constellations, of affected relatives, extending to third-degree. **METHODS:** A population-based resource that includes a computerized genealogy linked to statewide cancer records was used to identify genetic relationships among CRC cases and their first-, second-, and third-degree relatives (FDRs, SDRs, and TDRs). FRRs were estimated by comparing the observed number of affected persons with a particular family history constellation to the expected number, based on cohort-specific CRC rates. **RESULTS:** A total of 2,327,327 persons included in ≥3 generation family histories were analyzed; 10,556 had a diagnosis of CRC. The FRR for CRC in persons with ≥1 affected FDR = 2.05 (95% CI, 1.96–2.14), consistent with published estimates. In the absence of a positive first-degree family history, considering both affected SDRs and TDRs, only 1 constellation had an FRR estimate that was significantly >1.0 (0 affected FDRs, 1 affected SDR, 2 affected TDRs; FRR = 1.33; 95% CI, 1.13–1.55). The FRR for persons with 1 affected FDR, 1 affected SDR, and 0 affected TDRs was 1.88 (95% CI, 1.59–2.20), increasing to FRR = 3.28 (95% CI, 2.44–4.31) for probands with 1 affected FDR, 1 affected SDR, and ≥3 affected TDRs. **CONCLUSIONS: Increased numbers of affected FDRs influences risk much more than affected SDRs or TDRs. However, when combined with a positive first-degree family history, a positive second- and third-degree family history can significantly increase risk.**

*Keywords:* Colorectal; Relative Risk; UPDB.

Colorectal cancer (CRC) is the third most common cancer diagnosed in the United States and the second leading cause of death among cancers. It is estimated that in 2008 alone 148,810 people will be diagnosed with CRC and 49,960 will die of the disease.[1] Screening strategies such as fecal occult blood test, flexible sigmoidoscopy, and colonoscopy, among others, have been proven effective in reducing the incidence and mortality of the disease.[2] Although CRC incidence and mortality rates have been declining, most US adults are still not being screened or receiving regular screenings appropriate for their age or risk status.[3,4] Knowledge of increased risk can be a motivating factor in making decisions about screening.[5] If a greater proportion of adults received regular screenings appropriate for their risk, it is likely that additional reductions in CRC incidence and mortality could be achieved.

Family history is a well-established risk factor for CRC.[6] Many studies have estimated familial risk of CRC and consistently shown that having ≥1 affected first-degree relative (FDR) doubles a person's risk of CRC.[6–10] A recent random-effects analysis that pooled relative risk estimates for CRC from multiple published reports was presented by Butterworth et al in 2006.[11] Although this meta-analysis comprehensively included the relevant published research to date, limitations in the source studies bring to light additional questions. Most studies focused on categorizing ≥1 affected FDR. Additional risk factors in other studies included affected second-degree relatives (SDRs), age of onset in affected FDRs, sex, and relationship type (ie, parent or child, sister or brother). Although data on FDRs are the easiest to obtain from patients and may be the most clinically relevant, it is currently not known what impact more distant affected relatives have on risk. Just as important, the risk stemming from various combinations, or "constellations," of affected relatives has not been adequately explored. Patients frequently present to physicians reporting multiple relatives affected with colorectal cancer. These often include affected relatives from first-, second-, and third-degree relationships and might include positive family history from both parents (Figure 1 contains a pedigree diagram to show family relationships and degrees). Physicians presently have little if any data on how to estimate

*Abbreviations used in this paper:* CRC, colorectal cancer; E, expected; FDR, first-degree relative; FRR, familial relative risk; O, observed; SDR, second-degree relative; TDR, third-degree relative; UCR, Utah Cancer Registry; UPDB, Utah Population Database.



0016-5085/10/$36.00
doi:10.1053/j.gastro.2009.11.044

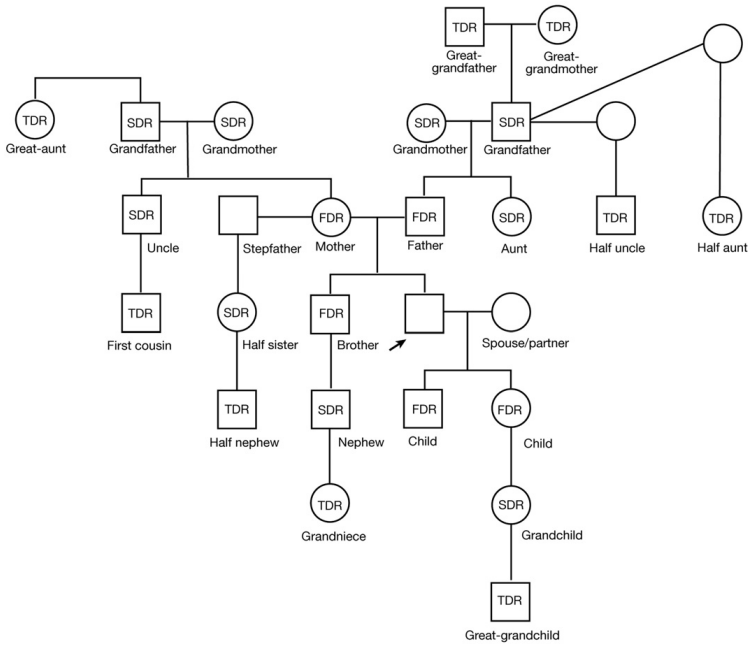


**Figure 1.** Sample pedigree structure illustrating some example relationships including first-, second-, and third-degree relatives (FDR, SDR, TDR). The proband is indicated with an arrow.

risk for such patients and thereby to determine appropriate screening.

The objective of our study was to expand the scope of previous CRC familial risk research to measure and report familial relative risk (FRR) estimates for a variety of specific constellations of family history. This investigation allows risk levels to be assigned for most combinations of affected relatives, thus assisting the physician in making more appropriate screening recommendations. To accomplish this we examined first-, second-, and third-degree risks in a population-based resource with a computerized genealogy linked to statewide cancer registry records. This resource, the Utah Population Database (UPDB), provides an unusual opportunity to investigate the relative risk of CRC in relatives in a large population at a more detailed level than has been previously published.

## Data

The UPDB is a population-based, computerized genealogic resource for Utah containing multiple record-linked data sources, including cancer registry records, birth and death certificates, and driver's license data, that has also been linked to inpatient and outpatient records from the University of Utah Health Sciences Center. The UPDB contains genealogies for the original Utah pioneers (members of The Church of Jesus Christ of Latter-day Saints, or Mormons) and their modern-day descendants and was created in the early 1970s with data from the Utah Family History Library.[12,13] The original Utah Genealogy included records for 1.6 million persons who were part of 6- to 7-generation pedigrees.[14] Today, the UPDB includes information for approximately 7 million persons, although not all have linked genealogic data. Some pedigrees are now >11 generations deep.

Of particular interest for this study were Utah Cancer Registry (UCR) records linked to the Utah genealogy. The UCR is a statewide cancer registry established in 1966 that includes records dating back to 1952. Since 1973 the UCR has been part of the Surveillance, Epidemiology, and End Results network of National Cancer Institute registries. Ninety-four percent of persons with cancer link to ≥1 records in the UPDB and 64.2% have family information. The UCR cancer records are coded by disease site according to the International Classification of Diseases of Oncology and include information on site, stage, grade, age at diagnosis, histology, and patient survival.[15] The UCR is careful to report only independent primary

sites of cancer in cases in which a person has multiple cancers.

Previous demographic and genetic analyses have shown that the population recorded in the UPDB is genetically representative of US white and northern European populations[16–19] with a low-to-normal level of inbreeding.[20] Most Utahns are members of The Church of Jesus Christ of Latter-day Saints, which has religious proscriptions against the use of coffee, tea, alcohol, and tobacco. Utah is among the states with the lowest rates of cancer,[1] and much lower smoking rates may play a role.[21]

Although the UPDB contains records for approximately 7 million people, this study used a subset of 2.3 million persons who were part of ≥3 generations of Utah genealogy data and descendants of original Utah pioneers.

Access to the UPDB is governed by the Utah Resource for Genetic Epidemiology which was created in 1982.[22] The Utah Resource for Genetic Epidemiology and University of Utah Institutional Review Board approvals were obtained to access these data and to conduct this research. Names and other identifying information were not available to the authors to protect the privacy of persons in the UPDB.

## Materials and Methods

FRR in relatives represents the ratio of the risk for a disease among relatives of probands to the risk for the disease in the general population. It is estimated as the number of observed (O) cases among relatives of probands divided by the number of expected (E) cases among the relatives (ie, FRR = O/E). The expected number of cases among relatives is estimated by using population rates. This ratio, also known as a "standardized morbidity ratio," is considered a reasonable approximation of true relative risk when the prevalence of the disease and the true relative risk in the population are low.[23] This method of estimating FRR in relatives has been used in previous UPDB analyses of familial risk in cancer[24,25] and other diseases.[26,27]

All persons in the UPDB who are part of ≥3 generations of Utah genealogy data (2.3 million) were assigned to 1 of 264 cohorts based on characteristics that may influence the quality and quantity of genealogic data: birth year (5-year groups), sex, amount of ancestral genealogy (>6 ancestors or not), and birthplace (Utah or not Utah). Internal, cohort-specific CRC rates were calculated by summing the number of CRC cases in each cohort and dividing by the total number of UPDB persons in that cohort. In this study a proband was defined as a person who has a particular constellation pattern of affected relatives (eg, 1 affected FDR, 0 affected SDRs, and ≥3 third-degree relatives [TDRs]), whether the proband is a CRC case himself or herself. All persons among the 2.3 million with a particular constellation pattern of affected relatives were considered probands for the corresponding FRR calculation.

After selecting a constellation pattern and determining the group of probands who fit the pattern, the number of observed CRC cases (O) in the group of probands is counted by cohort, without duplication. The expected number of cancers (E) among the defined set of probands is estimated by using the following formula: $E = \Sigma\ P_i \times C_i/N_i$ (for $i$ between 1 and 264), where $P_i$, $C_i$, and $N_i$ are the number probands, the number of CRC cases in the UPDB, and the number of persons in the UPDB, respectively, in the $i$th cohort group. This method assumes that the morbidity and migration rates for a given cohort of probands is on average the same as that for an equivalent cohort of persons in the UPDB.

For FRR = O/E, $P$ values were calculated, based on the null hypothesis FRR = 1.0 and the alternative hypothesis FRR >1.0. An assumption was made that the number of observed cases followed a Poisson distribution with the mean equal to the expected number of cases. Confidence intervals for the FRRs were estimated by the method given by Agresti.[28]

The selection of constellation patterns for which to calculate FRRs was based on common analyses in previous studies[11] and consensus of the authors. The first group of analyses performed in this study was based on systematically increasing numbers of relatives within each degree, for instance, calculating FRR for those with 1 affected FDR (irrespective of affected SDRs and TDRs), then 2 affected FDRs, then 3, and so forth. Because in a clinical setting individual patients are expected to have various degrees of family history knowledge, we calculated relative risks for the following situations: (1) only first-degree family history known, (2) only first- and second-degree family history known, and (3) first-, second-, and third-degree family history known. Additional constellation patterns and associated FRR estimates are contained in Supplementary Tables A and B (see supplemental material). Variations in age at diagnosis of CRC for the affected FDRs and SDRs (Supplementary Tables C and D) were also considered. FRRs for ≥1 affected FDRs by type of FDR (eg, mother/father, brother/sister, parent/sibling/child, offspring, and male/female) were estimated as well as the FRR for having both an affected mother and an affected father (Supplementary Table E). We also estimated FRRs based on whether a proband's family history was concentrated solely on one side of a family versus both sides (Supplementary Table F).

## Results

A total of 2,327,327 persons included in ≥3 generation family histories were included in this analysis. Among these persons included in the study, 10,556 were identified with a primary diagnosis of CRC. On the basis of consensus among the authors, an FRR estimate ≥2.0 with the lower confidence interval >1.0 was considered a

**Table 1.** Selected Familial Relative Risk (FRR) Estimates for Probands Considering Only First-Degree Relative (FDR) Family History

| No. of affected FDRs | No. of probands | FRR (95% CI) |
|---|---|---|
| 0 | 2,232,396 | 0.89 (0.87–0.91) |
| 1 | 87,089 | 1.91 (1.82–2.00) |
| ≥1 | 94,931 | 2.05 (1.96–2.14) |
| 2 | 6966 | 3.01 (2.66–3.38) |
| 3 | 762 | 4.43 (3.24–5.90) |
| 4 | 92 | 7.74 (3.71–14.24) |
| ≥5 | 22 | 19.86 (7.29–43.24) |

clinically relevant cutoff for elevated risk. All of the constellations producing FRR estimates meeting this clinically relevant criteria are presented in Tables 1–5, along with selected others that are presented for comparison.

### *First-Degree Risk*

The FRR estimates for probands with increasing numbers of affected FDRs, without respect to SDRs or TDRs, are shown in Table 1. The most commonly published FRR is for ≥1 affected FDR. We estimated FRR = 2.05 (95% CI, 1.96–2.14) for probands with ≥1 affected FDR, similar to the meta-analysis FRR of 2.07 (95% CI, 1.89–2.26) that was adjusted to account for suspected publication bias among source studies.[11]

### *Second-Degree Risk*

The FRR estimates for constellations with 0 or 1 affected FDRs and increasing numbers of affected SDRs are shown in Table 2. A positive second-degree family history (in the absence of a positive first-degree family history) can be associated with increased risk, but does not appear to be of the same magnitude as a positive first-degree family history. Second-degree family history does appear to affect risk when combined with first-degree family history. The FRR for 1 affected FDR with 1 affected SDR was 2.12 (95% CI, 1.90–2.35), significantly higher than the FRR for 1 affected FDR and 0 affected SDRs (FRR, 1.82; 95% CI, 1.72–1.93; $P = .007$). The FRR estimate for 1 FDR and ≥3 affected SDRs is even higher; FRR was 3.37 (95% CI, 2.20–4.93) and in fact exceeds the estimated FRR for 2 affected FDRs.

**Table 2.** Familial Relative Risk (FRR) Estimates for Probands With 0 or 1 Affected First-Degree Relatives (FDRs) and Increasing Numbers of Affected Second-Degree Relatives (SDRs)

| No. of affected FDRs | No. of affected SDRs | No. of probands | FRR (95% CI) |
|---|---|---|---|
| 0 | 0 | 1,965,853 | 0.86 (0.84–0.88) |
| 0 | 1 | 224,609 | 1.05 (0.99–1.11) |
| 0 | 2 | 33,407 | 1.20 (1.05–1.38) |
| 0 | ≥3 | 8527 | 1.48 (1.11–1.93) |
| 1 | 0 | 65,192 | 1.82 (1.72–1.93) |
| 1 | 1 | 16,760 | 2.12 (1.90–2.35) |
| 1 | 2 | 3776 | 2.31 (1.80–2.93) |
| 1 | ≥3 | 1361 | 3.37 (2.20–4.93) |

**Table 3.** Selected Familial Relative Risk (FRR) Estimates for Probands With 0 or 1 Affected First-Degree Relatives (FDRs) and Various Combinations of Affected Second-Degree Relatives (SDRs) and Third-Degree Relatives (TDRs)

| No. of affected FDRs | No. of affected SDRs | No. of affected TDRs | No. of probands | FRR (95% CI) |
|---|---|---|---|---|
| 0 | 0 | 0 | 1,470,367 | 0.83 (0.81–0.86) |
| 0 | 0 | ≥3 | 44,662 | 1.08 (0.97–1.20) |
| 0 | 1 | 2 | 20,321 | 1.33 (1.13–1.55) |
| 0 | 1 | ≥3 | 13,858 | 1.21 (0.98–1.48) |
| 0 | 2 | ≥3 | 4061 | 1.48 (0.98–2.16) |
| 0 | ≥3 | ≥3 | 2120 | 1.02 (0.41–2.09) |
| 1 | 0 | 0 | 41,369 | 1.76 (1.63–1.89) |
| 1 | 0 | 2 | 5560 | 1.90 (1.59–2.25) |
| 1 | 0 | ≥3 | 3255 | 2.01 (1.61–2.47) |
| 1 | 1 | 0 | 8836 | 1.88 (1.59–2.20) |
| 1 | 1 | 2 | 1882 | 2.50 (1.87–3.28) |
| 1 | 1 | ≥3 | 1357 | 3.28 (2.44–4.31) |
| 1 | 2 | 0 | 1669 | 2.37 (1.58–3.43) |
| 1 | 2 | 1 | 1006 | 1.98 (1.15–3.17) |
| 1 | 2 | 2 | 523 | 2.70 (1.44–4.62) |
| 1 | 2 | ≥3 | 578 | 2.38 (1.19–4.26) |
| 1 | ≥3 | 0 | 453 | 2.79 (1.12–5.76) |
| 1 | ≥3 | 2 | 206 | 5.32 (2.14–10.96) |
| 1 | ≥3 | ≥3 | 322 | 5.20 (2.24–10.24) |

### *Third-Degree Risk*

For those probands with no affected relatives (FDRs, SDRs, TDRs) (N = 1,460,367), FRR was 0.83 (95% CI, 0.81–0.86). Selected FRR estimates for constellations with 0 or 1 affected FDRs and various combinations of affected SDRs and TDRs are shown in Table 3. A positive

**Table 4.** Selected Familial Relative Risks (FRRs) for Probands With Affected First-Degree Relatives (FDRs) or Second-Degree Relatives (SDRs) Diagnosed at Certain Ages

| Proband | No. of probands | FRR (95% CI) |
|---|---|---|
| ≥1 affected FDR diagnosed <50 y of age | 6291 | 3.31 (2.79–3.89) |
| ≥1 affected FDR diagnosed between 50 and 59 y of age | 12,094 | 2.53 (2.24–2.85) |
| ≥1 affected FDR diagnosed ≥50 y of age | 89,340 | 2.02 (1.93–2.11) |
| ≥1 affected FDR diagnosed between 60 and 69 y of age | 25,084 | 2.22 (2.04–2.40) |
| ≥1 affected FDR diagnosed ≥60 y of age | 78,629 | 1.99 (1.90–2.09) |
| ≥1 affected FDR diagnosed between 70 and 79 y of age | 32,445 | 1.97 (1.83–2.12) |
| ≥1 affected FDR diagnosed ≥70 y of age | 56,065 | 1.97 (1.86–2.08) |
| ≥1 affected SDR diagnosed <50 y of age | 19,616 | 1.84 (1.61–2.09) |

**Table 5.** Selected Familial Relative Risk (FRR) Estimates for Probands With Various Affected First-Degree Relationship Types as Well as Male and Female Probands With ≥1 affected First-Degree Relative (FDR)

| Specific affected FDRs | No. of probands | FRR (95% CI) |
|---|---|---|
| ≥1 (parent) | 31,619 | 1.96 (1.77–2.16) |
| ≥1 (sibling) | 47,272 | 1.96 (1.86–2.07) |
| ≥1 (offspring) | 18,644 | 3.06 (2.76–3.38) |
| Mother and father both affected | 450 | 4.97 (2.72–8.34) |
| ≥2 (probands with both mother and father affected excluded) | 7392 | 3.21 (2.87–3.58) |
| ≥1 (male probands) | 48,751 | 1.96 (1.84–2.09) |
| ≥1 (female probands) | 46,180 | 2.12 (2.00–2.26) |

third-degree family history, in the absence of positive first- and second-degree family histories, does not confer a significant increased risk. For example, the FRR estimate for probands with 0 affected FDRs, 0 affected SDRs, and ≥3 affected TDRs was 1.08 (95% CI, 0.97–1.20). However, in combination with positive first- and second-degree family histories, third-degree family history can make a contribution to the total risk that is significant. As an example, the FRR for probands with 1 affected FDR, 1 affected SDR, and 0 affected TDRs was 1.88 (95% CI, 1.59–2.20), increasing to a significantly higher FRR of 3.28 (95% CI, 2.44–4.31) for probands with 1 affected FDR, 1 affected SDR, and ≥3 affected TDRs ($P = .004$). In the absence of a positive second-degree family history, presence of affected TDRs does not appear to significantly change FRR (which is significantly >1 for both constellations). For probands with 1 affected FDR, 0 affected SDRs, and ≥3 affected TDRs, FRR was 2.01 (95% CI, 1.61–2.47) compared with FRR of 1.76 (95% CI, 1.63–1.89) for probands with 1 affected FDR, 0 affected SDRs, and 0 affected TDRs ($P = .125$).

### *Age-Related Risk*

Elevated FRR estimates based on the age of diagnosis of affected FDRs are shown in Table 4. Typically disease onset <50 years is considered to be early for CRC, but we analyzed ages at diagnosis of 60 and 70 years as cutoffs as well. We estimated that for persons with ≥1 affected FDR diagnosed <50 years of age FRR was 3.31 (95% CI, 2.79–3.89); this is significantly higher than the estimate for ≥1 affected FDR when the age of diagnosis was ≥50 (FRR = 2.02; 95% CI, 1.93–2.11; $P < .001$). However, when the diagnosis age (of affected FDRs) was limited to between 60 and 69 years of age, the FRR estimate of 2.22 (95% CI, 2.04–2.40) was still elevated above the chosen cutoff. In fact, it was significantly higher ($P = .045$) than the FRR for probands with ≥1 affected FDR when the age of diagnosis was not considered (FRR = 2.05; 95% CI, 1.96–2.14), although for counseling purposes these numbers are not dissimilar.

When considering SDRs, age of diagnosis of the affected relative also affects risk. The FRR estimate for ≥1 affected SDR (without respect to FDRs and TDRs) was 1.27 (95% CI, 1.22–1.33). The estimate for ≥1 affected SDR diagnosed <50 years of age (FRR = 1.84; 95% CI, 1.61–2.09) was significantly higher ($P < .001$).

### *Relationship Type- and Sex-Related Risks*

We also estimated FRRs for specific FDR relationship types, shown in Table 5. No difference was found between FRRs for those with an affected parent versus an affected sibling. Differences between FRRs estimated for persons with affected brothers and sisters and for persons with affected mothers and fathers were also not significant. A statistically significant difference (of small magnitude) was observed between female and male probands with ≥1 affected FDR, with such females having FRR = 2.12 (95% CI, 2.00–2.26) versus FRR = 1.96 for males (95% CI, 1.84–2.09; $P = .04$). Of particular interest, and not previously published, was the risk estimate for those with both an affected mother and an affected father (FRR = 4.97; 95% CI, 2.72–8.34). This is increased, although not quite significantly ($P = .07$), over the FRR for ≥2 affected FDRs when probands with both an affected mother and father are excluded (FRR = 3.21; 95% CI, 2.87–3.58). This rather rare occurrence may represent the situation of >1 predisposing genes segregating in the offspring of the 2 affected parents. Therefore, we also investigated FRRs for a pattern of family history involving cases on both sides of the family versus cases on one side of the family only. To examine this issue, we compared FRRs for probands with ≥2 affected SDR relatives, separately for the case of ≥2 affected relatives on the same side of the family and then for ≥1 affected relative on each side of the family. We similarly made the comparison for probands with ≥2 affected TDRs, again separately for the case of ≥2 affected relatives on the same side of the family and then for ≥1 affected TDR on each side of the family. Results are shown in Supplementary Table F. We excluded all probands with any affected FDRs. In neither analysis was there any significant difference in the FRR estimates, whether the family history was one-sided or both-sided.

## Discussion

The purpose of this study was to define risk estimates for CRC, based on family history data from a homogeneous, well-characterized population, to better assist physicians in determining CRC risk in their patients. We used a large genealogic and cancer registry resource, the UPDB, to calculate FRRs for various constellations of family history risk of CRC. Characteristics of the Utah population represented in the UPDB data include extended relationships, large family sizes, low outmigration, low-to-normal inbreeding, and a genetic composition similar to the US white population. Cancer

records from the UCR, which are included in the UPDB data set, strictly define the disease of interest. Because the UPDB includes comprehensive statewide cancer records from the UCR, it is free of ascertainment and recall bias that might affect other studies that rely on interviews with probands to assess cancer in relatives. This lack of ascertainment bias is a particular strength of this resource.

Many previous studies have shown increased CRC risks for relatives of affected persons. Although the more common FRR estimates such as ≥1 affected FDR that we report here are consistent with studies included in the meta-analysis of Butterworth et al,[11] generally our estimates are lower with tighter confidence intervals. These differences may be due to the larger number of persons included in our analysis and perhaps differences in the incidence of disease between our population and those in other studies. Utah has the lowest incidence of CRC in the United States for both men and women, 47.5 and 35.2 per 100,000, respectively.[1]

In the meta-analysis of Butterworth et al[11] sibling risk (relative risk [RR] = 2.79; 95% CI, 2.36–3.29) was reported to be higher than parent-offspring risk (RR = 2.07; 95% CI, 1.83–2.34), and the authors suggested that this may indicate the presence of recessive genetic factors causing a susceptibility to CRC. We found no such difference between sibling and parent-offspring risk, even though the numbers of probands with an affected parent or affected sibling included in the analysis were >31,000 and >47,000, respectively. The Utah study's population-based approach avoids the challenges of combining studies with different methods and ascertainment bias as well as accumulating higher numbers of patients for analysis, thus making this study's results more robust.

Although we were limited in our ability to explore increased risk associated with whether ≥1 CRC predisposition genes had an opportunity to segregate in a pedigree, we only observed a clearly significant effect for persons with both mother and father affected. Having both an affected mother and affected father confers a higher degree of risk (FRR = 4.97; 95% CI, 2.72–8.34) than having ≥2 affected FDRs (FRR = 3.26; 95% CI, 2.92–3.63). Only a small number of probands had 2 affected parents (n = 450), but this is an interesting result that may be worth exploring in other large population sets. The elevated FRR in persons with both parents affected could result from gene–gene interaction, gene–environment interaction, or a combination of both. Our other FRR comparisons for second- and third-degree family history from one side of the family versus from both sides of the family did not show any significant differences.

On the basis of the FRR estimate for having no affected FDRs, SDRs, or TDRs (FRR = 0.83; 95% CI, 0.81–0.86), persons with no known family history have a mild but significant protection, as might be expected (given that our overall risk estimates for all groups considered must average 1.0). Because this risk estimate relies on information on current age of all FDRs, SDRs, and TDRs (information which would not be typically available in a clinical setting), clinical recommendations for persons with no known first-, second-, or third-degree positive family history should include standard risk estimates and screening recommendations based on their current age.

Increased numbers of affected FDRs influence risk much more than affected SDRs or TDRs. In fact, in the absence of a positive first-degree family history (ie, 0 affected FDRs) and considering both affected SDRs and TDRs (but not age of diagnosis in affected relatives) only 1 constellation had an FRR estimate that was significantly >1.0 in the CI (0 affected FDRs, 1 affected SDR, 2 affected TDRs; FRR = 1.33, 95% CI, 1.13–1.55). However, we previously noted that when combined with positive first-degree family history, the presence of positive second- and third-degree family history can significantly increase risk.

Age at diagnosis of CRC in affected relatives contributes significantly to risk estimates. Although an age at diagnosis <50 years typically has been used as a cutoff for early onset, we have shown that even diagnosis between 60 and 69 years of age in affected FDRs increases risk equivalent to the level of an affected FDR without respect to age at diagnosis. Therefore, older age of onset in an FDR should not be viewed as reassuring to the patient. In addition, even the age of onset in SDRs (<50 years) can have an effect on the proband's risk (FRR = 1.84; 95% CI, 1.61–2.09 versus FRR = 1.27; 95% CI 1.22–1.33 for ≥1 affected SDR without respect to age at diagnosis).

Precisely how our findings can be applied to CRC screening recommendations has yet to be determined, but extrapolation from current guidelines appears to be of some benefit. In the most current CRC screening recommendations, family histories that represent a 2- to 3-fold increased risk (usually any FDR with CRC diagnosed at age >60 years) suggest that CRC screening, as recommended for the general population, is indicated.[29,30] Specifically, this includes any one of the screening tools now used (annual fecal occult blood testing, sigmoidoscopy every 5 years, combination of the first 2, barium enema or computed tomographic colonography every 5 years, or colonoscopy every 10 years). The only difference in recommendations for persons with this level of familial risk is that they should start at age 40 years, rather than at age 50 years. This is because this group exhibits the same risk at age 40 years as the general population at age 50 years. Persons with a risk of ≥3-fold compared with the general population because of family history (included are those with an FDR diagnosed at age <60 years or 2 FDRs with CRC) are now recommended to have colonoscopy as the screening tool of choice,

**Table 6.** Family History Constellations That Will Produce Approximately 2-Fold and Approximately ≥3-Fold Familial Relative Risk Estimates

| Approximately 2-fold risk | Approximately ≥3-fold risk |
|---|---|
| 1 affected FDR diagnosed ≥50 y of age<br>*or*<br>1 affected FDR *and* 1 or 2 affected SDRs | ≥1 affected FDR diagnosed <50 y of age<br>*or*<br>≥1 affected FDR *and* ≥3 affected SDRs<br>*or*<br>≥2 affected FDRs (regardless of age at diagnosis)<br>*or*<br>≥1 affected FDR, ≥1 affected SDR, ≥3 affected TDRs<br>*or*<br>≥1 affected offspring |

FDR, first-degree relative; SDR, second-degree relative; TDR, third-degree relative.

starting at age 40 (or 10 years younger than the earliest diagnosis in the family) and have repeat colonoscopy every 5 years thereafter. Colonoscopy findings may alter these recommendations.

In view of these widely accepted guidelines, we would suggest provisionally that constellations of family risk that result in approximately 2-fold increased risk or approximately ≥3-fold risk be screened accordingly. A brief set of rules that specify constellations that meet these criteria is presented in Table 6. Those persons with a strong family history should always be considered for one of the inherited syndromes of CRC. Physicians should encourage persons with increased, but <2-fold, risk to be screened according to guidelines for average risk.

This study has provided evidence that the existence of affected extended relatives increases the risk of CRC in probands. However, clinicians may question whether many patients typically have valid family history information for relatives more distant than first-degree and whether the effort required to document and use this information in clinical practice is cost- and time-effective. With more people taking an interest in family history, and a growing number of electronic tools and standards for documenting and sharing family health histories, the collection and clinical use of data from patients on family health histories beyond the first-degree may be reasonable in the near future.[31,32]

With regard to the limitations of the study, these results may not be generalizable to other populations with different racial or ethnic compositions. The Utah population has been shown to be representative of the US white and Northern European populations. Other potential limitations include the reliance on appropriate cancer diagnosis coding and inability to capture relatives not represented in the UPDB genealogy or with cancer diagnosed outside the state or outside the UCR time period. We have not excluded persons from our analysis with familial forms of CRC such as hereditary nonpolyposis colorectal cancer because they may not be reliably identified; one may wonder if pedigrees containing persons with these conditions have skewed the risk estimates. However, in a previous UPDB study the number of persons meeting the Amsterdam I criteria was estimated to be small (65 of 9458 cases or 0.7% of the cases), and none of these persons had a histology indicating familial adenomatous polyposis syndrome.[24] Finally, we observed certain constellations in which the corresponding FRR estimates did not follow the anticipated trend. As an example from Table 3, FRR was 2.37 (95% CI, 1.58–3.43) for probands with 1 affected FDR, 2 affected SDRs, and 0 affected TDRs. For those with 1 affected FDR, 2 affected SDRs, and 2 affected TDRs, FRR was 2.70 (95% CI, 1.44–4.62). However, for probands with 1 affected FDR, 2 affected SDRs, and 1 affected TDR, FRR was 1.98 (95% CI, 1.15–3.17). Although there is clearly a pattern of increasing FRR for increasing numbers of affected relatives, individual estimates were not always consistent with the trend, and small sample size may be a factor.

## Conclusion

In summary, this study is unique in providing definitions of CRC risk based on first-, second-, and third-degree family history constellations that have not been reported previously. These risk estimates were based on computerized genealogy and cancer registry data for large numbers of persons from a well-defined population. We have demonstrated that, although influencing risk to a lesser extent than first-degree family history, positive second- and third-degree family histories can have a significant effect on a person's risk of CRC. We have also demonstrated how the age of cancer onset (estimated by age at cancer diagnosis) in relatives affects risk. We have provided a comprehensive set of supplemental tables that accommodate various degrees of family history knowledge, which can be used to more precisely define CRC risk.

With respect to future work, producing absolute risk calculations in real time from a person's family history constellation and current age based on the FRR estimates presented could be automated. A computerized CRC family history risk prediction tool could be created as part of a personal health record application or as a decision support component in an electronic health record. Although family history is an important risk factor for CRC, clinical, environmental, and behavioral factors are also important, but how they affect genetic susceptibility is uncertain. We are currently working to create a more comprehensive CRC risk prediction model based on a combined set of family history and clinical data for a subset of the persons included in this current study. It is hopeful that this will provide additional

insight on the contributions of family history as well as other factors on total CRC risk.

## Supplementary Material

Note: To access the supplementary material accompanying this article, visit the online version of *Gastroenterology* at www.gastrojournal.org, and at doi: 10.1053/j.gastro.2009.11.044.

### References


1. Jemal A, Siegel R, Ward E, et al. Cancer statistics, 2008. CA Cancer J Clin 2008;58:71–96.
2. Levin B, Lieberman DA, McFarland B, et al. 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. CA Cancer J Clin 2008; 58:130–160; or Gastroenterology 2008;134:1570–1595.
3. Smith RA, Cokkinides V, Eyre HJ. Cancer screening in the United States, 2007: a review of current guidelines, practices, and prospects. CA Cancer J Clin 2007;57:90–104.
4. Meissner HI, Breen N, Klabunde CN, Vernon SW. Patterns of colorectal cancer screening uptake among men and women in the United States. Cancer Epidemiol Biomarkers Prev 2006;15:389–394.
5. Palmer RC, Emmons KM, Fletcher RH, et al. Familial risk and colorectal cancer screening health beliefs and attitudes in an insured population. Prev Med 2007;45:336–341.
6. Fuchs CS, Giovannucci EL, Colditz GA, et al. A prospective study of family history and the risk of colorectal cancer. N Engl J Med 1994;331:1669–1674.
7. Johns LE, Houlston RS. A systematic review and meta-analysis of familial colorectal cancer risk. Am J Gastroenterol 2001;96: 2992–3003.
8. St John DJ, McDermott FT, Hopper JL, et al. Cancer risk in relatives of patients with common colorectal cancer. Ann Intern Med 1993;118:785–790.
9. Goldgar DE, Easton DF, Cannon-Albright LA, Skolnick MH. Systematic population-based assessment of cancer risk in first-degree relatives of cancer probands. J Natl Cancer Inst 1994;86: 1600–1608.
10. Hemminki K, Li X. Familial colorectal adenocarcinoma from the Swedish Family-Cancer Database. Int J Cancer 2001;94:743–748.
11. Butterworth AS, Higgins JP, Pharoah P. Relative and absolute risk of colorectal cancer for individuals with a family history: a meta-analysis. Eur J Cancer 2006;42:216–227.
12. Skolnick MH. The Utah genealogical data base: a resource for genetic epidemiology. In: Cairns J, Lyon JL, Skolnick MH, eds. Banbury Report No 4; Cancer. Incidence in defined populations. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory, 1980: 285–297.
13. Skolnick MH. Prospects for population oncogenetics. In: Mulvihill JJ, Miller RW, Fraumeni JF, eds. Genetics of human cancer. New York: Raven Press, 1977:19–25.
14. Cannon Albright LA. Utah family-based analysis: past, present and future. Hum Hered 2008;65:209–220.
15. World Health Organization. International classification of diseases for oncology. 3rd ed. Geneva, Switzerland: World Health Organization, 2000.
16. Cannon-Albright LA, Thomas A, Goldgar DE, et al. Familiality of cancer in Utah. Cancer Res 1994;54:2378–2385.
17. McLellan T, Jorde LB, Skolnick MH. Genetic distances between the Utah Mormons and related populations. Am J Hum Genet 1984;36:836–857.
18. Jorde LB, Shortsleeve PA, Henry JW, et al. Genetic analysis of the Utah population: a comparison of STR and VNTR loci. Hum Biol 2000;72:927–936.
19. Cannon-Albright LA, Farnham JM, Thomas A, Camp NJ. Identification and study of Utah pseudo-isolate populations—prospects for gene identification. Am J Med Genet A 2005;137A:269–275.
20. Jorde LB. Inbreeding in the Utah Mormons: an evaluation of estimates based on pedigrees, isonymy, and migration matrices. Ann Hum Genet 1989;53(Pt 4):339–355.
21. Lyon JL, Gardner JW, Klauber MR, Smart CR. Low cancer incidence and mortality in Utah. Cancer 1977;39:2608–2618.
22. Wylie JE, Mineau GP. Biomedical databases: protecting privacy and promoting research. Trends Biotechnol 2003;21:113–116.
23. Chaturvedi AK, Mbulaiteye SM, Engels EA. Underestimation of relative risks by standardized incidence ratios for AIDS-related cancers. Ann Epidemiol 2008;18:230–234.
24. Maul JS, Warner NR, Kuwada SK, et al. Extracolonic cancers associated with hereditary nonpolyposis colorectal cancer in the Utah Population Database. Am J Gastroenterol 2006 Jul;101(7): 1591–1596.
25. Allen-Brady K, Camp NJ, Ward JH, Cannon-Albright LA. Lobular breast cancer: excess familiality observed in the Utah Population Database. Int J Cancer 2005;117:655–661.
26. Cannon-Albright LA, Camp NJ, Farnham JM, et al. Genealogical assessment of heritable predisposition to aneurysms. J Neurosurg 2003;99:637–643.
27. Horne BD, Camp NJ, Muhlestein JB, Cannon-Albright LA. Identification of excess clustering of coronary heart diseases among extended pedigrees in a genealogical population database. Am Heart J 2006;152:305–311.
28. Agresti A. Categorical data analysis. New York: Wiley, 1990.
29. Winawer S, Fletcher R, Rex D, et al. Colorectal cancer screening and surveillance: clinical guidelines and rationale—Update based on new evidence. Gastroenterology 2003;124:544–560.
30. Burt RW. Colon cancer screening. Gastroenterology 2000;119: 837–853.
31. Simon C, Acheson L, Burant C, et al. Patient interest in recording family histories of cancer via the Internet. Genet Med 2008;10: 895–902.
32. Feero WG, Bigley MB, Brinner KM; Family Health History Multi-Stakeholder Workgroup of the American Health Information Community. New standards and enhanced utility for family health history information in the electronic health record: an update from the American Health Information Community's Family Health History Multi-Stakeholder Workgroup. J Am Med Inform Assoc 2008;15:723–728.


**Received July 27, 2009. Accepted November 6, 2009.**

***Reprint requests***

**Address requests for reprints to: David P. Taylor, MS, Department of Biomedical Informatics, University of Utah, School of Medicine, Health Sciences Education Building, 26 South 2000 East, Suite 5700, Salt Lake City, UT 84112-5750. e-mail: d.p.taylor@utah.edu; fax: (801) 581-4297.**

***Acknowledgments***

**The authors thank Steve Backus for database and software support and Jim Farnham and Kristina Allen–Brady for statistical analysis assistance.**

***Conflicts of interest***

**Randall Burt is a consultant for Myriad Genetics, but has no conflict. The remaining authors disclose no conflicts.**

***Funding***

**Research was supported by the Utah Cancer Registry, which is funded by contract N01-PC-35141 from the National Cancer**

Institute's Surveillance, Epidemiology, and End Results program with additional support from the Utah State Department of Health and the University of Utah. Partial support for all data sets within the Utah Population Database was provided by the University of Utah Huntsman Cancer Institute. Additional support was from R01, National Library of Medicine grant LM009331 (L.A.C.-A.), National Cancer Institute grants R01-CA40641 and PO1-CA73992 (R.W.B.), and an Intermountain Healthcare Homer Warner Center for Informatics Research Fellowship (D.P.T.).

# CHAPTER 3

# COMPARISON OF COMPLIANCE FOR COLORECTAL CANCER SCREENING AND SURVEILLANCE BY COLONOSCOPY BASED ON RISK

David P Taylor, Lisa A Cannon-Albright, Carol Sweeney, Marc S Williams, Peter J Haug, Joyce A Mitchell, and Randall W Burt

Conditionally accepted by *Genetics in Medicine*

## 3.1 Abstract

Purpose: To compare colonoscopy screening/surveillance rates by level of risk for colorectal cancer (CRC) based on age, personal history of adenomatous polyps or CRC, or family history of CRC.

Methods: Participants were aged 30-90 years, were seen within 5 years at Intermountain Healthcare, and had family history in the Utah Population Database. Colonoscopy rates were measured for those with/without risk factors.

Results: Among those aged 60-69, 48.4% had colonoscopy in the last 10 years, with rates declining after age 70. Percentages of those having had a colonoscopy in the last 10 years generally increased by risk level from 38.5% in those with a familial relative risk < 1.0 to 47.6% in those with a familial relative risk >3.0. Compared to those with no family history, the odds ratio for being screened according to guidelines was higher for those with 1 first-degree relative diagnosed with CRC ≥60 years or 2 affected second-degree relatives (1.54, 95% CI: 1.46-1.61) than those with 1 affected first-degree relative diagnosed <60 years or ≥2 affected first-degree relatives (1.25, 95% CI: 1.14-1.37).

Conclusions: Compliance with colonoscopy guidelines was higher for those with familial risk, but did not correspond with the degree of risk.

## 3.2 Introduction

Colorectal cancer (CRC) is the second leading cause of cancer-related death in the United States. In 2010 it is estimated that 142,570 cases were diagnosed and 51,370 deaths were caused by the disease.[1] Having a positive family history such as

a single affected first-degree relative essentially doubles an individual's risk for the disease.[2-5] Other important risk factors for CRC include age, with or without a personal history of CRC, adenomatous polyps or inflammatory bowel disease (IBD).[1, 6, 7]

CRC is often preventable through screening because precancerous polyps can be identified and removed.[8, 9] Findings from the National Polyp Study suggest that 76-90% of CRC occurrences could be prevented through periodic colonoscopy.[2, 10] Updated screening and surveillance guidelines were published in 2008-2009 by the United States Preventive Services Task Force (USPSTF), the American College of Gastroenterology, the American Cancer Society, the Multi-Society Task Force on CRC, and the American College of Radiology (ACS–MSTF–ACR).[8, 9, 11, 12] While current guidelines support the use of fecal occult blood testing (FOBT) or flexible sigmoidoscopy as screening options for those at average risk, colonoscopy is considered to be more sensitive.[13-15] Consequently, guidelines recommend colonoscopy as the screening/surveillance tool for those with significantly elevated risk arising from a personal history of adenomatous polyps, surgically resected CRC, inflammatory bowel disease (IBD), or a family history of CRC. Based on the 2003 National Health Information Survey (NHIS), percentages of men and women reporting colonoscopy (32.2% and 29.8%, respectively) were higher than those reporting FOBT (16.1% and 15.3%, respectively) or sigmoidoscopy (7.6% and 5.9%, respectively).[16] Current practice trends confirm that physicians most frequently recommend colonoscopy rather than one of the other test options prescribed in the guidelines.[17]

Screening guidelines recommend that average-risk individuals begin screening at 50 years of age. According to data from the 2005 NHIS and the 2006 Behavioral Risk Factor Surveillance System (BRFSS), between 50% and 60% of adults ≥50 years of age reported having had an FOBT in the past year and/or endoscopy (colonoscopy or sigmoidoscopy) in the past 10 years.[18, 19] According to 2008 BRFSS data for the state of Utah where this study was conducted, 67.2% of individuals aged ≥50 years report having ever had endoscopy versus 61.8% nationwide.[20]

Individuals who have undergone a surgical resection for CRC are at increased risk for a recurrence of CRC. Current surveillance guidelines recommend colonoscopy to be performed 1 year after resection based on reports of a high incidence of apparently metachronous second cancers (i.e., originating separately from the original cancer) within 2 years after resection.[21] It has been reported that between 55% and 61.2% of patients have had ≥1 colonoscopy or other complete colon examination within 18 months after resection, and between 52.4% and 73.6% have had ≥1 colonoscopy within 3 years after resection.[22-25]

For those with a family history of CRC or adenomatous polyps in first-degree relatives, it is recommended that screening start at age 40, or 10 years before the earliest age of diagnosis of a family member diagnosed with CRC, whichever is younger, with follow-up screening every 5 years. Several studies have compared CRC screening rates (colonoscopy and/or FOBT) between persons with a positive family history and those without.[26-31] In general, those with a positive family history of CRC were significantly more likely to be in compliance with screening

recommendations compared to those without a family history. But overall the prevalence of screening was found to be low, and the majorities of both groups had not been screened. Limitations of these studies include small sample sizes, highly selected populations, self-reported family history, and most importantly, self-reported CRC screening. Only one study used an electronic medical record (EMR) for documentation of CRC screening.[31] This study found that those with a positive family history were appropriately screened, based on risk-specific guidelines, less frequently than those with no family history of CRC. This may be due to the fact that screening guidelines are more stringent for those with a family history than those at average risk. Therefore compliance rates among those at increased risk are lower than those at average risk because more screenings are recommended in shorter time frames, meaning that there are more opportunities to be out of compliance.

We previously reported a comprehensive set of familial relative risk estimates based on 2.3 million individuals in Utah with various constellations of first-, second-, and third-degree relatives affected with CRC.[4] These estimates were produced using the Utah Population Database (UPDB), a population-based resource with a computerized genealogy linked to statewide cancer registry records. The UPDB was created in the early 1970s and contains genealogies for the original Utah pioneers (members of the Church of Jesus Christ of Latter-day Saints) and their modern-day descendants.[32, 33] A high proportion of Utah residents (approximately 60%) receive care through Intermountain Healthcare, an integrated healthcare system, and are represented in Intermountain electronic records. The UPDB

resource and the linkage to Intermountain records have provided unprecedented opportunities for records-based research on cancer screening behavior in relation to family history of cancer.

The objective of this study was to compare colonoscopy screening/surveillance rates among those with various levels of risk based on family history and other factors, in a large sample using electronic family history and EMR data. Using these risk factors linked to data on colonoscopy procedures performed, we measure the numbers in and out of compliance with adapted guidelines.

## 3.3 Materials and methods

The UPDB, described above, includes information for more than 7 million individuals, although not all have linked genealogical data. [4, 34] Cancer history information for these individuals is obtained from the Utah Cancer Registry (UCR; a National Cancer Institute Surveillance, Epidemiology, and End Results program registry since 1973) and death certificates. A linkage has been created between the UPDB and clinical records from Intermountain Healthcare; 3.2 million individuals have records in both sources.

Intermountain Healthcare is a community-owned, nonprofit health care system that serves the health needs of Utah and southeastern Idaho residents. Intermountain electronically integrates data for all aspects of care, including inpatient and outpatient clinical and administrative data for diagnoses, procedures, lab results, billing codes and information, and pathology reports.

The Resource for Genetic Epidemiology (RGE) at the University of Utah governs access to the UPDB.[35] RGE, University of Utah Institutional Review Board (IRB), and Intermountain Healthcare IRB approvals were obtained to conduct this research.

The individuals in this study were drawn from a pool of 357,208 CRC cases, matched controls (matched on age and sex), and relatives of cases and controls. Inclusion criteria for the current study were: (1) no record of death, (2) currently between the ages of 30 and 90, (3) part of ≥3 generations of Utah genealogy data and a descendant of original Utah pioneers, (4) seen as an inpatient or outpatient at Intermountain between December 2004 and December 2009, and (5) evidence of an Intermountain encounter 10 years previous to the most recent encounter. The encounter criteria help to exclude individuals who are not current Intermountain patients and those who have not been patients in the system long enough to adequately assess screening/surveillance compliance.

The following data were collected for study individuals (data source in parentheses): Date of last colonoscopy (Intermountain inpatient and outpatient Current Procedural Terminology/Healthcare Common Procedure Coding System and International Classification of Diseases 9 procedure codes [ICD-9]), diagnosis of CRC and surgical resection (Intermountain cancer registry), removal of an adenomatous polyp (contained in findings of pathology report), diagnosis of IBD (inpatient and outpatient ICD-9 codes and free text problem list entries), and dates of outpatient visits and inpatient hospital stays (Intermountain billing data). Numbers of first-, second-, and third-degree relatives affected with CRC were

obtained from UPDB genealogy and UCR cancer data. Among individuals with cancer recorded in the UCR, 94% link to ≥1 records in the UPDB and 64.2% have family information. The type of relation and age at diagnosis of affected relatives were also collected from the UPDB. The familial relative risk for each study individual was obtained by comparing their unique constellation of relatives affected with CRC, with familial relative risk estimates for various constellations previously published.[4] An example of a family history constellation for a proband (considering CRC in the first- through third-degree relatives) is 0 affected first-degree relatives, 1 affected second-degree relative, and 3 affected third-degree relatives. Familial relative risk based on these extended constellations provides more quantitative and precise risk estimates than the guideline-based family history risk categories and is presented to provide an additional perspective on risk.

While the total study population included individuals aged 30 to 90 years, we first evaluated the numbers of study individuals between 50 and 90 years of age who had evidence of colonoscopy in the past 10 years according to Intermountain data, stratified by age, risk factors, and also by familial relative risk level (e.g., <1.0, 1.0-1.99, 2.0-2.99, >3.0). These numbers are irrespective of how long individuals have had risk factors or whether colonoscopies were for screening, surveillance, diagnostic, or treatment purposes, due to the difficulty of distinguishing the reason for colonoscopy when using only coded data.

We also evaluated compliance with guidelines for those at normal risk or increased risk based on age, positive family history of CRC, a personal history of surgically resected CRC, or a personal history of adenomatous polyps.[8] Study

individuals aged 30-90 years with at least one risk factor were included. We used adapted guideline criteria (Table 3.1) to assign each individual to a status of compliant or not compliant with risk factor-specific CRC screening (age, family history) or surveillance (polyps or surgically resected CRC) guidelines using Intermountain colonoscopy data. Screening and surveillance guidelines from the

Table 3.1: Adaptation of American Cancer Society, the Multi-Society Task Force on CRC, and the American College of Radiology (ACS–MSTF–ACR) joint screening/surveillance guidelines for early detection of colorectal adenomas and cancer, for measuring compliance in study individuals.

| **Risk factor** | **Inclusion criteria for measurement of compliance** | **Compliance criteria** |
|---|---|---|
| Personal history of CRC | CRC resection between 18 and 191 months ago* | Colonoscopy within 18 months after resection |
| Personal history of adenoma | Polypectomy ≥6 years ago | Colonoscopy within 6 years after polypectomy |
| 1 CRC affected first-degree relative <60 years or ≥2 affected first-degree relatives, any age | Proband ≥40 years of age<br>OR<br>Proband age > (youngest affected relative dx age – 10) | Colonoscopy within last 6 years |
| 1 CRC affected first-degree relative ≥60 years or 2 affected second-degree relatives | Proband ≥40 years of age | Colonoscopy within last 11 years |
| None | Proband ≥50 years of age<br>AND<br>No history of IBD | Colonoscopy within last 11 years |

*Resection dates are available electronically farther back in time than colonoscopies. Patients with resections >191 months ago have been excluded.

ACS–MSTF–ACR for individuals at high or increased risk were simplified based on authors' expert opinion in order to measure compliance in a practical way using available electronic sources of data. Additional time periods beyond those specified in the guidelines were provided in order to count colonoscopies that occurred shortly after the due date. For example, according to the guidelines those ≥50 years of age with no other risk factors are to be screened every 10 years. Eleven years were provided in our adaptation of the guideline to measure compliance in order to allow capture of procedures occurring within the 10$^{th}$ year. For those with a personal history of surgically resected CRC, only 1 year surveillance compliance (adapted as 18 months) was measured because of the complexity of screening intervals past the first post-resection screening. Due to surgical resection data being available electronically before colonoscopy dates were available, individuals with resections earlier than 1994 were not included in the analysis. Compliance in those with IBD was also not assessed due to complexity; however, IBD diagnoses were used to exclude individuals from the group ≥50 years of age at "normal" risk (i.e., have no other risk factors considered in this analysis). Multivariate logistic regression was used to quantify the relative odds that individuals with each particular risk factor would be compliant with guidelines through colonoscopy, compared with those at normal risk, adjusting for age and sex.

## 3.4 Results

### 3.4.1 Colonoscopy within the last 10 years among those aged 50-90

There were 71,446 individuals aged 50 to 90 years included in Table 3.2. Among other risk factors for CRC, 5836 (8.2%) had at least one adenoma documented and 2738 (3.8%) had a history of an advanced adenoma. Individuals with advanced adenoma (defined here as multiple adenomas, villous adenoma, or high-grade dysplasia) are a subset of those identified with an adenoma. Among those with ≥1 CRC affected first-degree relative, 8.7% had a first-degree relative diagnosed under the age of 50 years. There were 55,646 (77.9%) considered at normal risk, having no CRC, IBD, or adenoma and having 0 CRC affected first-degree relatives and ≤1 CRC affected second-degree relatives.

Evidence of colonoscopy in the last 10 years was found for 34.1% of individuals at normal risk and 57.8% of those with one or more risk factors. By age group, the percentage of individuals with colonoscopy within 10 years was highest in the 60-69 year age range (48.4%), and lowest in the 80-90 age range (28.5%). Colonoscopy within the last 10 years was detected for most individuals with a history of CRC (65.7%), adenoma (83.9%), or advanced adenoma (93.9%). Regardless of risk defined by familial relative risk, the 60-69 year age group remained the group with the highest colonoscopy rates.

We evaluated evidence of colonoscopy in the last 10 years in this same sample aged 50-90 according to familial relative risk level (Table 3.3). The most

common risk category was familial relative risk < 1.0 (77.2%). There were 16.9% with a familial relative risk between 1.0 and 1.99, 4.6% with a familial relative risk

Table 3.2: Baseline characteristics for a sample of study individuals, aged 50-90, and the fractions with evidence of colonoscopy in the last 10 years.

| | n (%) | % with colonoscopy in last 10 years | Odds ratio | 95% Confidence Interval |
|---|---|---|---|---|
| **Ages 50-90** | 71,446 | 39.3 | | |
| Ages 50-59 | 25,374 | 39.0 | 1.00 | (Reference) |
| Ages 60-69 | 17,877 | 48.4 | 1.46 | 1.41 – 1.52 |
| Ages 70-79 | 15,577 | 38.2 | 0.96 | 0.93 – 1.00 |
| Ages 80-90 | 12,618 | 28.5 | 0.62 | 0.59 – 0.65 |
| **Sex** | | | | |
| Male | 31,609 (44.2) | 41.0 | 1.00 | (Reference) |
| Female | 39,837 (55.8) | 38.1 | 0.89 | 0.86 – 0.92 |
| **History of CRC** | | | | |
| No | 70,673 (98.9) | 39.1 | 1.00 | (Reference) |
| Yes | 773 (1.1) | 65.7 | 3.49 | 3.00 – 4.06 |
| **History of adenoma** | | | | |
| No | 65,610 (91.8) | 35.4 | 1.00 | (Reference) |
| Yes | 5836 (8.2) | 83.9 | 9.75 | 9.07 – 10.48 |
| With 0 affected FDRs | 4974 (85.2) | 84.2 | 9.86 | 9.11 – 10.66 |
| With ≥1 affected FDR | 862 (14.8) | 82.5 | 9.09 | 7.61 – 10.86 |
| With 0 dx < age 50 | 775 (89.9) | 82.1 | 8.89 | 7.38 – 10.71 |
| With ≥1 dx < age 50 | 87 (10.1) | 86.2 | 11.29 | 6.12 – 20.84 |
| **History of advanced adenoma** | | | | |
| No | 68,708 (96.2) | 37.2 | 1.00 | (Reference) |
| Yes | 2738 (3.8) | 93.9 | 26.40 | 22.54 – 30.91 |
| With 0 affected FDRs | 2311 (84.4) | 94.5 | 29.14 | 24.35 – 34.87 |
| With ≥1 affected FDR | 427 (15.6) | 90.9 | 17.31 | 12.43 – 24.10 |
| With 0 dx < age 50 | 377 (88.3) | 90.2 | 16.15 | 11.48 – 22.73 |
| With ≥1 dx < age 50 | 50 (11.7) | 96.0 | 38.41 | 9.31 – 158.42 |
| **Affected FDRs** | | | | |
| 0 affected FDRs | 64,159 (89.8) | 38.6 | 1.00 | (Reference) |
| ≥1 affected FDR | 7287 (10.2) | 45.6 | 1.40 | 1.33 – 1.47 |
| With ≥1 dx < age 50 | 632 (8.7) | 48.9 | 1.53 | 1.31 – 1.79 |
| With ≥1 dx age 50-59 | 1155 (15.9) | 50.6 | 1.64 | 1.46 – 1.84 |
| With ≥1 dx age 60-69 | 2064 (28.3) | 46.9 | 1.47 | 1.35 – 1.61 |
| **Risk level** | | | | |
| "Normal" risk: no CRC, no IBD, no adenoma, 0 affected FDRs, and ≤1 affected SDRs | 55,646 (77.9) | 34.1 | 1.00 | (Reference) |
| "Increased" or "high" risk: CRC, IBD, adenoma, ≥1 affected FDR, or ≥2 affected SDRs | 15,800 (22.1) | 57.8 | 2.74 | 2.64 – 2.84 |

Odds ratios are adjusted for age and sex.

Table 3.3. Sample of study individuals, aged 50-90, with evidence of colonoscopy in the last 10 years, according to levels of familial relative risk (FRR) and age.

| | (n = 71,446)(%) | % with colonoscopy in last 10 years |
|---|---|---|
| **FRR < 1.0** | 55,138 (77.2) | 38.5 |
| Ages 50-59 | 20,470 (37.1) | 37.6 |
| Ages 60-69 | 13,938 (25.3) | 47.2 |
| Ages 70-79 | 11,739 (21.3) | 37.6 |
| Ages 80-90 | 8991 (16.3) | 27.9 |
| **1.0 ≤ FRR < 2.0** | 12,070 (16.9) | 40.5 |
| Ages 50-59 | 3727 (30.9) | 42.1 |
| Ages 60-69 | 2933 (24.3) | 51.1 |
| Ages 70-79 | 2793 (23.1) | 38.6 |
| Ages 80-90 | 2617 (21.7) | 28.6 |
| **2.0 ≤ FRR < 3.0** | 3306 (4.6) | 47.5 |
| Ages 50-59 | 955 (28.9) | 53.7 |
| Ages 60-69 | 811 (24.5) | 57.8 |
| Ages 70-79 | 802 (24.3) | 42.8 |
| Ages 80-90 | 738 (22.3) | 33.3 |
| **FRR ≥ 3.0** | 932 (1.3) | 47.6 |
| Ages 50-59 | 222 (23.8) | 55.0 |
| Ages 60-69 | 195 (20.9) | 55.4 |
| Ages 70-79 | 243 (26.1) | 49.8 |
| Ages 80-90 | 272 (29.2) | 34.2 |

Based on a logistic regression model adjusted for age and sex, the trend in compliance with increasing FRR was significant at $p < 0.001$.

between 2.0 and 2.99, and 1.3% with a familial relative risk ≥ 3.0. Colonoscopy rates generally increased by risk level, from 38.5% in those with a familial relative risk < 1.0 to 47.6% in those with a familial relative risk ≥ 3.0.

### 3.4.2 Screening/surveillance compliance rates using colonoscopy according to adapted guidelines

We summarized colonoscopy screening compliance (Table 3.4) in a sample at normal or increased risk according to criteria presented in Table 3.1. A total of

Table 3.4. Colorectal cancer screening and surveillance compliance using colonoscopy according to risk-specific guidelines, and odds ratios for screening compliance through colonoscopy for individuals with CRC risk factors compared to those at normal risk.

| | (n = 73,912) | % compliant with screening/surveillance by colonoscopy | Odds Ratio‖ | 95% Confidence Interval |
|---|---|---|---|---|
| Normal risk * | 55,646 | 35.0 | 1.00 | |
| CRC surgical resection ** | 529 | 40.1 | 1.27 | 1.06 – 1.53 |
| Adenoma † | 1273 | 58.4 | 2.57 | 2.29 – 2.89 |
| Higher familial risk ‡ | | | 1.25 | 1.14 – 1.37 |
| Probands ≥40 years of age | 2518 | 38.6 | | |
| Probands <40 years of age | | | | |
| Proband current age > (earliest affected first-degree relative diagnosis - 10 years) | 66 | 33.3 | | |
| Proband current age ≤ (earliest affected first-degree relative diagnosis - 10 years) | 237 | NA¶ | | |
| Lower familial risk § | | | 1.54 | 1.46 – 1.61 |
| Probands ≥40 years of age | 9002 | 42.4 | | |
| Probands <40 years of age | 501 | NA¶ | | |

* ≥50 years of age, no CRC, no adenoma, no IBD, 0 affected first-degree relatives, and ≤1 affected second-degree relatives. Compliance: Colonoscopy within last 11 years.

** History of surgical resection for CRC ≥18 months ago and ≤191 months ago. Compliance: Colonoscopy ≤18 months after resection.

† History of adenoma ≥6 years ago. Compliance: Colonoscopy ≤6 years after polypectomy.

‡ 1 affected first-degree relative diagnosed <60 years or ≥2 first-degree relatives any age. Compliance: Colonoscopy within last 6 years.

§ 1 affected first-degree relative diagnosed ≥60 years or 2 affected second-degree relatives. Compliance: Colonoscopy within last 11 years.

¶ Screening not indicated until age 40

‖ Each risk factor was modeled separately, but in combination with variables for sex and age (<50 years, 5 year age groups from 50 to 90).

73,912 individuals ≥30 years of age with ≥1 risk factor were included in this analysis. Among 55,646 considered at normal risk, 35.0% underwent colonoscopy within the last 11 years. There were 529 with a history of a surgical resection for CRC ≥18 and ≤191 months ago and among these 40.1% had undergone colonoscopy within 18 months after the resection date. Among 1273 individuals with a history of adenoma documented ≥6 years ago, 58.4% had colonoscopy within 6 years after polypectomy. The higher set of criteria for increased risk based on a positive family history of CRC is "1 affected first-degree relative diagnosed <60 years or ≥2 first-degree relatives of any age." For comparison with the view of risk presented in Table 3.3, individuals with ≥1 affected first-degree relative diagnosed <60 years of age would have a familial relative risk of 2.69 (95% CI: 2.43-2.96) and for those with 2 first-degree relatives of any age the familial relative risk is 3.01 (95% CI: 2.66-3.38). In the 2518 individuals ≥40 years of age meeting the criteria, 38.6% were compliant with the guideline within the last 6 years. For those younger than 40 years of age who met the criteria, 33.3% underwent colonoscopy within the last 6 years. The lower set of criteria for increased risk based on family history of CRC is "1 affected first-degree relative diagnosed ≥60 years or 2 affected second-degree relatives." Once again, for comparison with Table 3.3, for those with ≥1 affected first-degree relative diagnosed ≥60 years the familial relative risk is 1.99 (95% CI: 1.90-2.09) and for those with 2 affected second-degree relatives (in the absence of any affected first-degree relatives) the familial relative risk is 1.20 (95% CI: 1.05-1.38). Among the 9002 who met the increased risk criteria and who were also ≥40 years of age, 42.4% underwent colonoscopy within the last 10 years.

We estimated the relative odds that individuals with a particular risk factor would be compliant with guidelines specific to their risk level, compared to compliance in those at normal risk with their appropriate guidelines, taking into account age and sex (Table 3.4). The highest odds ratio was observed for those with a history of adenoma ≥6 years ago (2.57, 95% CI: 2.29-2.89) and the lowest was observed in those meeting the higher familial risk criteria (1 CRC affected first-degree relative diagnosed <60 years or ≥2 first-degree relatives affected at any age; OR = 1.25, 95% CI: 1.14-1.37). All groups at higher risk had significant improvement in screening/surveillance compared to the referent group.

## 3.5 Discussion

We report an analysis of colonoscopy rates in those with, and without, specific risk factors for CRC. Colonoscopy rates in the last 10 years in a sample of individuals aged 50-90 years are reported (Table 3.2). Almost half of individuals aged 60-69 had evidence of colonoscopy in the last 10 years, which is consistent with national self-reported screening behavior statistics.[18, 19] Colonoscopy rates declined in those >70 years of age. Comparisons between differences in 10-year colonoscopy rates in those with and without risk factors in Table 3.2 should be interpreted with caution as some colonoscopies may be for surveillance (in those with surgically resected CRC or adenoma), others for screening, and others as part of diagnostic or treatment processes. Based on data presented in Table 3.3, it is clear that colonoscopy rates increase and then decline with age, peaking in the 60-69 age range. Published guidelines recommend that individuals with a positive family

history have more frequent screening, compared to those with no familial risk.[8, 11, 12] Our analysis shows that while rates generally increase with familial relative risk level, indicating that a positive family history has some effect on screening behavior in this population, the increase does not reflect the increased frequency recommended by the guidelines. For example, less than 40% of those with 1 affected first-degree relative diagnosed <60 years or ≥2 first-degree relatives any age had evidence of colonoscopy within the last 6 years. In those with 1 affected first-degree relative diagnosed ≥60 years or 2 affected second-degree relatives the percentage with evidence of colonoscopy within the last 11 years was just over 40%. Screening tests perform better in populations where the prior probability of having disease is higher. It would be expected that failure to comply with screening guidelines in populations at increased risk would have a disproportionate negative effect on CRC prevention. If higher priority were assigned to high risk individuals out of compliance with screening, the impact on prevention could be increased with a more efficient outlay of scarce resources.

Colonoscopy rates are reported for a sample of individuals aged 30-90 years according to risk factor-specific screening recommendations adapted from well-accepted guidelines (Table 3.4). The rate of colonoscopies within 18 months after CRC surgical resection (40.1%) and accompanying odds ratio for surveillance compliance compared to those at normal risk (1.27, 95% CI: 1.06-1.53) are lower than ideal. However, some patients with stage III and IV CRC may not have further surveillance after resection because cancer care is a higher priority. The colonoscopy rate within 6 years for those with polypectomy (58.4%) and the accompanying odds

ratio (2.57, 95% CI: 2.29-2.89) are higher but still leave room for improvement. Considering the influence of family history, the odds ratio for the lower level of familial risk (1.54, 95% CI: 1.46-1.61) is higher than the odds ratio for the higher level of risk (1.25, 95% CI: 1.14-1.37). Based on a logistic regression model comparing the higher level of family risk to the lower family risk and adjusting for age and sex, $p = 0.025$. This difference contradicts the assumption that those at higher risk would be screened more frequently. The central finding that people at the highest levels of risk based on family history are not being screened according to the frequency specified by guidelines is particularly concerning. Some studies have suggested increased compliance with family history is observed, yet in this large dataset the opposite is observed. There does not appear to be a study design reason to explain this finding and we believe it may well be real and should have impact on how physicians and others view family history in gaining screening compliance.

At present there are no standardized system-wide efforts within Intermountain to obtain and analyze family history and communicate risk to providers or patients. This lack of awareness of risk would interfere with appropriate application of risk-based screening guidelines. At least one study has shown an inverse correlation between the number of affected relatives and the accurate documentation of family history in the medical record by the provider.[36] The implication is that a large family history requires more time to collect taking up an unacceptable amount of the visit. Lastly, there may be a lack of provider awareness about the enhanced screening recommendations for those at high risk. Our study was unable to analyze the causes. In summary, colonoscopy rates within

the last 10 years in those with risk factors for CRC such as a previous diagnosis of CRC, adenoma, or a positive family history, were higher than those without these risk factors. While colonoscopy rates in those with risk factors according to guidelines were higher compared to rates for those without the risk factors, efforts to improve compliance are still warranted.

An important limitation of this analysis is that we are unable to ascertain colonoscopies that were performed outside the Intermountain system, which may have led to underestimating screening rates by colonoscopy. The study attempted to mitigate this by identifying individuals who had been seen in the system recently and who also had evidence of being long term users of the Intermountain system. However this does not guarantee that all colonoscopies were performed within the system. Despite this limitation, the numbers of individuals in this study were sufficiently large that the results are nonetheless meaningful. Although overall screening rates may have been improved by considering other screening tests such as FOBT or sigmoidoscopy, our analysis focused exclusively on colonoscopy, as is recommended by the majority of the guidelines we utilized. Considering current practice trends, and the fact that those with the risk factors considered in this study are more likely to undergo this procedure for screening/surveillance than other tests such as FOBT, we believe this is justified. In addition, we had concerns about completeness of FOBT and sigmoidoscopy data based on the fact that they are often performed in the outpatient setting and may not be as reliably documented electronically at Intermountain as colonoscopies. As previously noted, it was difficult to distinguish underlying reasons for colonoscopy. In terms of the potential

for bias this would tend to overstate compliance rates with guidelines. Screening and surveillance guidelines provide a recommendation for each risk factor separately and this is how compliance was measured. Therefore, individuals with more than one risk factor (e.g., a family history of CRC and a personal history of the disease and resection) would be represented in more than one category in the analysis. Also, individuals with familial forms of the CRC such as Lynch Syndrome have not been excluded from our dataset because it is difficult to reliably identify them. In a previous UPDB study the number of individuals meeting the Amsterdam I criteria was estimated to be small (65 of 9458 cases or 0.7% of the cases), and none had a histology indicating familial adenomatous polyposis syndrome.[37]

This analysis of screening and surveillance behavior through colonoscopy addresses the limitations of similar studies including small sample sizes, highly selective populations, and self-reported family history and CRC screening. The quality and depth of electronically available data on colonoscopy and risk factors, and particularly the integration of electronic family history data and cancer registry data in this study, are particular advantages contributing to this area of research.

This study demonstrates the feasibility to use data from EMRs in combination with coded family history information to assess risk. This has the potential to provide point-of-care clinical decision support and 'just in time' education to patients and providers. Future efforts are being directed to create a CRC family history risk algorithm within a patient-entered family history tool deployed in our electronic patient portal. Once deployed, this could allow combination of personal and family history risk factors and facilitate the delivery of

individualized risk-based screening recommendations to both patients and providers, and the impact on compliance with recommended screening could be assessed.

## 3.6 Acknowledgments

Research was supported by the Utah Cancer Registry, which is funded by contract N01-PC-35141 from the National Cancer Institute's Surveillance, Epidemiology, and End Results program with additional support from the Utah State Department of Health and the University of Utah. Partial support for all data sets within the Utah Population Database was provided by the University of Utah Huntsman Cancer Institute and the Huntsman Cancer Institute Cancer Center Support grant, P30 CA42014 from the National Cancer Institute. Additional support was from R01, National Library of Medicine grant LM009331 and the Huntsman Cancer Foundation (L.A.C.-A.), National Cancer Institute grants R01-CA40641 and PO1-CA73992 (R.W.B.), NIH National Center for Research Resources grant 5UL1-RR025764 (G.J.S.), and an Intermountain Healthcare, Homer Warner Center for Informatics Research Fellowship (D.P.T.).

## 3.7 References

1. Jemal A, Siegel R, Xu J, Ward E. Cancer statistics, 2010. CA Cancer J Clin 2010;60:277-300.

2. Winawer S, Fletcher R, Rex D, et al. Colorectal cancer screening and surveillance: clinical guidelines and rationale-Update based on new evidence. Gastroenterology 2003;124.

3. Lynch HT, de la Chapelle A. Hereditary colorectal cancer. N Engl J Med 2003;348.

4. Taylor DP, Burt RW, Williams MS, Haug PJ, Cannon-Albright LA. Population-based family history-specific risks for colorectal cancer: a constellation approach. Gastroenterology 2010;138:877-85.

5. Fuchs CS, Giovannucci EL, Colditz GA, Hunter DJ, Speizer FE, Willett WC. A prospective study of family history and the risk of colorectal cancer. N Engl J Med 1994;331.

6. Schatzkin A, Freedman LS, Dawsey SM, Lanza E. Interpreting precursor studies: what polyp trials tell us about large-bowel cancer. J Natl Cancer Inst 1994;86:1053-7.

7. Bernstein CN, Blanchard JF, Kliewer E, Wajda A. Cancer risk in patients with inflammatory bowel disease: a population-based study. Cancer 2001;91:854-62.

8. Levin B, Lieberman DA, McFarland B, et al. 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;134:1570-95.

9. Whitlock EP, Lin JS, Liles E, Beil TL, Fu R. Screening for colorectal cancer: a targeted, updated systematic review for the U.S. Preventive Services Task Force. Ann Intern Med 2008;149:638-58.

10. Winawer SJ, Zauber AG, Ho MN, et al. Prevention of colorectal cancer by colonoscopic polypectomy. The National Polyp Study Workgroup. N Engl J Med 1993;329.

11. Screening for colorectal cancer: U.S. Preventive Services Task Force recommendation statement. Ann Intern Med 2008;149:627-37.

12. Rex DK, Johnson DA, Anderson JC, Schoenfeld PS, Burke CA, Inadomi JM. American College of Gastroenterology guidelines for colorectal cancer screening 2009 [corrected]. Am J Gastroenterol 2009;104:739-50.

13. Rockey DC, Paulson E, Niedzwiecki D, et al. Analysis of air contrast barium enema, computed tomographic colonography, and colonoscopy: prospective comparison. Lancet 2005;365.

14. Lieberman DA, Weiss DG, Bond JH, Ahnen DJ, Garewal H, Chejfec G. Use of colonoscopy to screen asymptomatic adults for colorectal cancer. Veterans Affairs

Cooperative Study Group 380. N Engl J Med 2000;343.

15. Schoenfeld P, Cash B, Flood A, et al. Colonoscopic screening of average-risk women for colorectal neoplasia. N Engl J Med 2005;352:2061-8.

16. Meissner HI, Breen N, Klabunde CN, Vernon SW. Patterns of colorectal cancer screening uptake among men and women in the United States. Cancer Epidemiol Biomarkers Prev 2006;15.

17. Klabunde CN, Lanier D, Nadel MR, McLeod C, Yuan G, Vernon SW. Colorectal cancer screening by primary care physicians: recommendations and practices, 2006-2007. Am J Prev Med 2009;37:8-16.

18. Shapiro JA, Seeff LC, Thompson TD, Nadel MR, Klabunde CN, Vernon SW. Colorectal cancer test use from the 2005 National Health Interview Survey. Cancer Epidemiol Biomarkers Prev 2008;17.

19. (CDC) CfDCaP. Use of colorectal cancer tests--United States, 2002, 2004, and 2006. MMWR Morb Mortal Wkly Rep 2008;57.

20. (CDC) CfDCaP. Behavioral Risk Factor Surveillance System Survey Data. In. Atlanta, Georgia: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention; 2010.

21. Rex DK, Kahi CJ, Levin B, et al. Guidelines for colonoscopy surveillance after cancer resection: a consensus update by the American Cancer Society and the US Multi-Society Task Force on Colorectal Cancer. Gastroenterology 2006;130:1865-71.

22. Elston Lafata J, Cole Johnson C, Ben-Menachem T, Morlock RJ. Sociodemographic differences in the receipt of colorectal cancer surveillance care following treatment with curative intent. Med Care 2001;39:361-72.

23. Elston Lafata J, Simpkins J, Schultz L, et al. Routine surveillance care after cancer treatment with curative intent. Med Care 2005;43:592-9.

24. Cooper GS, Yuan Z, Chak A, Rimm AA. Geographic and patient variation among Medicare beneficiaries in the use of follow-up testing after surgery for nonmetastatic colorectal carcinoma. Cancer 1999;85:2124-31.

25. Cooper GS, Kou TD, Reynolds HL, Jr. Receipt of guideline-recommended follow-up in older colorectal cancer survivors : a population-based analysis. Cancer 2008;113:2029-37.

26. Clavel-Chapelon F, Joseph R, Goulard H. Surveillance behavior of women with a reported family history of colorectal cancer. Prev Med 1999;28.

27. Thrasher JF, Cummings KM, Michalek AM, Mahoney MC, Moysich KB, Pillittere DM. Colorectal cancer screening among individuals with and without a family history. J Public Health Manag Pract 2002;8:1-9.

28. Seeff LC, Nadel MR, Klabunde CN, et al. Patterns and predictors of colorectal cancer test use in the adult U.S. population. Cancer 2004;100.

29. Longacre AV, Cramer LD, Gross CP. Screening colonoscopy use among individuals at higher colorectal cancer risk. J Clin Gastroenterol 2006;40.

30. Murff HJ, Peterson NB, Greevy RA, Shrubsole MJ, Zheng W. Early initiation of colorectal cancer screening in individuals with affected first-degree relatives. J Gen Intern Med 2007;22.

31. Palmer RC, Emmons KM, Fletcher RH, et al. Familial risk and colorectal cancer screening health beliefs and attitudes in an insured population. Prev Med 2007;45.

32. Skolnick MH. The Utah genealogical data base: a resource for genetic epidemiology. In: Cairns J LJ, Skolnick MH, eds. Banbury Report No 4; Cancer Incidence in defined populations. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory; 1980:285-97.

33. Skolnick MH. Prospects for population oncogenetics. In: Mulvihill JJ, Miller RW, Fraumeni JF, eds. Genetics of human cancer. New York: Raven Press; 1977:19-25.

34. Cannon Albright LA. Utah family-based analysis: past, present and future. Hum Hered 2008;65:209-20.

35. Wylie JE, Mineau GP. Biomedical databases: protecting privacy and promoting research. Trends Biotechnol 2003;21.

36. Grover S, Stoffel EM, Bussone L, Tschoegl E, Syngal S. Physician assessment of family cancer history and referral for genetic evaluation in colorectal cancer patients. Clin Gastroenterol Hepatol 2004;2:813-9.

37. Maul JS, Warner NR, Kuwada SK, Burt RW, Cannon-Albright LA. Extracolonic cancers associated with hereditary nonpolyposis colorectal cancer in the Utah Population Database. Am J Gastroenterol 2006;101.

## CHAPTER 4

# HOW WELL DOES FAMILY HISTORY PREDICT WHO WILL GET COLORECTAL CANCER? IMPLICATIONS FOR CANCER SCREENING AND COUNSELING

David P Taylor, Gregory J Stoddard, Randall W Burt, Marc S Williams, Joyce A Mitchell, Peter J Haug, and Lisa A Cannon-Albright

*Genetics in Medicine* 2011; 13(5): In press

ARTICLE

# How well does family history predict who will get colorectal cancer? Implications for cancer screening and counseling

*David P. Taylor, MS[1], Gregory J. Stoddard, MPH[2], Randall W. Burt, MD[3,4], Marc S. Williams, MD[1,5], Joyce A. Mitchell, PhD[1], Peter J. Haug, MD[1,6], and Lisa A. Cannon-Albright, PhD[7,8]*

**Purpose:** Using a large, retrospective cohort from the Utah Population Database, we assess how well family history predicts who will acquire colorectal cancer during a 20-year period. **Methods:** Individuals were selected between ages 35 and 80 with no prior record of colorectal cancer diagnosis, as of the year 1985. Numbers of colorectal cancer-affected relatives and diagnosis ages were collected. Familial relative risk and absolute risk estimates were calculated. Colorectal cancer diagnoses in the cohort were counted between years 1986 and 2005. Cox regression and Harrell's C were used to measure the discriminatory power of resulting models. **Results:** A total of 431,153 individuals were included with 5,334 colorectal cancer diagnoses. Familial relative risk ranged from 0.83 to 12.39 and 20-year absolute risk from 0.002 to 0.21. With familial relative risk as the only predictor, Harrell's C = 0.53 and with age only, Harrell's C = 0.66. Familial relative risk combined with age produced a Harrell's C = 0.67. **Conclusion:** Family history by itself is not a strong predictor of exactly who will acquire colorectal cancer within 20 years. However, stratification of risk using absolute risk probabilities may be more helpful in focusing screening on individuals who are more likely to develop the disease. ***Genet Med* 2011:xx(x): 000–000.**

**Key Words:** *colorectal cancer, family history, risk prediction, genetic epidemiology, Utah Population Database*

Colorectal cancer (CRC) is the second leading cause of death among cancers in the United States. In 2009, it was estimated that 147,000 cases would be newly diagnosed and that 50,000 deaths would be caused by the disease.[1] Because CRC often develops from precancerous polyps that can be identified and removed, it is one of the few cancers that can be prevented through appropriate screening.[2,3] It has been estimated that more than half of deaths from CRC could be prevented through early detection.[4] Increased surveillance in those at elevated risk may lead to the detection of more cases and, therefore, a potentially greater mortality reduction than general surveillance of the population.[5] However, based on known risk factors, predicting who will develop CRC is still a challenge.

Family history has often been cited as an important risk factor for CRC based on evidence that those with a positive family history for CRC have elevated risk compared with those with no family history of the disease,[6–9] and evidence that a stronger family history results in even higher risks.[10] The most commonly used measure for family history of CRC is "≥1 affected first-degree relative," and in a recent, large, population-based study, the associated familial relative risk was estimated as 2.05 (95% confidence interval: 1.96–2.14).[10] Current screening guidelines are informed by these types of familial relative risk studies and typically recommend that individuals with a positive family history be screened earlier and more frequently than those without.[2,11,12] In addition, clinicians tend to rely on familial relative risk estimates rather than other types of risk representations such as absolute risk, even though absolute risk may be more easily interpretable and understood.[9,13,14]

Although it has been established that a positive family history is associated with increased risk, an important question is whether increased familial relative risk is actually a clinically significant predictor of who will develop CRC. There are few published large prospective studies that assess familial risk and subsequent CRC diagnosis. The studies that do exist rely on self-reported family history and are limited to first-degree relatives.[15,16] The primary purpose of this study is to assess how well family history predicts who will get CRC over a period of 20 years, using a large, retrospective cohort from the Utah Population Database (UPDB). Such information is critical for health policy organizations that address screening strategies and similarly important for practitioners who recommend screening, and genetic counselors who advise persons of cancer risk.

## MATERIALS AND METHODS

The UPDB is a population-based, electronic genealogical resource that contains multiple linked data sources including statewide cancer registry records.[17] It was created in the early 1970s with data from the Utah Family History Library and contains genealogies for the original Utah pioneers and their modern day descendants.[18,19] While the original Utah Genealogy included records for 1.6 million persons,[17] today the UPDB includes information for approximately 7 million persons, with some pedigrees >11 generations deep, although not all persons have linked genealogic data. The UPDB also includes a link to the Utah Cancer Registry (UCR), a statewide cancer registry established in 1966, which since 1973 has been part of the Surveillance, Epidemiology, and End Results (SEER) network of National Cancer Institute registries. Among those with cancer in the UCR, 94% link to ≥1 records in the UPDB, and 64.2% have family information. Cancer records are coded by disease

From the [1]Department of Biomedical Informatics, University of Utah School of Medicine; [2]Division of Epidemiology, Department of Internal Medicine, University of Utah School of Medicine; [3]Huntsman Cancer Institute; [4]Department of Internal Medicine, University of Utah School of Medicine; [5]Clinical Genetics Institute, Intermountain Healthcare; [6]Homer Warner Center for Informatics Research, Intermountain Healthcare; [7]Division of Genetic Epidemiology, Department of Internal Medicine, University of Utah School of Medicine; and [8]George E. Wallen Department of Veterans Affairs Medical Center, Salt Lake City, Utah.

David P. Taylor, MS, Department of Biomedical Informatics, University of Utah School of Medicine, Health Sciences Education Bldg, 26 South 2000 East, Suite 5700, Salt Lake City, UT 84112-5750. E-mail: d.p.taylor@utah.edu.

Disclosure: The authors declare no conflict of interest.

Submitted for publication September 13, 2010.

Accepted for publication November 9, 2010.

DOI: 10.1097/GIM.0b013e3182064384

**Table 1** Examples of colorectal cancer (CRC) family history constellations and corresponding familial relative risk estimates[a]

| No. affected first-degree relatives | No. affected second-degree relatives | No. affected third-degree relatives | Familial relative risk (95% CI) |
|---|---|---|---|
| 0 | 0 | 0 | 0.83 (0.81–0.86) |
| 0 | 0 | ≥3 | 1.08 (0.97–1.20) |
| 0 | 1 | 2 | 1.33 (1.13–1.55) |
| 1 | 0 | 0 | 1.76 (1.63–1.89) |
| 1 | 0 | ≥3 | 2.01 (1.61–2.47) |
| 1 | 1 | 0 | 1.88 (1.59–2.20) |
| 1 | 1 | ≥3 | 3.28 (2.44–4.31) |
| 2 | 0 | 0 | 2.96 (2.41–3.60) |
| 2 | 0 | ≥3 | 4.82 (3.18–7.02) |
| 2 | 1 | 1 | 1.80 (0.82–3.41) |
| 2 | 1 | ≥3 | 4.67 (2.72–7.47) |
| ≥3 | 0 | 0 | 2.96 (1.42–5.44) |
| ≥3 | 0 | 1 | 4.21 (1.82–8.30) |
| ≥3 | 0 | ≥3 | 9.63 (5.26–16.15) |
| ≥3 | 1 | 0 | 12.39 (7.08–20.12) |
| ≥1 (dx age <50 yr) | NA | NA | 3.31 (2.79–3.89) |
| ≥1 (dx age 50–59 yr) | NA | NA | 2.53 (2.24–2.85) |
| ≥1 (dx age 60–69 yr) | NA | NA | 2.22 (2.04–2.40) |

[a]Full estimates are reported in Ref. 10.
CI, confidence interval.

site according to the International Classification of Diseases of Oncology.[20] Information on site, stage, grade, age at diagnosis, histology, and patient survival are included. The UCR only reports independent primary cancers.

In contrast to many religious populations, individuals in the UPDB have been shown to be genetically representative of US white and northern European populations[21–24] with a low-to-normal level of inbreeding.[25] Also, many are members of The Church of Jesus Christ of Latter-day Saints, which has religious proscriptions against the use of coffee, tea, alcohol, and tobacco. Consequently, much lower smoking rates may play a role in Utah being among states with the lowest rates of cancer.[26]

This project used a subset of UPDB records representing a group of 2.3 million persons. These individuals were part of ≥3 generations of Utah genealogy data and descendants of original Utah pioneers. To protect the privacy of the study individuals, identifying information was not available to the authors. The Utah Resource for Genetic Epidemiology, created in 1982, governs access to the UPDB.[27] The Utah Resource for Genetic Epidemiology and the University of Utah Institutional Review Board granted approvals to conduct this research.

A retrospective cohort study design was used for this research. Considering that the latest cancer diagnosis information available in the dataset was from 2005, an observation period of 20 years was selected, with enrollment and family history assessment determined for the year 1985. The following were selection criteria: (1) individual's record was available in the UPDB before 1986, (2) no record of death before 1986, (3) no record of CRC diagnosis before 1986, and (4) as of December 31, 1985, individual was between the ages of 35 and 80 years. Data were collected on the individual's age and family history as of December 31, 1985. Numbers of CRC-affected first-degree relatives, second-degree relatives, and third-degree relatives were gathered and numbers of CRC-affected first-degree relatives diagnosed between ages 50 and 69 years.

Previously, we reported familial relative risks for probands with various combinations, or constellations, of affected relatives with CRC, using the group of 2.3 million persons described earlier.[10] Examples of constellations and their corresponding familial relative risks are listed in Table 1. A familial relative risk for each proband in this study as of the assessment date (December 31, 1985) was produced based on the proband's constellation of affected relatives in 1985.

During the observation period from 1986 to the end of 2005, data were collected on the years of occurrence of CRC diagnoses and deaths (from any cause) within the cohort. When an individual was diagnosed with CRC or died during the observation period, their record was censored during the year the event occurred. Cox regression was used to analyze the dataset, with CRC diagnosis as the dependent variable. The concordance statistic (or area under a receiver operating characteristic curve) is often used to assess the discriminatory power of a prediction model.[28–31] Discriminatory power measures the ability of a model to distinguish between those individuals having a particular outcome and others without the outcome. It corresponds to the probability that a randomly selected individual who develops the disease has a higher predicted risk than that of a

**Table 2** Baseline characteristics and observed and expected CRC diagnoses for 20-yr observation period, 1985–2005

| | *n* (%) | No. w/CRC in observation period (%) | Expected no. CRC cases[a] |
|---|---|---|---|
| Total individuals | 431,153 | 5,334 (1.2) | |
| Age 35–49 yr | 163,277 (37.8) | 886 (0.5) | |
| Age 50–59 yr | 87,828 (20.4) | 1,249 (1.4) | |
| Age 60–69 yr | 104,420 (24.2) | 1,840 (1.8) | |
| Age 70–80 yr | 75,628 (17.5) | 1,359 (1.8) | |
| Familial relative risk | | | |
| $0 \leq$ Familial relative risk $< 1.0$ | 402,317 (93.3) | 4,660 (1.2) | |
| $1.0 \leq$ Familial relative risk $< 2.0$ | 19,299 (4.5) | 402 (2.1) | |
| $2.0 \leq$ Familial relative risk $< 3.0$ | 8,238 (1.9) | 226 (2.7) | |
| $3.0 \leq$ Familial relative risk $< 4.0$ | 1,250 (0.3) | 40 (3.2) | |
| $4.0 \leq$ Familial relative risk $< 9.0$ | 41 (<0.1) | 4 (9.8) | |
| Familial relative risk $\geq 9.0$ | 8 (<0.1) | 2 (25.0) | |
| Twenty-year absolute risk prediction | | | |
| $0 \leq$ Absolute risk $< 0.01$ | 173,655 (40.3) | 990 (0.6) | 808 |
| $0.01 \leq$ Absolute risk $< 0.03$ | 247,438 (57.4) | 3,991 (1.6) | 3,822 |
| $0.03 \leq$ Absolute risk $< 0.05$ | 8,550 (2.0) | 286 (3.3) | 326 |
| $0.05 \leq$ Absolute risk $< 0.07$ | 1,465 (0.3) | 61 (4.2) | 83 |
| $0.07 \leq$ Absolute risk $< 0.13$ | 33 (<0.1) | 3 (9.1) | 3 |
| Absolute risk $\geq 0.13$ | 12 (<0.1) | 3 (25.0) | 2 |

[a]From absolute risk predictions based on 1981–1985 UPDB-specific incidence and mortality data.

randomly selected individual who does not develop the disease. The probability can range from 0.50, representing essentially a coin toss, to 1.00, representing perfect discrimination. A concordance statistic $\geq 0.70$ is generally considered a threshold for a potentially useful model, but a value $\geq 0.80$ may be a more reasonable level to provide adequate clinical utility.[32]

An equivalent of the concordance statistic for use with Cox regression, Harrell's C,[33] was calculated for each model developed and compared. Because of the resource-intensive nature of the Harrell's C calculation, for models that included more than 100,000 individuals, Harrell's C was averaged across 10 random samples of 10% of the individuals in the model. Models were developed for familial relative risk as the sole predictor, age as the sole predictor, and familial relative risk and age included together. Familial relative risk and age were modeled as categorical variables.

As an alternative to using familial relative risk as the predictor in a Cox regression, absolute risk was also used. For each study individual, the absolute risk of developing CRC in the next 20 years was estimated using the individual's age and familial relative risk in 1985, according to the method by DuPont and Plummer.[34] This method also requires age-adjusted CRC morbidity rates and age-adjusted all-cause mortality rates to estimate the absolute risk. These rates were created directly from the UPDB population individually for the years 1981–1985 and then averaged. The purpose was to simulate risk estimates in 1985 as if it were a prospective study. Absolute risk was also modeled using categorical variables. The expected numbers of individuals to develop CRC within the observation period among different levels of risk were estimated by summing the predicted absolute risk probabilities in each risk category.

Subgroup analyses were also performed by dividing the cohort into familial relative risk deciles, absolute risk deciles, and age groups. For each, Cox regression was performed, and Harrell's C was calculated for the highest risk decile, or age group, when compared with the lowest.

## RESULTS

There were a total of 431,153 individuals included in the cohort. Baseline characteristics of these individuals are listed in Table 2, as well as numbers of CRC diagnoses. The range of familial relative risk was 0.83–12.39. The majority of individuals in this cohort (93.3%) had a familial relative risk $<1.0$. Less than 0.4% had a familial relative risk $\geq 3.0$. The range of 20-year absolute risk was 0.002–0.21, and the majority (57.4%) had a probability between 0.01 and 0.03. More than 2% had a 20-year absolute risk probability $\geq 0.03$. During the observation period, 5,334 individuals developed CRC. The age category (measured at baseline) with the most CRC diagnoses was 60–69 (1,840/5,334 = 34.5%). The percentages of observed CRC cases out of total individuals in each absolute risk category ranged from 0.6% ($0 \leq$ absolute risk $< 0.01$) to 25% (absolute risk $\geq 0.13$).

Table 3 contains results of Cox regression and Harrell's C analyses. When familial relative risk was the only predictor, Harrell's C = 0.53. When age was the only predictor, the age

**Table 3** Results of Cox regression to predict diagnosis of colorectal cancer (CRC) based on (a) familial relative risk, (b) age, (c) familial relative risk and age, (d) absolute risk, (e) familial relative risk comparing the highest with lowest decile, and (f) absolute risk comparing the highest with lowest decile

| | Hazard ratio (95% CI) | Harrell's C |
|---|---|---|
| Familial relative risk | | |
| $0 \leq$ Familial relative risk $< 1.0$ | Reference category | 0.53 |
| $1.0 \leq$ Familial relative risk $< 2.0$ | 1.88 (1.70–2.08) | |
| $2.0 \leq$ Familial relative risk $< 3.0$ | 2.50 (2.19–2.86) | |
| $3.0 \leq$ Familial relative risk $< 4.0$ | 2.82 (2.06–3.85) | |
| $4.0 \leq$ Familial relative risk $< 9.0$ | 10.65 (4.00–28.40) | |
| Familial relative risk $\geq 9.0$ | 38.24 (9.56–152.94) | |
| Age (yr) | | |
| 35–49 | Reference category | 0.66 |
| 50–59 | 2.81 (2.58–3.06) | |
| 60–69 | 3.91 (3.61–4.24) | |
| 70–80 | 5.12 (4.70–5.57) | |
| Familial relative risk and age | | |
| $0 \leq$ Familial relative risk $< 1.0$ | Reference category | 0.67 |
| $1.0 \leq$ Familial relative risk $< 2.0$ | 1.67 (1.51–1.85) | |
| $2.0 \leq$ Familial relative risk $< 3.0$ | 2.27 (1.98–2.59) | |
| $3.0 \leq$ Familial relative risk $< 4.0$ | 2.79 (2.04–3.81) | |
| $4.0 \leq$ Familial relative risk $< 9.0$ | 6.67 (2.50–17.77) | |
| Familial relative risk $\geq 9.0$ | 28.43 (7.11–113.73) | |
| Age 35–49 yr | Reference category | |
| Age 50–59 yr | 2.76 (2.54–3.01) | |
| Age 60–69 yr | 3.84 (3.54–4.16) | |
| Age 70–80 yr | 4.99 (4.58–5.44) | |
| Absolute risk | | |
| $0 \leq$ Absolute risk $< 0.01$ | Reference category | 0.64 |
| $0.01 \leq$ Absolute risk $< 0.03$ | 3.48 (3.25–3.74) | |
| $0.03 \leq$ Absolute risk $< 0.05$ | 7.64 (6.70–8.72) | |
| $0.05 \leq$ Absolute risk $< 0.07$ | 9.52 (7.35–12.33) | |
| $0.07 \leq$ Absolute risk $< 0.13$ | 21.85 (7.03–67.86) | |
| Absolute risk $\geq 0.13$ | 80.08 (25.78–248.73) | |
| Familial relative risk (highest to lowest comparison) | | |
| Lowest decile (familial relative risk = 0.83) | Reference category | 0.54 |
| Highest decile ($1.02 \leq$ Familial relative risk $\leq 12.39$) | 2.17 (2.00–2.35) | |
| Age (oldest to youngest comparison) | | |
| 35–49 yr | Reference category | 0.69 |
| 70–80 yr | 5.08 (4.66–5.53) | |
| Absolute risk (highest to lowest comparison) | | |
| Lowest decile ($0.002 \leq$ absolute risk $\leq 0.003$) | Reference category | 0.78 |
| Highest decile ($0.02 \leq$ absolute risk $\leq 0.21$) | 12.21 (10.47–14.24) | |

CI, confidence interval.

group 35–49 years was used as the reference and Harrell's C = 0.66. The age group with the highest hazard ratio was 70–80 (5.12, 95% confidence interval: 4.70–5.57). Combining age and familial relative risk as predictors produced a Harrell's C = 0.67. Using absolute risk as the predictor, which is based on age, familial relative risk, and population-specific CRC incidence and all-cause mortality rates, produced a Harrell's C = 0.64.

In the subgroup analysis, when the highest decile of familial relative risk ($1.02 \leq$ familial relative risk $\leq 12.39$) was compared with the lowest (familial relative risk = 0.83), Harrell's C = 0.54. For age, comparing those in the 70–80 years age group with those in the 35–49 years group, Harrell's C = 0.69. Harrell's C = 0.78 for the analysis comparing the highest absolute risk decile ($0.02 \leq$ absolute risk $\leq 0.21$) with the lowest ($0.002 \leq$ absolute risk $\leq 0.003$). Harrell's C statistic estimates the probability that, of two randomly chosen patients, the patient with the higher prognostic score will remain free of CRC longer than the patient with the lower prognostic score from the Cox regression model.[33] That is, the model result and the actual patient outcome were concordant, where the model correctly discriminated, 78% of the time in this Cox regression model.

## DISCUSSION

We have described a retrospective cohort study that included 431,153 individuals aged 35–80 years at the beginning of the 20-year observation period. We are not aware of any other retrospective cohort or prospective studies of family history and CRC that have followed up this many individuals over this length of time. In addition, family histories of CRC were available electronically through a population-based electronic medical data resource as opposed to typically self-reported data.

Numerous studies have demonstrated increased familial relative risk for CRC in those with affected relatives.[9] According to our analysis, however, family history as represented by a familial relative risk estimate is by itself not a good predictor (Harrell's C = 0.53) of exactly who will develop CRC in the next 20 years. Even when comparing the highest familial relative risk decile in the cohort ($1.02 \leq$ familial relative risk $\leq$ 12.39) with the lowest (familial relative risk = 0.83), Harrell's C was 0.54. When familial relative risk cutoffs were set even higher for the comparison, Harrell's C continued to decline, perhaps due to fewer numbers of cases. Familial relative risk is commonly used to communicate risk levels in the literature and among physicians and genetic counselors. However, with a disease such as CRC, it may not be commonly understood that although a familial relative risk estimate may be elevated, the corresponding absolute risk may still not be high. For example, hypothetically if a disease affects 10 of 1000 people with a particular risk factor and affects 1 of 1000 people without the risk factor, the relative risk is 10.0. Despite the seemingly large relative risk, the absolute risk for those with the risk factor is still only 10 of 1000 (1%).

In contrast to familial relative risk, age is a stronger predictor for CRC (Harrell's C = 0.66). Including familial relative risk in addition to age only improves the discriminatory power by 0.01, to 0.67. Absolute risk, which combines both familial relative risk and age, produced a Harrell's C = 0.64. Considering absolute risk uses the same variables and takes into account population-specific CRC incidence rates; it is not clear why this statistic was not higher. In the age subgroup analysis comparing those in the 70–80 years age group with those in the 35–49 years group, Harrell's C = 0.69. These findings illustrate that using familial relative risk in combination with age, or alternatively absolute risk, has moderate predictive value for CRC. However, in the absolute risk subgroup analysis where the highest decile was compared with the lowest, Harrell's C improved substantially to 0.78. Although one may question the clinical utility of this particular subanalysis, it is worth noting that the highest decile of absolute risk includes those with a 20-year risk of 0.02 and greater. For illustration, 0.02 is essentially the 20-year risk of a 50-year old with ≥1 CRC affected first-degree relative, so the highest decile of absolute risk includes more than just those at the extreme high end of risk based on family history.

To provide additional perspective on the levels of Harrell's C found in this study based on family history or family history in combination with age, a recent comprehensive risk prediction model for CRC that included a range of risk factors including family history produced a concordance statistic of 0.61.[28] This was based on validation in a population independent of the one used to build the model.

Despite the moderate Harrell's C value of models taking into account family history and age, the potential clinical value of a predictive model based solely on these risk factors is doubtful. However, it may be useful to consider aspects of the analyses presented in Tables 2 and 4 for decisions about appropriate screening. At the very highest levels of risk (familial relative risk ≥ 4.0 or absolute risk ≥ 0.07), relatively large percentages of individuals (e.g., 1 in 4 or 1 in 10) categorized by both familial relative risk and absolute risk end up developing CRC. However, at more moderate levels of risk, absolute risk tends to stratify individuals more appropriately. As an example, there were 4,660 CRC cases among 402,317 individuals with familial relative risk <1.0 (1.2%). Absolute risk roughly divides the same number of individuals into two categories, absolute risk < 0.01 and $0.01 \leq$ absolute risk < 0.03, where 3,991 cases were classified among 247,438 individuals (1.6%) in the latter category. In the former category, at the lowest level of absolute risk, there were 990 cases out of 173,655 individuals (0.6%). At higher levels of risk, there were 272 CRC cases that developed in 9,537 individuals with familial relative risk ≥2.0 (2.9%). There were 353 cases that developed in the 10,060 with absolute risk >0.03 (3.5%). Absolute risk also has the benefit of facilitating the prediction of expected numbers of cases, and based on expected numbers of CRC cases in Table 2, it predicts fairly well how many individuals in each risk category are going to develop the disease in a 20-year time period.

This method of using absolute risk, based on age and family history, may be a reasonable way to quantify and stratify CRC risk, and these results demonstrate the possible utility. Particularly in the higher absolute risk categories, the numbers of cancers or precancers discovered through screening could potentially be much higher than in the general population. Considering the costs of screening, particularly colonoscopy, there are financial benefits in targeting and screening a smaller segment of the population and detecting a greater number of potential cases. It may also be easier to motivate those who are at increased risk to undergo screening, especially using more understandable absolute risk probabilities. In addition, the potential yield by percentage of appropriate screening increases as risk level increases.

Additional insights from absolute risk estimates may be gained from Table 4, which presents an adaptation of National Comprehensive Cancer Network (NCCN) screening guidelines for those with a positive family history of CRC.[35] Twenty-year absolute risk estimates are provided for 5-year age increments from 35 to 80 years according to notable family history patterns

**Table 4** Twenty-year absolute risk estimates by age and history of CRC-affected first- and second-degree relatives (FDRs and SDRs)

| Patient age (yr) | "Average" risk | ≥1 affected FDR dx age 50–59 yr | ≥1 affected FDR dx <age 50 yr | ≥1 affected FDR dx ≥age 60 yr | Two affected FDRs dx any age | Two affected SDRs dx any age |
|---|---|---|---|---|---|---|
| 35 | 0.002 | 0.006 | 0.008 | 0.005 | 0.007 | 0.003 |
| 40 | 0.004 | 0.011 | 0.015 | 0.009 | 0.013 | 0.005 |
| 45 | 0.007 | 0.018 | 0.023 | 0.014 | 0.021 | 0.009 |
| 50 | 0.010 | 0.026 | 0.034 | 0.021 | 0.031 | 0.013 |
| 55 | 0.014 | 0.036 | 0.046 | 0.028 | 0.042 | 0.017 |
| 60 | 0.018 | 0.045 | 0.059 | 0.036 | 0.054 | 0.022 |
| 65 | 0.021 | 0.052 | 0.067 | 0.041 | 0.061 | 0.025 |
| 70 | 0.021 | 0.052 | 0.067 | 0.041 | 0.061 | 0.025 |
| 75 | 0.018 | 0.045 | 0.058 | 0.035 | 0.053 | 0.021 |
| 80 | 0.013 | 0.032 | 0.042 | 0.025 | 0.038 | 0.015 |
| Colonoscopy recommendation[a] | Every 10 yr starting at age 50 yr | Every 5 yr starting at age 40 yr | Every 3–5 yr starting at age 40 or 10 yr before the earliest CRC dx | Every 5 yr starting at age 50 yr | Every 3–5 yr starting at age 40 or 10 yr before the earliest CRC dx | Every 5 yr starting at age 50 yr |

[a]Adapted with permission from The NCCN 3.2010 Colon Cancer Clinical Practice Guidelines in Oncology. National Comprehensive Cancer Network, 2010. Available at: http://www.nccn.org. Accessed September 3, 2010. Most recent and complete version of the guideline is available at www.nccn.org.

and for those at average risk (defined as having a familial relative risk of 1.0). NCCN colonoscopy recommendations for each family history category are noted as well. One may consider the absolute risk of a 50-year old with average familial relative risk as estimated from our dataset (20-year absolute risk = 0.01) as a reference point. Generally, there is consistency between the NCCN recommendations and the absolute risk patterns in that the most aggressive screening recommendations are associated with the highest levels of absolute risk. However, increased screening based on affected second-degree relatives may not be justified based on these data.

Although the absolute risk estimates in this research are based on familial relative risks that consider extended CRC-affected relatives (second- and third-degree relatives) in addition to affected first-degree relatives, our previous work has shown that the influence of extended relatives is relatively small; risk estimates are available that consider only first-degree relatives.[10] Because patients often are not aware of the cancer history (or get it wrong) in their extended relatives, this may be important to consider.[36] In addition, considering the limited time clinicians have to obtain family history, not having to collect data on second-degree relatives would be beneficial to some degree. Further research could address the impact of limiting the familial relative risk estimates to only affected first-degree relatives in the absolute risk estimates and in the expected/observed numbers of cases by risk category.

The limitations of this study include the fact that we were unable to determine whether some individuals moved out of state during the 20-year observation period and, therefore, should have been censored for analysis; this limitation would have served only to lower our estimates of diagnosis rates from truly higher rates and does not change our conclusions. Similar to other UPDB-based studies,[10] these results are generalizable to other populations of northern European origin but may not be generalizable to populations with very different racial and ethnic compositions. There is a reliance on appropriate cancer coding, but the source of cancer data was a National Cancer Institute's Surveillance, Epidemiology, and End Results registry. Not all relatives of individuals may be represented in the UPDB genealogy, but we anticipate no bias in such representation. In addition, the incidence of CRC in Utah is the lowest in the United States.[37] Although CRC incidence rates are slightly different in men and women and also between sites (e.g., colon versus rectum), we did not distinguish by sex or by site in this study, consistent with the level of granularity of the previously generated familial relative risk estimates. It is also not known what screening may have occurred in the cohort, particularly in those at increased familial risk, and what effect this may have had in preventing CRC that would have otherwise occurred. This remains a possible minimal confounding factor based on observations in a yet to be published parallel study. Also, individuals with familial forms of CRC such as hereditary nonpolyposis CRC have not been excluded from this study because they may not be reliably identified and one may question whether this could skew the analysis. However, based on a previous study using individuals in the UPDB, only a small number met the Amsterdam I criteria (65 of 9458 cases or 0.7% of the cases), and none had a histology indicating familial adenomatous polyposis syndrome.[38] Despite the limitations identified, this study adds considerable definition and specifics as to how the relative risks, which to date have been used to establish screening strategies for those with a family history of this disease, actually play out over a 20-year period. These results should be carefully considered by health policy organization as they establish screening guidelines and by clinicians and genetic counselors as they deal with persons and families with familial colon cancer risk.

In conclusion, although previous studies have demonstrated increased relative risk among those with a family history of CRC, this large retrospective cohort study has demonstrated that

family history, without respect to age, is not a strong predictor of exactly which individuals will acquire CRC in the next 20 years, based on Cox regression and a measurement of concordance. It is important to keep in mind that even if a relative risk estimate may seem large, absolute risk may still be small if the incidence of a disease is low. When combined with age in an absolute risk estimate, family history does seem to improve concordance in a subgroup analysis to compare those at higher risk with those at very low risk. However, it is doubtful that a clinically useful statistical model for predicting who will acquire CRC at an individual level can be produced using just age and family history. Despite this, absolute risk predicts fairly well how many individuals in particular risk categories will develop the disease over a period of 20 years. Stratification of risk using absolute risk in a clinical setting could help target screening on those individuals who are more likely to develop the disease. Future work would include validating these absolute risk estimates in an independent population, performing a cost/benefit analysis to determine optimal screening recommendations based on risk levels, and providing a web-based tool for clinicians to estimate absolute risk based on a patient's current age and family history.

## ACKNOWLEDGMENTS

The authors thank Steve Backus for database and software support and Jim Farnham for additional statistical guidance.

Research was supported by the Utah Cancer Registry, which is funded by contract N01-PC-35141 from the National Cancer Institute's Surveillance, Epidemiology, and End Results program with additional support from the Utah State Department of Health and the University of Utah. Partial support for all data sets within the Utah Population Database was provided by the University of Utah Huntsman Cancer Institute and the Huntsman Cancer Institute Cancer Center Support Grant, P30 CA42014 from the National Cancer Institute. Additional support was from R01, National Library of Medicine Grant LM009331 and the Huntsman Cancer Foundation (L.A.C.-A.), National Cancer Institute Grants R01-CA40641 and PO1-CA73992 (R.W.B.), NIH National Center for Research Resources Grant 5UL1-RR025764 (G.J.S.), and an Intermountain Healthcare, Homer Warner Center for Informatics Research Fellowship (D.P.T.).

## REFERENCES

1. Jemal A, Siegel R, Ward E, Hao Y, Xu J, Thun MJ. Cancer statistics, 2009. *CA Cancer J Clin* 2009;59:225–249.
2. Levin B, Lieberman DA, McFarland B, et al. 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.
3. Pignone M, Rich M, Teutsch SM, Berg AO, Lohr KN. Screening for colorectal cancer in adults at average risk: a summary of the evidence for the U.S. Preventive Services Task Force. *Ann Intern Med* 2002;137:132–141.
4. Colditz GA, Atwood KA, Emmons K, et al. Harvard report on cancer prevention volume 4: Harvard Cancer Risk Index. Risk Index Working Group, Harvard Center for Cancer Prevention. *Cancer Causes Control* 2000;11:477–488.
5. Hunt LM, Rooney PS, Hardcastle JD, Armitage NC. Endoscopic screening of relatives of patients with colorectal cancer. *Gut* 1998;42:71–75.
6. Johns LE, Houlston RS. A systematic review and meta-analysis of familial colorectal cancer risk. *Am J Gastroenterol* 2001;96:2992–3003.
7. Goldgar DE, Easton DF, Cannon-Albright LA, Skolnick MH. Systematic population-based assessment of cancer risk in first-degree relatives of cancer probands. *J Natl Cancer Inst* 1994;86:1600–1608.
8. Hemminki K, Li X. Familial colorectal adenocarcinoma from the Swedish Family-Cancer Database. *Int J Cancer* 2001;94:743–748.
9. Butterworth AS, Higgins JPT, Pharoah P. Relative and absolute risk of colorectal cancer for individuals with a family history: a meta-analysis. *Eur J Cancer* 2006;42:216–227.
10. Taylor DP, Burt RW, Williams MS, Haug PJ, Cannon-Albright LA. Population-based family history-specific risks for colorectal cancer: a constellation approach. *Gastroenterology* 2010;138:877–885.
11. Screening for colorectal cancer: U.S. Preventive Services Task Force recommendation statement. *Ann Intern Med* 2008;149:627–637.
12. Rex DK, Johnson DA, Anderson JC, Schoenfeld PS, Burke CA, Inadomi JM. American College of Gastroenterology guidelines for colorectal cancer screening 2009 [corrected]. *Am J Gastroenterol* 2009;104:739–750.
13. Gaissmaier W, Gigerenzer G. Statistical illiteracy undermines informed shared decision making. *Z Evid Fortbild Qual Gesundhwes* 2008;102:411–413.
14. Gigerenzer G, Edwards A. Simple tools for understanding risks: from innumeracy to insight. *BMJ* 2003;327:741–744.
15. Fuchs CS, Giovannucci EL, Colditz GA, Hunter DJ, Speizer FE, Willett WC. A prospective study of family history and the risk of colorectal cancer. *N Engl J Med* 1994;331:1669–1674.
16. Murphy G, Shu XO, Gao YT, et al. Family cancer history affecting risk of colorectal cancer in a prospective cohort of Chinese women. *Cancer Causes Control* 2009;20:1517–1521.
17. Cannon Albright LA. Utah family-based analysis: past, present and future. *Hum Hered* 2008;65:209–220.
18. Skolnick MH. Prospects for population oncogenetics. In: Mulvihill JJ, Miller RW, Fraumeni JF, editors. Genetics of human cancer. New York: Raven Press, 1977:19–25.
19. Skolnick MH. The Utah genealogical data base: a resource for genetic epidemiology. In: Cairns J, Lyon JL, Skolnick MH, editors. Banbury report no 4; cancer incidence in defined populations. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory, 1980:285–297.
20. World Health Organization. International classification of diseases for oncology, 3rd ed. Geneva, Switzerland: World Health Organization, 2000.
21. Cannon-Albright LA, Thomas A, Goldgar DE, et al. Familiality of cancer in Utah. *Cancer Res* 1994;54:2378–2385.
22. McLellan T, Jorde LB, Skolnick MH. Genetic distances between the Utah Mormons and related populations. *Am J Hum Genet* 1984;36:836–857.
23. Jorde LB, Shortsleeve PA, Henry JW, Vanburen RT, Hutchinson LE, Rigley TM. Genetic analysis of the Utah population: a comparison of STR and VNTR loci. *Hum Biol* 2000;72:927–936.
24. Cannon-Albright LA, Farnham JM, Thomas A, Camp NJ. Identification and study of Utah pseudo-isolate populations-prospects for gene identification. *Am J Med Genet A* 2005;137A:269–275.
25. Jorde LB. Inbreeding in the Utah Mormons: an evaluation of estimates based on pedigrees, isonymy, and migration matrices. *Ann Hum Genet* 1989;53:339–355.
26. Lyon JL, Gardner JW, Klauber MR, Smart CR. Low cancer incidence and mortality in Utah. *Cancer* 1977;39:2608–2618.
27. Wylie JE, Mineau GP. Biomedical databases: protecting privacy and promoting research. *Trends Biotechnol* 2003;21:113–116.
28. Park Y, Freedman AN, Gail MH, et al. Validation of a colorectal cancer risk prediction model among white patients age 50 years and older. *J Clin Oncol* 2009;27:694–698.
29. Kurz DJ, Bernstein A, Hunt K, et al. Simple point-of-care risk stratification in acute coronary syndromes: the AMIS model. *Heart* 2009;95:662–668.
30. Elmore JG, Fletcher SW. The risk of cancer risk prediction: "What is my risk of getting breast cancer"? *J Natl Cancer Inst* 2006;98:1673–1675.
31. Freedman AN, Seminara D, Gail MH, et al. Cancer risk prediction models: a workshop on development, evaluation, and application. *J Natl Cancer Inst* 2005;97:715–723.
32. Ohman EM, Granger CB, Harrington RA, Lee KL. Risk stratification and therapeutic decision making in acute coronary syndromes. *JAMA* 2000;284:876–878.
33. Harrell FE Jr, Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. *JAMA* 1982;247:2543–2546.
34. Dupont WD, Plummer WD Jr. Understanding the relationship between relative and absolute risk. *Cancer* 1996;77:2193–2199.
35. NCCN 3.2010 Colon Cancer Clinical Practice Guidelines in Oncology. National Comprehensive Cancer Network, 2010. Available at: http://www.nccn.org. Accessed September 3, 2010.
36. Douglas FS, O'Dair LC, Robinson M, Evans DG, Lynch SA. The accuracy of diagnoses as reported in families with cancer: a retrospective study. *J Med Genet* 1999;36:309–312.
37. Jemal A, Siegel R, Ward E, et al. Cancer statistics, 2008. *CA Cancer J Clin* 2008;58:71–96.
38. Maul JS, Warner NR, Kuwada SK, Burt RW, Cannon-Albright LA. Extracolonic cancers associated with hereditary nonpolyposis colorectal cancer in the Utah Population Database. *Am J Gastroenterol* 2006;101:1591–1596.

# CHAPTER 5

# DISCUSSION

In this research we explored the role of family history in CRC using a combination of unique data resources. In this section of the dissertation the principal findings of each phase of the research will be presented. Insights that were gained in the process and future directions for research in this area will also be discussed.

## 5.1 Family-history-based constellation risk for CRC

The first phase of research explored family history-based risk for CRC in more depth than has been published to date, by not only examining the contribution of CRC-affected first-degree relatives, but from constellations of affected relatives from the first, second, and third degrees. We measured risk in a population of more than 2.3 million individuals with electronically available family history and cancer history data, and provided very precise individual risk estimates based on an individual's specific "constellation" of affected relatives. We hypothesized that the influence of extended relatives' CRC histories might have a substantial impact on risk to a proband, an effect that may have been missed in other studies that typically look only at first-degree relatives because of the difficulty of obtaining

accurate extended family histories. This study confirmed the role that affected first-degree relatives play in risk. For individuals with ≥1 affected first-degree relative, the relative risk was 2.05 (95% CI, 1.96 –2.14), which was consistent with previously published estimates. Our estimates were generally slightly lower than other similar published estimates and had tighter confidence intervals, which may be due to the large numbers of individuals in our study as well as lower rates of cancer in the state. However, when there was no positive first-degree history and risk was only estimated considering the second and third degrees, we did not find a substantial increase in risk compared to those at average risk (RR = 1.0). (As a reminder, we considered a two-fold increase in risk or a relative risk of 2.0 to be a convenient cutoff for a "substantial" or "clinically significant" increase in risk.) When combined with a positive first-degree family history, however, the influence of affected extended relatives can increase risk to a degree that may be clinically significant for certain constellations. In general though, this research confirmed the conventional wisdom that first-degree relatives are the most important to consider in clinical assessments of family history for CRC risk evaluation.

Other findings from this phase of the research were that when a proband's mother and father are both affected with CRC, risk is much higher (FRR=4.97; 95% CI: 2.72-8.34) than in those with ≥ any other 2 affected first-degree relatives (FRR=3.21; 95% CI,2.87–3.58). While the difference is not quite statistically significant at the 0.05 level (p = 0.07), it is an interesting finding that has not been previously published and could result from gene–gene interaction, gene–environment interaction, or a combination of both. In addition, we found that

although an age at diagnosis <50 years typically has been used as a cutoff for early onset, even diagnosis between 60 and 69 years of age in affected first-degree relatives increases risk to a degree that is clinically significant and perhaps has not been highlighted before. One of the most useful aspects of this research may have been the comprehensive set of relative risk estimates for constellations of affected family members from the first to the third degree that were produced. We also provided estimates for different levels of known family history (e.g., just first degree or first and second degree) to accommodate clinicians with varying levels of knowledge of a patient's family history. This first phase of the research provided a solid foundation for the other two phases in the project and also makes a contribution to the literature concerning family history-based risk for CRC.

## 5.2 Screening and surveillance compliance through colonoscopy in those at increased risk for CRC

In this second phase of research we combined data from the UPDB and Intermountain Healthcare to compare colonoscopy screening and surveillance rates by level of risk for CRC based on age, personal history of adenomatous polyps or CRC, and family history of CRC. Common guidelines for those at increased risk based on these risk factors were adapted to set standards for appropriate screening frequency. Previous studies have compared screening rates in those with, and without, a positive family history, but the sample sizes were generally low, highly selected, and family history and CRC screening data were self-reported. Our study

improved on each of these limitations, making use of the recently created linkage between the UPDB and Intermountain Healthcare data.

We found that colonoscopy rates generally increase with familial relative risk level, which indicates that a positive family history tends to have some effect on screening behavior. However, according to recommended guidelines, there is still much room for improvement in how frequently those at increased risk (as well as those at average risk) are screened based on their family history. This finding also applies to those at increased risk based on a history of adenoma or personal history of CRC. An important limitation of this study is that we were not able to measure colonoscopies that occurred outside the Intermountain system, which could have led to underestimation of the rates. A study such as this was not necessary to recognize that screening rates are not at recommended levels, even in those at increased risk. However, this study was unique in being able to quantify these rates in such a large population with electronically available family history, colonoscopy, adenoma, and surgically resected CRC data. It may help other researchers to explore innovative uses of combined data resources such as these. In the ideal research world a patient would receive care in a single-payer, single healthcare network where a record of all health events and interactions with the system would be electronically documented in a structured and coded fashion. In the real world where patients have interactions with multiple health networks and data are stored in various formats and levels of structure, it would still be worthwhile to integrate records across systems. With the rise of health information exchanges and projects such as the Federated Utah Research and Translational Health e-Repository (FURTHeR), the

informatics infrastructure for the Center for Clinical and Translational Science at the University of Utah, this is becoming more and more of a reality.

## 5.3 How well does family history predict who will get colorectal cancer?

The final phase of the research intended to explore the value of family history in clinical practice through a retrospective cohort study to determine how well knowing a patient's risk at a point in time helps predict whether they will develop CRC in the next 20 years. This retrospective study design is essentially the same as a prospective study leading some to characterize these types of studies as "pseudo-prospective" trials. Few prospective studies have been reported for family history's role in CRC, and those that exist have small sample sizes and self-reported family history data. No retrospective cohort study has been published in this area. This study, following more than 430,000 individuals over 20 years, is just one of many examples of how the UPDB can facilitate research that could not be conducted in many, if any, other places in the world.

The main finding of this study was that family history, by itself, is not a strong predictor of exactly who will develop CRC within 20 years. Despite the elevated relative risk estimates found in the first phase of the project and in the literature, this result was not surprising based on the fact that CRC is a relatively rare disease. In order for a risk factor to have strong predictive value by itself, it would need to produce relative risk estimates many times greater than what we found. Absolute risk, although not mentioned in the literature as frequently, is much easier to interpret and understand. However, the estimates are often much

smaller and perhaps less impressive than relative risk estimates, which may be factors in why they are not as commonly used.

Although overall family history did not appear to have strong predictive value at the individual level, in those at the highest levels of risk it has the capability to predict the percentage of individuals among particular risk categories who will develop the disease. In other words, stratification using a tool like absolute risk can identify those who are at very high risk, and a good percentage of those at very high risk end up developing the disease. Focusing screening efforts in those at higher levels of risk, for instance based on absolute risk stratification, would result in detecting more cases than uniform screening in the entire general population.

In comparing screening guidelines from the National Comprehensive Cancer Network for those with a positive family history of CRC, we found consistency in the guidelines with absolute risk patterns we observed; the most aggressive screening recommendations are associated with the highest levels of absolute risk.

Our conclusions were that it is unlikely that a clinically useful model can be created to predict who will develop CRC at an individual level, with a reasonable degree of accuracy, using just family history and age. However, absolute risk predicts fairly well the numbers of individuals in certain risk categories that will develop the disease over a 20-year period. Using estimates such as these in a clinical setting could help identify those who are more likely to develop the disease, and therefore where screening efforts are more likely to be productive.

## 5.4 Additional insights and lessons learned

In addition to the findings already presented there were other insights gained from this research that may be beneficial for those exploring similar paths. First, the data resources available in Utah are incredibly unique. There are few places with so much clinical, family history, and cancer registry data available electronically for research. Ties between the UPDB and health care institutions such as Intermountain Healthcare and the University of Utah Health Sciences Center facilitate rich research opportunities. Luckily projects such as FuRTHER will help to tie these unique data resources even more closely together to facilitate more advanced research and requiring less administrative overhead.

While the data resources allowed us to explore interesting questions posed in this research, it is difficult to predict risk in cancer. The known risk factors are generally not strong enough for accurate prediction in typical clinical practice. It is interesting to note that other risk prediction models with very modest predictive values (e.g., AUC = 0.60) are commonly used in clinical practice for risk prediction. They may have their places in clinical practice, but many may not be aware that their predictive value is so limited.

While family history has been cited as one of the most important risk factors for CRC, on an absolute scale, it is still not a strong predictor. However, as previously described, even though it is not a strong predictor overall, family history still appears to have some value in stratifying those at risk. Better tools to collect and make use of family history in a clinical setting should be explored for CRC and other conditions where family history may have an even greater impact.

Although clinical data resources in the state, such as those at Intermountain Healthcare, are some of the most complete in the nation, using EMR data in a research project such as this can be challenging. Although many of the challenges were known at the beginning of the project, their extent was not fully appreciated. An important objective early in the project was to build a comprehensive risk model for CRC that not only utilized family history data, but other risk factors based on Intermountain clinical and administrative data. We believed that a matched case-control design was most appropriate, based on previous UPDB research involving family history. In addition, we approached the development of a risk model from the perspective of the data mining/machine learning world more than from epidemiology. However, we found these techniques from machine learning inadequate to deal with certain issues, for instance to appropriately handle data based on a matched case-control design.

The problems continued and deepened. In the initial comprehensive model study design we did not select controls with appropriately matched exposure windows. After going through the labor to collect and analyze data, we realized we had some made some "textbook" epidemiological study design errors. With the proper help of an epidemiologist we created a modified case-control study design. Even with the modified study design we encountered a number of challenges. It is very easy to make study design mistakes that will cause bias or produce unreliable results in the analysis. Considering that we were selecting Intermountain patients as controls (seen as either inpatients or outpatients), rather than selecting them from the general community, we potentially were creating "Berkson's bias," where

selected controls are actually sicker than those individuals in the general community.

These issues highlighted the need for proper epidemiological and statistical assistance in the early planning stages of the project. In retrospect it seems simple, but research objectives evolve and finding assistance that is well-matched to accomplish these objectives can be challenging. Opportunities are becoming increasingly common because of the availability of data resources and conducting this type of research effectively is facilitated by collaboration between those with expertise in fields such as biomedical informatics, epidemiology, and statistics. How to effectively build cross-functional teams to answer research questions such as these may be a worthwhile area for future exploration.

Perhaps the most significant problem we encountered, which was a factor that produced such a complicated study design, was the issue of missing clinical data. Although Intermountain has a longer history of electronically captured data than most institutions in the US, data were not reliably available before 2001 for the risk factors considered. There were certain data elements available long before that, but not being able to measure all the relevant risk factors reliably for the same period of time caused difficulty. Other problems included multiple sources of the same basic data that were not always in sync, multiple master patient identifiers for the same patient, and missing or erroneous clinical data. Also, it was difficult to determine if certain data were missing, for example, because the patient did not have a particular condition in the past, or simply because it was just not documented (i.e., the “none” versus “unknown” problem). Intermountain provides a substantial

amount of care in the state, but it was impossible to determine if a patient had care provided outside the system and therefore documentation might be missing. Some of the known risk factors for CRC are just not available as part of EMR data (e.g., red meat consumption), are not documented reliably (e.g., alcohol use), or are not contained in structured data that we could easily collect. With regard to medications, it was impossible based on the data available to determine if a patient filled a prescription, let alone actually took the medication.

At the most basic level, we could not determine the lifetime exposure for the risk factors, or even for a reasonable period of time, based on the EMR data. Although considerable time was spent in trying to address these issues and make the best of the situation, the problems halted this track of the research. EMR data may be useful depending on the research question, but for many studies there are substantial limitations. Understanding the limitations of the data and working with individuals who have deep experience in study design to mitigate or work around them is essential.

Another lesson learned was that how risk is communicated is important. The literature around risk factors in CRC focuses on relative risk, which is not as intuitive as methods such as absolute risk. As previously described, relative risk may appear to be elevated based on a particular risk factor but because of the overall low incidence of a disease, the actual risk (as reflected by absolute risk—a probability over a defined period of time) may still be low. While these limitations of relative risk are documented, it appears that even among researchers and clinicians, these issues are not well understood.

## 5.5 Relevance for biomedical informatics

The Department of Biomedical Informatics at the University of Utah has three academic tracks: (1) Clinical research and translational informatics, (2) Health care/clinical informatics, and (3) Public health informatics. This research project has helped advance the field of biomedical informatics in each of these areas. First, with respect to clinical research and translational informatics, we extended genetic epidemiological tools and techniques to measure risk more comprehensively and with more granularity than reported in the past. We explored how disparate data resources (both clinical- and research-oriented) could be combined to answer clinically-relevant questions. This involved investigating the quality of clinical data resources for a secondary research use as well. We also explored the feasibility of machine learning techniques in disease risk prediction.

In the health care/clinical informatics area, this research provides foundational knowledge for other risk prediction-related projects involving clinical decision support, human-computer interaction, patient portals and personal health records, and ambulatory EMR order entry. It particularly raises the question of how risk factors can be better captured as part of an electronic health record, particularly through more direct patient involvement.

Last, in public health informatics we conducted a population-based retrospective cohort study to measure whether a risk factor was clinically significant, which was a novel use of the data resource. We also applied absolute risk on a population level to predict how many individuals at certain levels of risk will develop a disease over time. In summary, this research has advanced the field of

biomedical informatics through the innovative application of tools and techniques on unique data resources to explore risk for disease.

## 5.6 Future directions

It has been noted that a unique set of data resources are available in Utah, without which this research would not have been possible. However, the UPDB has strict guidelines around appropriate uses of the data. Although it would be feasible to identify individuals or families at increased risk for a disease such as CRC based on risk estimates based on UPDB family history and cancer data, contacting these individuals (i.e., using the data for clinical care purposes) is prohibited. Therefore, other methods must be explored for capturing and utilizing family history data in clinical practice. In actuality, according to HIPAA, a physician may disclose family history information about a patient to another health care provider for the purpose of treating another patient (e.g., a family member). In a health system such as Intermountain Healthcare where many patients may be related, one might foresee an opportunity to collect family history relationships and hopefully with the patient's consent, use that information and existing medical records to build a network of family health histories that could be used in clinical care.

In addition, several other opportunities have been identified for future research. One would be to validate the absolute risk estimates produced in the third phase of the project in an independent population to determine whether our techniques are sound and how well the Utah population compares to other populations with different characteristics. A second opportunity would be to

perform a cost/benefit analysis using our findings on risk levels to determine optimal screening recommendations, for instance based on a technique like colonoscopy. Another direction would be to build a web-based tool for clinicians to use to estimate absolute risk for CRC based on a patient's age and family history. A further step along this path would be to integrate a tool such as this into an EMR along with actionable recommendations. More generally, further work could explore how to communicate risk most effectively among clinicians and patients. Finally, further exploration would be valuable to determine what types of research questions EMR data are best suited for and how the quality of clinical data sources may be improved to facilitate research.

# APPENDIX

# SUPPLEMENTARY TABLES FOR POPULATION-BASED FAMILY HISTORY-SPECIFIC RISKS FOR COLORECTAL CANCER: A CONSTELLATION APPROACH

A.1: Familial relative risks for probands considering only first- and second-degree relative family histories. 'NA' means this degree was not considered.

| Number of affected FDRs | Number of affected SDRs | Number of affected TDRs | N (probands) | FRR | Lower CI (95% level) | Upper CI (95% level) |
|---|---|---|---|---|---|---|
| 0 | 0 | NA | 1,965,853 | 0.86 | 0.84 | 0.88 |
| 0 | 1 | NA | 224,609 | 1.05 | 0.99 | 1.11 |
| 0 | 2 | NA | 33,407 | 1.20 | 1.05 | 1.38 |
| 0 | ≥3 | NA | 8527 | 1.48 | 1.11 | 1.93 |
| 1 | 0 | NA | 65,192 | 1.82 | 1.72 | 1.93 |
| 1 | 1 | NA | 16,760 | 2.12 | 1.90 | 2.35 |
| 1 | 2 | NA | 3776 | 2.31 | 1.80 | 2.93 |
| 1 | ≥3 | NA | 1361 | 3.37 | 2.20 | 4.93 |
| 2 | 0 | NA | 4699 | 2.78 | 2.39 | 3.22 |
| 2 | 1 | NA | 1644 | 2.59 | 1.93 | 3.40 |
| 2 | 2 | NA | 433 | 6.26 | 4.16 | 9.05 |
| 2 | ≥3 | NA | 190 | 9.63 | 4.97 | 16.82 |
| ≥3 | 0 | NA | 509 | 4.41 | 3.03 | 6.19 |
| ≥3 | 1 | NA | 221 | 8.63 | 5.59 | 12.74 |
| ≥3 | 2 | NA | 94 | 1.77 | 0.21 | 6.39 |
| ≥3 | ≥3 | NA | 52 | 4.24 | 0.51 | 15.32 |

A.2: Familial relative risks for probands considering first-, second-, and third-degree relative family histories.

| Number of affected FDRs | Number of affected SDRs | Number of affected TDRs | N (probands) | RR | Lower CI (95% level) | Upper CI (95% level) |
|---|---|---|---|---|---|---|
| 0 | 0 | 0 | 1,470,367 | 0.83 | 0.81 | 0.86 |
| 0 | 0 | 1 | 356,128 | 0.86 | 0.82 | 0.91 |
| 0 | 0 | 2 | 94,696 | 0.95 | 0.87 | 1.03 |
| 0 | 0 | ≥3 | 44,662 | 1.08 | 0.97 | 1.20 |
| 0 | 1 | 0 | 132,580 | 0.95 | 0.88 | 1.03 |
| 0 | 1 | 1 | 57,850 | 1.09 | 0.98 | 1.21 |
| 0 | 1 | 2 | 20,321 | 1.33 | 1.13 | 1.55 |
| 0 | 1 | ≥3 | 13,858 | 1.21 | 0.98 | 1.48 |
| 0 | 2 | 0 | 15,756 | 1.18 | 0.97 | 1.43 |
| 0 | 2 | 1 | 9199 | 1.12 | 0.84 | 1.45 |
| 0 | 2 | 2 | 4391 | 1.25 | 0.84 | 1.79 |
| 0 | 2 | ≥3 | 4061 | 1.48 | 0.98 | 2.16 |
| 0 | ≥3 | 0 | 2938 | 1.48 | 0.91 | 2.26 |
| 0 | ≥3 | 1 | 2157 | 1.56 | 0.89 | 2.53 |
| 0 | ≥3 | 2 | 1312 | 1.93 | 0.93 | 3.55 |
| 0 | ≥3 | ≥3 | 2120 | 1.02 | 0.41 | 2.09 |
| 1 | 0 | 0 | 41,369 | 1.76 | 1.63 | 1.89 |
| 1 | 0 | 1 | 15,008 | 1.90 | 1.70 | 2.12 |
| 1 | 0 | 2 | 5560 | 1.90 | 1.59 | 2.25 |
| 1 | 0 | ≥3 | 3255 | 2.01 | 1.61 | 2.47 |
| 1 | 1 | 0 | 8836 | 1.88 | 1.59 | 2.20 |

A.2: continued

| Number of affected FDRs | Number of affected SDRs | Number of affected TDRs | N (probands) | RR | Lower CI (95% level) | Upper CI (95% level) |
|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 4685 | 1.97 | 1.59 | 2.41 |
| 1 | 1 | 2 | 1882 | 2.50 | 1.87 | 3.28 |
| 1 | 1 | ≥3 | 1357 | 3.28 | 2.44 | 4.31 |
| 1 | 2 | 0 | 1669 | 2.37 | 1.58 | 3.43 |
| 1 | 2 | 1 | 1006 | 1.98 | 1.15 | 3.17 |
| 1 | 2 | 2 | 523 | 2.70 | 1.44 | 4.62 |
| 1 | 2 | ≥3 | 578 | 2.38 | 1.19 | 4.26 |
| 1 | ≥3 | 0 | 453 | 2.79 | 1.12 | 5.76 |
| 1 | ≥3 | 1 | 380 | 1.69 | 0.46 | 4.34 |
| 1 | ≥3 | 2 | 206 | 5.32 | 2.14 | 10.96 |
| 1 | ≥3 | ≥3 | 322 | 5.20 | 2.24 | 10.24 |
| 2 | 0 | 0 | 2613 | 2.96 | 2.41 | 3.60 |
| 2 | 0 | 1 | 1197 | 2.24 | 1.59 | 3.06 |
| 2 | 0 | 2 | 528 | 1.77 | 0.97 | 2.96 |
| 2 | 0 | ≥3 | 361 | 4.82 | 3.18 | 7.02 |
| 2 | 1 | 0 | 717 | 2.54 | 1.57 | 3.89 |
| 2 | 1 | 1 | 439 | 1.80 | 0.82 | 3.41 |
| 2 | 1 | 2 | 241 | 1.57 | 0.51 | 3.67 |
| 2 | 1 | ≥3 | 247 | 4.67 | 2.72 | 7.47 |
| 2 | 2 | 0 | 172 | 6.71 | 3.67 | 11.26 |
| 2 | 2 | 1 | 95 | 5.64 | 1.83 | 13.16 |
| 2 | 2 | 2 | 88 | 4.63 | 0.96 | 13.53 |

A.2: continued

| Number of affected FDRs | Number of affected FDRs | Number of affected FDRs | Number of affected FDRs | Number of affected FDRs | Number of affected FDRs | Number of affected FDRs |
|---|---|---|---|---|---|---|
| 2 | 2 | ≥3 | 78 | 7.05 | 2.59 | 15.34 |
| 2 | ≥3 | 0 | 40 | 3.31 | 0.08 | 18.42 |
| 2 | ≥3 | 1 | 48 | 11.96 | 3.26 | 30.62 |
| 2 | ≥3 | 2 | 46 | 16.37 | 4.46 | 41.91 |
| 2 | ≥3 | ≥3 | 56 | 8.21 | 1.69 | 24.00 |
| ≥3 | 0 | 0 | 246 | 2.96 | 1.42 | 5.44 |
| ≥3 | 0 | 1 | 127 | 4.21 | 1.82 | 8.30 |
| ≥3 | 0 | 2 | 52 | 1.33 | 0.03 | 7.39 |
| ≥3 | 0 | ≥3 | 84 | 9.63 | 5.26 | 16.15 |
| ≥3 | 1 | 0 | 98 | 12.39 | 7.08 | 20.12 |
| ≥3 | 1 | 1 | 36 | 2.41 | 0.06 | 13.45 |
| ≥3 | 1 | 2 | 28 | 10.04 | 2.07 | 29.35 |
| ≥3 | 1 | ≥3 | 59 | 5.61 | 1.82 | 13.08 |
| ≥3 | 2 | 0 | 41 | 0.00 | NA† | NA† |
| ≥3 | 2 | 1 | 16 | 22.05 | 2.67 | 79.64 |
| ≥3 | 2 | 2 | 18 | 0.00 | NA† | NA† |
| ≥3 | 2 | ≥3 | 19 | 0.00 | NA† | NA† |
| ≥3 | ≥3 | 0 | 11 | 15.90 | 0.40 | 88.58 |
| ≥3 | ≥3 | 1 | 11 | 7.19 | 0.18 | 40.05 |
| ≥3 | ≥3 | 2 | 20 | 0.00 | NA† | NA† |
| ≥3 | ≥3 | ≥3 | 10 | 0.00 | NA† | NA† |

†When the observed and expected numbers of affected probands are both 0, a confidence interval is not applicable.

A.3: Familial relative risks for probands with affected first-degree relatives diagnosed with CRC at various ages. 'NA' means this degree was not considered.

| | Number of affected FDRs | Number of affected SDRs | Number of affected TDRs | N (probands) | RR | Lower CI (95% level) | Upper CI (95% level) |
|---|---|---|---|---|---|---|---|
| ≥1 affected FDR diagnosed <50 years of age | ≥1 (dx age <50) | NA | NA | 6291 | 3.31 | 2.79 | 3.89 |
| ≥1 affected FDR diagnosed between 50 and 59 years of age | ≥1 (dx age 50-59) | NA | NA | 12,094 | 2.53 | 2.24 | 2.85 |
| ≥1 affected FDR diagnosed ≥50 years of age | ≥1 (dx age ≥50) | NA | NA | 89,340 | 2.02 | 1.93 | 2.11 |
| ≥1 affected FDR diagnosed <60 years of age | ≥1 (dx age <60) | NA | NA | 18,199 | 2.69 | 2.43 | 2.96 |
| ≥1 affected FDR diagnosed between 60 and 69 years of age | ≥1 (dx age 60-69) | NA | NA | 25,084 | 2.22 | 2.04 | 2.40 |
| ≥1 affected FDR diagnosed ≥60 years of age | ≥1 (dx age ≥60) | NA | NA | 78,629 | 1.99 | 1.90 | 2.09 |
| ≥1 affected FDR diagnosed <70 years of age | ≥1 (dx age <70) | NA | NA | 42,452 | 2.33 | 2.18 | 2.48 |
| ≥1 affected FDR diagnosed between 70 and 79 years of age | ≥1 (dx age 70-79) | NA | NA | 32,445 | 1.97 | 1.83 | 2.12 |
| ≥1 affected FDR diagnosed ≥70 years of age | ≥1 (dx age ≥70) | NA | NA | 56,065 | 1.97 | 1.86 | 2.08 |

A.4: Familial relative risks for probands with affected second-degree relatives diagnosed with CRC at various ages. 'NA' means this degree was not considered.

| | Number of affected FDRs | Number of affected SDRs | Number of affected TDRs | N (probands) | FRR | Lower CI (95% level) | Upper CI (95% level) |
|---|---|---|---|---|---|---|---|
| ≥1 affected SDR diagnosed <50 years of age | NA | ≥1 (dx age <50) | NA | 19,616 | 1.84 | 1.61 | 2.09 |
| ≥1 affected SDR diagnosed ≥50 years of age | NA | ≥1 (dx age ≥50) | NA | 275,779 | 1.24 | 1.18 | 1.30 |
| ≥1 affected SDR diagnosed <60 years of age | NA | ≥1 (dx age <60) | NA | 58,351 | 1.56 | 1.43 | 1.70 |
| ≥1 affected SDR diagnosed ≥60 years of age | NA | ≥1 (dx age ≥60) | NA | 245,140 | 1.24 | 1.18 | 1.30 |
| ≥1 affected SDR diagnosed <70 years of age | NA | ≥1 (dx age <70) | NA | 136,228 | 1.37 | 1.28 | 1.45 |
| ≥1 affected SDR diagnosed ≥70 years of age | NA | ≥1 (dx age ≥70) | NA | 178,263 | 1.25 | 1.18 | 1.32 |

A.5: Familial relative risks for probands with various affected first-degree relationship types. 'NA' means this degree was not considered.

| | Number of affected FDRs | Number of affected SDRs | Number of affected TDRs | N (probands) | FRR | Lower CI (95% level) | Upper CI (95% level) |
|---|---|---|---|---|---|---|---|
| ≥1 affected parent | ≥1 (parent) | NA | NA | 31,619 | 1.96 | 1.77 | 2.16 |
| ≥1 affected sibling | ≥1 (sibling) | NA | NA | 47,272 | 1.96 | 1.86 | 2.07 |
| ≥1 affected brother | ≥1 (brother) | NA | NA | 24,117 | 1.99 | 1.84 | 2.14 |
| ≥1 affected sister | ≥1 (sister) | NA | NA | 25,048 | 2.04 | 1.90 | 2.19 |
| ≥1 affected offspring | ≥1 (offspring) | NA | NA | 18,644 | 3.06 | 2.76 | 3.38 |
| Affected mother | 1 (mother) | NA | NA | 15,589 | 2.03 | 1.78 | 2.30 |
| Affected father | 1 (father) | NA | NA | 16,480 | 1.96 | 1.68 | 2.27 |
| Mother and father both affected | Mother and father both affected | NA | NA | 450 | 4.97 | 2.72 | 8.34 |
| ≥1 affected male first-degree relative | ≥1 (male) | NA | NA | 49,716 | 2.08 | 1.96 | 2.21 |

A.5: continued

| | Number of affected FDRs | Number of affected SDRs | Number of affected TDRs | N (probands) | FRR | Lower CI (95% level) | Upper CI (95% level) |
|---|---|---|---|---|---|---|---|
| ≥1 affected female first-degree relative | ≥1 (female) | NA | NA | 49,460 | 2.14 | 2.02 | 2.27 |
| Males with ≥1 affected first-degree relative | ≥1 (probands male) | NA | NA | 48,751 | 1.96 | 1.84 | 2.09 |
| Females with ≥1 affected first-degree relative | ≥1 (probands female) | NA | NA | 46,180 | 2.12 | 2.00 | 2.26 |

A.6: Sidedness analysis to compare risk between situations where affected relatives are concentrated on one side of family versus both sides.

| Number of affected maternal FDRS | Number of affected paternal FDRs | Number of affected 'both/neither' FDRs | Number of affected maternal SDRs (M) | Number of affected paternal SDRs (P) | Number of affected 'both/neither' SDRs | Number of affected maternal TDRs (M) | Number of affected paternal TDRs (P) | Number of affected 'both/neither' TDRs | N (probands) | FRR | Lower CI (95% level) | Upper CI (95% level) | Notes |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 0 | (≥2M and 0P) OR (0M and ≥2P) | | NA | NA | NA | NA | 11,597 | 1.41 | 1.13 | 1.73 | |
| 0 | 0 | 0 | ≥1 | ≥1 | NA | NA | NA | NA | 5414 | 1.47 | 1.06 | 1.98 | |
| | | | | | | | | | | | | | |
| 0 | 0 | 0 | 0 | 0 | 0 | (≥2M and 0P) OR (0M and ≥2P) | | NA | 66,969 | 1.00 | 0.91 | 1.09 | N=66,994 before removing those with looped relatives |
| 0 | 0 | 0 | 0 | 0 | 0 | ≥1 | ≥1 | NA | 36,048 | 0.87 | 0.78 | 0.98 | N=36,072 before removing those with looped relatives |

We estimated similar FRRs under both conditions and compared them. Probands with family history restricted to one side were defined as having ≥2 affected maternal relatives and 0 affected paternal relatives or 0 affected maternal relatives and at ≥2 affected paternal relatives. Probands with family history on both sides were defined as having ≥1 number of affected maternal relatives and ≥1 affected paternal relatives from a particular degree. These estimates allowed us to compare family histories that included equal numbers of affected relatives of a particular degree, where in one group the affected relatives were strictly on one side and in the second were distributed on both sides. Affected relatives were categorized as maternal, paternal, or both/neither. Both/neither contained relative types such as children and siblings (who have both maternal and paternal genetic contributions) and these were excluded from the FRR calculation. Also, for probands who had a loop in their pedigree (e.g., the grandparents of the proband included 2 sisters who married 2 brothers) we also did not consider the proband.