ABSTRACT

CUBBAGE, DANIEL FREDERICK. SIMULATION OF COLORECTAL CANCER: THE NATURAL HISTORY OF THE DISEASE. (Under the direction of Dr. Stephen D. Roberts.)

This thesis presents a comprehensive, fully verified, and validated model of the natural history of Colorectal Cancer (CRC). CRC is the fourth most common type of cancer and the second leading cause of cancer death among both men and women. Individuals who develop CRC often fail to detect symptoms until the cancer is in an advanced stage. There are a number of screening methods designed to detect CRC in its early stages or prevent CRC by identifying and removing adenomous polyps, a pre-malignant form. However, because of the long latency of CRC and the time needed for clinical trials, it is not practical to provide clinical trials of all the screening and treatment strategies for CRC. Models offer an alternative means to analyze of screening/surveillance recommendations. Before considering any CRC medical interventions with a model, a model of the natural history of CRC is of fundamental importance.

A model of the natural history of CRC requires a compromise of knowledge of CRC and data describing it. These compromises are described by modeling assumptions regarding the actual process of CRC development. To summarize the outcomes, the two primary measurements are the costs associated with the treatments and the years of life, or life-years. These measurements can be modified in several ways, by discounting or adjusting life-years to reflect the quality of life based upon different states of health. Within the medical decision-making community, two primary types of models for CRC have been developed, Markov models and discrete-event simulations. While the Markov models are easy to build and provide a basic analysis of the impact of screening, a more flexible, but more complex, approach is the discrete-event simulation model. One discrete-event simulation is the Vanderbilt Model that is the predecessor to the Vanderbilt-NC State model presented in this thesis.

The Vanderbilt-NC State model improves the original Vanderbilt model with enhanced features such as database storage of inputs and Excel outputs. It also models additional important factors such as race, family history, reference year, risk effects, and histology. The object-oriented design allows the discrete-event simulation to follow the adenomas and people through the system, rather than forcing these objects into a process flow. When an individual is created, his natural death and first adenoma development are scheduled. The adenoma object is then responsible for its own progression up to cancer and potential cancer death. It also schedules the next adenoma, which will then follow its own timeline through the simulation.

Once the model was constructed, it had to be verified and validated against cancer information from clinical studies. A detailed calibration procedure was implemented to match the model output with the cancer incidence, adenoma prevalence, and people with adenomas. In the process of validation, the Vanderbilt-NC State model outcomes are compared to data from other sources that were not used in the fitting process. The model's output was compared to the cumulative risk of getting cancer obtained from a national CRC database (SEER). The model was also compared to a previous simulation that sampled from an adjusted lifetime that had the risk of CRC eliminated. This comparison was performed to validate the life-year gain associated with the elimination of CRC. Finally, the Vanderbilt-NC State model was compared to the previous Vanderbilt model. In each of the comparisons, the Vanderbilt-NC State model was consistent with the patterns and magnitudes of outcomes found in the external data.

Once the model matched the literature results, analysis could be performed to determine the impact of CRC. According to the model, CRC reduces average lifespan by 0.24 years. This average loss of life comes from approximately 2.5% of the population losing an average of over 10 years of life. The average cost associated with the diagnosis and treatment of the disease is $2,188 per person.

The model is the first comprehensive model of CRC and can be extended to consider screening and other medical interventions without direct experimentation using patients.

SIMULATION OF COLORECTAL CANCER: THE NATURAL HISTORY OF DISEASE

By

DANIEL CUBBAGE

A thesis submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Master of Science

Industrial Engineering

Raleigh

2004

APPROVED BY:

_____ _____

_____
Chair of Advisory Committee

Biography

Daniel Cubbage is a graduate student with a major in Industrial Engineering at North Carolina State University, where he will receive his Master of Science degree in August 2004. He also received his bachelor's degree in Industrial Engineering from North Carolina State in May of 2000, graduating Magna Cum Laude. He is from Apex, North Carolina where his parents reside.

Acknowledgements

I would like to thank Dr. Steve Roberts for his support and guidance through the long process of completing this thesis. I would also like to express my gratitude to Dr. Reid Ness for the extensive medical insight he has provided. I would also like to thank Bob Klein for his work on the previous model as well as the help with questions on the current model. Additional thanks are due to Lijun Wang who has provided extensive help along the way. Ann Holmes is also due thanks for the help with the discounting portions of the thesis.

Table of Contents

List of Tables

List of Figures

1. Introduction to Colon Cancer

Colorectal cancer (CRC) is second only to lung cancer as a cause of cancer-related death in both men and women. The lifetime risk of colon cancer in the general population is about 6 percent. Men are at only slightly greater risk than women (only rectal cancer in men is significantly higher than women). The vast majority of colon cancer cases occur in people over age sixty. If the person is a first-degree relative (parent, sibling, or child) of someone who had a colon cancer, then their risk is increased two to three fold. Furthermore, heredity also affects the tendency to have colon cancer.

There is widespread belief and some evidence that diet and environment significantly influence the incidence of colon cancer. High-fat and low-fiber diet along with excessive weight gain, obesity, and sedentary lifestyle all seem to be influential factors (Lieberman et al. 2003). People with the highest incidence of colon cancer in the world are those living in the United States or Europe, almost regardless of their ethnic origin, which tends to support this conclusion.

The purpose of this thesis is to model and understand the natural course of colon cancer (CRC). However it is first important to understand the colon and how colon cancer can form.

1.1 A Brief Orientation to the Colon

The colon is the last segment of the human digestive system. After food is digested in the stomach, it enters the small intestine where the nutrients are absorbed through digestion. The indigestible part is then passed to the large intestine and eventually expelled from the body through the rectum using the specialized muscles and nerves in the anus which acts as a valve. See Figure 1 below (from the individual's vantage):

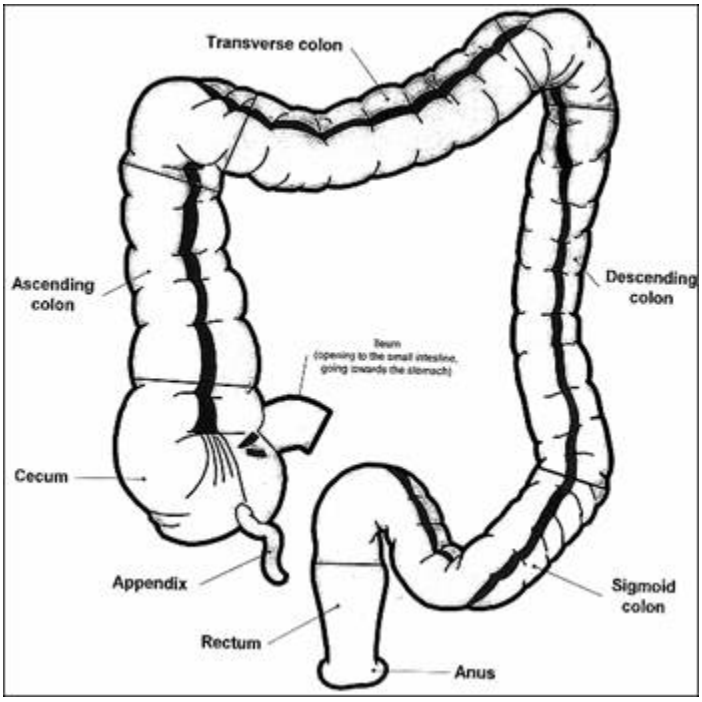


Figure 1: A diagram of the colon

The large intestine consists of the colon and the rectum (called the terminal extraperitoneal segment). It is called "large" because its diameter is roughly 2 to 2.5 inches in diameter in the cecum and right colon (although it narrows to about an inch at the end of the rectum). In general, the purpose of the colon is to absorb the water and mineral salts from undigested food, with the residue passing as feces towards the rectum where it is to be excreted. The length of the colon is typically about five feet and is composed of five sections: ascending colon, transverse colon, descending colon, sigmoid colon, and rectum. Unlike the small intestine, which is almost sterile, the colon has significant bacteria (which have some very beneficial effects). In fact, about 90% of the dry weight of the stools is bacteria and not undigested food.

The colon is somewhat like a corrugated tube. It is elastic, flexible and can expand and move. There are several named junctures in the colon, however the two main "kinks" are the right colic flexure (between the ascending and transverse colon) and the left colic flexure (between the transverse and descending colon, and also called the splenic

flexure). The sigmoid maintains more of an "s" shape. Figure 2 below gives an inside view of a portion of the transverse normal colon.

Figure 2: An inside view of the transverse normal colon

Although the colon wall contains several layers of tissue, the inner lining or epithelium is of greatest interest relative to colon cancer because that is where most colon cancers begin. The colonic epithelium has a glandular appearance from the inside and acts principally to absorb water and secrete mucus. It is characterized by the long, thin pits called crypts which contain special cells.

The epithelial cells are sacrificial, surviving only for a brief period of time. Androuny (Androuny 2002) gives the following description: "There is an amazing turnover of epithelial cells, with the columnar absorptive and goblet cells having a lifetime of only a few days. These cells begin as undifferentiated (meaning without any recognizable distinguishing characteristics) cells in the deeper zones of the crypts and they move up to the surface where they take on their recognizable features. When cells reach the flat surface of the epithelium they begin to degenerate and eventually slough off into the lumen and become part of waste eliminated. The life cycle of a cell in this process is four to six days. The enteroendocrine cells probably survive a little longer than their columnar absorptive and goblet cell counterparts and probably migrate up the crypt (independently of them) as well. The epithelium lining the colon is thus replaced completely every four

to six days and because of the constant renewal process that goes on in the bowel, the epithelium is particularly sensitive to noxious substances."

In a person who is 60 years old, this means that a five day life cycle has occurred about 4000 times. When multiplied by the number of such cells, the total number is staggering. So it is not unexpected, that in the reproduction of the cells there will be errant genetic mutations. However, it is believed that most of these mutations are naturally disposed of through the body's natural policing at the cellular level. Nonetheless, the existence of polyps in the colon is fairly common in adults.

Inside or luminal views of the colon can be obtained from endoscopes, which are flexible, fiberoptic tubes about the thickness of a finger that are attached to a light source and have digital view components present at the tip. The endoscope is inserted into the anus to examine the interior portions of the colon. Views can be displayed on a monitor. Air can be expelled into the colon in front of the scope to improve the navigation and view. A colonoscopy employs an endoscope to examine the entire colon while a flexible sigmoidoscopy (shorter endoscope) only examines the rectum and most of the sigmoid portion of the colon. Sigmoidoscopy can be performed in a doctor's office with local anesthetic whereas the colonoscopy is performed in an endoscopy suite in a hospital with a more complete anesthetic. These scopes have played an important role in understanding the colon.

Both of these tools are used both to diagnose a problem once symptoms develop, as well as for screening. It is not the purpose of this thesis to consider the various screening, surveillance, and treatment options for colon cancer. Instead the focus here is on the natural course of colorectal cancer. This simulation will only look at the use of colonoscopy for diagnostic purposes once symptoms develop.

1.2 Adenomas and Colorectal Cancer (CRC)

1.2.1 Adenoma Incidence and Progression

The exact cause of colorectal cancer, like most cancers, is not fully understood. However, evidence has accumulated over several years that CRC develops from the adenomas present in the colon. A colon adenoma is typically an "unusual" collection of cells on the interior lining of the colon forming a tumor. Sometimes invisible, even to a trained eye, this collection usually begins as benign (not cancerous). Somehow these adenoma cells gain a "growth advantage" over other cells in the colon and the collection enlarges to the point they become polypoid, creating a protrusion inward from the lining of the colon. Polypoid adenomas are sometimes loosely referred to as polyps. However, polyps are more technically considered to be "normal" tissue, while adenomas are the cell collections leading to CRC whether polypoid or not. While it is possible for early, non-visible adenomas to become cancerous, the generally accepted view is that the adenoma tends to grow before becoming cancerous. Figure 3 is of a polypoid adenoma in the colon.

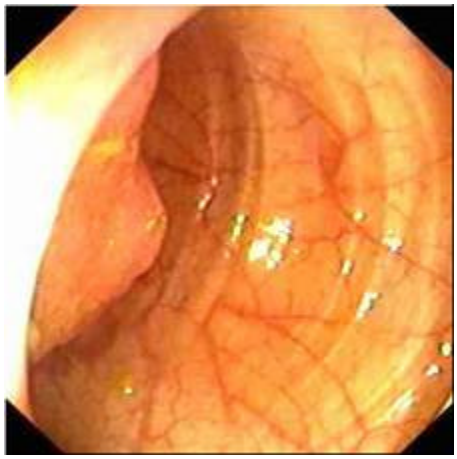


Figure 3: A polypoid adenoma in the colon

Adenomas do not produce symptoms but should be removed, because, if undetected, will typically continue to grow and become invasive. Since the adenomas are in direct contact with the fecal stream in which many carcinogens are present, these immortalized cells are now exposed to this non-sterile environment for a long period of time. It is this exposure which may contribute to their becoming malignant. Whatever the cause, if the adenoma survives long enough, it will become cancerous, even if it grows slowly in size. The larger adenomas are more at risk of becoming cancerous than small adenomas. Also correlated with increased cancer risk is the histology of the adenoma.

1.2.2 Histology of Adenomas

Along with size, the other main feature that is used to characterize an adenoma is its histology. Histology refers to the microscopic structure of the adenoma. Adenoma histology is classified based on the proportion of villous versus tubular architecture. Tubular histology is characterized by an alteration of the normal glandular structure so that crypts are lengthened and branching with lining cells that have activated metabolic and replicate function. Villous histology is characterized by densely arrayed, lengthened crypts that give the appearance of long fingerlike fronds. A microscopic view of these characteristics is shown below in Figure 4.

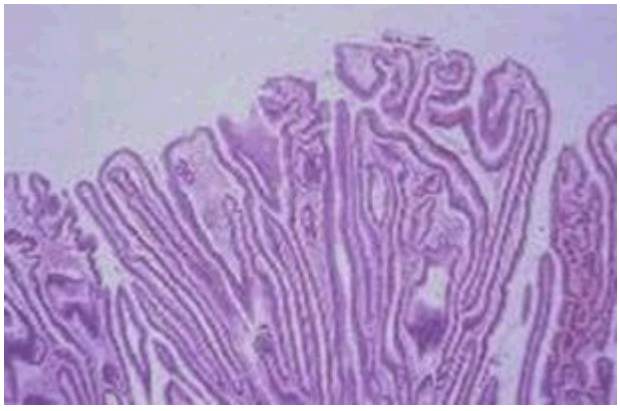


Figure 4: A microscopic view of the polyp glands

There are three main histological groupings:

- Tubular – Characterized by having 0-25% villous characteristics. This group makes up 70-85% of all polyps.
- Tubulovillous – Characterized by 25-75%, or 25% to 50% depending on the source, villous characteristics. This category makes up 10-25% of all polyps. A microscopic view of this is shown in Figure 5 below.

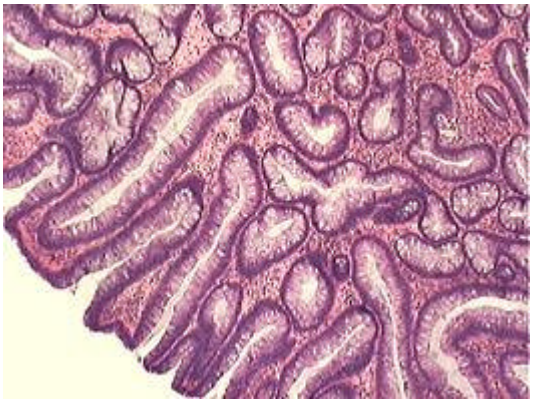


Figure 5: A microscopic view of a Tubulovillous polyp

- Villous – This category is 5% of all polyps and is made up of polyps containing more than 75% villous characteristics. A photograph of a polyp determined to be villous is shown in Figure 6. It is important to know that in general, histology can not always be determined by gross features, and requires microscopic analysis to determine.

Figure 6: A photograph of a villous polyp

These are arranged in order of increasing risk since the risk of malignancy is associated closely with the degree of villous characteristics.

1.2.3 Development of CRC

Interestingly, adenomas in the colon on the left side of the body have a lower chance of being cancerous than those in the right colon. Figure 7 is a picture of cancer in the colon:

Figure 7: A photograph of Colon Cancer

The dogma that most CRCs develop from preexisting colorectal adenomas has become established through clinical and molecular biologic observations. CRC is more than twice as prevalent among those patients with a family history in a first-degree relative, but the majority of CRC cases occur in patients without any identifiable heredity predisposition. CRC is most likely initiated by carcinogens in the fecal flow causing genetic alterations or mutations in the natural life cycle of the cells. It is important to understand that the process is not an infection, but a genetic damage to cells. The accumulated genetic changes cause cells to lose their normal control of reproduction and to gain a growth advantage that is perpetuated in the cell offspring. These accumulated genetic changes lead to adenomas and in certain cases to CRC.

Sporadic CRC is CRC that develops in someone without a specific family history or congenital predisposition. Current thought is that there are two molecular pathways that lead to colon cancer. The most probable is the chromosomal instability in which the

DNA damage is not repaired and abnormal cells continue to proliferate. The second pathway is the microsatellite instability pathway in which mutagens in the stool of the bowels pass by the cells in the colon surface and damage the DNA of these cells from which mutations lead to tumor development.

Almost all persons with colorectal adenoma and most with cancer experience no symptoms until after metastasis has occurred. Blood in the stool is the symptom most often associated with CRC and usually is associated with advanced disease. Changes in bowel habits and abdominal discomfort can also be symptoms of colon cancer. Anemia, phlebitis, and pulmonary embolism can be signs of cancer, including colon cancer.

1.3 Treatment of CRC

The stage of colon cancer at diagnosis determines how it will be treated. Considerable attention has been given to meaningful stages so that treatment is appropriate. There are different methods of staging, the most common being the Modified Astler-Coller Duke's classification and the TNM system of staging (where T is tumor size and extent, N is whether lymph nodes are involved and the number of nodes and M is whether cancer has spread to other parts of the body). A combination of these is given below, along with the typical treatment (from http://www.cancercare.org).

Stage 0 / Tis (carcinoma in situ) N0 M0 / N0 Dukes' stage

In Stage 0, the cancer is at a very early stage and is located only in the inner lining of the colon. The recommended treatment for Stage 0 colon cancer is surgical removal of the tumor, along with parts of the colon on either side of the tumor site. If detected early, colon cancer is highly curable and has a low risk for recurrence.

Stage I (T1 N0 M0, or T2 N0, M0) / Dukes' A

In this stage, the cancer has grown down to (inside out) but not through several layers of the colon (muscularis propria). The cancer is still confined to the wall of the colon and

has not spread to nearby organs as yet. Surgery alone is the recommended treatment at Stage I. Stage I is also highly curable, with a low risk for recurrence.

Stage II (T3 N0 M0 or T4 N0 M0) / Dukes' B

In Stage II, the cancer is more significant but has not spread (metastasized) to organs or lymph nodes. Lymph nodes are small, bean-shaped structures which serve as sites where bodily immune responses can be organized against infection and are common sites for cancer spread. The recommended treatment for Stage II is surgical removal of the tumor. Adjuvant therapy (chemotherapy and radiation therapy) is also recommended for some Stage II patients with disease in the rectum.

Stage III (Any T N1, 2 or 3 M0) / Dukes' C

In this stage, the cancer has spread outside the large intestine to lymph nodes, but not to other organs. Treatment for Stage III colon cancer includes surgical removal of the affected section of the colon. Chemotherapy is recommended for all patients with colon cancer and the combination of chemotherapy and radiation therapy for all patients with Stage III rectal cancer. Studies have shown that the number of lymph nodes involved affects the outcome. Patients with 1-3 nodes involved have significantly greater survival rates than those with 4 or more nodes involved.

Stage IV (Any T Any N M1) / Dukes' D

Stage IV is the most advanced stage of colon cancer. The cancer has spread beyond the colon, rectum or regional lymph nodes to distant organs or tissue (such as liver, ovaries and lungs). Although cancer is not usually curable at this stage, surgery is still the recommended treatment. Surgical resection of the colon and reconnection of the large intestine is done so as to remove blockage of the colon and any other local complications. Chemotherapy and/or radiation are generally given for palliative purposes.

Surgery or “resection” is when the diseased portion of the colon is cut away and the two parts reconnected (remaining colon is said to be anastomosed). As will be describe, surgery removes “modular” sections of the colon. A graphical display of the colon is

shown again in Figure 8 to aid the discussion (remember this is the view from the person).

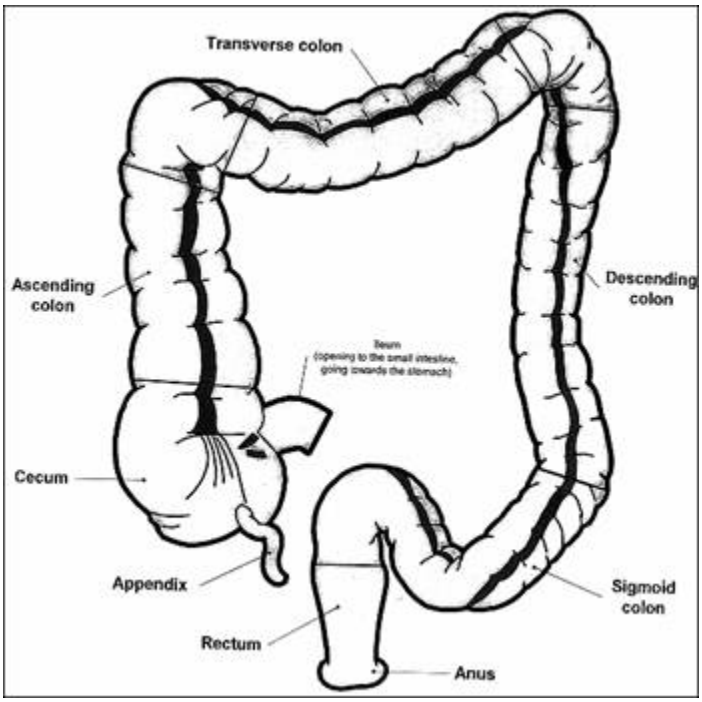


Figure 8: Cancer Resection Diagram

Before further comments, it needs to be noted that the resection margins actually represent areas which are perfused by a common large artery. Thus, the entire vascular bed can be removed at the same time, lessening post-op bleeding and making the procedure easier in general. Cancer in the ascending, or right colon can be resected with only the removal of that section. The descending, or left colon can only be resected by removing both the sigmoid and the left colon. For cancers in both the upper rectum and sigmoid sections of the colon, both the sigmoid and the upper rectum require removal. A lower rectum cancer is the most serious, and requires the removal of both sections of the rectum along with the sigmoid section of the colon. If reconnection is not possible a temporary or permanent colostomy is performed which creates an opening in the abdomen into the colon to provide a new path for waste material to leave the body into a special bag.

Chemotherapy uses drugs to treat the cancer. Usually a combination of drugs is used and the drugs are taken over the course of several weeks or months. Chemotherapy can reduce the size of the cancer, possibly slow its growth, maybe prevent it from spreading, and may relieve the symptoms of the cancer, enabling the patient to live longer and in more comfort (there are increases in survival in patients with Stage III disease).

Radiation therapy is used to slow or stop tumor growth through radiation. It can be used either before or after surgery. Radiation and chemotherapy are often used together with surgery in patients with rectal cancer. Radiation is almost never used alone for treatment of colon cancer.

If left unchecked, cancer cells do not remain where they originate. The cancer cells will proliferate and accumulate, with each new generation having the "growth advantage" that prevents their destruction. The life of the cancer cells makes them prone to the further accumulation of genetic damage, which produces more toxic cancer that is more likely to invade and spread. Once firmly established in the lining of the colon, the cells not only spread along the lining but also begin to penetrate into the colon layers. Deeper into the layers of the colon and farther from the lining, there are the lymphatic and blood vessels. When cancer cells invade the lymph system, they are first transported to the lymph node where lymph node function is highly degraded (lymph nodes filter the lymph fluid and produce antibodies that ward off infections). The lymph system interacts with the blood system which supplied the fresh blood and returns the "used" blood. Eventually, the cancer cells spread from their original site being transported via the bloodstream and the lymph system. Cancer cells also and usually do simply invade the blood stream from the lining colonic vasculature itself.

Cancer is said to metastasize when it becomes invasive or when it becomes involved with the lymph and/or blood systems. With colon cancer there are five ways of "spreading": (1) lymphatics, (2) blood, (3) direct extension to other tissue, (4) free intra-abdominal spread, and (5) spread within the lumen of the bowel. Once metastasis takes place, the cancer begins a "cascade," which usually kills the person.

1.4 Modeling the Natural History of CRC and Organization of this Thesis

Because CRC directly affects individuals, it is not economically feasible or morally justified to arbitrarily intervene in the lives of people who might become victims of CRC. Thus, using models to study various means to identify and treat people for CRC is an attractive alternative to direct experimentation. With valid models it may be possible to predict the course and outcome, in terms of costs and life-years gained, of screening, surveillance, and treatment methods.

Mathematical, statistical, and computer modeling has been previously applied to medical decisions with varying degrees of success. In Chapter 2, the various models of CRC and modeling approaches are reviewed. Of special interest to this work is the prior discrete event model of CRC developed by the Dr. Reid Ness and his associates at the Vanderbilt Medical School (Ness et al. 2000). The work in this thesis builds on that earlier model using support from a grant to Vanderbilt University from the Department of Health and Human Services, National Institute of Health, National Cancer Institute, grant number 1 R01 CA92653 01A1, titled "Simulation Modeling of Colorectal Cancer." More specifically, a subcontract from Vanderbilt to North Carolina State University, named "Simulation Model for Colorectal Cancer," grant number VUMC CA#9195 supported the work for this thesis.

Using the previous Vanderbilt model as a beginning, this thesis reconsiders the modeling approach and the specification of the input data to create a new object-oriented simulation of the natural history of CRC. The natural history model is fundamental to the study of screening, surveillance, and treatment strategies because it not only validates the underlying biomedical basis for the modeling activity, but also creates the numerical specifications for relationships between key variables.

Chapter 3 of this thesis describes what is being called the "Vanderbilt-NC State" model. Unlike the prior model which was written in a "typical" queuing simulation language (INSIGHT), this model is written directly using objects and events for the Windows environment. The object-oriented approach greatly increases model flexibility, which is critical in the highly volatile CRC environment where theory is unproven and data difficult to interpret. Furthermore the approach greatly improves execution speed and integrates well with Windows-based databases and spreadsheets.

Chapter 3 provides a careful delineation of the modeling assumptions. It describes the objects and events used in the simulation, along with their relationships and pathways. Chapter 3 also describes the data sources for the model construction and details the difficult task of modeling natural lifetimes without CRC.

Chapter 4 of this thesis focuses on the analysis of the model. This chapter contains the usual simulation concerns of verification and validation. However, there is a special need to synthesize data for the CRC model because it requires information about the course of CRC that cannot be observed. This need greatly complicates the verification and validation because observed results from clinical studies must be matched by model outcomes. Finally, the simulation model must produce the ultimate outcomes, namely the costs and life-years. Since this thesis focuses only on the natural history of CRC, the costs emanate from symptoms of CRC and their treatment. No screening or other interventions are considered. The consequence is that persons affected by CRC have increased costs and shortened lives.

Chapter 5 provides conclusions and recommendations for further study. The primary conclusion is that the Vanderbilt-NC State model is a valid representation of the natural history of CRC. There are, naturally, limits to this claim. While the model is believed to be a legitimate description of the natural course of CRC, it is most accurate where data are most available. Fortunately this region of accuracy also falls into the ages and patient groupings for which screening and surveillance are possible. Clearly, the next phase of

related work should concentrate on the best ways to intervene in the course of CRC using the work provided by this thesis as a basis for comparison and extension.

2. Modeling of Natural History of Disease

A model is a "working hypothesis." It is a detailed explanation or description of a phenomenon or problem. In this case, the phenomenon is colon cancer and the working hypothesis in this thesis is how colon cancer develops. Mathematical, statistical, and computer models are more than theory or supposition. In addition to the identification of key variables, the model inter-relates the variables quantitatively. In this chapter we examine the context of constructing models of medical history and describe the kinds of outcomes that models attempt to predict. We survey the models of CRC to discuss their strengths and weaknesses.

2.1 General Modeling of Medical History

Models of medical history can be categorized by the collection of persons being considered. The two prominent groupings are cohort models and population models. A cohort model is one that attempts to model a homogenous group of people to determine the effect of a disease. This homogenous group can have specific levels of detail. At its most basic, a cohort model can specify a population with a given age, gender, and race. This additional level of detail provides for medical interventions, like screening for CRC, to be studied.

The second grouping is a population model. This considers the population as a whole to determine the effect of a medical intervention. For example, a population study could examine the effect of a one-time colonoscopy being performed on everyone in the North Carolina who is over 50. A cohort model might be generalized to the population either by running many different models and then aggregating them together using a percent of the population at a given age or by treating the population characteristics as random variables.

2.2 Examining Outcomes

Medical decision-making models has tended to employ a standard set of outcome measures (Gold et al. 1996). These standard outcome measures provide generally accepted criteria to make informed decisions affecting policy and procedure guidelines for disease screening and treatment. However, these measures rely heavily on the perspective used in the model. Should a model be examined from the perspective of a patient, for whom additional years of life could be purchased with insurance covering expenses? Or from the perspective of a Health Maintenance Organization (a type of health insurance plan), where the chosen approach is to perform interventions such as screening only when it costs less than the treatment of the disease if treated later? Neither of these seems to be the ideal perspective, especially for a societal organization trying to set treatment guidelines. Instead, these organizations focus on the societal cost of the treatment and screening. This focus examines both the additional years of life – but from an aggregate perspective of the whole population – and the cost to society for the treatment and the intervention. This approach has become the generally accepted method of evaluating different treatment and intervention options (Gold et al. 1996).

2.2.1 Life-years

A key measure affecting policy decisions is the number of years lost due to the disease. In global terms, these are called life-years. Traditionally, this measure is computed for the number of years of life lost to the disease for every 100,000 people. For example, suppose the elimination of a specific type of cancer adds an additional 10 years of life to everyone who would have otherwise gotten that cancer. However, if the cancer only affects the lifetimes of 2.5% of the population then the average additional life-years gained per person is only ¼. So if the average life span were reduced by ¼ of a year by a specific cancer, the cost in life-years would be 25,000 life-years for the cohort of 100,000 people. Once a new treatment or screening procedure is modeled, the cost of the disease in life-years is recalculated. This new measurement of cost can then be compared to the

original cost in life-years to determine the additional life-years a given treatment or screening option will grant.

2.2.2 Quality-Adjusted Life-years (QALYs)

QALYs are a modification to the overall life-years calculation based upon the quality of life. The quality of life affects the utility a person experiences at any given time. The utility of life tends to fall as people age due to declining health. For example, a patient who is bedridden has a lower utility than a completely healthy patient. These utilities may be determined by the use of the "standard gamble" instrument which measures a person's willingness to sacrifice additional years of life for the relief of the ailments associated with the disease (Gold et al. 1996).

One way a QALY may be computed is by multiplying a life year by its utility. So the time spent bedridden is multiplied by the lower utility to compute the QALY for the patient. The total QALYs for a person are simply the sum of all quality-adjusted life-years. The multiplication is an approximation to calculating the integral of the utility function over time. This simplification is acceptable because the utility is assumed to change at discrete time intervals. Using this measure tends to make the benefits of reducing or eliminating a disease more pronounced because, in addition to life years gained, the measure also includes improvements in quality of life associated with treatment.

2.2.3 Costs

Since an objective of most medical decision models is to direct policy of disease treatment or screening or to better understand the impact of a disease, actual costs are a main factor to investigate. The cost most properly reported is the societal cost of the disease (Gold et al. 1996). The first key component of this cost is the cost of treating the disease including hospitalization costs, medicine costs, and doctor costs. Secondary costs include the costs of screening procedures and new treatments. Any new treatment or screening must prove that it is more cost-effective than existing procedures if it hopes to be implemented by Health Management Organizations (HMO) or as public policy. Tertiary costs, which include the costs of missed work and other time, can be even larger than the secondary costs. Depending on your perspective some of these costs do not necessarily apply. Loss of work is of concern if the perspective is that of a large company looking at treatment recommendations for employees, but is of little concern if the perspective is that of an HMO. Again, the most traditional approach for medical modeling considers all costs because it considers these costs from the perspective of the whole society.

2.2.4 Discounting

Discounting is a means of valuing earlier events and rewards more than later ones. A simple example of this view is that most people would rather have an additional $10,000 now than $11,000 ten years from now, simply because $10,000 is worth more to them now than $11,000 in ten years. Within discounting there are two methods of discounting, continuous and discrete. Discrete discounting is the simplest method to understand. An analogy is financial compounding, when a bank gives interest on a Certificate of Deposit (CD) at the end of every year. A similar method is used for medical models, where the discount on the life-years or cost is calculated at end of the year. The formula for discrete discounting is shown below.

$$P = \frac{F}{(1+i)^n}$$

P = Present Worth

F = Future Worth

i = Discount Rate

n = Number of Years

For example, consider a problem where there is a discount rate (i) of 5%, and a $10000 medical charge ($F$) that will occur 10 years in the future (n). The present value (P) of the charge would be as seen below.

$$P = \frac{\$10000}{(1+0.05)^{10}} \approx \$6139$$

The other form of discounting is continuous discounting. Again using a financial analogy, consider a bank that charges you 5% interest on a loan, but they compute the interest owed every day. Now extend that scenario to re-computing interest all the time, and it is continuous compounding. While in some cases appropriate, in medical decision models there are usually discrete costs, so continuous compounding is not as applicable. Even in cases where the continuous form is the correct descriptor, such as energy costs, the discrete approximation is close enough to yield adequate results with much less computational effort (Blank and Tarquin 2002).

The issue of discounting in medical decision-making is more troublesome since both costs and life years (QALYs) are being considered. One possible way to combine costs and life years into a single cost-effectiveness measure is to compute a cost per quality-adjusted life year. If cost-effectiveness is used to evaluate alternative medical interventions, then discounting can be used to create a common basis to compare the different streams of costs and quality adjustments generated by each alternative. The discounting of cost streams for purposes of comparison is commonly accepted. While slightly less intuitive, this discounting can also be applied to QALYs (Nord 1999). The currently recommended discount rate in modeling medical decisions is 3 % for both costs and QALYs (Gold et al. 1996). Using the same discount rate for both costs and QALYs avoids the problem of having different discount rates change the cost-effectiveness in different years (called “intertemporal inconsistencies”).

Since this thesis only addresses the course of CRC, not decisions regarding medical interventions like screening, there is no defined decision point or reference point for discounting. Thus neither costs nor QALYs are discounted.

2.3 Modeling of CRC

There have been two primary means of creating a model of CRC. First is the Markov chain approach. In general, these models have identified health states and rely on transition functions to cause the model to step between states. Based either on computing the probability matrix changes or a Monte Carlo simulation, the goal of these models is to estimate the proportion of time in each state. The second method of modeling CRC is a discrete-event simulation. This approach follows a cohort or population of people from birth to death. The differing models of each type are discussed below. Pignone (Pignone et al. 2002) has provided an excellent starting point for the analysis of CRC modeling by performing a meta-analysis of published CRC screening literature.

2.3.1 General CRC Data Sources

Before discussing the different models used to represent and study CRC, it is important to understand the data sources that are common to most or all of the models. The first piece of information any model needs is the natural life expectancy of a person. While numerous life tables are available, the most commonly used source is http://www.demog.berkeley.edu/wilmoth/mortality/ (Wilmoth 2003). Models must then duplicate the effect of CRC on an individual by matching outputs to the cancer incidence and survival reported. All general information about cancer can be found in the SEER database available online (National Cancer Institute 2003). Adenoma incidence is another important component of CRC models. This information comes from autopsy studies by Eide, Arminski, Blatt, and others (Eide 1986; Arminski and McLean 1964; Blatt 1961).

Once a model has been constructed and parameterized, the model must then be externally validated against numbers that were not used to fit the model. In the case of CRC, there are several screening intervention studies that have been done which are used to validate the models. The first and most commonly used study is the National Polyp Study (Winawer et al. 1992). This study tracked a large set of individuals and compared screening with colonoscopy to screening using barium enemas. The other primary study that is used to compare model outputs is the Minnesota Fecal Occult Blood screening Test (FOBT) trial (Mandel et al. 1993). This trial observed the reduction in CRC and CRC mortality caused by screening for CRC using an FOBT.

2.3.2 Markov Medical History Models

The first primary method of modeling the course of the disease is the use of Markov models. These models enumerate health states a person will experience during the course of the disease. The changes in state are described using transition diagrams very similar to flow charts. Each transition has a certain rate (in the case of continuous-time Markov process) or probability (in the case of discrete-time Markov process) associated with it. In a continuous-time process, this rate describes an exponential holding time that determines the time the person stays in a given state. It is then possible to perform either a Monte Carlo simulation or either iterate or solve the transition matrix mathematically to determine the steady-state behavior of the model. These steady-state percentages are then used to determine the time spent in each state.

An alternative approach to modeling a Markov process is to use a spreadsheet process. In these models, rather than using a true Markov chain analysis software package, a spreadsheet is used and transitions are checked every year (or other fixed interval). This discrete-time process uses simulated people in a fashion closer to another type of model, a discrete-event simulation. The problem with these models is the huge number of states that can develop. Some models of this type have up to 5400 states. Since only a completely homogeneous population can be run through the simulation (no variability in the risk value), risk can only be handled by adding extra states (Davies, Roderick, and

Raftery 2003). In addition these models are often deterministic, with a fixed amount of time being spent in key states.

Now let us examine a simple discrete-time Markov model as a review. Figure 9 below shows a simple Markov model that has everyone start in a "well" state (Naimark et al. 1997).

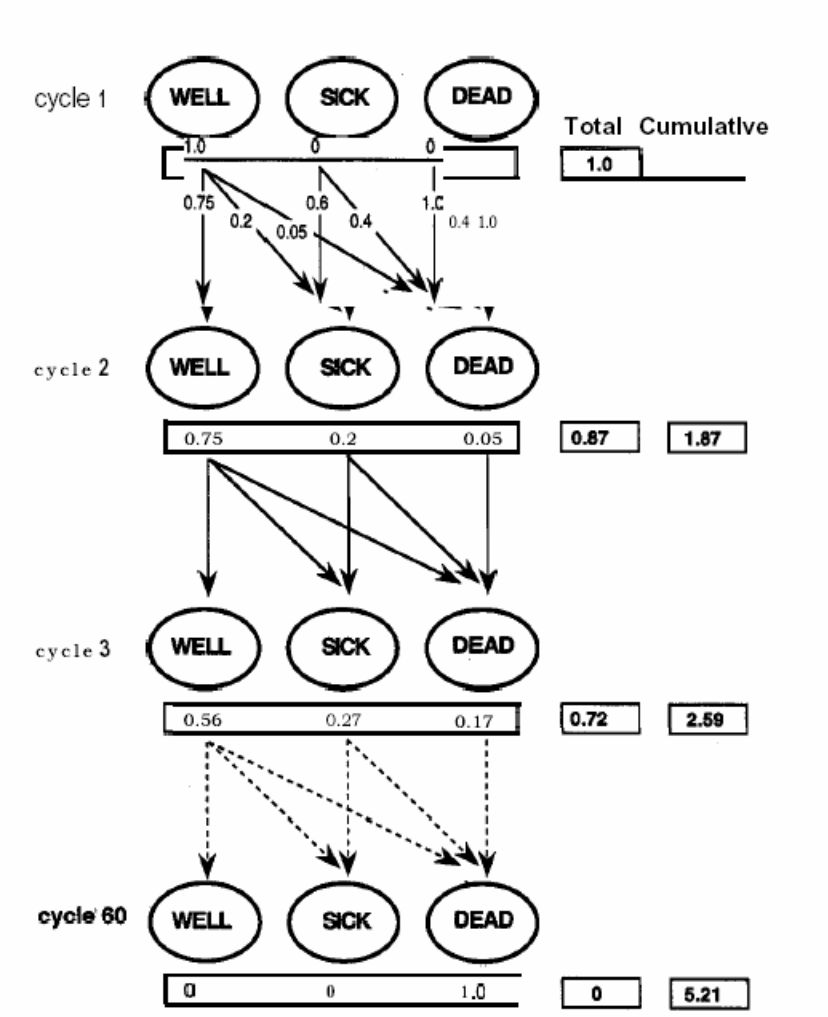


Figure 9: Example of a Markov Model

Based upon the transition probabilities, new numbers of people enter each state after the first iteration (in this case 0.75 are now healthy, 0.2 sick, and 0.05 dead). The second step is to iterate again to determine the fraction of people in each state. Note how in each step, the number of healthy individuals is constantly decreasing and the percent dead is

increasing as it "absorbs" people from the other cells. In this model, it takes until the 60th step for all people to die. Utilities, costs, and times are calculated base upon the percentage of people in each state for a given year.

Below are some of the main models that take a Markov approach to modeling CRC. In general, model details reported are sparse, so the depth of discussion for the model is sometimes limited.

Harvard Model

Like the entire current group of CRC models, this model is a cohort model. Specifically, the model follows a 50-year-old cohort until death. The model is a Markov model containing 60 different health states, and is analyzed by the SMLTREE software (Hollenberg 1985). The model has yearly transitions in which people will change states based upon the transition probabilities (Frazier et al. 2000)

Model Structure

A simplified version of the state transition diagram of the model is shown below in Figure 10. There are several states within each of these main categories.

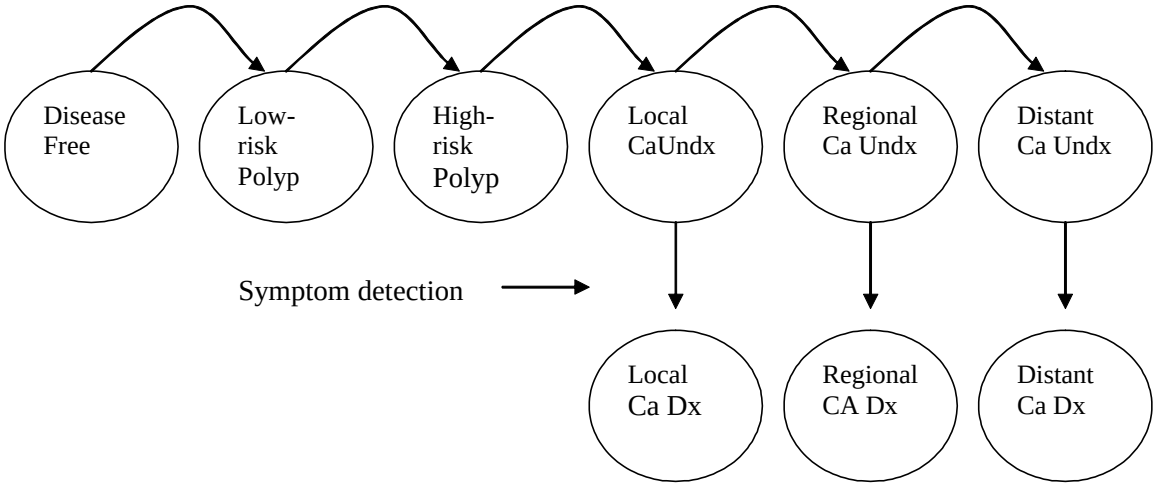


Figure 10: Outline of the Harvard model of CRC

As can be seen from the picture, this model breaks polyp development into two sections.

The distinction between a low and high-risk polyp is accomplished by defining a high-risk polyp as one with a diameter of greater than 1 cm. This model is a representation of 50-year-old people going on to eventual death from CRC or other causes. Starting with the introduction of polyps, this model has cancer first becoming local, and then transitioning through regional cancer to distant cancer. At any point, symptoms can develop and be detected. In this model, all cancers develop from polyps.

Initial numbers of people were placed in each state to "preload" the model. This preloading is required because by age 50 some people will have already had cancer, some will have undetected adenomas of different risks, and some will be completely healthy. Because the model starts at time 50 though, the number of people in each state is set at the start of the model. Each year a "step" is taken, where new numbers of individuals are in each state. Statistics are recorded every year, and the process is repeated until all individuals end in one of the absorbing states of either natural death or CRC death. These states are called absorbing states because eventually all people transition into those states and do not leave.

Data Sources

This model uses the data from six autopsy studies to determine the adenoma prevalence, and then fit the probabilities of transitioning to different states to match these results. These results of the studies also determine the initial polyp prevalence for the cohort. Screening studies were used to determine the rate for transition to a high risk polyp based upon repeat screenings to determine which adenomas had progressed to the next size and risk category. As with other models, the transitions between cancer progression and symptoms were varied to create output that closely resembles cancer incidence as reported by SEER.

Verification and Validation

To validate this model, a patient cohort similar to the Minnesota Trial was run. This trial was conducted to determine the effect of screening on patients. It tested the reduction in CRC mortality by performing a fecal occult blood test (FOBT). The test checks stool for the presence of abnormal DNA, an indicator of cancer. A comparison of the simulation results to the Minnesota Trial revealed that the model validated well, and the results are shown below in Table 1.

	Mortality Reduction	**Incidence Reduction**
Minnesota Trial	33	11
Model	31	7

Table 1: Comparison of Harvard Model to Minnesota Trial

Synopsis

This model is a characteristic Markov example of modeling CRC. It uses the standard procedure of calculating state populations every year, and then recalculating the number of people in each state. The predication relative to the Minnesota Trial, in particular, provides validity to the model. A simplifying assumption of this model was not to model individual patient risk and patient history.

UCSF Model

Like the Harvard Model, this model is also a cohort based Markov model. It also models the cohort from age 50 until death. The model is implemented in DataPro (Ladabaum et al. 2001)

Model Structure

The state transition diagram of the model is shown below in Figure 11:

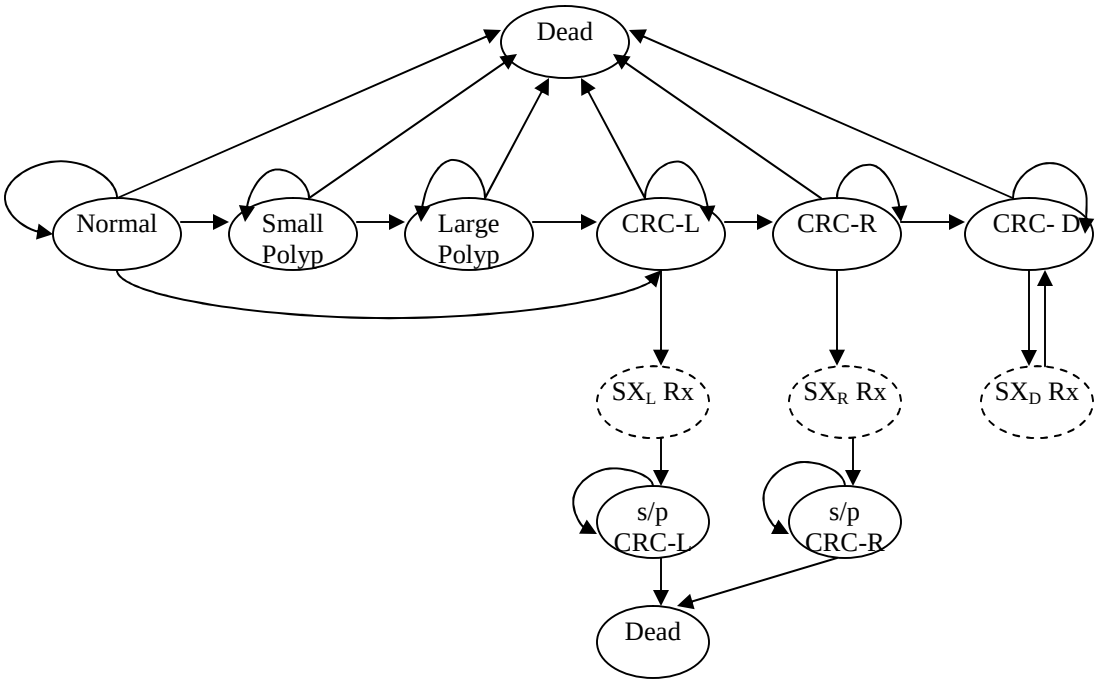


Figure 11: Model Structure for UCSF Model

The first distinguishing feature of this model is that the first step of the transition matrix is used to determine which screening strategy is performed. This model examines natural history as well as 5 screening strategies. This model has the same basic structure as the Harvard model, but does not compute the state probabilities that determine the number of people in each state. Instead, this model is a Monte Carlo simulation that runs through the model 3500 times to determine approximate values for the percent of people in each state at a given time. This method is computationally very slow, but allows for solutions where an exact transition calculation cannot be determined (Bong 2004). One new feature of this model is that it has a small probability for cancer to develop without developing from an adenoma. The percentage of all cancers that develop this way is 15%.

Data Sources

Transition time from local cancer to regional cancer and from regional to distant cancer is fixed at two years each. The mortality for each year is fixed at a specific percentage based upon whether the cancer is regional or local. These values are 1.74%/year for local cancers, and 8.6%/year for regional cancers. These cancers develop symptoms at fixed rates per year based upon the stage as well. Consequently the farther progressed the cancer, the more likely it is that symptoms will develop that year. In general, all transition probabilities were varied until the outputs met the following criterion for the literature: 1) cancer incidence, 2) cancer states at time of diagnosis, 3) polyp prevalence, and 4) % of polyps that are large.

Verification and Validation

The UCSF model was verified by matching the SEER incidence by age and stage along with the adenoma prevalence and size data. The validation is performed by comparing the total cancer cases and total cancer care costs from the model with observed values from the SEER database (National Cancer Institute 2003).

Synopsis

This model is currently only validated on the overall population. However, different populations having a differing personal risk of developing CRC or differing race or gender population groups can be modeled, as long as they have been verified previously (Ladabaum et al. 2001). The model uses a fixed dwell time (time in state) for the time from adenoma incidence to cancer. Also, the location of the adenoma is not modeled, which limits the ability to accurately model sigmoidoscopy screening, since it can only reach certain sections of the colon. Finally, like all the other Markov models of CRC, this model tracks the most advanced lesion, and does not allow for competing adenomas and reoccurrence of disease due to new adenomas. The main distinctions of this model when compared to other Markov models is the use of Monte Carlo simulation and the allowance of cancer to arise without an adenoma precursor.

Michigan Model

The Michigan model is another use of a Markov model to represent CRC. The model itself is based on US population overall rates, so it does not include population-based breakdowns.

Model Structure

As with the other Markov models of CRC, this model also tracks the most advanced lesion. As such there is a single pathway from normal through polyps to cancer. Like the Harvard Model, this model has the adenoma progress through the low-risk and then on to the high-risk polyp before transitioning to cancer. Within cancer, there are three stages – two local stages that correspond to stage one and two discussed in Section 1.3, then regional and distant cancer. However, the dwell time, or time from adenoma to cancer, is fixed so there is no variability in the time to later stages (Vijan et al. 2001). Once symptoms develop, the cancer can either cause death or the patient becomes a survivor and will live to his natural life. Figure 12 shows the state transition.

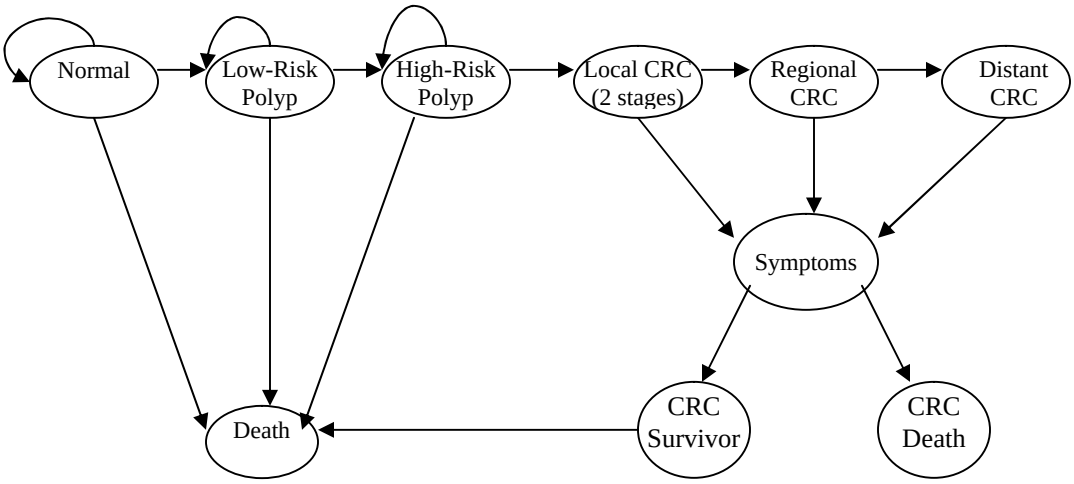


Figure 12: Model Structure for Michigan Model

Data Sources

The prevalence of polyps is preloaded into the system based upon autopsy studies performed by Correa and others (Correa et al. 1977). This preloading is done by setting initial numbers of patients into differing health states such as undiagnosed CRC, high risk polyp already present, and full health. To get the results, a fixed polyp to cancer dwell time of 10 years was chosen. This dwell time is the shortest dwell time of all the CRC models. Like the UCSF model, this model initially had some cancer develop from sources other than polyps. This value was initially set to 25%, but the assumption has been revised. The new assumption is that 100% of all cancers develop from polyps, so the model now matches assumptions of the Harvard model. Time in cancer stages is a fixed number for this model. So once again, there is no variability in how long a cancer will stay as local cancer before becoming regional or distant cancer. The cancer will spend two years as a local cancer, and 1 year as regional cancer. Like the MISCAN model discussed later in Section 2.3.3, mortality rates from cancer are based on older data to match prescreening information. Prior to 1982, no systematic colorectal screening was performed, and the disease was treated only when symptoms arose. After 1982, screening began to be implemented. This screening initially increased the number of cancers found, but also improved survival because more cancers were being caught at earlier stages. Using data from the period prior to 1982 is a good representation of a natural history model, and also allows for the most correct modeling of screening effects.

Verification and Validation

As validation of this model, the Minnesota FOBT Trial was used (see Section 2.3.1). A special model with modified population characteristics was used to represent the trial's populations. Because the original Markov model was designed using the overall population, new random input parameters for the random variables needed to be created to reflect the different population characteristics. The model predicted that screening causes a 39% reduction in CRC mortality, while the study predicted a reduction of 33%.

Synopsis

In general, this model is very similar to the other Markov models previously explored. It too has the benefit of having a simple-to-understand state transition diagram. However, like all Markov models, there is limited flexibility. The model has fixed dwell times, and is limited by variables that are all independent. Also, any variation in polyp or cancer risk requires a full model recalibration. This model found that colonoscopy represented a cost effective means of screening, and colonoscopy was found to be the preferred screening method in situations where compliance was less than perfect (Sonnenberg, Delco, and Inadomi 2000).

2.3.3 Discrete Event Simulation Models

Discrete-event simulation models are used to represent a medical life history by producing timelines of the health events in individuals. Because the events occur randomly, a simulation of a large group of people is used to estimate the results. However, the specification of exactly how these events occur depends on the input to a model and the execution of the model.

To illustrate the fundamental concerns, consider a "simplified" sequence of events in the timeline of an individual who dies from causes related to colorectal cancer (CRC) – (Ness et al. 2000) shown in Figure 13.

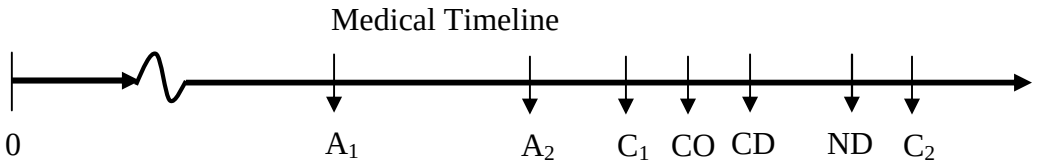


Figure 13: Discrete Event Timeline

In this example, as the person ages he develops an undetected adenoma (A_1) in his colon. Then a second adenoma in the colon develops (A_2), also undetected. Eventually, the first adenoma produces invasive cancer at that site (C_1). The cancer becomes symptomatic

and a colonoscopy (surgery) is performed (CO) to remove the cancer. However the complications of the cancer results in the patient's death at CD. The person does not experience natural death (ND), nor does the person live long enough for the cancer to develop at the second site (C_2).

The various markers in the medical timeline are generally called "events." Were the processes producing this timeline to be simulated, there are several modeling issues that affect how these events are to be generated. Implicit in our examination is the recognition that these events occur "randomly," namely their occurrence cannot be predicted with certainty. Therefore proper representation of these input random variables is critical.

To construct a model that can simulate timelines such as that seen earlier, we will view the timeline creation as a discrete-event simulation. Events, as they are known, are stored in a calendar of future events. The simulation executes by removing the next event, updating the age of the patient to the time of that event, and then executing whatever processes are associated with that event, including collecting any statistics and creating new events.

The key to the simulation is to know when future events will occur so they can be maintained in the event calendar. To "know" when future events occur requires a "model" of their occurrences. The model of occurrences can either be explicit in the form of input or implicit in that it is a consequence of other actions within the simulation.

The advantages of this approach to modeling is that there is a much greater ability to accurately reflect the characteristics of a healthcare system, but conversely they are more difficult to understand and need specialized developers to create the models (Davies, Roderick, and Raftery 2003)

MISCAN

The MISCAN model is the first example of a discrete event simulation model. The model itself is programmed in Delphi as a discrete event simulation. This approach allows for the setting of a starting age using a birth cohort instead of using a fixed starting age. Using such a starting age is not done in any of the Markov models, because of the number of additional states that would be needed to represent the system. Consider the Harvard model as an example. It currently has 60 states and only models from age 50 to 80. Imagine the number of states that will be needed to represent different rates for all the different ages. Thus no adenomas need to be "preloaded" into the system as is the case with Markov models. MISCAN starts at birth. MISCAN also accounts for current population estimates and screening performance, so it has been used to perform capacity studies for different screening practices (Brown, Klabunde, and Mysliwiec 2003). These capacity studies are conducted to determine the number of doctors and staff needed to perform different screening practices.

Model Structure

MISCAN is a multi-disease model that allows for colorectal cancer to strike an individual multiple times as well as have multiple adenomas develop. In addition, adenomas can arise in multiple locations within the colon and all develop independently of each other. The incidence rate is combined with an individual risk factor to determine incidence and growth. This technique adds flexibility and face validity when compared to a population based risk adjustment within Markov models. The patient flow of the model is shown below in Figure 14:

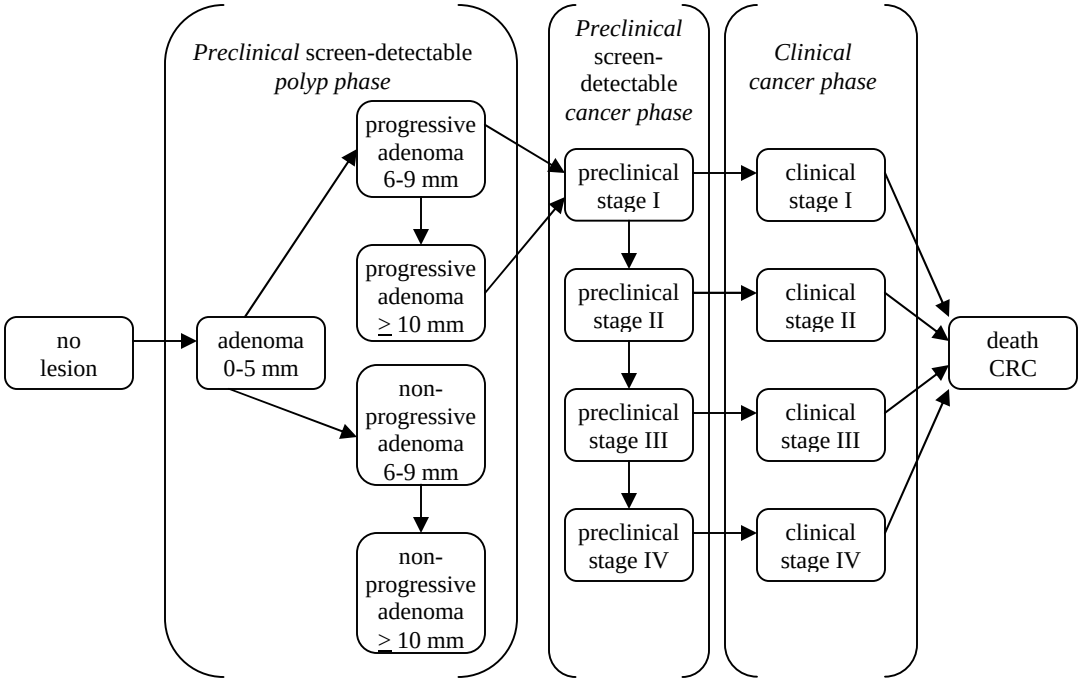


Figure 14: Diagram of MISCAN Model

In contrast to the three stages of cancer seen in previous models, this model has cancer transitioning through 4 stages. Stages 1 and 2 of cancer in this model are analogous to local cancer in the other models, and represent the Duke's A and B stages of CRC discussed previously. Risk in this model is used to modify the transition times between states, instead of the alternate modification to adenoma incidence used in the Vanderbilt model. In other words, the risk affects progression instead of affecting incidence as it does in the Vanderbilt model.

Data Sources

This model is also the first model to include some random variable valuations set by expert opinions. This adds to the face validity of the model, but also makes fitting the output to target values more complicated. The input parameters were adjusted until the model output matched target values- the SEER cancer incidence, survival, and stage specific distributions, along with the adenoma prevalence from the autopsy studies. An important distinction of this model's targets is that they are based on data collected in 1978, which is prior to screening. This data allows for any intervention that is later

modeled to impact the system as observed before screening, instead of having a model that attempts to impact data that represents cancer once screening began.

The model's input parameters can be broken down into three areas, as shown below: direct estimates, assumptions from experts, and input from fit procedure (Loeve et al. 1999) as seen in Table 2.

Direct Estimates	**Assumptions**	**Resulting from fit procedure**
Demography	Duration distribution in preclinical states	Probability of adenoma being progressive
Distribution of lesions over large bowel	Transition probabilities from preclinical non-invasive states	Individual risk index
Survival after clinical diagnosis	Correlation between durations in subsequent states	Incidence rate of adenomas
Sensitivity, specificity, and reach of screening tests	Dependency of test outcomes	Transition probabilities from preclinical invasive states
	Survival after screen detected diagnosis	

Table 2: Source of Parameter Values for MISCAN model

Verification and Validation

This model is the most thoroughly validated model of those surveyed. It currently has been validated based upon comparison to data from the National Polyp Study, the Kaiser Flexible Sigmoidoscopy study, and the Minnesota Trial. In addition, work is in progress to validate the model against data from the Funen Trial and the Nottingham Trial.

Synopsis

The key strength of this approach is the ability to model multiple adenomas within an individual person. Also, through weighting, a population-based model is possible by aggregating the results of different populations to create a larger population. An interesting finding of the model was that FOBT screening would be cost-effective but not cost-savings (Loeve et al. 2000). So while the cost for the incremental increase in life was below an accepted level, making the screening cost effective, the screening did not actually reduce the societal cost of treating CRC.

Vanderbilt Model

This model is the predecessor to the Vanderbilt-NC State simulation model that is the subject of this thesis. The Vanderbilt model is designed as a discrete event simulation model, and is implemented in INSIGHT 5.4. The model is a homogenous cohort model and can be used to perform population analysis through use of multiple simulations. Like the MISCAN model, this model simulates individual patients starting at birth..

Model Structure

The changes in health states that a patient can undergo in this model are shown below in Figure 15.

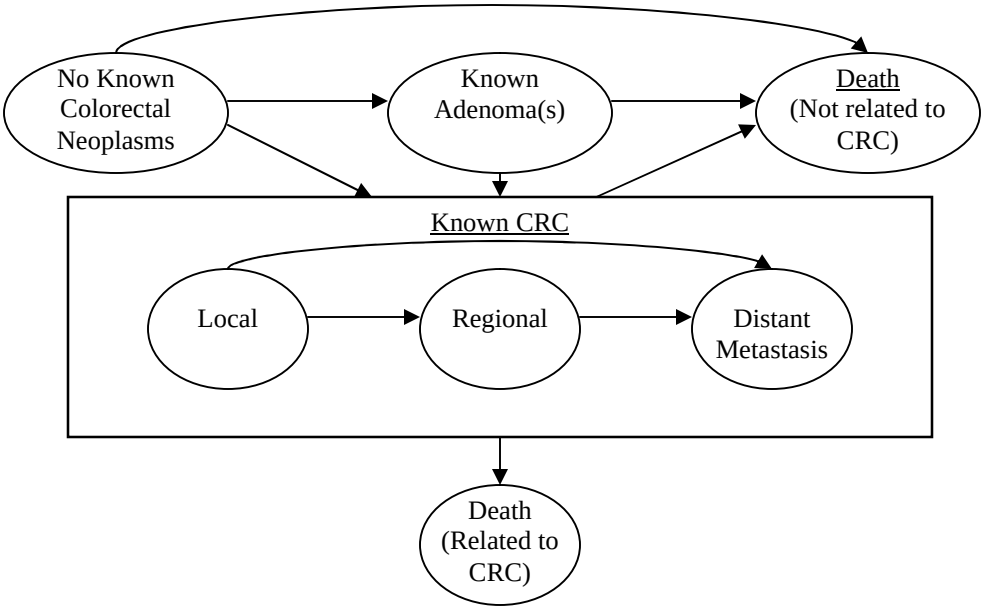


Figure 15: Basic Structure of the Vanderbilt model

Like the MISCAN model, this model is a multi-disease model. So multiple adenomas will develop and progress towards cancer and potentially multiple incidences of cancer will occur. There is also a risk factor that affects the incidence of adenomas. This helps to represent the wide range of adenomas incidences seen in patients observed in clinical practice. The main events in the model are transitions to the next stage for an adenoma or cancer, and the development of new adenomas for the person. Events can also arise from tests or surgery that alters the clinical parameters. As described by Ness et al (Ness et al. 2000) , a unique characteristic of the model is the breakdown of progressive adenomas into two speeds. These speeds represent the slower progressing adenomas as well as the faster progressing adenomas. The model defines means for the two speeds (26 years and 75 years respectively) with upper bounds double the means..

Data Sources

Because few sources describe CRC developing from flat lesions, this model has 100% of all cancers developing from the adenomas. The dwell time of adenomas in this model is a JohnsonSB random variable with only 2.5% of all adenomas becoming cancer in 10

years. This distribution was chosen to fit both the autopsy data on adenoma prevalence and SEER cancer incidence. The model data is based only on the white race.

Verification and Validation

This model was internally validated by its ability to duplicate the target values of adenoma prevalence and CRC incidence that are derived from the literature. Finally, the model was externally validated by comparing the model to the National Polyp Study. This study tracked 1400 people with adenomas for six years following polypectomy. The comparison of cancer incidence is shown below. The model confidence interval from the simulation was 3±3 cancers from a cohort of that makeup, while the actual study showed an average of 5 cancers. So the observed mean was within the confidence interval, suggesting that the model was correct (although the prediction interval would have been sufficient).

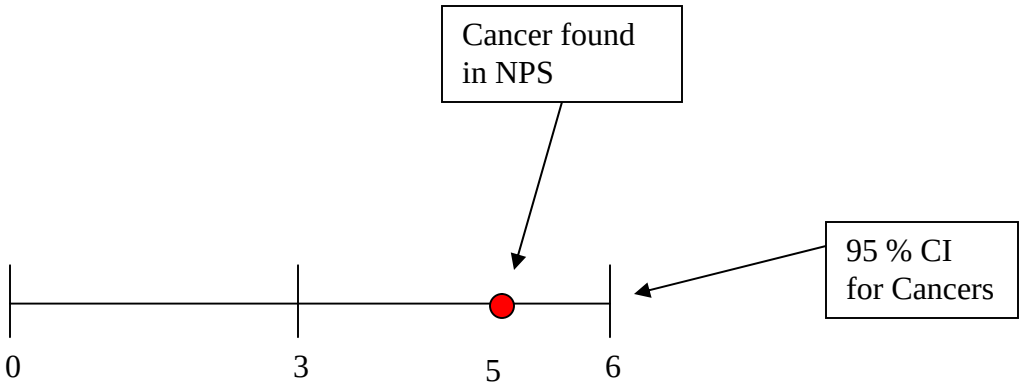


Figure 16: National Polyp Study cancer confidence interval

Synopsis

As a discrete event simulation, this model has the benefit of being very flexible. It allows for the assignment of attributes to patients, and can also accept quality of life adjustments from cancer and treatments. In addition, it also allows for the testing of alternative screening methods. The main drawback of this model is the implementation language. The language is a queuing language with no extensibility. So the flow of the disease within a patient had to be fit into the queuing frameset. Because of the lack of

extensibility, if a section of the language did not operate in the desired manner, there was no ability to "drop down" a layer and create your own object. In addition, the language itself is an interpretative language, so instead of compiling the program, it runs each person by interpreting the model statements. The net effect of this interpretation is easy adjustment to the model, but slow running of many runs. Despite the language's shortcomings, the model itself works remarkably well, and provides a foundation for the current Vanderbilt-NC State model. The Vanderbilt-NC State model also incorporates other details of CRC progression not present in the Vanderbilt model.

2.4 Summary of Chapter 2

This chapter began with a brief overview of the common measures used for many medical decision models. Current Markov models of CRC were then discussed to provide a background for the simulation. These models all had many deterministic (fixed) times associated with the states, and needed recalibration if the population shifted or another population needed to be examined. The previous simulation models created prior to the Vanderbilt-NC State model were then examined as the first models to bring the flexibility of discrete-event simulation to CRC modeling. While powerful, a new model is needed that shifts modeling to the more robust object-oriented design. In addition, prior models both lack modern interface tools, so a new model that utilizes a database for inputs and a better output presentation tool is necessary. Finally, we observe that existing models, including the Vanderbilt model, do not address the issues of multiple races, the birth year of individuals, the role of histology, and the risk affects on progression as well as incidence.

3. Simulation of the Natural History of CRC

The previous chapter reviewed the existing models of CRC and found that the most flexible and descriptive models of CRC were discrete event simulations. However the existing simulation models have some significant deficiencies relative to comprehensiveness, flexibility, and efficiency. This chapter describes the design and development of the Vanderbilt-NC State model of the natural history CRC, a new discrete event model based on current simulation technology. The focus of this model is on only natural history of CRC and not the screening for it. Other work within this overall project will focus on screening and surveillance strategies.

3.1 Basic Design of the Simulation

The Vanderbilt-NC State model is a discrete event simulation whose object-oriented design is implemented in the .NET programming environment (largely using VB.NET). A Microsoft Access database is used to store the simulation inputs and save output results. Microsoft Excel spreadsheets display the immediate results.

The object-oriented design distinguishes this simulation from most other simulations. As Joines and Roberts (Joines and Roberts 1998) have argued, discrete event simulation needs the additional versatility provided by the object-oriented approach to expand its usefulness beyond the queuing environment. Due to the uncertain nature of CRC, an object-oriented approach is necessary to avoid complications when dealing with the limited data and poorly understood causal pathways. Since there is no clear queuing system in CRC, the object-oriented approach allows the primary objects, the people, to experience disease and treatment pathways through the model, based upon the person's own unique attributes and the random processes affecting the outcomes. In the case of CRC, these pathways may be influenced by the experience of the secondary objects such as the birth and development of adenomas. If the patient does not die of other causes, these adenomas travel through various stages until becoming symptomatic cancer. Each

affected person will have a possible collection of adenomas that will determine their CRC history.

The Vanderbilt-NC State CRC model is constructed on top of a general object-oriented, open-source discrete-event simulation platform, not described in this thesis. This basic simulation platform is written in the .NET programming environment and provides basic simulation object classes. The components (object classes) include basic facilities for random number generation, random variate generation, event construction and handling, and statistics collection. Of special importance to the CRC model is that this package contains an entity object class and an event object class. Each entity object has its own event calendar and has an inheritable "activate" function which causes the entity to remove events from its event calendar and process those events. The specific events in the event calendar are defined by creating instances from the inherited event classes and specifying the process of updating the state of the simulation based on that event.

The addition of object-oriented design and extensibility is one key area where the Vanderbilt-NC STATE model provides improvements to the existing Vanderbilt model. Another key feature that improves upon the prior model is the addition of family history to the model. This addition allows more specific screening strategies to be tested to determine cost-effectiveness as well as adding validity to the model. The Vanderbilt-NC State model also allows different birth years to be examined, which adds robustness for future use of the model. The model also incorporates risk, also used in the MISCAN model and the existing Vanderbilt models, but extends its use by allowing risk to affect both incidence and progression, better reflecting the myriad pathways of the disease. In addition, the model is the first simulation model to utilize histology instead of adenoma size. While size does correlate with histology, histology is a better predictor of the risk an adenoma poses of becoming cancerous.

Finally, it should be observed that many of the random variables within the Vanderbilt-NC State CRC natural history model are often represented by JohnsonSB distributions and the time-dependent variables are usually modeled by non-homogeneous Poisson

processes (NHPP). The choice of the JohnsonSB distribution(Roberts 2003) was made because of the lack of data and the need to represent a wide variety of non-symmetrical shapes, both characteristics of biomedical variables. A JohnsonSB distribution has excellent flexibility because of its four parameters. The presence of bounds in biomedical variables adds to the utility of the "bounded" Johnson. The NHPP was chosen because of the belief that the Poisson rate applies to medical incidence and that the rate function varies with time. In fact, all incidence data is given in terms of rates per 100,000 people subdivided by year or age groupings and thus is easily modeled with the NHPP. The NHPPs in this thesis use piece-wise linear rate functions.

3.1 Individuals and Adenomas Objects

In the context of an object-based approach to simulation, the individual person is an object (or instance) from the person class, which inherits the base level entity class. Any adenoma is an object from the adenoma class. The adenomas belong to the person. Each adenoma can have several events associated with it and these events are stored in separate event calendars. This design facilitates the removal of adenoma related events when the adenoma is removed from the person. The simulation time update mechanism is expanded to include the earliest adenoma event along with earliest general simulation events. It is useful to describe the assumptions of the model in terms of the properties of a person and the properties of an adenoma.

3.1.1 Person Object

Relative to the simulation of the natural history of CRC, each individual in the simulation has the characteristics listed below. The specification of these characteristics defines a "population at risk" for colon cancer. These characteristics are specifications of the scenario being simulated and create a homogeneous cohort population:

Reference Year: the current year of the simulation. This determines the birth year to be used for the population.

Age: *current* age can range from 0 to 110 years. Within the population at risk, the *study* age is the age at which individuals enter the study and statistics begin to be recorded. While everyone starts at age 0, those who die before the study age are not included.

Gender: male or female

Race: white or black

Family History: first degree or none. This tells whether a person has a first degree relative with a history of CRC. A first-degree relative is defined as a sibling or parent who has had CRC.

In addition, all individuals contain the following object properties that are unique to that individual. These are obtained either by sampling from random variate generators according to the scenario inputs or by following the simulation pathways.

Natural Death: expected time of death without CRC. The distribution for this comes from the modified Berkeley Mortality Database which is a function of reference year, age, gender, and race as discussed in general in Section 2.3.1, and is specifically modified as described in Section 3.6.

Overall Risk: determines the individual's risk for CRC as discussed in the subsequent assumption in Section 3.2.

Number of Adenomas: a counter that tracks the number of adenomas that an individual has. Each adenoma maintains a collection of events, so the number of adenomas corresponds to the number of adenoma event collections.

Accumulated Cost: the total discounted cost to date for CRC care for the individual. People who reside in the various known cancer states may incur costs to the heath care system and may have their quality of life changed. In this thesis

we will treat medical sector costs as "costs to society." (True individual costs would need to include waiting, travel costs, lost work, etc.)

Accumulated Quality of Life: the total discounted quality-adjusted life-years (see Utility below) for the individual, as affected by CRC.

Utility: the utility associated with the current health state. As the cancer becomes worse, the quality of life decreases, yielding a lower utility. The quality of life will be applied to the time the individual spends in the health state. Quality of life will be measured using a "utility" that can be used to adjust the number of life-years for an individual who losses quality due to colon cancer by multiplying the years spent in a specific health state by its utility to yield a "quality-adjusted life-year" or QALY.

Health State: the state of the person relative to their colon cancer health is described by the following four basic states:

No CRC neoplasm(s)
Known CRC neoplasm(s): the health care system knows
Unknown CRC neoplasm(s)
Death unrelated to CRC
Death related to CRC

Cancer State: both the "known" and "unknown" CRC neoplasms can reside in the following cancer states:

Local CRC: Stage I and II cancer
Regional CRC: Stage III cancer
Distant Metastasis: Stage IV cancer
Death related to CRC

Symptomatic: cancer states may be either "non-symptomatic" or "symptomatic." People with symptoms (like blood in the stool) may seek treatment. Again, since

our model excludes screening and models only the natural history of CRC, only those people who are symptomatic will receive treatment.

ID: each individual has a unique ID that is used for bookkeeping as well as associating adenomas with individuals.

Resection: contains the areas of the colon which have been removed for each individual. Once an area has undergone a resection, no new adenomas can develop in that area.

3.1.2 Adenoma Object

The characteristics of the adenoma are more complex and can also depend on the population specifications such as age, gender, race, and family history

All objects of the adenoma class possess the following characteristics:

Histology: represents the type of the adenoma within the person. All adenomas begin as non-visible adenomas for purpose of the model. They then immediately progress into either advancing or non-advancing adenomas based upon a percent progressing variable. The histology is a combination of shape and size of the adenoma, as described in Section 1.2.2. An advanced adenoma is defined as one that is larger than one cm in diameter or is either tubulovillous or villous. Non-advanced adenomas compose all other adenomas. The categories of histology are: Non-Visible, Non-Advanced Adenoma, Advanced Adenoma, and Cancerous.

Adenoma Location: stores the specific location of the adenoma. An adenoma can be in one of five locations: Right Colon, Left Colon, Sigmoid, Upper Rectum, and Lower Rectum. When viewed from the person, these locations can be seen in the picture below in Figure 17 (note these designations are different from those in

Figure 8). These locations are defined because they may affect screening sensitivity or cancer treatment.

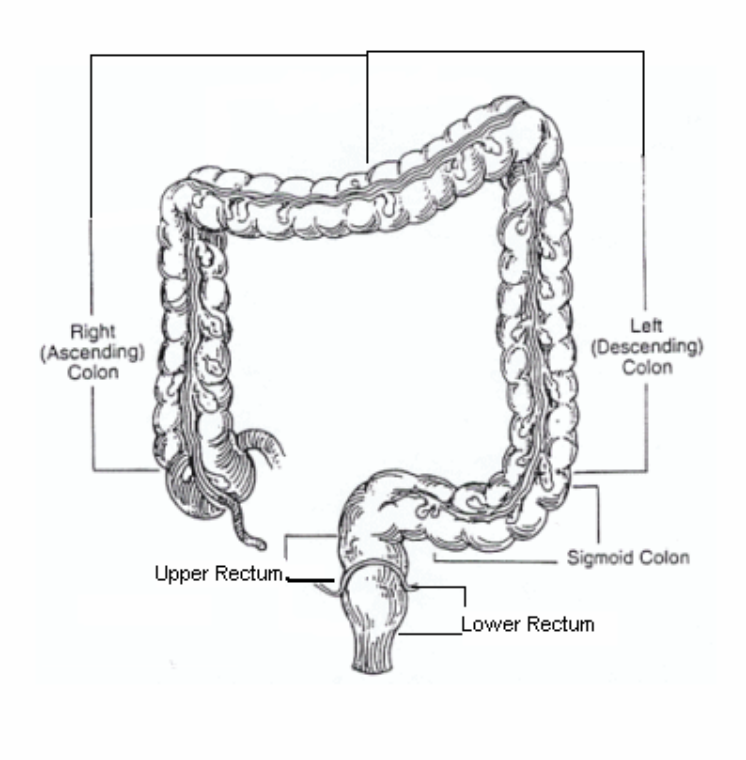


Figure 17: Locations within the Colon

Adenoma ID: each adenoma has a unique ID number associated with it. This ID will be used to keep track of adenomas and their states when performing screening and treatment, as well as tracking which adenoma caused a specific cancer.

3.2 Assumptions

Since the full course of CRC is not completely understood, the model was formulated on the basis of assumptions regarding the process of CRC, namely our "grand hypothesis." Below are the key assumptions made regarding individual CRC development. Frequent reference is made to the use of "fitting" to obtain the values for the variables described. The fitting procedures used will be discussed later in Section 4.2.

The creation of the CRC model was a compromise of available knowledge and data. Clinical knowledge was, in the main, supplied by Dr. Reid Ness of Vanderbilt University who is a Principal Investigator in the overall project. He is also co-author of several relevant publications that reported on the original Vanderbilt model (see Section 2.3.3). Dr. Ness furthermore has created and cataloged a database of CRC literature which informed much of the data found in the database for the Vanderbilt-NC State model. Earlier in this project Cindy Leibsch (Liebsch 2003) wrote a Master's Thesis at NC State in which she surveyed fifteen "experts" in CRC to determine key opinions about the development of CRC. The results of her thesis, which included statistical distributions of development elements, will be referred to as coming from the "expert panel."

3.2.1 Individual Risk Related to Colon Cancer

The "risk" related to colon cancer in an individual affects both the chances of acquiring adenomas and the inevitable progression of these adenomas to cancer.

Assumption 1: *Risk is a characteristic of individuals, dependent on family history, race, and gender and influences both the rate of adenoma appearance and the progression of the adenoma to cancer.*

All individuals have different risks of acquiring adenomas. Apparently the most important characteristic affecting risk is the presence of CRC in a first-degree relative (i.e., parent, sibling, or child). Risk also appears to be influenced to a lesser degree by race and gender. Specific risks to individuals may include insufficient fibrous diet and lack of exercise, but the exact affect of these characteristics is not known and therefore not included in this model. Also excluded from the model is risk associated with Familial adenomatous polyposis and hereditary nonpolyposis CRC which are a fundamentally different causes of CRC from those affecting adults over age 40.

Risk is modeled with two JohnsonSB distributions both having a minimum of 0.0 and a maximum is 1.0. In both cases, the distribution is highly positively skewed and the mode approaches zero. With family history, the mean is 0.11 while without any family history, the mean is 0.056. These values are similar to the results obtained by Johns, et al. (Johns and Houlston 2001), who performed a meta-analysis to determine the appropriate risk values for both with and without family history

The JohnsonSB parameters corresponding to these risk values were derived using the VIM program developed by Roberts (Roberts 2004). This program allows users to fit a desired distribution to certain statistical characteristics such as percentiles and moments.

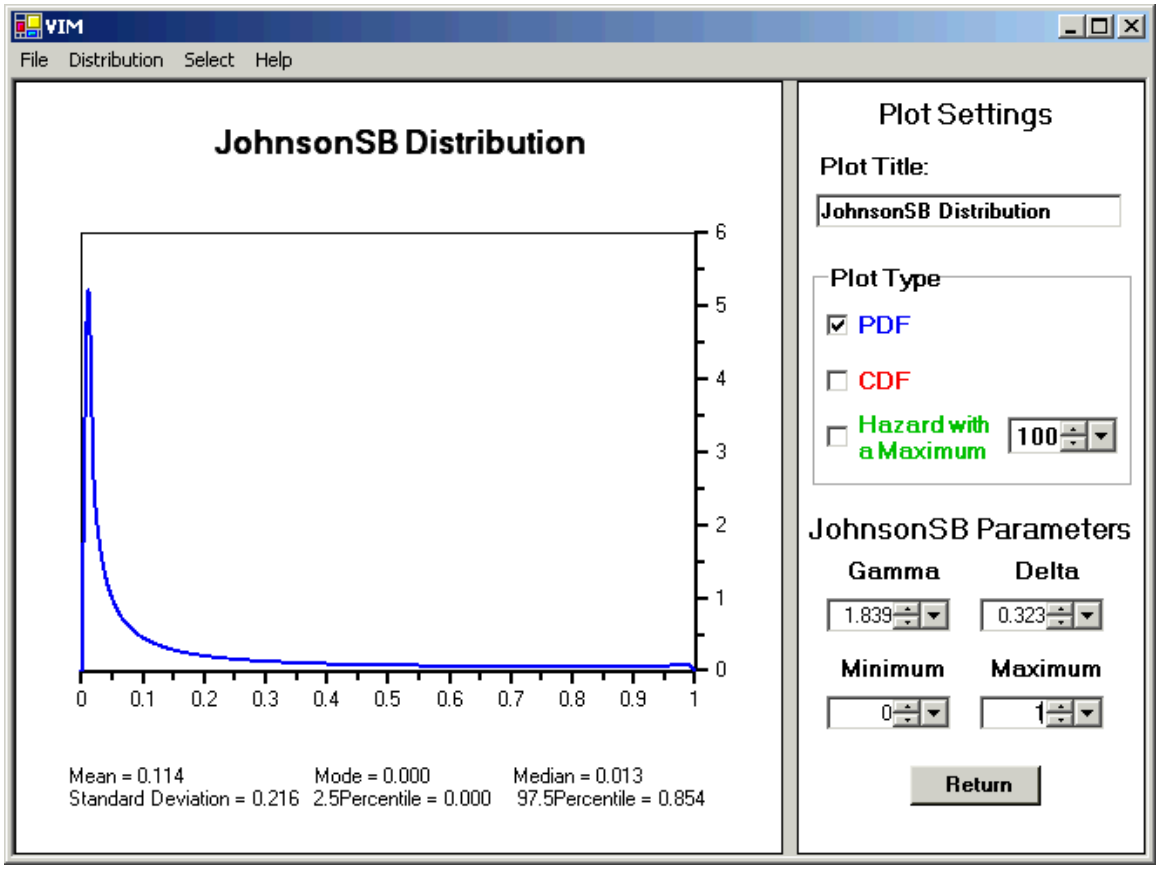


Figure 18: Absolute Risk without Family History

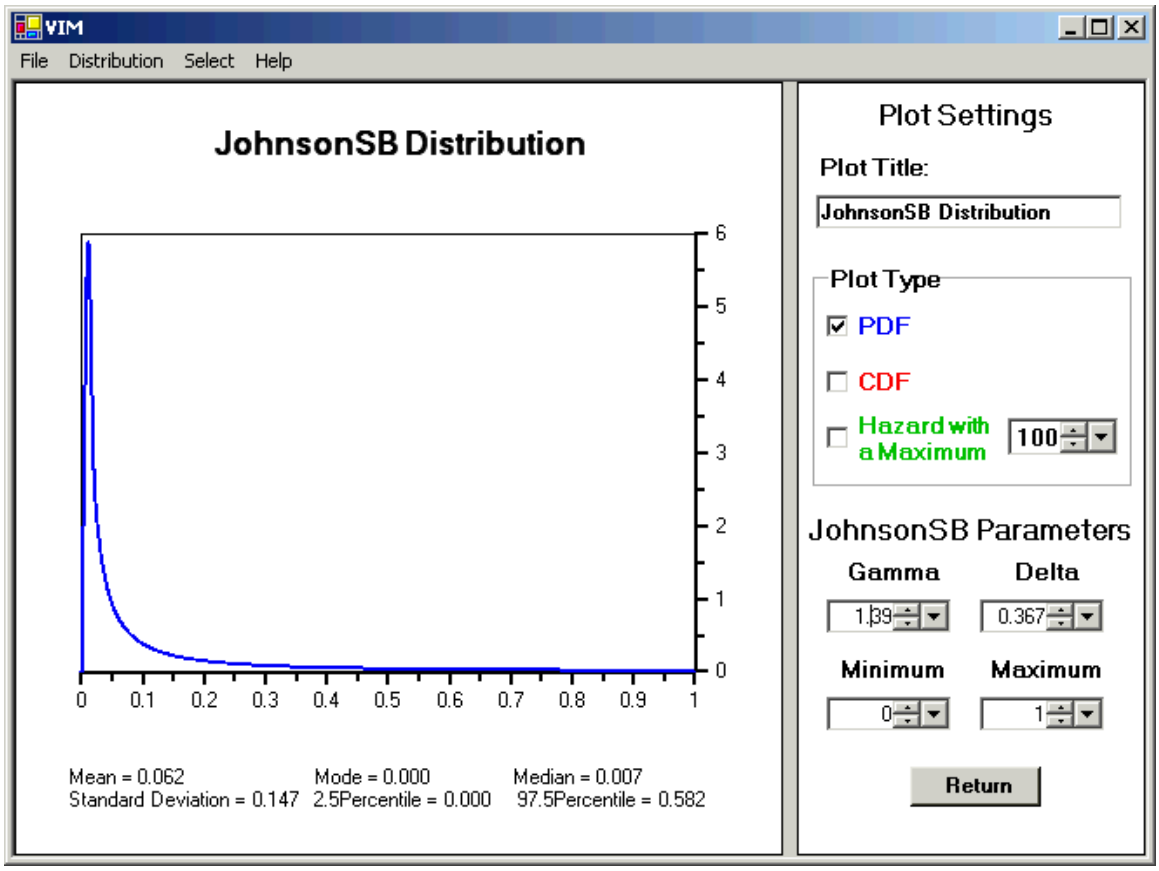


Figure 19: Absolute Risk with Family History

The prior distributions in Figures 18 and 19 represent the absolute individual risk on a scale between 0.0 and 1.0. The implication is that the lifetime risk for the average person without a family history of colon cancer is .056. Therefore relative risk would be the absolute risk divided by .056 and the distribution would range from 0.0 to 17.85 with a mean of 1.0 as shown below in Figure 20:

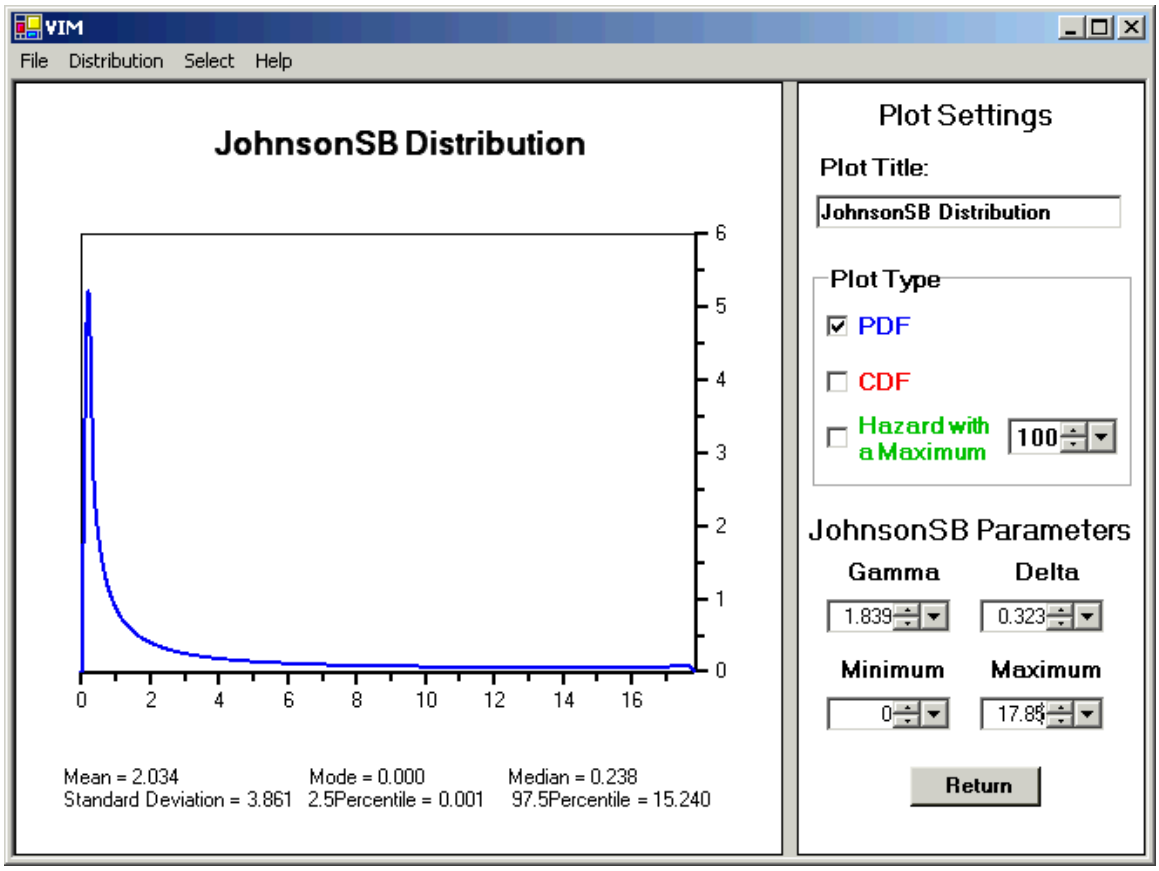


Figure 20: Relative Risk without Family History

The mode remains close to 0.0, and 50 percent of the people have a relative risk of less than 0.060. Similar relative risk can be obtained for those with a family history of colon cancer as seen in Figure 21.

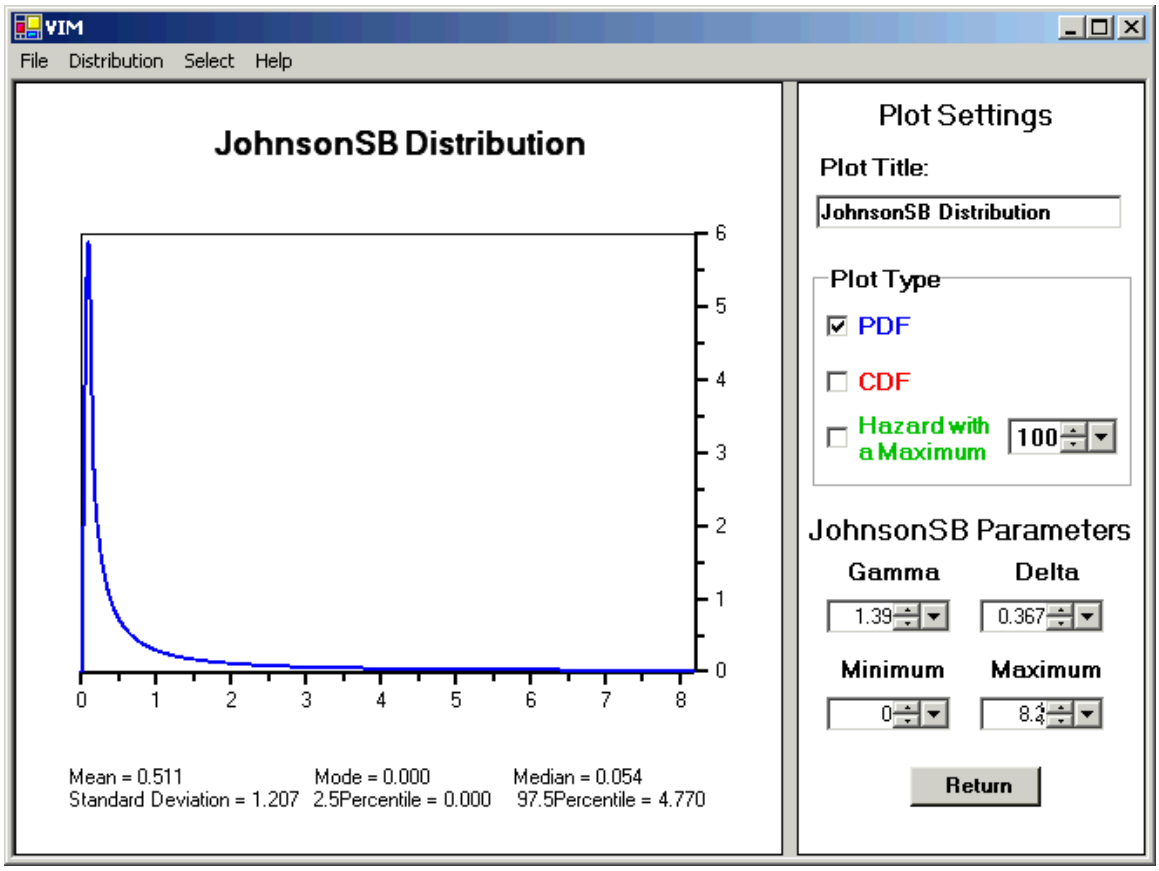


Figure 21: Relative Risk with Family History

These base risk values are then modified by a data fitting procedure (described in Section 4.2) to accurately model each race and sex combination's cancer incidence.

Assumption 2: *Colon Cancer begins as a non-visible, benign adenoma.*

Although it may be possible for "spontaneous" colon cancer to appear, there is growing biomedical evidence that colon cancer begins as a benign adenoma, not visible with current technology.

The base incidence of non-visible adenomas is determined by fitting the observed cancer incidence, adenoma incidence, and percent of people with adenomas at different ages to numbers found in several studies. The fitting procedure is explained more fully in Chapter 4.

Assumption 3: *A non-visible, benign adenoma transitions immediately to the next stage.*

From this base non-visible adenoma, there are three states to which the adenoma may transition. It can become a progressive adenoma that will eventually become cancer, transition to a non-progressing adenoma, or transition immediately to cancer. These transitions are determined by an adenoma type variable that is fit to match target values for cancer incidence.

Assumption 4: *Risk affects different people in one of three modes.*

Risk affects each individual in a slightly different way. The two main ways that risk can affect a person are by affecting the adenoma incidence rate, and by affecting the adenoma progression rate. Described below are the three modes that risk can affect these two pathways. Which pathway is taken is determined by sampling from the risk affects variable from the input database.

Incidence Only: The first and primary method that risk affects a person is by changing the incidence rate of adenomas. Someone with a lower relative risk will have the time to next adenoma appearance increased by some amount. Adenoma incidence is modeled with Non-homogeneous Poisson Process (NHPP) whose rate function is described in the inputs database. The model implements the change in incidence by multiplying the relative risk for that person by the incidence rate for that period. So a base rate of 1 per year will become a rate of 2 per year for someone with a personal risk of 2.

Progression Only: In this case, the person's risk does not affect the incidence of adenomas in any fashion. Instead, time between stages of adenoma growth and cancer incidence is changed by the person's relative risk. For example, a personal risk of 2 would change the time until an adenoma becomes advanced from 8 years to 4 years, and the time to cancer from 10 years to 5 years.

Both Incidence and Progression: This is the most complicated case. In this person, the risk affects both the overall adenoma incidence and the progression time. The motivation for the following process is based on a Leibsch survey (Liebsch 2003) in which the respondents felt that risk was not a uniform process that affected all individuals in the same manner. The experts felt that risk affected some people by increasing the adenomas incidence rate, while in others the effect showed as a decreased time to cancer. The amount it affects each is determined by the risk affects incidence variable in the database. Specifically, her study showed that the expert panel felt that risk affected only incidence in 50 % of the cases, only progression in 25 % of the cases, and both in 25 % (Liebsch 2003).

Consider the person described above who has a personal risk of 2, and assume that the risk affects incidence variable is 2/3 weighted to affect incidence. First, a modified risk value is determined for the influence on incidence. The new risk is determined using the following formula:

New Risk = (Personal Risk - 1) * Risk Affects Incidence + 1

= (2-1) * 2/3 +1

= $1\ ^{2}/_{3}$

The incidence rate function is then multiplied by the new risk to create the new incidence rate function for that individual. Finally it would yield a new rate value of $1^{2}/_{3}$ per year in this case.

Now, the new time to progression needs to be determined. Again, it is started with a new risk value for progression. The formula is the same as for incidence, except risk affects incidence is changed to (1 – risk affects incidence) to account for the percentage of the risk that affects progression. The original progression time is then divided by the new risk value to yield the new time to progression. Following the previous example of a base time to advanced adenoma of 8 years, the following computes the progression time. First, the new risk becomes:

New Risk = (Personal Risk - 1) * (1-Risk Affects Incidence) + 1

= (2-1)*1/3 +1

$= 1\ ^{1}/_{3}$

The progression time then becomes:

Progression Time = $8/ (1^{1}/_{3})$

$= 6$ years

3.2.2 Progression Types and Pathways to Cancer

These assumptions relate to the types of adenoma progression as well as how these adenomas transition to cancer.

Assumption 5: *Adenomas are of three progression types: non-progressive, progressive, and immediate progressive.*

These three progression types compose all of the adenomas that arise. Non-progressive adenomas have no chance of developing to cancer, but can become advanced to match the data on the percent of detected adenomas that are seen to be advanced. Progressive adenomas develop into cancer sometime between 0 and 60 years as a base time. The Johnson SB for this distribution is based upon the results of the expert panel (Liebsch 2003). Progressive adenomas account for 85% of all cancer incidences. Immediate progressing adenomas become cancer as soon as the adenoma develops, and should account for 15 % of the observed cancer incidence at any given age. Since these adenomas progress immediately, there is only the 0 to 10 years time period until the cancer becomes symptomatic and treatment begins.

Assumption 6: *The distribution of adenomas to progression type is dependent on age.*

As the body repair mechanisms wear down, the ability of the body to deal with anomalous cells begins to decline. This decline creates an increased incidence of adenomas overall as a person ages. It is further assumed that the percent of adenomas that are progressive generally should increase as time passes.

Assumption 7: *Incident adenoma location is gender, age, and race dependent.*

The data from the polyp studies (Winawer, Fletcher, and Rex 2003) indicate that adenomas arise in differing sections of the colon depending on the age of the person. The five sections of the colon that the adenomas can develop in are the low rectum, high rectum, sigmoid, left colon, and the right colon (recall Figure 17). This distribution of the adenomas when they develop has also been shown to depend on race and gender. Lastly, if a section of the colon is removed, any adenomas that would have been distributed to that section simply never occur.

Assumption 8: *Cancer stage progression and symptom development are lesion-specific and independent of personal risk.*

This assumption states that progression through the stages of cancer is independent of other variables. So the personal risk does not affect the transition times by any causal means. In addition, the stage progression and the time to symptoms are independent of each other.

Assumption 9: *Regional and distant metastasis rates are independent processes.*

This assumption makes it clear that the transitions from a local cancer to regional or distant cancer are modeled as competing processes. So it is not necessary for a cancer to become regional before it transitions to distant cancer. At time of symptoms, 50% of cancers should be local, 20% of should be regional, and 30% should be distant. This assumption allows for a much simpler fit of these valuations.

3.2.3 Process Assumptions

The following set of assumptions relates to the incidence of adenomas and adenoma progression to cancer.

Assumption 10: *Adenoma incidence is described by a non-homogeneous Poisson process (NHPP) whose rate function is described by a piecewise linear function.*

This function has values at discrete ages and the NHPP (linearly) interpolates values between those points. A sample of the rate function for females is shown below in Figure 22.

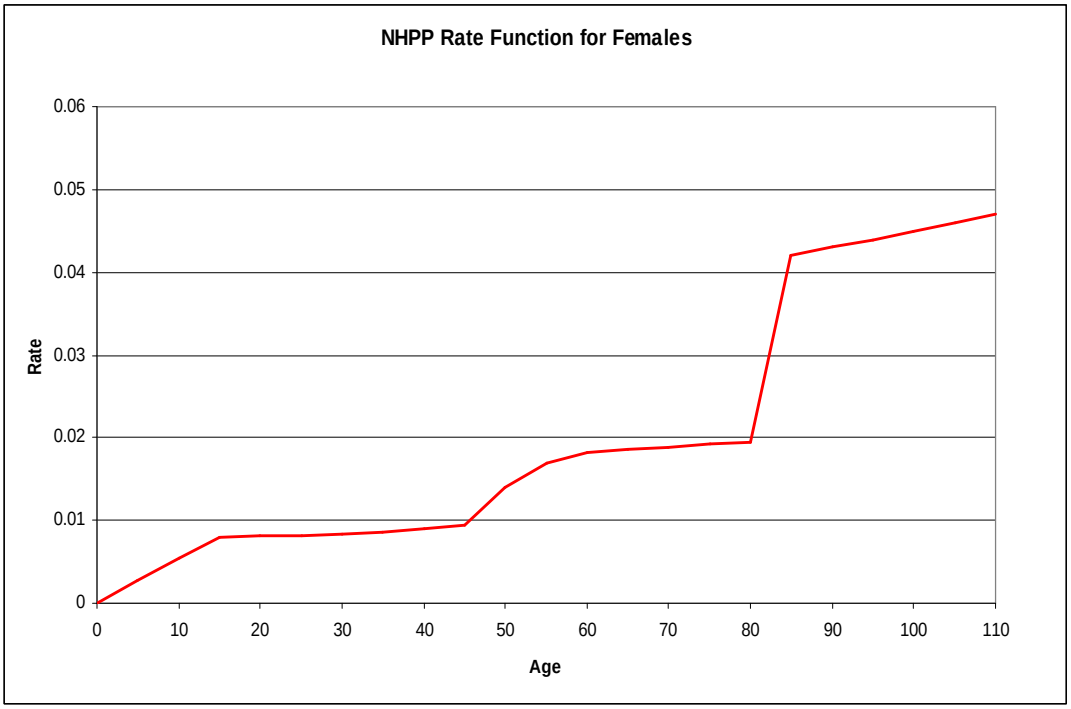


Figure 22: Incidence Function for Females

Assumption 11: *The time to cancer incidence is described by a Johnson SB distribution whose mean is 22 and mode is 20.*

These two values were determined by the expert panel (Liebsch 2003) and represent the consensus as to when cancer will appear. A graph of the distribution can be seen below in Figure 23.

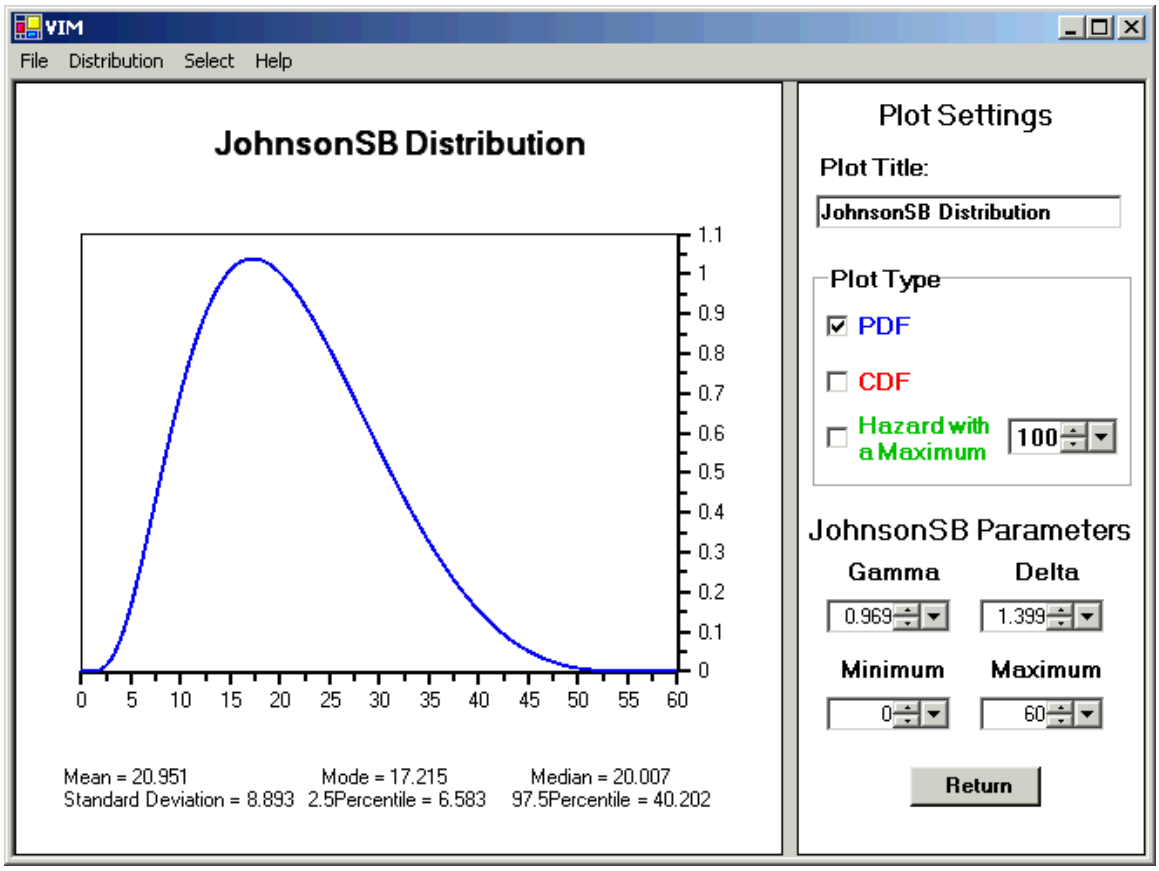


Figure 23: Time to Cancer for Progressive Adenomas

Assumption 12: *Time to local and regional cancer is determined by a Johnson SB distribution.*

This distribution allows for direct modeling of the targets for cancer stage at time of symptoms. In addition, as pointed out by Roberts (Roberts 2003a), the JohnsonSB is often preferable to the triangular distribution or the beta distribution in the absence of data. In this case the desired distribution is slightly skewed to the left (closer to 0), but otherwise is reminiscent of a normal curve.

3.2.4 Other Assumptions

The following additional assumptions describe other aspects of the model not previously discussed.

Assumption 13: *Rate of progression to death, and potential survival (from CRC) is determined by cancer stage at time of symptoms.*

Because treatment cannot begin until symptoms are detected and the cancer is diagnosed, the cancer stage dictates both the treatment and the potential survival associated with becoming symptomatic. A corollary to this is that no one can die from CRC until symptoms have developed.

Assumption 14: *Cost is associated with diagnosis and treatment of CRC*

The main costs examined by this model are the cost of a diagnostic colonoscopy upon symptoms developing and the costs of treatment. At time of symptoms, there is an initial colonoscopy done to determine the extent (stage) of the cancer. Afterwards, there is a one-time charge for the surgery and other treatment associated with the first part of cancer care. Then there is an annual continuing care cost for follow-up treatment associated with the cancer. This cost continues until the patient dies of CRC, or survives for five years. Lastly, there is a larger one-time cost associated with the patient dying from CRC (we are considering only the extra expense of dying from CRC than other "natural causes"). For purposes of the model, the terminal event occurs when the patient has 18 months left to live with CRC, or immediately upon symptoms if they will survive less than 18 months.

Assumption 15: *Utility changes are associated with health state changes and testing.*

The person's utility is a measure of the quality of life a person experiences as a result of their health state. By the time screening starts, the utility is already slightly diminished

based upon the process of aging and minor chronic ailments. The utility changes further upon detection of cancer or adenomas. In addition, temporary changes in utility can occur due to performing of tests such as colonoscopy or complications associated with these tests.

3.3 Modeling Adenoma Incidence and Progression

3.3.1 Incidence

Along with overall mortality for an individual, the incidence of adenomas is the primary driver of the model. Immediately upon creation of a person, the time to the first adenoma is determined. This adenoma, along with subsequent adenomas that develop after the initial one, is the primary factor in determining cancer incidence. Adenoma incidence is determined by a combination of two factors. The first is the non-homogeneous Poisson process (NHPP) that describes the incidence rate. This incidence rate is specific to each race and gender, but within the race-gender group it is the same for populations with and without family history. A sample of the rate function is shown below in Figure 24.

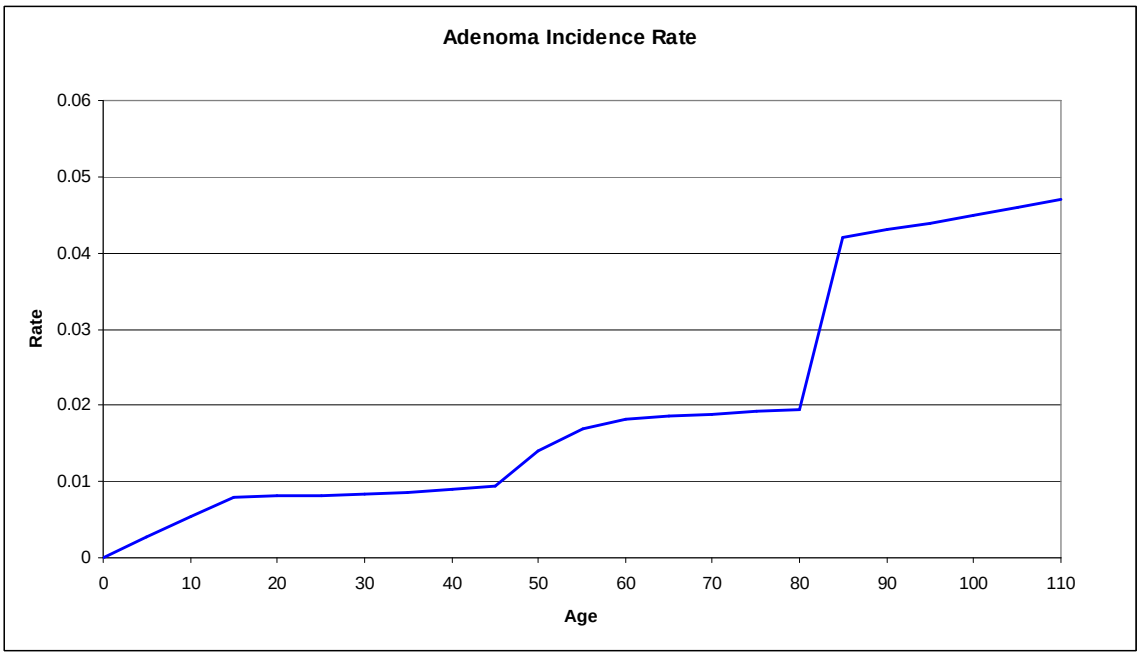


Figure 24: Adenoma incidence for white females

The base incidence is then modified by the individual's risk which is determined by his attributes. Depending on a person's characteristics, risk can:

- Affect only incidence;
- Affect only progression; or
- Affect both the incidence and progression.

These effects occur in accordance with the method described in Section 3.2.1.

Once the initial adenoma develops, the time to the next adenoma is immediately obtained. Upon generation of the following adenomas, the cycle repeats, setting the time for the next adenoma immediately upon creation. In this way there is a steady stream of new adenomas for the entire life of each person in the population.

3.3.2 Progression

Once an adenoma has been generated, it must then follow a progression pattern. This pattern can take several pathways including: immediate transition to cancer, transitioning to a non-progressive cancer, and becoming a slow progressing adenoma that will eventually become cancer. A visual representation of these pathways is shown below in Figure 25.

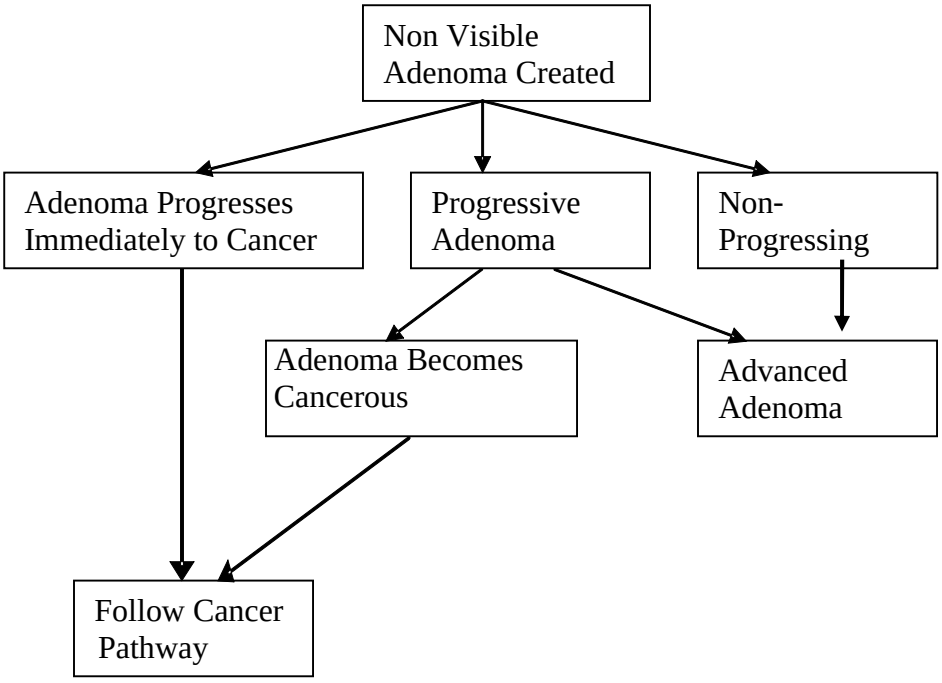


Figure 25: Pathway from Adenoma to Cancer

As can be seen from the flowchart, all adenomas start as non-visible adenomas once they are created. From that state, there is an instantaneous transition into one of three types of adenoma progression modes. The transition to these three modes is determined by the progression type variables. These variables are age dependent, with the percent that become progressive growing as the person gets older reflecting the body's inability to defend itself against defective cells deteriorates.

Immediately Progressing: This category of adenoma immediately progresses to local cancer upon generation. From there it follows the usual cancer pathways.

Progressive Adenomas: The bulk of the progression modeling occurs in this category of adenomas. There are two steps that an adenoma of this type can take. It can become an advanced adenoma as defined by its histology or become cancerous. Most adenomas will become advanced adenomas before becoming cancerous, but not quite all. So these two

processes are modeled as competing processes. Once the adenoma becomes cancerous, it follows the usual cancer pathway. However, once an adenoma becomes cancerous, the event to become an advanced adenoma is removed from the event calendar since the adenoma is already at a more advanced state.

Non-progressive Adenoma: This adenoma has no chance of ever becoming cancerous. This category is the dominant type of adenoma for all age groups, with between 90%-99% of all adenomas falling into this category. If necessary, to meet the model's requirements for the number of advanced adenomas, it is possible to allow these adenomas to progress to advanced, but not to progress to cancer. However, this function is currently not enabled since data fitting has not shown this to be necessary.

3.4 Event Modeling

As a discrete event simulation model, changes to the state of the person and their adenomas occur at events. These events cause statistics to be collected and new events to be created. Below are the events that comprise this model. When a person object is created several procedures are executed within its new object initialization. A simple event graph of the model is shown on the following page (Figure 26). The event graph portrays all events in the simulation and, when relevant, what conditions are required for an event to be scheduled. If no conditions are stated, the event is always scheduled.

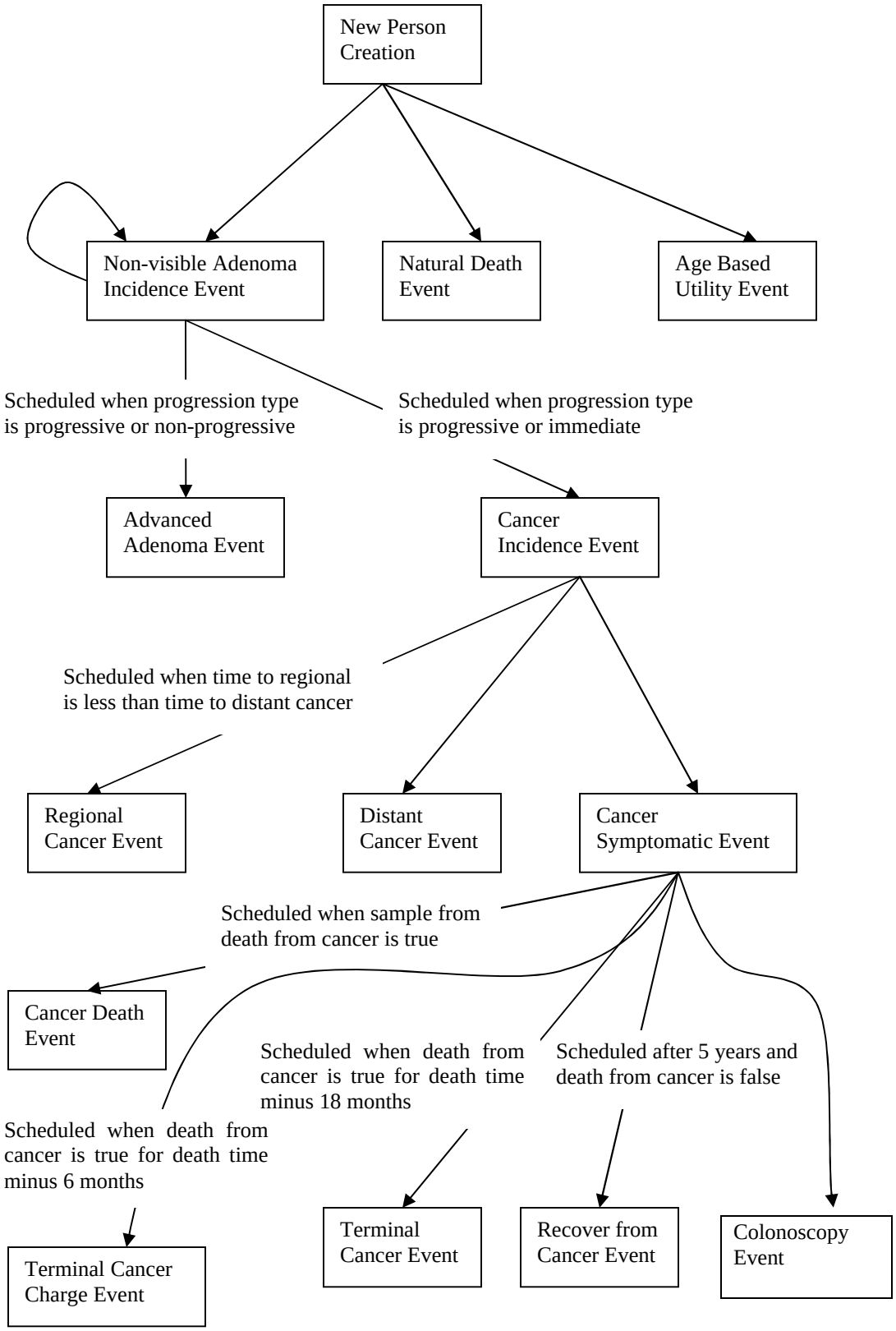


Figure 26: Event Graph of the Simulation

3.4.1 New Person Creation

Description: The creation of a new person begins the next replication for the scenario. A replication of the simulation consists of the experience of one person. This creation is the initiator of all future events.
Predecessor: None
Events Scheduled: The first event scheduled is the natural death event. This event is based on the natural lifetime of the patient without the influence of CRC. The time for this event comes from sampling the natural lifetime variable. The second event that is scheduled is the first adenoma incidence. The time to this event is set based upon the modified incidence function. This function is created by adjusting the base incidence variable by the individual risk of the person. Lastly, the creation schedules an age based utility event. This event lowers the utility of the individual to reflect chronic illness and the accumulation of minor injuries that occur with age. The time of this event is determined by a variable set in the input database.
Statistics Collected: None

3.4.2 Natural Death Event

Description: This event is the end of the natural life of a person. The person either has not experienced CRC or has survived the disease and lived until their natural life ends.
Predecessor: New Person Creation
Events Scheduled: At the completion of this event, the event calendars for the person and their adenomas are cleared and that replication is complete.
Statistics Collected: Statistics are collected about the lifetime of the person and compiled both for individuals without CRC and all people. For overall individual statistics, quality adjusted life-years (QALYs) and cost for that person are also recorded. In order to later calculate population statistics, other statistics are collected as well. These include the number of adenomas the person had at the time of death, and a binary statistic on whether that person had adenomas at all. This event represents how long a person will live if CRC

does not impact their lifetime. If a person has CRC but dies of natural causes, the continuing care costs are collected at this event. Costs associated with this can be found in the recover from cancer event description.

3.4.3 Non-visible Adenoma Incidence Event

Description: This event occurs when an adenoma develops. The event assigns the adenoma type that determines the future events and schedules the next adenoma.
Predecessor: New Person Creation, Non Visible Adenoma Event
Events Scheduled: The adenoma is immediately assigned one of three categories: non-progressing, progressive, and immediately progressing. An immediate progressing adenoma schedules a cancer incidence event immediately. If the adenoma is progressive, then two events are scheduled. First, an advanced adenoma event is scheduled. This event progresses the adenoma to the next stage of development. The second event that is scheduled is the cancer incidence event. The cancer incidence event will cause the adenoma to progress to local cancer. If an adenoma is non-progressive, then only the advanced adenoma event is scheduled. Lastly, all adenoma types then schedule the next adenoma to develop.
Statistics Collected: A counter that records the number of adenomas is incremented to reflect the presence of a new adenoma.

3.4.4 Advanced Adenoma Event

Description: This event changes the histology of the adenoma to advanced. This increase in histology increases the likelihood of discovery during a screening process.
Predecessor: Adenoma Incidence Event
Events Scheduled: None
Statistics Collected: Increments counter for number of advanced adenomas developed.

3.4.5 Cancer Incident Event

Description: This event is the transition point for an adenoma. At this point the adenoma has become cancer and starts to damage the body. The cancer will progress to regional and distant cancer in time, and also develop symptoms that tell the person to seek treatment.
Predecessor: Adenoma Incidence Event
Events Scheduled: Upon processing, this event potentially schedules three events. First, the progression time to both regional and distant cancer are determined, which are used to schedule the Regional Cancer Event and the Distant Cancer Event. The regional cancer event is not scheduled if the distant cancer event occurs first. Finally, it sets the time for progression to symptoms, causing the scheduling of the Cancer Symptomatic Event.
Statistics Collected: None

3.4.6 Regional Cancer Event

Description: This event moves the cancer stage to regional cancer within the person. This transition to regional cancer changes the person's health state to regional cancer, and will affect their survival once cancer is diagnosed. The treatment for regional cancer is more invasive, and will generate lower utilities once diagnosed.
Predecessor: Cancer Incident Event
Events Scheduled: None
Statistics Collected: None

3.4.7 Distant Cancer Event

Description: This event moves the cancer stage to distant cancer within the person. This transition to distant cancer changes the person's health state to distant cancer, and will affect their survival once cancer is diagnosed. Upon diagnosis, the treatment options are more limited, and the outlook for future survival is low.
Predecessor: Cancer Incident Event

Events Scheduled: None
Statistics Collected: None

3.4.8 Cancer Symptomatic Event

Description: It is at this event that the person first becomes aware of the cancer and seeks treatment. A colonoscopy is performed to determine the extent of the cancer, and treatment of the cancer begins. Affected sections of the colon are resected in most cases, and chemotherapy is given to all patients who have regional or distant cancer.
Predecessor: Cancer Incident Event
Events Scheduled: The first event scheduled is a Symptomatic Colonoscopy Event. This Symptomatic Colonoscopy Event is scheduled as an event to facilitate the easy interface with the screening options within the code. This event determines the extent of the damage to the colon and treatment options. As part of the colonoscopy, adenomas may be found and removed. In that case, all events associated with the removed adenomas are deleted. Should surgery prove necessary, sections of the colon are resected, thus eliminating all adenomas in those sections. As with the adenomas removed by screening, all events associated with those adenomas are removed from their event calendars. We sample from a random distribution that determines if the person will survive this cancer based upon the stage of cancer. If they do, then a Recover From Cancer Event is scheduled. If not, a Cancer Death Event is scheduled, whose time to death is based upon the stage of the cancer at symptom occurrence. When a cancer is fatal two other events are scheduled, the Terminal Cancer Event and the Terminal Charge Event, both of which are associated with the terminal stages of cancer. Utility is lowered further by scheduling the Terminal Cancer Event (currently at 18 months from death), and a final terminal cost is accrued 6 months before the patient dies with the Terminal Cancer charge event. In addition, a person can die in surgery, which results in the scheduling of a cancer death event (because the surgery was performed to treat the cancer) one month after the surgery.

Statistics Collected: At this time, the counter for cancer incidence is incremented for that specific age. This breakdown counter tracks the number of cancers that develop in 5-year periods (0-5, 5-10, etc). An initial treatment cost is also incurred, as well as a continuing care cost for the follow-up treatment of the disease. The new lowered utility of the person is based upon the stage of the cancer and the location. The costs associated with this event are a one-time charge of $20,323 for local cancer, $23,368 for regional cancer, and $26,708 for distant cancer. Expected chemotherapy (and chemotherapy complication) costs and associated utility reductions are included in the main utilities and costs associated with the differing stages of cancer, so if complications arise during treatment, no further utility adjustment is required. In addition, no further events are needed to raise utility since there is not a recovery to the pre-complication utility. Complications from the colonoscopy are covered in the colonoscopy event code.

3.4.9 Colonoscopy Event

Description: This event occurs immediately after (zero time) a Cancer Symptomatic Event and is a regular colonoscopy with no follow-up screening. It is important to note that the colonoscopy is 100% sensitive to cancer presence. Again, this section is modeled as an event to make the code portable, and allow the same colonoscopy event to be called for all screening or treatment events that require a colonoscopy in later studies.
Predecessor: Cancer Symptomatic Event
Events Scheduled: Any adenomas detected during the colonoscopy will be removed. This removal causes all events associated with these adenomas to be removed from their event calendars. There are also a small number of people who have complications with the colonoscopy. These people have a Complication Utility Event scheduled, which returns to the original utility after the complications pass. Because this is the only item that occurs in this event, a separate event description is omitted.
Statistics Collected: The main statistic collected is the cost of the colonoscopy. In addition any adenomas removed are counted in their respective age-based group.

3.4.10 Recover from Cancer Event

Description: This event is the time at which continuing care costs end. Should a person survive CRC, after 5 years of treatment, they are discharged and resume their normal lifetime, so no further costs are accrued. In addition, if a person survived distant cancer, their utility is raised to that of someone with regional cancer.
Predecessor: Cancer Symptomatic Event
Events Scheduled: None
Statistics Collected: This event represents the end of care for the cancer, so costs for the five years of treatment are collected. These costs are $539 per year for local cancer, $2,461 per year for regional cancer, and $26,855 per year for distant cancer. If the patient survives distant cancer, their utility improves to that of regional cancer (0.5-0.7 depending on cancer location). Otherwise, the patient retains the same post-cancer utility.

3.4.11 Cancer Death Event

Description: This event is the cancer death event of the person. At this time the person dies from CRC, which is prior to their scheduled natural death.
Predecessor: Cancer Symptomatic Event
Events Scheduled: Upon processing of this event, all future events are removed from all the event calendars and this replication is completed.
Statistics Collected: The main statistic that is recorded is the life span of the person. This death statistic is recorded in two collectors, one for only CRC victims, and one for the overall population. In addition, the death of the person is recorded in a separate collector, broken down by age, which is used to track the number of people who had adenomas who die in each age group. This same breakdown style is used in another collector that tracks when adenomas die. In other words, this collector counts the number of adenomas at the time of the person's death and records them into that time frame. Then the final costs for the individual are collected in the overall cost collector.

3.4.12 Terminal Cancer Event

Description: This event occurs when a patient reaches the final stages of CRC that will eventually kill them. The event occurs 18 months from death, or immediately if the person experiences symptoms and will not survive 18 months.
Predecessor: Cancer Symptomatic Event
Events Scheduled: None
Statistics Collected: The utility of the patient is lowered for the remainder of their life. This utility valuation is currently set at 0.25. The cost of continuing care is also recorded at this time. The valuations for the cost of care can be found in the recover from cancer event description.

3.4.13 Terminal Cancer Charge Event

Description: This event is the last event scheduled by the Cancer Symptomatic Event and represents the final treatments that occur during months before death from cancer.
Predecessor: Cancer Symptomatic Event
Events Scheduled: None
Statistics Collected: There is an additional one-time charge for the care given to the patient for the remainder of their life. This cost is $21,172 for the end of life care given to the patient.

3.4.14 Age Based Utility Event

Description: This event reflects the transition to a slightly lower quality of life based upon reaching a more advanced age. Currently, this time is set to occur at age 50, when accumulated minor ailments and injuries lower the average utility very slightly.
Predecessor: New Person Creation
Events Scheduled: None
Statistics Collected: The utility of the person is lowered from 1 to 0.91 if other causes have not already lowered it further.

3.5 Role of Data in the Simulation

Within the simulation, there are two distinct and separate uses of data. First, data is used as the input values for the simulation. All of this information is stored within an Access database for ease of use and portability. The second use of the data is as targets for verification and validation of the model in Chapter 4.

3.5.1 Input Database

The Access database attached to the simulation stores all data necessary for the simulation to run. Input distributions are fit to target guidelines using VIM (Roberts 2004) before being stored in the input database. The database itself has a component design. The primary table that describes the cohort in the simulation is the PopulationAtRisk table. A portion of this table can be seen in Figure 27 below.

PopulationAtRisk : Table

Scenario	Scenario Name	Reference Year	Age	Gender	Race	Family History	Population Size
11	WM Base - H	2003	40	Male	White	First Degree	210000
12	WM Base - N	2003	40	Male	White	None	240000
13	WF - H	2003	40	Female	White	First Degree	180000
14	WF - N	2003	40	Female	White	None	210000
16	BF - H	2003	40	Female	Black	First Degree	210000
17	BF- N	2003	40	Female	Black	None	240000
0	Base Case	2003	50	Female	Black	First Degree	10000

Figure 27: Screenshot of PopulationAtRisk Table

This table contains all the cohort data the simulation needs. The simulation will retrieve the values for the input variables and then run the simulation. The PopulationAtRisk table stores a scenario ID and a description of the scenario. Along with this information are columns identifying the race, gender, and family history of the population to be studied in the scenario. There is a population size column that specifies how many people it is to simulate. Scenario information can be constructed or edited from within the simulation user interface.

The basic simulation variables and their valuations are created and modified by the database administrator. These two tables are the InputVariables table and InputVariableValuations table. The InputVariables table (partially shown below in Figure 28) stores the variable name in a long text form such as "Time to Asymptomatic Cancer" along with a unique variable ID for that variable.

InputVariableValuations : Table

VariableID	Gender	Race	Family History	Input Distribution	Parameter Name	Valuation
1	Female	All	First Degree	JohnsonSB	delta	0.816
1	Female	All	First Degree	JohnsonSB	gamma	2.226
1	Female	All	First Degree	JohnsonSB	maximum	1
1	Female	All	First Degree	JohnsonSB	minimum	0
1	Female	All	None	JohnsonSB	delta	0.661
1	Female	All	None	JohnsonSB	gamma	2.6
1	Female	All	None	JohnsonSB	maximum	1
1	Female	All	None	JohnsonSB	minimum	0
1	Female	All	Both	johnsonSB	delta	0.323
1	Female	All	Both	johnsonSB	gamma	1.839
1	Female	All	Both	johnsonSB	minimum	0
1	Female	All	Both	johnsonSB	maximum	1
1	Male	All	First Degree	JohnsonSB	delta	1
1	Male	All	First Degree	JohnsonSB	gamma	2.6
1	Male	All	First Degree	JohnsonSB	maximum	1
1	Male	All	First Degree	JohnsonSB	minimum	0
1	Male	All	None	JohnsonSB	delta	0.75
1	Male	All	None	JohnsonSB	gamma	2.81
1	Male	All	None	JohnsonSB	maximum	1
1	Male	All	None	JohnsonSB	minimum	0

Figure 28: Screenshot of InputVariableValuations

To minimize error due to randomization, each variable uses its own unique random number seed, which is also stored within the database. Thus, the simulation employs the concept of "common random numbers" (Law and Kelton 2000). These common random numbers mean that multiple scenarios can be run on the same population, and the same random numbers will be assigned to each random variate. So each person will have the same natural life in all scenarios and have the same numbers sent to the adenoma incidence generators, and so on. In this manner the effects of the screening strategies become more apparent since they are the only source of change in the model.

Additionally, InputVariables stores the category and measure of the variable. The category groups the variables into loosely related clinical categories such as adenoma progression, utilities, or costs. Measure is a unit-specific breakdown that is chosen from

a dropdown menu. These describe the type of distribution associated with the variable and can specify if the variable is a probability, time distribution, or risk distribution among other choices.

InputVariableValuations (see Figure 11 for a sample of the table) is the table that specifies the values for the variables in the InputVariables table. For example if the Variable 1, "individual risk" is described by a Johnson SB distribution, the first section of InputVariableValuations is the variable ID along with the race, gender, and family history to which the valuation applies. Using these columns for each combination allows different populations to have their own valuations for each variable if necessary. For example, individual risk is dependent on race, gender, and family history; and so will have 8 sets of valuations within the table (2 races * 2 genders * 2 family history = 8 valuations). Next is the distribution name, which in this case is a JohnsonSB. The next column is the parameter name, followed by the valuation of that parameter. Since the JohnsonSB has four parameters associated with it, each of these valuations will have four rows of information; one for each parameter – Delta, Gamma, Min, and Max. Based upon the distribution type, the simulation knows which parameters to look for and finds the valuations associated with each of the parameter key words. To complete this example, the individual risk will have one row in the InputVariables table, and 32 rows in the InputVariableValuations table (the 8 valuations * 4 parameters). The simulation uses these tables to retrieve the distribution it needs based upon the previously defined scenario parameters.

Once all inputs have been retrieved, all variable valuations are then written to the ScenarioInputs table. This table cannot be changed since it represents a complete scenario. Once a simulation is run for the first time, the valuations for it are retrieved from the InputVariables table and written to the ScenarioInputs table. Every time that scenario is run from then on, the valuations are retrieved from the ScenarioInputs table. This switch of data sources provides two benefits. First, it permits scenarios to be repeated exactly. Furthermore, having two source methods allows the database administrator to update valuations as new information becomes available without the fear

of changing the values used to simulate a prior scenario. If a new study using new information is desired, the user needs only define a new scenario to see the effect of the new valuations.

All constants associated with the simulation are also stored within the database. This design has two important benefits. First, it provides for a central location for the database manager should any of these values change. Since an important part of the final analysis of the alternative screening methods is the sensitivity analysis that will be performed in the future, it is important to be able to easily change any values associated with the simulation. Consequently, keeping constants out of the code eliminates any "magic numbers" whose interpretation depends on the coded value. Instead these values are clearly labeled in the database. Thus values for these can be changed without changing the simulation code.

3.5.2 Data Sources for the Simulation

All of the distributions and constants that are stored in the database that composes the simulation run must have a reference source. Within this simulation, there are three primary sources. First are the distributions that are fit manually so that the simulation output matches certain established target parameters. This type of input will be discussed further in Chapter 4.

Second, there are distributions and constants with values based upon medical studies, or have a standard value agreed upon by an organization such as the National Cancer Institute (National Cancer Institute 2003). If set by such an organization, there are predefined ranges of valuations that are agreed upon, such as colonoscopy sensitivity values. This range defines the possible values for sensitivity analysis and also allows for standardization between different simulation models of CRC. The other alternative for this type of data is data that have a clinical research background. Different studies may observe slightly different results when studying the same phenomenon. This information

provides a range for future sensitivity analysis. This information was compiled and provided to the simulation team by Dr. Ness after a thorough literature review.

Lastly, the simulation includes data from an expert panel conducted by Liebsch (Liebsch 2003). As explained in Liebsch 2003,

> *fifteen experts from the areas of gastroenterology, epidemiology, and microbiology were recruited to serve on the expert panel. Three rounds of web-based surveys were conducted to reach consensus on four different study objectives related to adenoma development and cancer progression. The final simulation model inputs were developed using the estimates and the VIM distribution-fitting software.*

This process was followed to determine to what extent risk affected progression and incidence. The panel agreed that risk affects both incidence and progression, and also reached a consensus as to the extent of the effect. The final result of this section was that risk would affect only incidence in half of the population, only progression in another quarter, and would affect both for the remaining quarter. The remaining two distributions determined by this system were the time to cancer and time from cancer incidence to symptoms.

3.6 Modeling of Natural Lifetimes

To simulate the effects of colorectal cancer, it is first necessary to know the lifetimes of individuals without the effect of CRC. To determine these lifetimes, the lifetime probability functions for all possible age, gender, race, and family history groups are needed. Currently there are life tables for each race and gender with birth years 1968 to present (Wilmoth 2003), and tables for gender with birth years from 1900 to present within the Berkeley Mortality Database. Lifetimes are also given in both sets of tables for all groups. Also the probability of death is given for these same groups. Since the simulation needs to be run using an “at risk” population of a specific race as well as

gender that was born before 1968, it is necessary to adjust the life tables for birth years from 1900 to reflect the differences according to race.

3.6.1 Life Table Adjustments from Race

To adjust the life tables for race, a cumulative distribution function (CDF) for death was constructed for each year of life beginning in year 1900. The probability density function (PDF) was examined and discarded as a possible method of adjustment because the probability of dying in a given year would vary significantly from year to year, sometimes switching between increasing and decreasing the death probability. Figure 29 shows the PDF values for the three categories of white females for the birth year or 1968.

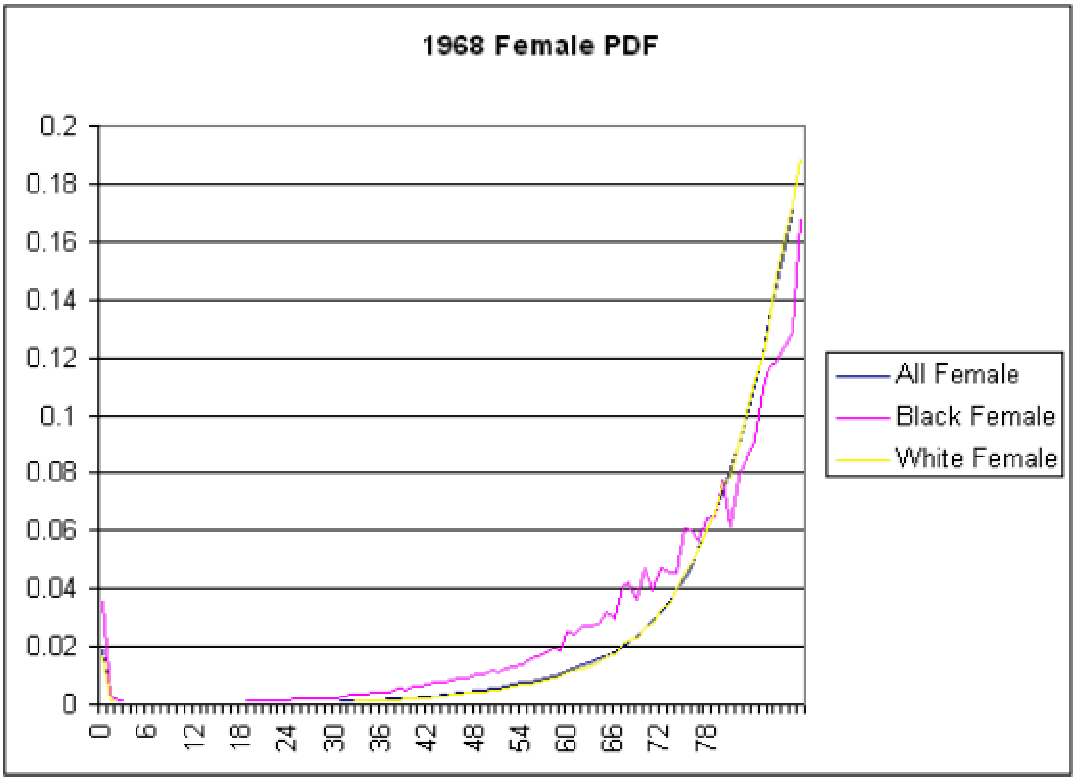


Figure 29: Instantaneous Probability of Death for Females Born in 1968

As can be seen the white female PDF continually oscillates above and below the overall rate and the black female adjustments are erratic, so there is no clear trend for the PDF values.

For this reason, the CDF was chosen to represent the death rates within the life table because of its clear trend from year to year. A graph of the CDFs for 1968 is shown below in Figure 30.

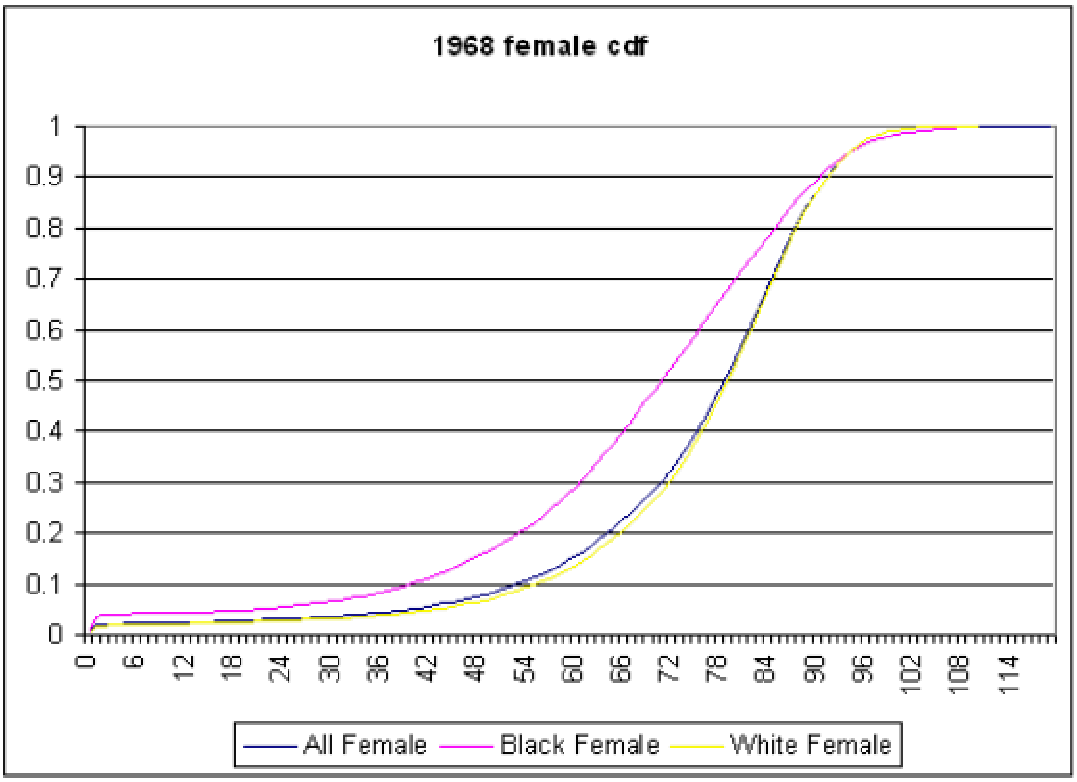


Figure 30: Probability of Dying For Females before Specific Ages

As can been seen from the graph, there is a clear trend for the each race to be different from the overall female population, which lends itself to modeling previous years using some adjustment scheme. This adjustment is close to constant over all birth years, but does show indications of a trend. This trend is discussed later in this section.

The method chosen is to adjust the difference between the CDF and either zero or one, depending on whether there is an increase or decrease to the CDF. By adjusting the

percentage change in this difference, it is possible to maintain the adjusted values as a CDF. Without such a method, it is possible that the adjusted function would go above one or below zero, depending on the specific year that was being adjusted.

The actual calculation proceeded as follows. First, the adjustment factor for race is computed to see if it will increase or decrease the CDF value.

If the adjustment will reduce the CDF, then:

- Given a specific age and birth year, reduce the CDF by the percentage value given for that specific race.

If the adjustment factor will increase the CDF:

- Calculate the difference between the CDF value and 1.0.
- Given a specific age and birth year, reduce that value by the percentage given for that race.
- Calculate the new CDF value

This method allows any given CDF value for a specific birth year to be adjusted for any death age, thus modeling the effect of race. The actual adjustment values are a straight average of the percent change for each birth year.

Figure 31 shows these trends for the white female adjustments. It is tracking the adjustment factors needed to create the white female data from the overall female data, and uses the current ages of 41, 55, 79, and 81 as examples. As can be seen from the graph, there is a slight trend in the white female population. However, the magnitude of the adjustment itself is very small, so the effects of the trend are minimal. The black female population, which has a larger absolute adjustment, does not show this trend. Therefore, a fixed adjustment is used for all populations instead of a floating adjustment factor based upon age.

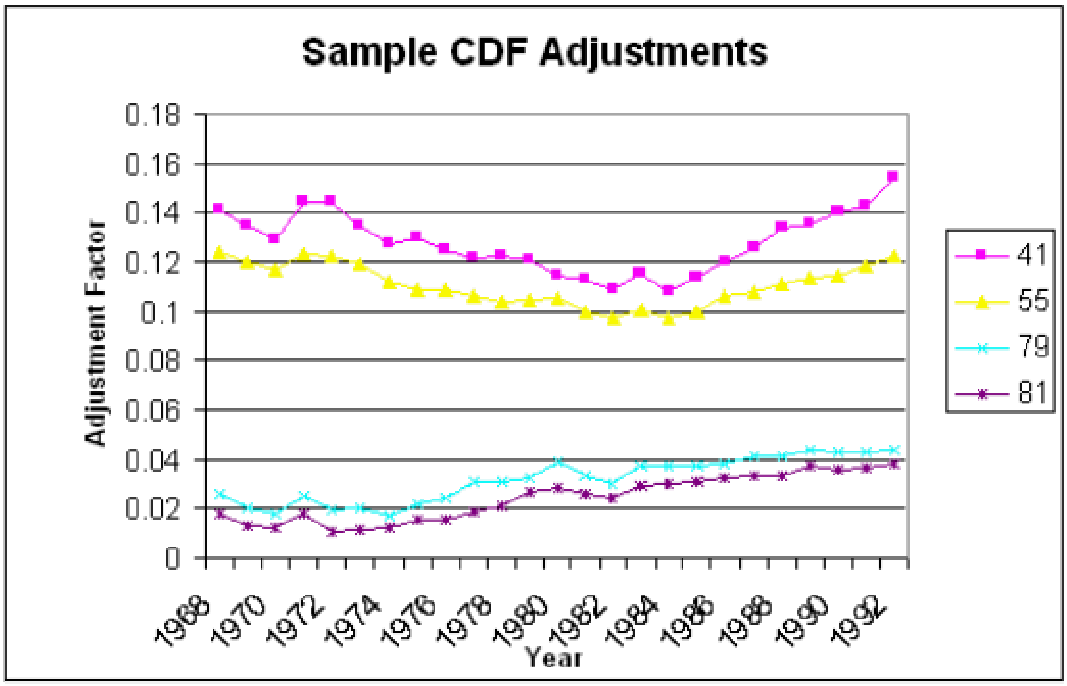


Figure 31: Race Adjustment Factor at Various Current Ages

3.6.2 Life Table Adjustments from Age

In order to assign polyps/cancers to persons before screening, the existing Vanderbilt model (Ness et al. 2000) starts everyone at age 0, but then excludes persons who died or had CRC detected before their start age, so no adjustment to the life tables are necessary due to age. Removing the segment of the population that has already died produces the same effect as shifting the CDF towards the starting age.

3.6.3 Life without CRC

There exists data on the probability of dying from CRC given a specific current age and possible death age (National Cancer Institute 2003).With these data, it becomes possible to compute the mortality due to CRC to find lifetime distribution of an individual without the effects of CRC. This computation can then be used to determine the effect that CRC has on the population, as well as providing a measure of the benefit of its removal. Since

cancer incidence and mortality vary by year, multiple years' data were chosen to yield the most accurate result. For the white population, the sample years of 1973 to 1982 were chosen. Cancer incidence in this population increases until about 1982 when screening comes into effect and then slowly declines for the remainder of the data sample. So these years best represent the cancer incidence and mortality prior to current screening methods, and thus will give the best impact of the benefit of CRC's elimination. All of the cancer mortality data is stored in 5-year increments, so it is necessary to interpolate to find the yearly mortality due to CRC.

3.6.4 CRC Mortality Rate Interpolation

Since the CRC mortality rates are listed in 5-year increments, some interpolation method is necessary to get yearly rates for all ages. The DevCan program (DevCan 2003), which is used to query the SEER database, allows the selection of age ranges, and does the interpolation for the user using the method discussed by Fay (Fay 2003). This program eliminated the need for manual interpolation for ages 0 to 95. However, there is only one mortality probability given for all ages above 95 (95+). Because the life tables used in the simulation go to 110, this age was selected as the value for the 95+ category. Linear interpolation was used to derive the values between 95 and 110.

It is important to note that the CRC rate data are grouped by race and gender, but not by birth year. The data are not split by birth year because the study (National Cancer Institute 2003) was examining cancer probabilities for 1975 and 2000 only. Consequently, people of all birth years were included when creating the CDF for colon cancer probabilities.

The standard linear approximation formula used to create the mortality rate for ages 96 - 109 is explained below:

- Determine cancer rates for the ages of 95 and 110
- Calculate a weighted average of two of probabilities using the death age's distance from the category values to determine the actual probability of death from CRC.

The following example using a white male with a death age of 106 illustrates this calculation using the 2000 cancer data:

- Suppose the death probabilities (PD_x) at age *x* from the cancer tables are as follows
    - $PD_{95} = 0.025203$ $PD_{110} = 0.025509$
- The weighted probability for age 106 would then be:
    - $PD_{106} = 0.025203+(106-95)/(110-95)*(0.25509-0.25203) = 0.02745$

3.6.5 Cancer Rate Adjustment to Life Tables

Given the adjusted life tables and the cancer rates by year, it is possible to construct a new distribution of lifetimes without cancer. A sample of this table is shown below in Table 3. The distribution itself is specified by an empirical continuous distribution.

race	gender	birth_year	age	rate
Black	Female	1900	0	0
Black	Female	1900	1	0.133254
Black	Female	1900	2	0.165403
Black	Female	1900	3	0.180728
Black	Female	1900	4	0.191588
Black	Female	1900	5	0.200078
Black	Female	1900	6	0.206308
Black	Female	1900	7	0.210798
Black	Female	1900	8	0.214009
Black	Female	1900	9	0.216489
Black	Female	1900	10	0.218639
Black	Female	1900	11	0.220779
Black	Female	1900	12	0.223109
Black	Female	1900	13	0.225629

Table 3: Sample of Cancer Adjusted CDF table

The calculations shown in the above section would be repeated for each age, race, and gender combination to create this table. This calculation will allow users to view the CDF for a given current age as well as increasing the speed of computing the CDF values since only data retrieval is necessary. This result could be important if multiple groups were run through the simulation, requiring the CDF to be recreated several times over the course of the simulation.

Using this method, the adjustment from an initial age is used. Specifically, the following steps are used for each age. First, subtract the CDF value for CRC death from the overall death CDF. Divide this value by one minus the CRC death CDF value to determine the new CDF value. This result brings the overall distribution back to a CDF since the decrease in death due to cancer will increase the probability of natural death at certain ages.

Using hypothetical values for a 50-year-old white male born in 1950 (since the simulation year is 2000), the computation steps are as follows:

- A cancer death probability of 0.03 is calculated for that age
- A death probability of 0.1 is calculated (after adjustment due to initial age)

- The new death probability without cancer is (0.1-0.03) / (1-0.03) = 0.0722

These steps are then followed for ages 51 through 110 to construct a revised CDF without cancer.

Within the data table, the death probability is calculated for each birth year in the desired range. The current simulation requires only birth years 1915 till 1980 need to be included within the database. This adjustment speeds computation as well as reduces the effort required in data entry.

3.7 Summary of Chapter 3

This chapter describes the simulation of the natural history of CRC. We reviewed the main objects used in the simulation, as well as the way they interact with the model. Key assumptions were explained. Simulation events were identified. In the process of creating this description, the pathways of these objects were discussed along with the events that that affect them. The input database structure and rationale were discussed to provide a framework for the inputs that are needed for each simulation run. Finally the daunting, but necessary task of determining natural life without CRC is presented.

4. Analysis of the Natural History Model

With the completion of the model construction as described in Chapter 3, the next task was to verify and validate the model, and then analyze the results. The verification actions were to ensure that all portions of the model worked as expected; the validation included the more time intensive task of making sure that the model output matched targets obtained from the medical literature. The analysis involved comparing the simulation output to the original Vanderbilt model outputs. In addition, the simulation output was compared to a simple lifetime analysis to externally validate the impact of CRC on a patient's life.

4.1 Verification

As the model is being created, verifying that it was working appropriately was integrated into the design process. There were three aspects of this verification process. They were: 1) stepping through the simulation source code line-by-line and routine-by-routine to verify the general program flow, 2) creating and examining an output trace file, and 3) watching the key variables within the execution of the code. Throughout the verification process, changes to the code to correct bugs were made as needed. The programming employed Microsoft Visual Studio .NET 2003, which provides an integrated development environment for constructing, compiling, testing, and executing code. This development environment contains numerous tools to construct robust and efficient programs for Microsoft Windows quickly.

4.1.1 General Flow Verification

Verifying the general program flow was the most cursory, and was done primarily to insure the correct execution of the code and the behavior of the event calendars. Several people were processed through the simulation using the Visual Studio .NET "debug" feature to watch the processing of the simulation from within the code. One intention

was to verify that the correct distributions were being retrieved from the database and sampled properly. When all variables were retrieved correctly, the next step was to look at the overall program flow and make sure that the people were advancing through the process correctly and following the expected decisions at each event.

4.1.2 Trace Output Analysis

Trace statements have been added to the code. These trace elements are placed at every event so that they output the event time, adenoma ID if appropriate, and the action the code is taking relative to that event (see example of trace in Figure 32). The trace statements are written to an external file whose content is then examined for discrepancies. Since the general flow of the model has usually been verified, the trace provides a summary of experience that is more easily seen as a whole, such as time between events and response to cancer symptoms. The time between adenomas can also be examined and compared to the expected rates from the distribution. Other time-based events can be observed. For example, an error in the NHPP calculation was identified when all adenomas after the first one occurred exactly 20 years later. In addition the trace identified an error that was missed during the general flow verification. The model initially scheduled both regional and distant cancers every time a cancer arose. If the time to regional cancer was long enough to occur later than the distant cancer event, the model was having the cancer switch from distant to regional cancer. This transition is obviously not a correct pathway for the disease, but was difficult to detect using only the flow. Once detected in the trace, a section of code was added that made sure to only to schedule a regional cancer event if it occurred before the distant cancer. This small check had the added benefit of reducing the run time slightly by eliminating an event in some cases.

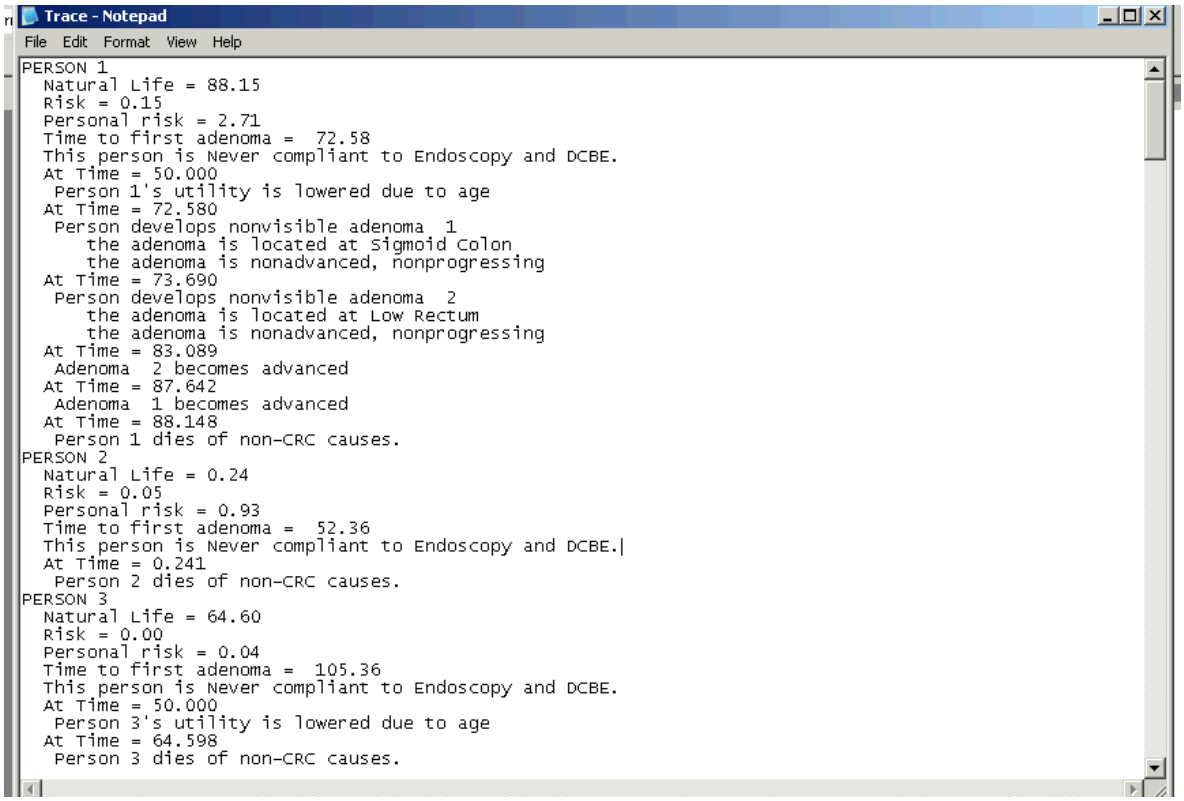


Figure 32: Trace of Simulation

4.1.3 Step-by-Step Processing Within Events

The most comprehensive verification occurred when the Visual Studio debugger was used to step through the more complicated events like the adenoma incidence event and watch the key variables within the event to ensure proper functioning. The methodology for verifying one event is given as an example for the other events. When the adenoma incidence event is triggered, the type of progression of the adenoma must to be determined. Here the simulation must sample from the proper distribution and correctly assign the adenoma to one of the three types. If the adenoma could progress to an advanced adenoma, then the simulation must correctly check the risk affects variable to see if risk affects progression in this individual. If it does, then the calculations were done by hand to verify that the time to next stage was sampled correctly once the adjustment due to family history was added. Lastly, the scheduling of the next adenoma was viewed to see if it correctly scheduled the next adenoma, and to further verify that

the simulation was sampling from the NHPP for incidence correctly. Similar step-by-step processing was done for the other larger events to ensure their proper functioning.

4.2 Output Targets

As described in Chapter 3, the Vanderbilt-NC State natural history CRC model is a "grand hypothesis." It consists of numerous assumptions about the natural history of CRC in individuals. More than the assumptions, the model connects variables influencing this history through causal numerical relationships, such as events. Input values for all variables in the model are incomplete, especially those with random components having age-dependent parameters.

Direct estimates of many of these input values are not present in the medical literature. Also some of these estimates, such as how long it takes cancerous adenomas to progress to more significant states are unobservable (it would be completely unacceptable and unethical to allow such adenomas to progress without immediate action). Therefore some input values within the model must be inferred from other known values.

4.2.1 Finding Missing Values

The problem of missing values in a data set is a well-researched problem in statistics. The problem occurs when some values are unreported. The most common application of this problem is in surveying, where some people may not respond to certain questions. The problem also occurs in clinical trials where a patient either will not or is unable to continue a trial. There are several approaches discussed in the literature (Committee for Proprietary Medical Products 2001). The first and most often employed method is to perform a complete case analysis that looks only at the complete sections of the information. The main alternative is to impute a "reasonable" alternative measure for the variable. The valuation is determined by attempting to average over the missing data to create a smoothed function (Schafer and Olsen 1998).

Unfortunately, the problem of missing data in this simulation is not as simple. While the process may be understood, numerical descriptions of the processes have not been developed in most cases. What this yields is an almost complete lack of input values, but output values that are observable through different studies. In addition, any changes to the input may not affect the simulation output for a number of years and thus create delayed results which further complicate the input specification.

Matching CRC results is not as simple as in a more observable processes with known data and distributions (Yorke-Smith and Gervet 2001). In those environments all the processes can be seen and tracked to get samples of the input values associated with each process. Because of the complex nature of CRC, simple substitution of missing values is not possible, regardless of the criteria. So to match CRC results, some form of input fitting must be accomplished to create a process that mimics what little is known about the real system.

4.2.2 General Methodology for Matching Output Targets

CRC incidence is the primary concern of the natural history model, and is the primary output target that is used to "fit" the model. After cancer incidence, the percent of people with adenomas is the next target that needs to be matched. Within the remaining degrees of freedom, the adenomas per 100 people are the last target that needs to be met by the simulation. Most of the variables within the simulation were predetermined based upon either expert opinion studies or established by comprehensive medical studies. The input variables where no distribution was known in advance were the adenoma incidence rate, the risk function for the individual, and the percent of adenomas that were progressing.

The most intuitive approach was to separate the fitting processes into stages focusing first on the most important adenomas leading to cancer. Hence progressing adenomas were considered first. Then non-progressing adenomas can be added.

Initially the goal is to make 100% of progressing adenoma fit cancer incidence by age group. Once an age group was fit, the next age could be fit by adding incidence to later age groups until all age groups had been fit. Upon completion of the cancer fit, non-progressing adenomas could be added back in until the percent of people with adenomas were met. Finally, the risk could be adjusted to make the adenomas per 100 people correct.

However, this approach produced very poor results. Since cancer takes an average of 20 years to develop, and 22 to become symptomatic, there is not a simple method to adjust the incidence upwards to meet the adenoma targets while correctly lowering the percent progressing to keep the cancer incidence the same. Even if this obstacle is overcome, the risk adjustment to meet both adenoma targets ruins any previous fit. Because personal risk affects the distribution of time to cancer, the end result is that any adjustment made in the risk function changes when cancers appear, and necessitates refitting the cancer incidence.

The final methodology used can be seen below in Figure 33. The first step was to fit the percent of people with adenomas first. Initially adenomas are set to non-progressing. This method provides a quick and efficient fit of that target with none of the difficulties involved with attempting to maintain a cancer incidence arising from unknown years. After an initial fit of people with adenomas, the risk function is adjusted to meet the adenomas per 100 people target. Once both of these are fit, the final step for the main targets is to increase the percent of adenomas progressing until the cancer incidence targets are met. This step completes the main fit, and immediate cancers are added in by adjusting the percent of adenomas that are immediate cancer. Finally, the target for advanced adenomas by age is fit once everything else is complete. Each step is described in detail in the following subsections.

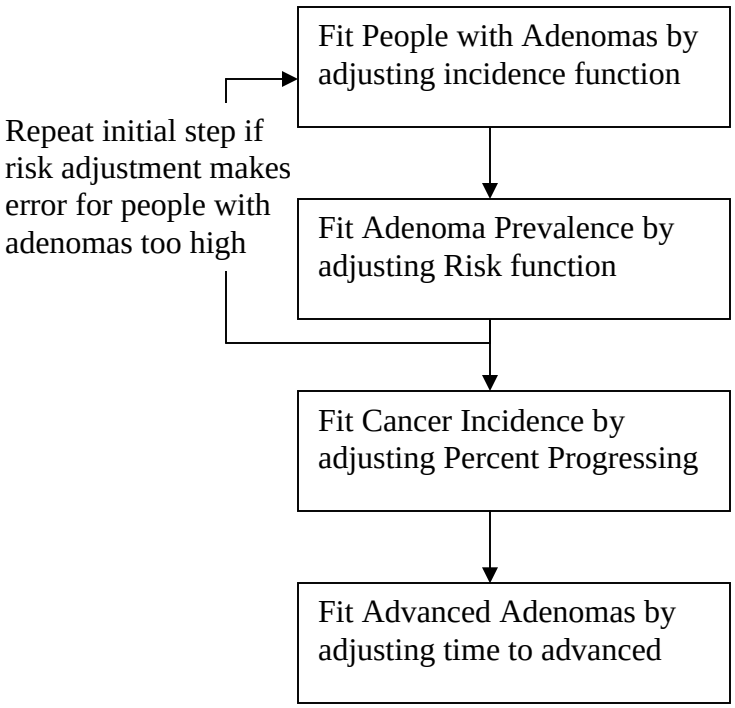


Figure 33: Flowchart of Fitting Procedure

To fit these targets, a special spreadsheet was developed that calculated all of the needed input values. A view of this spreadsheet and the definitions of the cells can be found in the Appendix H. There were several criteria used for the input fitting. In general, the average error and the maximum error of the potential fit were used to determine the fit for a given set of inputs. The average error was simply an average for the percent error for all the desired years. The maximum error was the maximum observed percent error over the desired ages. The average error was used to insure that the model fit the lower ages as well as the later ages where cancer and adenomas are both more common. The maximum error was used to keep the fit adequate for all values, even when average percent error was acceptable. In addition, an un-quantified visual "smoothness of the fit" was also examined. This smoothness considered the output to make sure that the output does not oscillate around the targets. This oscillation could yield an acceptable error, but still have a very poor visual fit, which would harm the face validity of the model. An

example of where smoothness was used to determine the lack of fit can be seen in the initial fit of the immediate cancer in Section 4.2.6.

In addition to these fit measurements, there was a restriction for the ages of the fits as well. The ages of 40 and above were emphasized in all fitting. A fit for ages above 80 were a "nice to have", but only ages to 80 were used to fit the model. This emphasis on the middle occurs because the target population for the screening and surveillance are the ages between 40 and 80. Prior to 40 there are not enough cancers to justify screening, and after 80 there are too few people to screen and the natural mortality is already so high as to make any kind of intervention much less effective.

In order to have some confidence in the fit, it was necessary to run large numbers of people through the simulation. The largest number of patients run through models described in Chapter 2 was 100,000 people. This provided a starting point for the analysis and fitting of the model. With only 100,000 people to run through the model, the cancer incidence curve was not smooth at the later ages. To smooth the curve further, the run length was increased to slightly over 200,000 people. Because the people with family history have more adenomas and cancers, the run time is slightly longer, so 210,000 people were run for that group. Groups of 240,000 people with no family history were simulated to provide for additional smoothing in the later years. Based upon the confidence intervals of the fitted cancer incidence, this replication length is sufficient for a ±10 precision for a confidence interval. Further discussion of the confidence intervals for the cancer incidence is discussed in their sections to provide further justification for the chosen run length.

4.2.3 Percent of People with Adenomas

The first target to be fit is the percent of people with adenomas (PPA). All adenomas are assigned as non-progressive adenomas to meet this target. This setting eliminates cancer from the population, so it overstates the survival of the patients. However, the fit from this step will be a preliminary fit. It is important to note that PPA is almost independent

of the personal risk function, and so it was fit first. The references for this target (Arminski and McLean 1964; Blatt 1961; Correa et al. 1977; Eide 1986; Eide and Stalsberg 1978; Vatn and Stalsberg 1982; Williams, Balasooriya, and Day 1982) did not differentiate between races, so the targets are only for males and females. The targets for the male population are shown in Table 4 below.

Age	Percent of People with Adenomas
40	20 %
60	37 %
80	50 %
100	60 %

Table 4: Percent of People with Adenomas

Following the convention of the previous Vanderbilt simulation model (Ness et al. 2000), the incidence rate value is broken down generally into five-year increments, but with some of the earlier groups combined, since the incidence is so low. The rate function for the NHPP is also linearly interpolated between the two valuations to create a smoothed rate function. This smoothing of the rate function makes fitting the later data easier because there are no unexpected jumps in incidence from one age range to another.

Since there are only four ages for the targets for percent of people with adenomas, the incidence valuations of the previous Vanderbilt model were used as a starting point for fitting. An iterative process was followed to fit the target. This process is described below.

1) Run the simulation for both scenarios with and without family history and copy the output to the appropriate page of the spreadsheet. The spreadsheet will calculate all of the values on a separate page for easy visual comparison to the targets.
2) If the target is within acceptable error (defined as within 5% error for this stage of fitting), stop. Otherwise adjust all incidence values above the previously fit

target. This means that initially all valuations are adjusted, and only those values above 40 are adjusted once the 40 target is met. The adjustment formula is:

New valuations = Old Valuations * Desired Incidence / Observed Incidence

The error computation is the standard percent error calculation. For reference it is shown below.

$$Error\% = \frac{|TargetValue - ModelOutput|}{TargetValue} * 100$$

If the target was a non-random process, the next set of simulations would return the desired target. However, the simulation is a random process with the additional complexity of running two groups (with family history and without) and combining them into an aggregate error measure. This approach does usually come closer to the desired target.

3) If the target is within error tolerance after rerunning scenarios both with and without family history, then stop and move to the next target value (step 4). If not, repeat step 2 until the valuation is within tolerances.
4) Return to step 2 using the new target values and observed results. These adjustments are only performed on the valuations above the last target. So incidence rates for ages forty and below remain untouched if that was the last completed section of the fit.
5) One of the constraints of the target is that the incidence function must never decrease. This condition follows from the accepted medical view that adenomas develop more frequently as the body's maintenance functions being to wear down. Should an adjustment bring the incidence lower than a previous one, that adjustment is undone. The previous value is lowered, and the value two age groupings previous is increased to compensate. These changes cause the previous target to be refit. For example, if during fitting the age 60 target, the age 45 incidence drops below the 40 incidence, that adjustment is undone. The incidence valuation for 40 is lowered slightly, and the incidence rate for 35 is raised some to

compensate. The age 40 target is then fit again before moving on to the age 60 target.

These steps are then repeated until all ages are fit. As long as the fit is reasonably close, it is acceptable to move to the next phases of fitting. Since both the risk function and the addition of cancer to the model will change this valuation slightly, an approximate fit is acceptable. A sample of the final fit of the white male population is in Table 5 below. Once a slight refit is completed due to the cancer incidence and risk adjustments, there will still be some differences between races and birth cohorts due to varying base lifetimes.

Age	Target	Output
40	20 %	20.92 %
60	37 %	37.10 %
80	50 %	51.30 %
100	60 %	60.45 %

Table 5: Percent of People with Adenomas Comparison

4.2.4 Adenomas per 100 People

The next phase of fitting is the adenomas per 100 people (APP). This fit is the most approximate of the fitting phases. Only the risk function can be changed to affect these targets. This target is also complicated further because the studies that provided this target (Arminski and McLean 1964; Blatt 1961; Correa et al. 1977; Eide 1986; Eide and Stalsberg 1978; Vatn and Stalsberg 1982; Williams, Balasooriya, and Day 1982) are not the same as the study that provided the percent of people with adenomas targets. As such, the targets often conflict. The decision of Dr. Ness was that the APP fit was the least important, and that a visual fit was all that was needed. When it proved impossible to fit all targets of both types, more weight would be put on the fit of the earlier age groups for APP since there are more surviving people in those groups, and thus more data points were used to create the targets.

An initial risk function for all races was determined by Dr. Ness, based on his clinical experience. This function was then adjusted as needed to fit the targets. If the percent of people with adenomas (PPA) is on target, but the APP is too high, then that suggests that those people who do get adenomas are getting too many. This result in turn indicates that the tail of the risk function is too thick, so the very high relative risk values (12 + times as likely to get adenomas and cancer) are appearing too often. VIM was used extensively to define new distributions to adjust this tail by specifying means and modes.

When adjusting the risk function, bringing the mode closer to zero would decrease the adenomas per 100 people valuations for quite a while. After passing a certain point – usually a model setting of around .01 – making the mode lower would then create an increase the adenomas per 100 people. This change occurs because the mode has been shifted so low in the JohnsonSB distribution, that while holding the mean constant, the median for the distribution begins to shift higher if the mode is shifted any lower. For the male population, this minimum point appears very sensitive. Any higher or lower modes would increase the later target values far beyond any acceptable visual fit. This tenuous value was not present for the female population, so a much tighter fit was achieved. The results of these fits are shown below in Table 6.

	Female		Male	
Age	Target	Model	Target	Model
40	20	18.94	30	26.16
45	24	24.46	36	33.24
50	29	31.88	44	42.13
55	35	40.03	54	52.36
60	42	48.55	65	63.11
65	49	57.29	78	73.60
70	58	66.12	94	84.35
75	68	74.51	111	95.16
80	79	82.98	133	105.61
85	92	94.84	159	120.64
85 +	108	117.93	188	147.38

Table 6: Targets for Adenoma Incidence per 100 people

Once these targets had been fit, the fitting procedure for percent of people with adenomas is repeated one more time since the risk function adjustments modify not just how many adenomas develop, but also when those adenomas will develop. Completing this step is fairly quick, and usually took less than two iterations before returning to an acceptable fit.

4.2.5 Cancer Incidence from Progressing Adenomas

As stated earlier in the section, the most important target is cancer incidence from progressing adenomas. Because a progressive adenoma that develops at the age of five can progress to cancer (become incident) any time from 0 to 60 years later, this incidence target is also the most difficult to fit. It is further complicated by the use of two separate populations. The low risk group will have longer time to get cancer, and the high risk (with family history) group will have a shorter than average time to cancer, so the time to cancer from an adenoma has a bimodal shape. In addition, because the risk can affect incidence and progression, those people who get adenomas early in life tend to have cancer in a shorter time than those people who have adenomas develop later in life. The assumed target was to use 85% of the SEER data as directed by Dr. Ness.

The basic procedure to meet this target was again an iterative one based on starting with the younger ages. The simplified procedure is described below.

1) Start with the percent progressing for the youngest group (15 – 25)
2) Increase the percent progressing until it approximates the target. In age groups before forty, this means that the model output is within one year of the target since there are so few instances of CRC at such an early age.
3) Repeat Step 2 for all age groups up to age 40.
4) Now, adjust all the percent progressing according to the following formula

   New Valuation = Old Valuation * Target / Model Output

   This step is rerun and repeated until the target is met. Unfortunately, the age 40 group is the only group addressed in this step whose target can be met using a formula approach. Many of the adenomas created in later stages will not develop into

cancer for many years to come, so a simple adjustment formula does not work as well for later targets.

Let us examine a hypothetical fitting of the age 45 cancer incidence to see how the procedure works for later years.

1) Most of the desired incidence for this range would already be met by the earlier progressive adenomas (up to age 40).
2) The percent progressing is adjusted for the 40 to 45 range until it is significantly larger than the 35 to 40 percent progressing. However, this has little effect on the target because very few adenomas will progress to cancer and become symptomatic within 5 years. As such, what ends up happening is that one of the later age ranges (suppose the 50 target for this example) already has a model output above the desired target.
3) The percent progressing for the 40 to 45 age grouping is lowered until there is still some opportunity for increase from the 45-50 progressive adenomas.
4) The earliest ages of 15-25 and 25-35 are slightly reduced to counter balance an increase in the percent progressing for the 35-40 age range.
5) Then, we delicately increase or decrease the values that would have the most effect on the targets. So if we are still too low for ages 45-50, we will increase 35-40 slightly. Assuming that our 40 target was met during the adjustment in step 4, we need to adjust ages 15-25 and 25-35 down very slightly (since 35-40 only has minimal effect on the 40 cancer incidence) to compensate for the drop in the percent for ages 35-40. This is repeated until both targets are met.

Once age 40 is fit again and age 45 is fit for the first time, we move on to fitting the age 50 cancer incidences. The procedure is the same as that performed for age 40 and 45, but it becomes more complicated. Now if the rate is too low or too high, values for all of the previous percentages can be changed. As a rule in this heuristic fitting, later age groups were given preference in adjustment because they required less refit work. If age 70 was being fit, adjustments would be made only bounded by incidence of the age 75 target – to the 60-65 and 65-70 percent progressing. While more impact would be obtained by adjusting the earlier ages such as 50-55, this adjustment would

also require refitting not just ages 60-65 and 65-70, but also make refitting ages 50-55 and 55-60 necessary.

Once all ages had a very loose fit, smoothing the fit became necessary. There was often a later age group whose fit was too high even with a zero percent progressing. The same idea used the later years was followed, but there were many more iterations compensating for changes as the maximum and average error was reduced.

The final decision criterion applied to the fit of the model was the average percent error associated with the age ranges of 40 to 80. Ages 40 and younger were of low interest since they would likely not be in any screening protocol because they had cancer prior to any possible start of screening. Ages 80 and above were again visually fit since there were not enough people remaining in any of the studies that created the targets to generate confidence in the target. In general, the goal was to maintain a steady or increasing percent of progressive adenomas. Few of the race-based progressive adenoma incidences were always increasing, but all came very close to meeting this goal. Figure 34 shows the percent progressing values used to create this fit for white males. The remaining percent progressing can be found in Appendix A, and fit of the cancer incidence can be found in Appendix B.

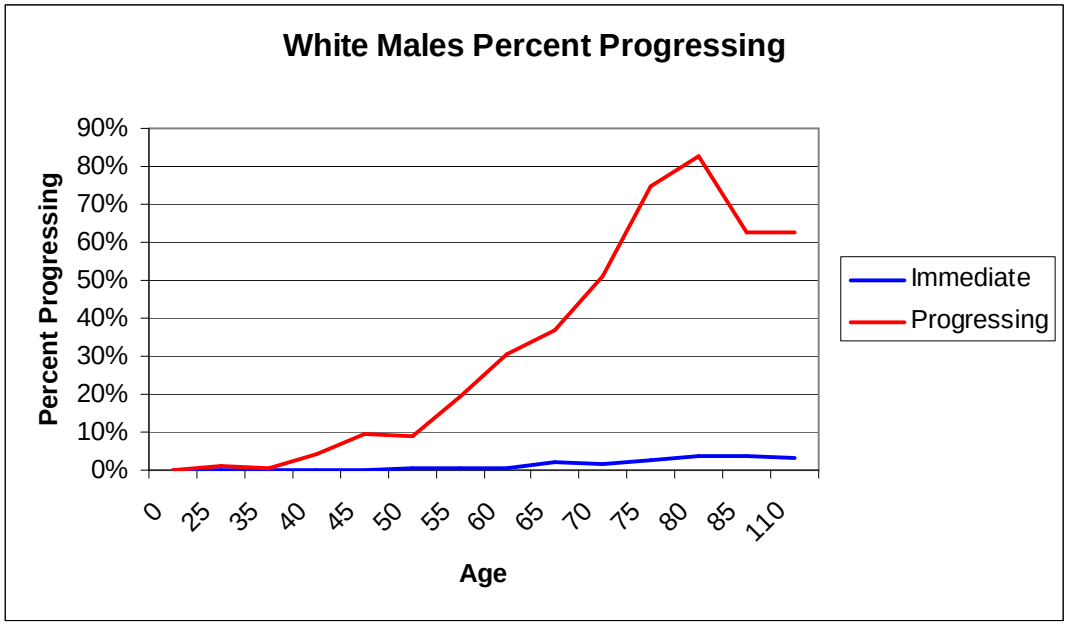


Figure 34: Percent of Adenomas Progressing

Figure 35 shows the fit for the simulation along with the confidence intervals for the cancer incidence fit. The half-width for the confidence interval for ages 75 to 80 is 11% of the sample mean. To get to a the desired half-width of a 95 % confidence interval, a simulation of approximately 1.2 million people would need to be run.

$$n = n_0 \frac{H_0^2}{H_{desired}^2} = 210000 \frac{37.53^2}{17.05^2} = 1017378$$

Due to time constraints associated with the fitting process and model creation, this level of precision was not feasible. Further justification for adding additional people to the simulation was to determine the effect they had on the estimates. With the addition of people up to a population of 400,000, the estimates changed by less than 1 %. Therefore it was deemed better to spend the effort improving the fit instead of reducing the confidence interval.

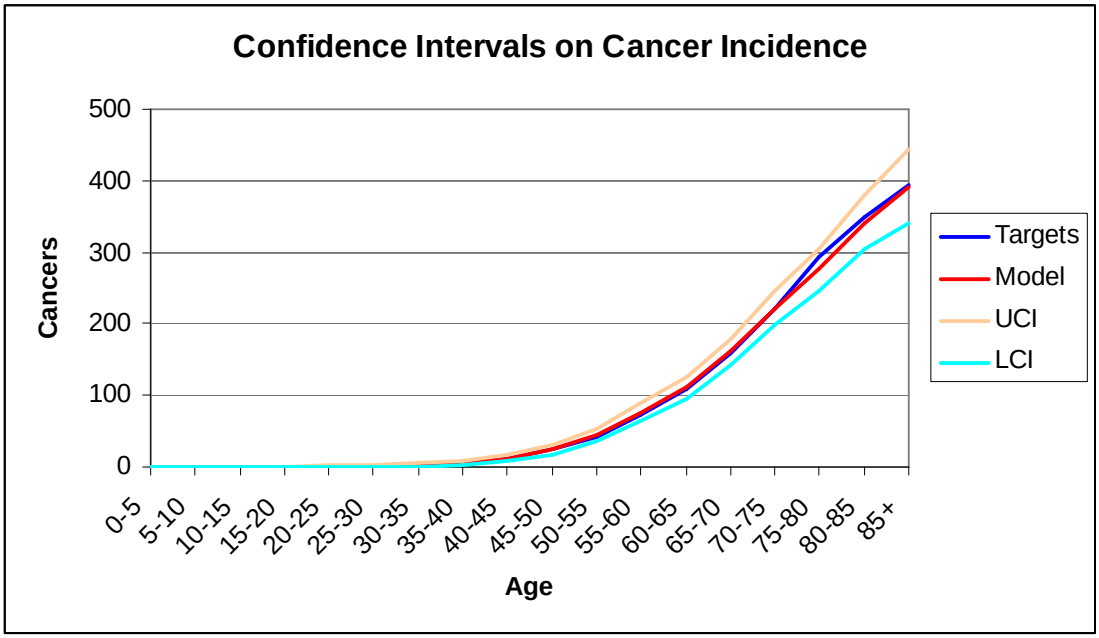


Figure 35: Confidence Intervals on Cancer Incidence from Progressive Adenomas

4.2.6 Immediate Cancers

The fitting of the immediate cancers employed a simplified version of the fitting of cancers that develop from progressive adenomas. Approximately 15% of the cancers should develop from immediate cancers, so the overall cancer incidence by age group multiplied by 0.15 yields our targets for each age. Table 7 shows the desired outputs for white males.

Age	Immediate Cancer Incidence
0-5	0.00
5-10	0.00
10-15	0.01
15-20	0.02
20-25	0.09
25-30	0.20
30-35	0.47
35-40	1.05
40-45	2.11
45-50	4.42
50-55	8.40
55-60	15.69
60-65	25.81
65-70	39.30
70-75	56.55
75-80	71.14
80-85	86.59
85+	94.86

Table 7: Immediate Cancer Targets

This output is met by increasing the percent of immediate cancers within the input database. Then when an adenoma occurs, a larger percentage will become cancer immediately. There is a short time lag for cancer to become symptomatic; so adenomas that become immediate cancers can take up to 10 years before they are recognized.

The fitting follows the same iterative steps as progressive cancers. However, each the percentage of immediate cancers will mostly affect that group and the subsequent group's number of cancers observed. The procedure used is demonstrated below.

1) Starting with the initial ages of 0-25, increase the percent of immediate cancers until a loose fit is returned. This fit can be very approximate, since there will be less than one cancer per 100,000 people at these ages. Also, anyone who gets a cancer during these ages is excluded from any potential screening studies, so there is no impact on screening comparisons.
2) Repeat the same slow increase for each of the following groups up to age 40. Again, these fits are not as important as the fits that will follow, so approximate fits are acceptable.
3) The next step is to slowly increase the percent of immediate cancers (or decrease if the target is passed) for all subsequent age groups. These groups have larger values as well as more importance to any screening model, so the error rates of all values 40 and above were averaged to determine the quality of fit. Some values might be set to zero and still be too high, in that case proceed to the next age group.
4) Once all age groups have an initial fit, smoothing needs to be done. If a percent immediate is zero and the model's output is still over the desired output, the previous value is lowered, and a cascading change is made to spread out the error from that value (and potentially eliminate it). For example, suppose ages 70-75 has an immediate cancer incidence that is too high, and a percent immediate of zero. The percent progressing of 65-70 is lowered to reduce the cancer incidence in the 70-75 ages. To compensate, 60-65 is slightly increased to bring the cancer incidence of 65-70 back to where it was. This cascades down until the error has been reduced as much as possible.

This procedure yields an acceptable fit, with an average error of 7.8% for the simulation for both genders and races. However, as can be seen from Table 8 and Figure 36 below, there were still a few very large errors that account for most of this average error. The percent progressing values used to yield these values are found in the Appendix E.

	White Males	White Females	Black Males	Black Females
Average Error	7.9	8.8	7.1	7.5
Max Error	29.9	19.4	23.3	17.7

Table 8: Initial Error for Immediate Fit

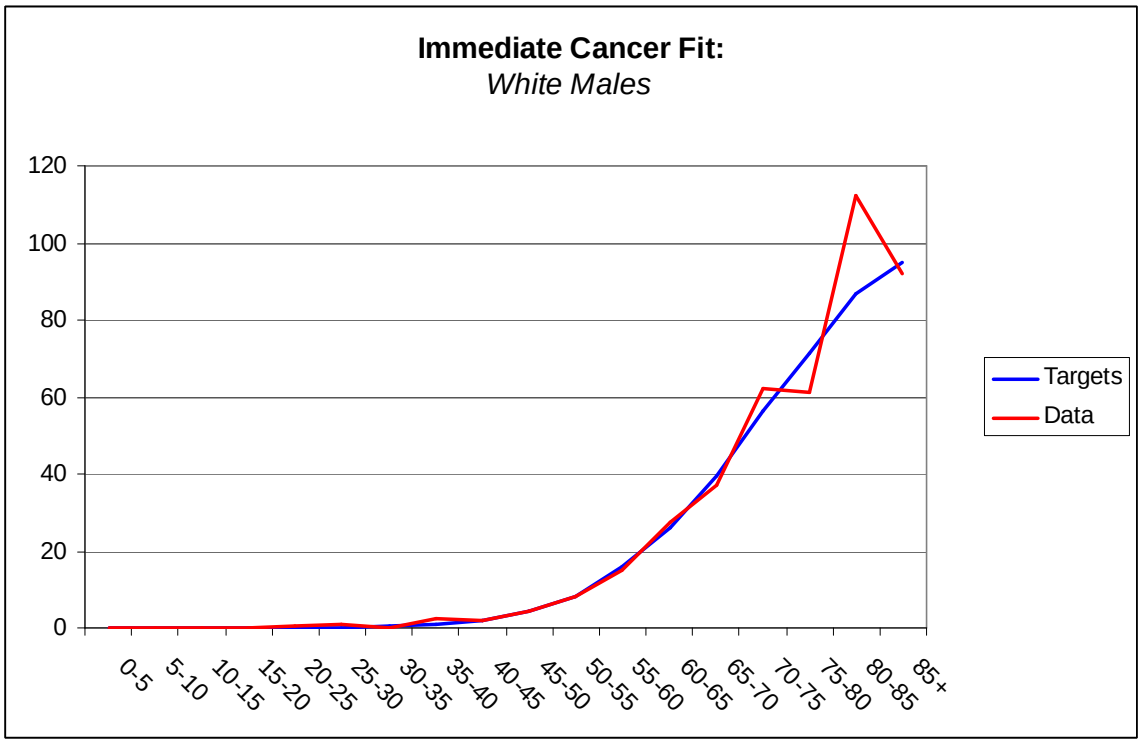


Figure 36: Immediate Cancer Fit for White Males

Since the fit is still too approximate, the five-year increments were reduced to 2.5 year age groups to reduce the variability in the error and improve the fit. The same procedure used previously was followed. The addition of the extra categories improved the fit dramatically. The fit for the white males can be seen below in Figure 37 and graphs for the remaining fits can be found in the Appendix D.

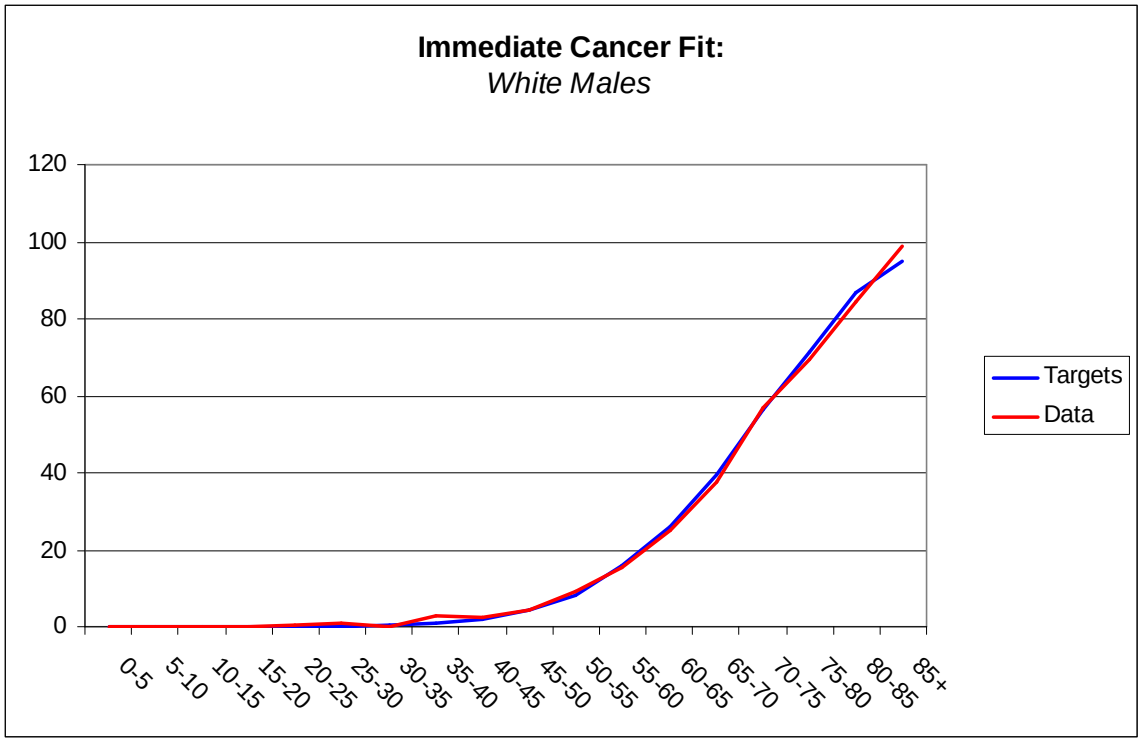


Figure 37: Improved Fit for Immediate Cancers for White Males

4.2.7 Cancer Stage Progression

In addition to fitting the cancer incidence, there are also outcomes on the stage of cancer at the time of symptoms. These data come from the diagnostic colonoscopies performed when symptoms develop. This target is especially important to match because it determines the survival of the patient once cancer is diagnosed. 50 % of all patients have local cancer, while 20 % have regional cancer and 30 % have distant cancer at diagnosis.

Because of the inadvisability of letting a cancer develop in order to retrieve the data, there are no other data on which to base the creation of the distribution for time to regional or distant cancer. As such, the same distribution shape parameters were chosen as were used in the time from incidence to symptoms. The percentages were increased or decreased by adjusting the maximum value of the distributions. There was a very simple iterative procedure used to fit this target. The simulation was run for both scenarios with family history and no family history, and then combined to form a weighted average of

the stage distribution at the time of cancer symptoms. A summary of the procedure is as follows:

1) Start with maximum times of 10 years for all distributions. This starting point will make the local cancer percentage well over 50%. So increase the maximum of both the time to regional and time to distant cancer until the local cancer percentage increases to fifty.
2) Shorten the maximum for time to distant cancer until 30 % of the cancers are now distant.
3) This will have disturbed the local cancer percentage some. So now adjust both maximum times for regional and distant cancers upward until the local cancer is back to fifty percent.
4) Now repeat Step 2 until distant cancer is back to 30 %.
5) Repeat Steps 3 and 4 until the error is small.

This procedure typically yielded a maximum error of 0.95 % within only a few iterations.

4.2.8 Advanced Adenomas

This target represents the percentage of all adenomas that are advanced at key ages. An advanced adenoma is thought to indicate a higher risk of cancer for that adenoma. In addition, because of adenoma histology, the adenoma becomes more detectable to screening. The histology of the adenoma has no bearing on the natural history model, but does affect the screening model because advanced adenomas are more detectable than non-advanced adenomas. The targets for this are shown in Table 9.

Age	Target	
	Male	Female
40	5.8%	13.5%
45	6.9%	15.1%
50	8.0%	17.3%
55	9.6%	19.4%
60	11.3%	22.1%
65	13.1%	24.7%
70	15.3%	27.4%
75	18.0%	30.3%
80	20.4%	33.6%
85	23.5%	36.8%
90	26.7%	40.1%

Table 9: Targets for Advanced Adenomas

This target will need further fine tuning once screening is implemented to make the screening match its validation targets based upon the reduction in cancer from screening observed in clinical trials. Because the fit on the advanced adenomas is approximate, an average percent error of less than five percent for the ages of 40 to 80 is used as to compute the error.

To fit the advanced adenomas, an initial JohnsonSB was chosen with delta of 1 and a gamma of 0, and whose max was arbitrarily set to 90 years. This decision created a distribution whose shape was somewhat normal, but whose distribution allowed for changes should they become necessary. The steps used to adjust this distribution to match the targets are described as follows:

1) Bring the maximum down until the percent error for age 40 is within target error ranges.
2) Examine the output distribution. There was a consistent flat spot where the percentage of advanced adenomas flattens out. To move this flat spot out past age 80, adjust gamma down until shape approximates the desire shape.

3) The previous adjustment will throw off the location of the graph, so adjust the maximum age down (or up if necessary) to bring the percent error for age 40 into target error ranges.
4) Since the general shape will not change by adjusting the maximum, this should bring the entire graph into acceptable error ranges so fitting is complete.

The results of this fitting for black females are shown below in Figure 38. All other fits had approximately the same fit characteristics.

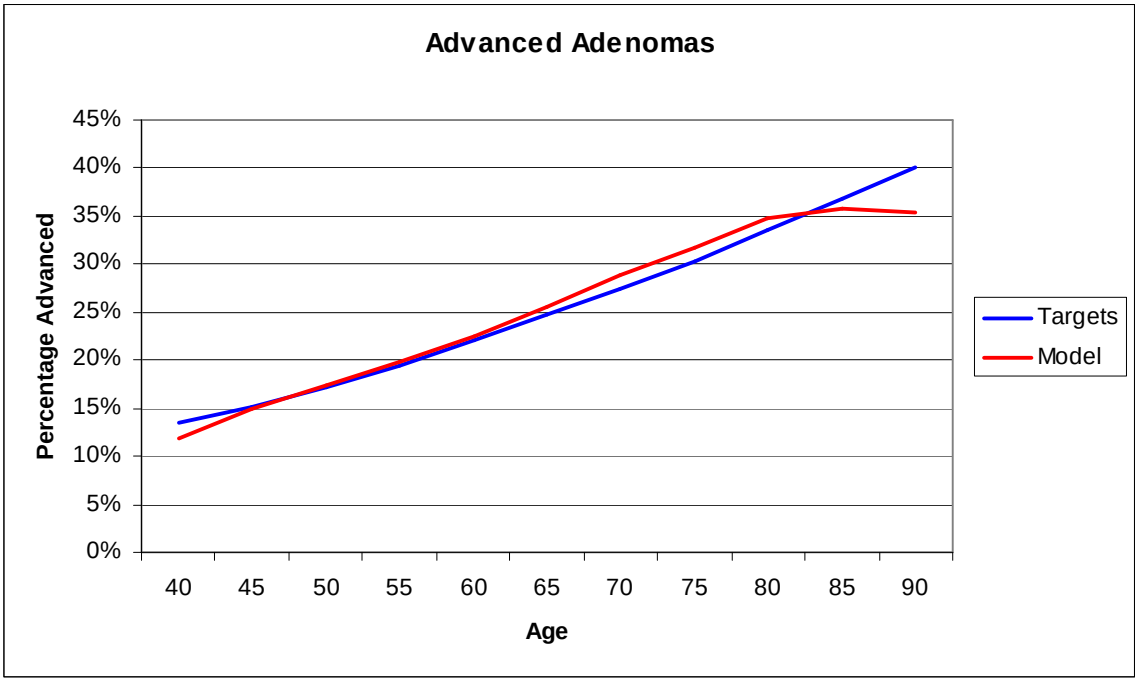


Figure 38: Advanced Adenoma Fit

4.2.9 Cancer Survival

The last of the targets to be fit did not need an iterative approach to match. SEER provides from its website, www.seer.cancer.gov, the survival of patients with cancer based upon their stage at diagnosis, race and gender for five years. If someone survives five years of cancer, they are assumed to be treated and continue on their remaining years. These targets are shown below in Table 10. The values in the table are the chances of surviving that long, given a particular cancer type.

		SURVIVAL RATES, BY RACE, SEX, STAGE AND YEAR					
			WHITES			BLACKS	
		TOTAL	MALES	FEMALES	TOTAL	MALES	FEMALES
1 YEAR	LOCALIZED	0.9552	0.9549	0.9554	0.9321	0.9327	0.9315
	REGIONAL	0.8823	0.892	0.8731	0.8797	0.8872	0.8732
	DISTANT	0.3893	0.4047	0.3735	0.3642	0.3632	0.3652
2 YEAR	LOCALIZED	0.9378	0.9389	0.9366	0.9016	0.901	0.9022
	REGIONAL	0.7776	0.7881	0.7676	0.7553	0.7609	0.7506
	DISTANT	0.193	0.2003	0.1856	0.1687	0.1741	0.1639
3 YEAR	LOCALIZED	0.9174	0.9189	0.9158	0.8702	0.8713	0.8693
	REGIONAL	0.6967	0.7033	0.6905	0.6615	0.656	0.6661
	DISTANT	0.12	0.1204	0.1197	0.1028	0.1066	0.0994
4 YEAR	LOCALIZED	0.8972	0.8972	0.8973	0.8444	0.8432	0.8457
	REGIONAL	0.6378	0.636	0.6393	0.5909	0.5794	0.6004
	DISTANT	0.0885	0.0849	0.092	0.0772	0.0803	0.0744
5 YEAR	LOCALIZED	0.8799	0.8788	0.881	0.8233	0.8172	0.8285
	REGIONAL	0.6006	0.5936	0.6069	0.5545	0.5353	0.5703
	DISTANT	0.071	0.066	0.0759	0.0645	0.0654	0.0636

Table 10: Cancer Survival by Stage

To transform this data into a distribution, an Excel spreadsheet was used. A screenshot of the spreadsheet can be seen in Appendix G. Each stage of cancer, race, and gender combination was fit using the same procedure on identical spreadsheets. The first step was to convert the survival probability into a probability of death within the specified time frame. Then conditional probability was used to calculate the cumulative probability of death given that cancer was fatal.

The next step was to create a function to generate the CDF value for the JohnsonSB distribution given the minimum, maximum, gamma, and delta. The JohnsonSB distribution was chosen to because of its applicability where data is limited, in addition to the fact that it matched the shape of the targets very well. Microsoft Excel's solver function was then used to calculate a gamma and delta that would minimize the percent error between the calculated JohnsonSB and the data from SEER given a maximum time of five years to death. This procedure returned an excellent fit for all stages, with local

cancer having the highest error of only 2.4%. The CDF for the JohnsonSB (in blue), and the data from SEER (in red) is shown in Figure 39 below. A sample graph for each stage can be found in Appendix F.

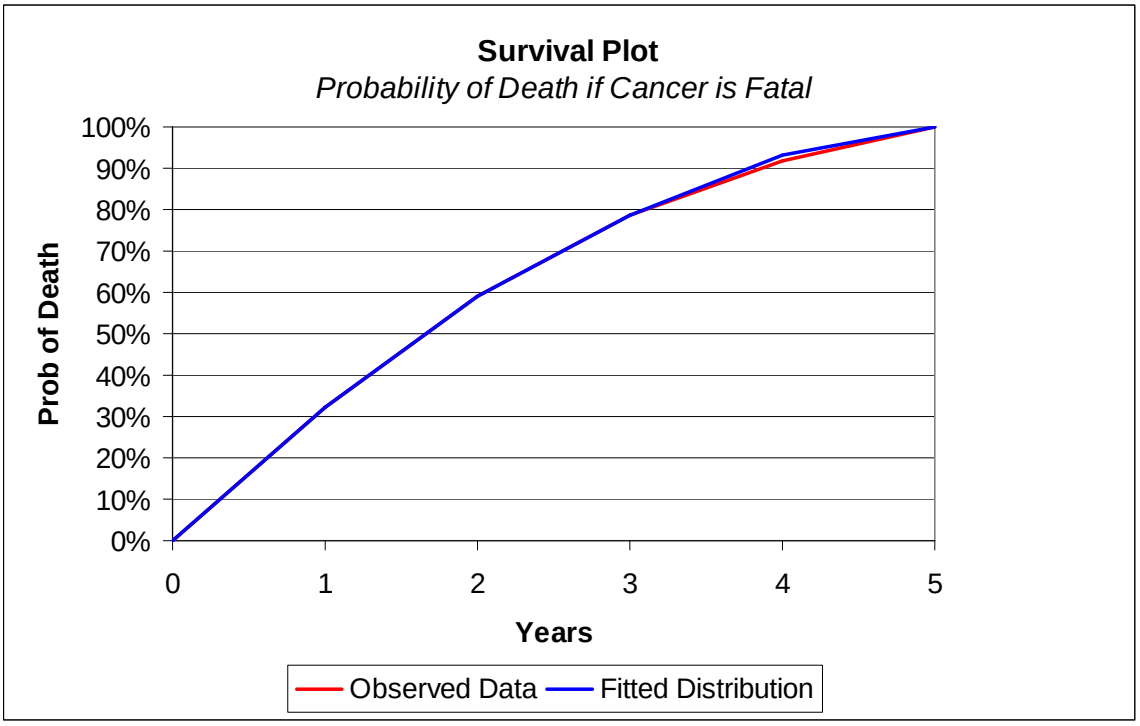


Figure 39: Plot of Model Survival Curve versus Observed Data

An important point is that this is the probability of death if cancer is fatal. The simulation first checks to see if cancer is fatal, so this distribution gets sampled only if the cancer is fatal. Fitting just one function that included surviving patients was examined, but was discarded because of the high error in the fit.

4.3 Model Results

With the basic model both created and fit to match the findings of the literature, the next step is to determine the effect of removing CRC completely and to determine the impact that CRC has to society. Before any analysis can be done, it is necessary to define the population of interest. For this simulation, data from the 2000 census was used to

determine the percentages of the population for each gender and race. Dr. Ness provided the figures for the percent of people with family history. The summary is that 19 % of the population has some family history of CRC, 49 % of the population is male, and 86 % of the population is white. The most obvious impact of CRC is loss of life from CRC. There are two main ways to examine this impact. First, the impact can be viewed from the point of a CRC victim to determine the years of life lost due to CRC for those people who die from it. The results can be seen in Table 11.

Group	Years of Life Lost
White Males	10.42
White Females	11.72
Black Males	10.87
Black Females	11.77
Overall	11.12

Table 11: Life-years Lost to CRC for Affected Patients

From a societal perspective, the more important piece of information is the total cost of the disease in life-years. These results can be seen in Table 12. Because of the large number of people who get CRC at a relatively young age, the overall impact of this disease is almost a quarter of a year per person.

Group	Years of Life Lost
White Males	0.22
White Females	0.27
Black Males	0.16
Black Females	0.21
Overall	0.24

Table 12: Average Life-years Lost for the Entire US Population

Along with life-years, the quality-adjusted life-years are an important measure of impact CRC has on the population. Without CRC, the average white female has a QALY total of 72.2. This number is derived from the utility of one for the first fifty years and a utility of 0.91 for the remaining years. With CRC included, the average QALY total for the population is only 65.6, so a significant reduction in the quality of life is caused by

CRC. The full results for QALY are shown below in table 13 along with the confidence intervals associated with the QALYs with cancer and without cancer.

Race/Gender	Family History	With CRC	UCI	LCI	Without CRC
Black Female	No History	60.92	61.01	60.84	65.63
Black Female	History	60.57	60.65	60.49	65.63
Black Male	No History	53.97	54.05	53.90	58.65
Black Male	History	53.70	53.78	53.63	58.65
White Male	No History	64.36	64.44	64.27	66.11
White Male	History	63.07	63.16	62.98	66.11
White Female	No History	65.82	65.91	65.73	72.2
White Female	History	64.60	64.69	64.51	72.2

Table 13: Comparison of QALYs with and without CRC

In addition to years of life, and the quality of those years, the other key metric for gauging the impact of a disease from society's standpoint, is the average cost of the disease per person. The average cost per person is $2348 for white females as a whole, and a much larger $4312 per person for the population portion with family history. These costs, especially for the group with family history, suggest that some form of screening or effective treatment option will almost definitely be worthwhile based upon the costs associated with the disease. The average costs per person for each group appear in Table 14 below.

Race/Gender	Family History	Average Costs per Person
Black Female	History	$3591.28
Black Female	No History	$1406.31
Black Male	History	$3068.56
Black Male	No History	$1168.30
White Male	History	$4339.89
White Male	No History	$1764.40
White Female	History	$4312.69
White Female	No History	$1840.18

Table 14: Average Costs of CRC

4.4 Comparison of Results

Beyond determining the impact of CRC is the comparison of this natural history model to other results. Other external validation and confirmation of the model will come in the later sections of the project once screening is added. Nevertheless, the simulation model can be compared to several sources. First, there are data on the cumulative risk of CRC available from the SEER database. While from the same source as the cancer incidence, this data was not used in the creation of the model, so it provides some measure of external validation. The second source of comparison is a simulation that samples from the life tables with and without CRC removed (see Section 3.6 for more on life table adjustments). This simulation provides a raw measure of the impact CRC has on a population. Lastly, the simulation can be compared to the previous Vanderbilt model. Within the Vanderbilt Model, one of the scenarios is no screening. This scenario does have follow-up screening once cancer has been detected, but the impact of this late screening is minimal. Since the model was developed independent of the Vanderbilt-NC STATE model, it provides a good comparison.

4.4.1 Comparison with SEER Cumulative Risk

The first external validation of the model is a check on the cumulative risk of cancer from the SEER database. This risk of developing cancer is related to the cancer incidence targets used to fit the model, but is slightly different. These numbers represent the cumulative risk for person to develop cancer. This risk compares to the cancer incidence that was seen during a different year for varying ages. A comparison of the male and female population fit can be seen below in Figure 40. The match between the model and the SEER data is exceptionally good, so it is safe to conclude that the model is valid based upon this comparison.

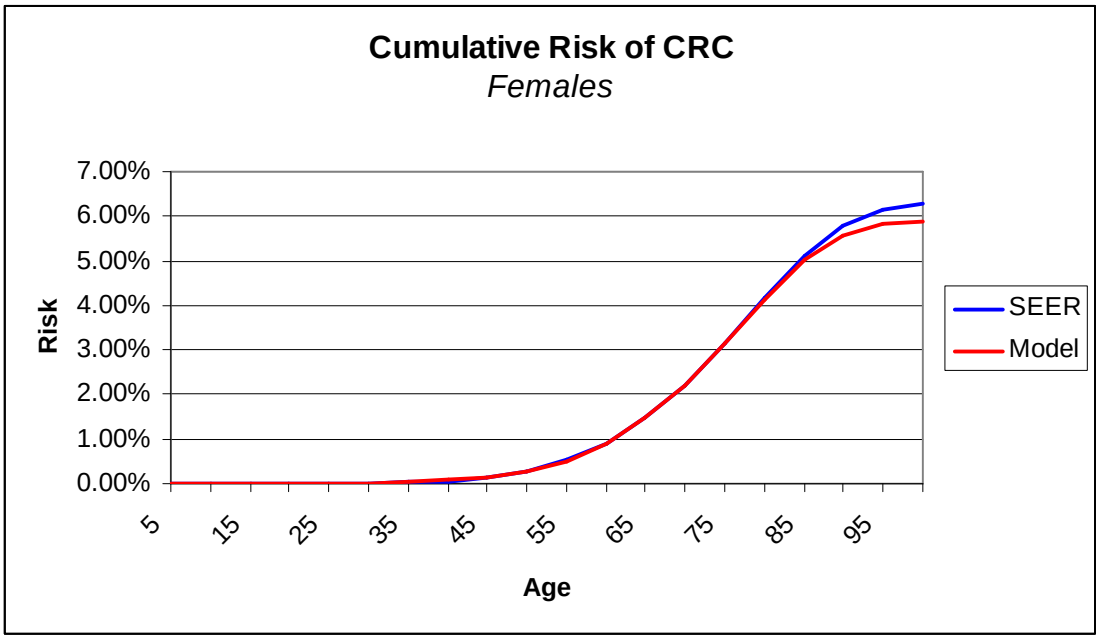


Figure 40: Comparison of Results to SEER Cancer Risk - Females

4.4.2 Comparison with Life Tables

This analysis provides the most basic comparison and external validation of the model. A simulation previously developed samples from life tables that have CRC included and life tables without CRC and then compares the lifetime of the person. The adjustments based upon CRC removal are described in Section 3.6. Because that simulation looks only at life tables, any analysis beyond life-years lost to CRC is not feasible. The results of this comparison can be seen in Table 13.

Group	Years of Life Lost	
	Vanderbilt-NC State	Life-table Adjustment
White Males	0.22	0.15
White Females	0.27	0.21
Black Males	0.16	0.09
Black Females	0.21	0.15
Overall	0.24	0.17

Table 15: Comparison to Life-Table Adjustment

Clearly, the model shows the same basic trend as the life-table adjustment, but is consistently higher than the life table adjustments. This discrepancy can be explained by examining the data used for both methods. The life table adjustment adjusts based upon cancer mortality from 1975. This mortality was chosen because it was prior to screening; however, treatment options have evolved since 1975. The simulation model uses incidence rates based upon the pre-screening years of 1973 to 1982, but all survival information is based upon modern treatment methods. So the increase comes from the improved treatment methods that have developed over the last 20 years. This slight increase from the life table adjustment is also seen in the Vanderbilt model used as a final comparison in the section below. Thus the model's validity is maintained for this test as well.

4.4.3 Comparison to Vanderbilt Model

With the main external validation complete, a comparison to the previous Vanderbilt model is now considered. First, cancer incidence is examined to see if the two models follow the same general trend. Because the cumulative risk function has already been matched by comparing the information to the SEER database, cancer incidence is only a means to provide more evidence to support the validity of the model. Figure 41 below shows the comparison between the cancer outputs for the two models. The general trend is the same, which is all that can be expected given the addition of different risk groups to the Vanderbilt-NC State model and the Vanderbilt model's use of slow and fast progressing adenomas (as discussed in Section 2.3.3). So the model's cancer output is further validated.

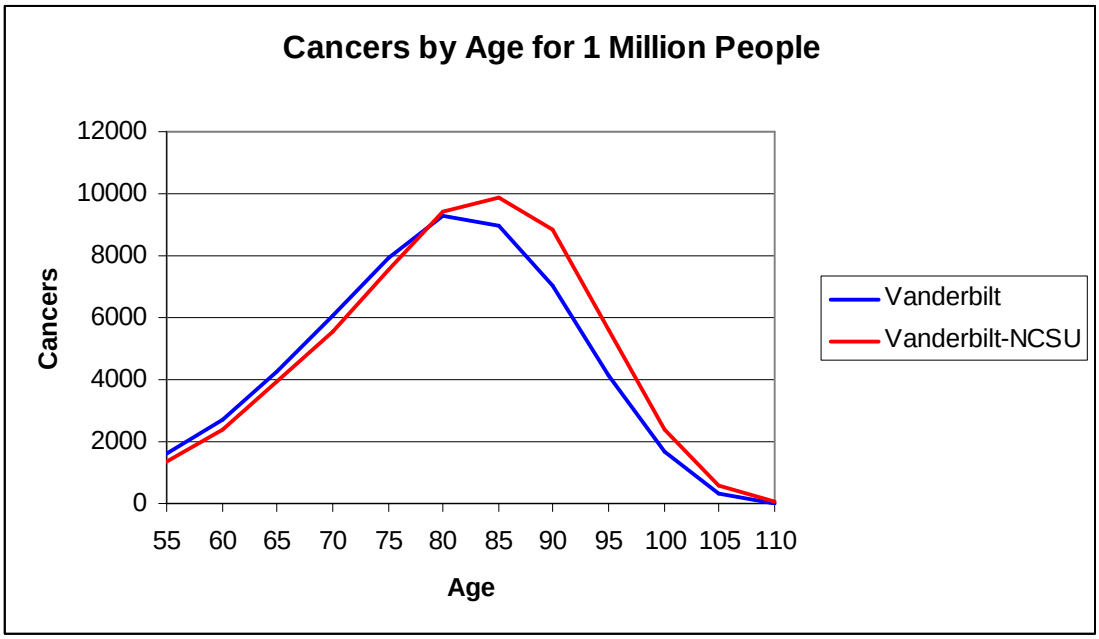


Figure 41: Comparison of Cancer Incidence to Vanderbilt Model

The last check for validity with the Vanderbilt model are the costs. The Vanderbilt model provides the only means of validating the cost of treating CRC. These results can be seen in Table 14 below. Accounting for the many differences between the models, including changes in the survival data used, the costs and discounted costs do match in magnitude and general trend.

	Vanderbilt	Vanderbilt-NC State
Costs	$2,611	$2,348
QALYs beyond 50	23.51	21.19

Table 16: Comparison of Costs with Vanderbilt Model

4.5 Summary of Chapter 4

This chapter described the verification and validation of the model. It also compared the results from the model to output from three different sources that were not used in the data fitting procedure. The sources for the targets for fitting the simulation were presented along with rationale for why the fitting was necessary. Given this background,

the actual fitting procedure was discussed along with the goodness of fit of the model. Lastly, analysis on the impact of CRC on individuals was examined and these results were compared to the results obtained through three other methods. Thus the Vanderbilt-NC State model is considered verified and validated and is believed to provide a legitimate means to reveal the natural history of CRC without intervention.

5. Conclusions and Recommendations

Since CRC is the second leading cause of cancer death among both males and females, the prevention and treatment of CRC is an important component of national healthcare policy and medical decision making. Before a rational policy can be set however, some method of evaluating alternative treatment and screening strategies is necessary. A discrete event simulation of the natural history of the disease such as the model in this thesis is fundamental for exploring alternative medical interventions. Because of the huge expense, and questionable ethics in performing large scale screening trials of medical interventions in real life, a simulated model of the disease is an important tool that allows the examination of alternative hypotheses. The creation of this simulation model is an important first step.

A simulation model for the natural course of CRC is a compromise of knowledge and data. A review of current models and data for CRC illustrated the general benefits of a discrete event simulation. It was concluded that there was opportunity to create a more comprehensive simulation of the natural course of CRC using modern computer technology.

The new simulation, called the Vanderbilt-NC State model, is a "grand hypothesis" that was built upon a series of identified assumptions. These assumptions were translated into a series of events describing the possible history of CRC in individuals. The events were related in an event graph along with a comprehensive description of each event. The inputs to the model were gathered from an extensive review of the medical literature and from an expert panel associated with the project. The simulation input was stored in a database to provide an easily manageable means of working the large variety of inputs while still providing the users the ability to change values within the simulation. The model itself was created in an object-oriented language, which makes even changes to the process flow straightforward. Because the object-oriented design makes the simulation person-focused instead of process-focused, the addition of screening and other medical

interventions can be studied without having to remap the flow the process. Once the simulation was constructed, extensive validation was performed to ensure that the simulation accurately and correctly modeled observed outcome. The simulation output closely matches studies tracking cancer incidence and mortality, adenoma incidence and prevalence, post cancer survival, and several other factors.

Matching all of these targets gives confidence in the model of the natural history of the disease. The original purpose of the project, namely to provide an updated and more robust model of CRC than the previous Vanderbilt model, was deemed a success. Some of the key factors that make this model successful are:

- The model is more robust than other similar models because it utilizes individual risk that is based on gender, race, and family history instead of an overall risk function.
- Risk can influence both the progression and incidence of adenomas
- The object-oriented approach makes the simulation more flexible and easier to change.
- The database interface within the simulation is an excellent method of storing the data in a convenient and easily manageable location, while still providing an intuitive interface for running the simulations.
- The level of model detail allows for many potential extensions to include test characteristics for future screening tests that may not have been developed or tested yet.
- The extensive validation provides for complete confidence the model, and in any screening analysis that is built upon it.

Although work for this thesis was a part of a larger project and the larger project encompasses this work, the specific contributions of this thesis were: (1) the construction of the natural history model with people and adenoma objects and events, (2) the programming of the interactions between the people, adenomas, and statistics at the events, (3) the incorporation of the data from the many sources into the model using the data base, (4) the creation of the adjusted life tables for additional ages and races without

CRC, (5) the verification of the simulation model, (6) the calibration of the simulation model by the fitting of unknown input parameters, (7) the validation through the external comparisons, and (8) the production of the outcomes, including costs and utilities.

5.1 Recommendations for Future Study

The main area for future research is the addition of screening to the CRC simulation. Right now, the simulation is a good model of the natural history of CRC, but no healthcare policy testing or intervention strategies have been studied. Adding screening strategies will allow for studies into the optimum screening type and surveillance interval. Moreover, the addition of screening will allow for further validation of the model using outcomes that were not used to fit the simulation

The second area for future research is to consider a population basis instead of a cohort basis. The current simulation takes a homogeneous group of people and runs them through the simulation. The National Cancer Institute would like to be able to simulate a diverse group of many different races and genders, as well as different screening strategies in various population subgroups. This is a non-trivial task since it requires a fundamental change in the way the inputs are acquired and people are created. One potential avenue is to have input distributions that determine the race, gender, and starting age of the person. Then instead of recording statistics if they pass a fixed starting age, record the statistic only when a person passes their individual starting age.

The last potential extension to the model is in performing sensitivity analysis. Because of the uncertain nature of the natural history of the disease, as well as variability in treatment costs and effectiveness, there are many variables that need to be examined closely. Many are modeled with random variables. This analysis would identify highly sensitive pieces of information that impact either patient survival or screening decisions. This approach would focus future medical research or epidemiology studies into the few key areas that have the most effect on the CRC outcomes.

Reference List

1. Androuny, FA. 2002. *Understanding Colon Cancer*.: University Press of Mississippi.

2. Arminski, TC and DW McLean. 1964. Incidence and distribution of adenomatous polyps of the colon and rectum based on 1,000 autopsy examinations. *Diseases of the Colon & Rectum* 7: 249-261.

3. Blank, L and A Tarquin. 2002. *Engineering Economy*. 5th ed. New York: McGraw Hill.

4. Blatt, LJ. 1961. Polyps of the colon and rectum: incidence and distribution. *Diseases of the Colon & Rectum* 4: 277-282.

5. Bong, D. 2004. "Monte Carlo Simulation." Available on-line via http://www.visionengineer.com/mech/monte_carlo_simulation.shtml

6. Brown, M, C Klabunde, and P Mysliwiec. 2003. Current Capacity for Endoscopic Colorectal Cancer Screening in the United States: Data from the National Cancer Institute Survey of Colorectal Cancer Screening Practices. *American Journal of Medicine* 115: 129-133.

7. Committee for Proprietary Medical Products. 2001. "Points to Consider on Misssing Data." Available on-line via http://www.emea.eu.int/pdfs/human/ewp/177699EN.pdf [accessed May 18,2004].

8. Correa, P, JP Strong, A Reif, and WD Johnson. 1977. The epidemiology of colorectal polyps: prevalence in New Orleans and international comparisons. *Cancer* 39, (5): 2258-2264.

9. Davies, R, P Roderick, and J Raftery. 2003. Evaulation of Disease Prevention and Treatment using Simulation Models. *European Journal of Operations Research* 150: 53-66.

10. DevCan Probability of Developing or Dying of Cancer Software. Ver. 5.0. Statistical Research and Applications Branch, National Cancer Institute.

11. Eide, TJ 1986. The age-, sex-, and site-specific occurrence of adenomas and carcinomas of the large intestine within a defined population. *Scandinavian Journal of Gastroenterology* 21, (9): 1083-1088.

12. Eide, TJ and H Stalsberg. 1978. Polyps of the large intestine in Northern Norway. *Cancer* 42, (6): 2839-2848.

13. Fay, MP. Estimating Age-Conditional Probability of Developing Cancer using a Piecewise Mid-Age Group Joinpoint for the Rates. 2003-01. 2003. Statistical Research and Applications Branch, National Cancer Institute.

14. Frazier, AL, GA Colditz, CS Fuchs, and KM Kuntz. Cost Effectiveness of Screening for Colorectal Cancer in the General Population. JAMA 284, 1954-1961. 2000.

15. Gold, M, J Siegel, L Russel, and M Weinstein. 1996. *Cost Effectiveness in Health and Medicine*. Edited by Gold, M, J Siegel, L Russel, and M Weinstein. New York: Oxford University Press.

16. Hollenberg, J. SMLTREE. 1987.

17. Johns, LE and RS Houlston. 2001. A systematic Review and Meta-Analysis of Familial Colorectal Cancer Risk. *American Journal of Gatroenterology* 96, (10): 2992-3003.

18. Joines, JA and SD Roberts. 1998. Fundamentals of object-oriented simulation. In *Proceedings of the 1998 winter simulation conference*, 141-151. Washington DC.

19. Ladabaum, U, C Chopra, G Huang, J Scheiman, and M Chernew. 2001. Aspirin as an Adjunct to Screening for Prevention of Sporadic Colorectal Cancer. *Annals Of Internal Medicine* 135, (9): 769-781.

20. Law, AM and WD Kelton. 2000. *Simulation Modeling and Analysis*. 3rd ed. Boston: McGraw Hill.

21. Lieberman, D, S Prindiville, D Weiss, and W Willett. 2003. Risk Factors for Advanced Colonic Neoplasia and Hyperplastic Polyps in Asymptomatic Individuals. *JAMA* 290, (22): 2959-2967.

22. Liebsch, C. 2003. Simulation Input Modeling in the Absence of Data. Master of Science Simulation Input Modeling in the Absence of Data, North Carolina State University.

23. Loeve, F, R Boer, G Oortmassen, M Ballegooijen, and J Habbema. 1999. The MISCAN-COLON Simulation Model for the Valuation of Colorectal Cancer Screening. *Coomputers and Biomedical Research* 32: 13-33.

24. Loeve, F, M Brown, R Boer, M Ballegooijen, G Oortmassen, and J Habbema. 2000. Endoscopic Colorectal Cancer Screening: a Cost Saving Analysis. *Journal of the National Cancer Institute* 92, (7): 557-563.

25. Mandel, JS, JH Bond, TR Church, DC Snover, GM Bradley, LM Schuman, and F Ederer. 1993. Reducing mortality from colorectal cancer by screening for

fecal occult blood. Minnesota Colon Cancer Control Study. *New England Journal of Medicine* 328, (19): 1365-1371.

26. Naimark, D, M Krahn, G Naglie, D Redelmeier, and A Detsky. 1997. Primer on Medical Decision Analysis: Working with Markov Processes. *Medical Decision Making* 17: 152-159.

27. National Cancer Institute. 2003. "SEER Cancer Registries." Available on-line via http://www.seer.cancer.gov/ [accessed February 23,2003].

28. Ness, R, A Holmes, R Klein, and R Dittus. 2000. Cost Utility of One-Time Colonoscopic Screening for Colorectal Cancer at Various Ages. *American Journal of Gastroenterology* 95: 1800-1811.

29. Nord, E. 1999. *Cost-Value Analysis in Health Care*. New York: Cambridge University Press.

30. Pignone, M, S Saha, T Hoerger, and J Madelblatt. 2002. Cost-Effectiveness Analyses of Colorectal Cancer Screening: A Systematic Review for the US Preventive Services Task Force. *Annals Of Internal Medicine* 137, (2): 96-104.

31. Roberts, SD. 2003. Input Models for the Simulation of Medical Life Histories.

32. Roberts, SD. 2004, Simulation Input without Data, *Proceedings of6th Annual Simulation Solutions Conference '04*, March 16, 2004, Orlando, Florida.

33. Schafer, J and M Olsen. 1998. "Multiple Imputation for multivariate missing-data problems: a data analyst's perspective." Available on-line via http://www.stat.psu.edu/~jls/mbr.pdf [accessed May 18,2004].

34. Sonnenberg, A, F Delco, and J Inadomi. 2000. Cost Effectiveness of Colonoscopy in Screening for Colorectal Cancer. *Annals Of Internal Medicine* 133, (8): 573-584.

35. Vatn, MH and H Stalsberg. 1982. The prevalence of polyps of the large intestine in Oslo: an autopsy study. *Cancer* 49: 819-825.

36. Vijan, S, E Hwang, T Hofer, and R Hayward. 2001. Which Colon Cancer Screening Test? A Comparison of Costs, Effectiveness, and Complaince. *American Journal of Medicine* 111: 593-601.

37. Williams, AR, BAW Balasooriya, and DW Day. 1982. Polyps and cancer of the large bowel: a necropsy study in Liverpool. *Gut* 23: 835-842.

38. Wilmoth, JR. 2003. "Berkeley Mortality Database." Available on-line via http://www.demog.berkeley.edu/wilmoth/mortality/ [accessed January 23,2003].

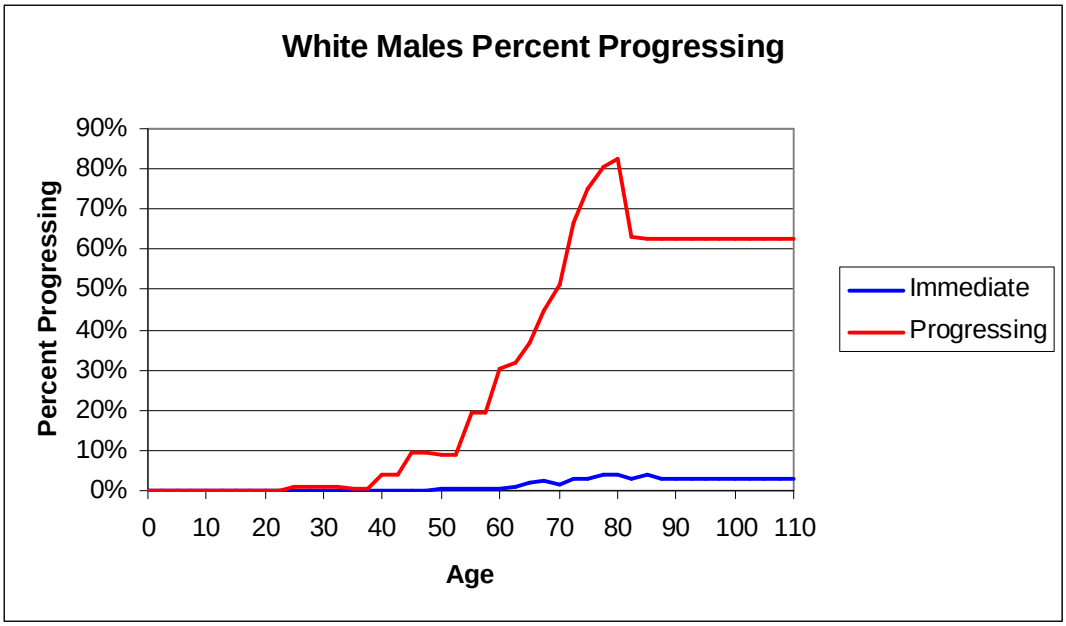
39. Winawer, S, R Fletcher, and D Rex. 2003. Colorectal Cancer Screening and Surveillance: Clinical Guidelines and Rational - Update Based on New Evidence. *Gatroenterology* 124: 544-560.

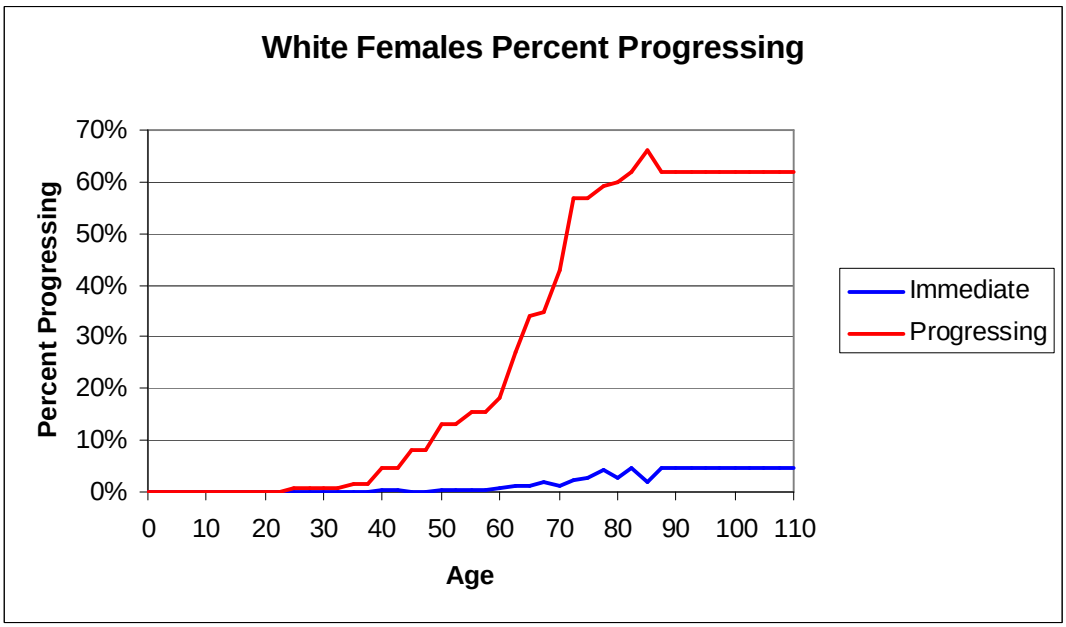
40. Winawer, SJ, AG Zauber, MJ O'Brien, LS Gottlieb, SS Sternberg, ET Stewart, JH Bond, J Schapiro, JF Panish, JD Waye, RC Kurtz, M Shike, and MH Ho. 1992. The national polyp study. *Cancer* 70, (5): 1236-1245.

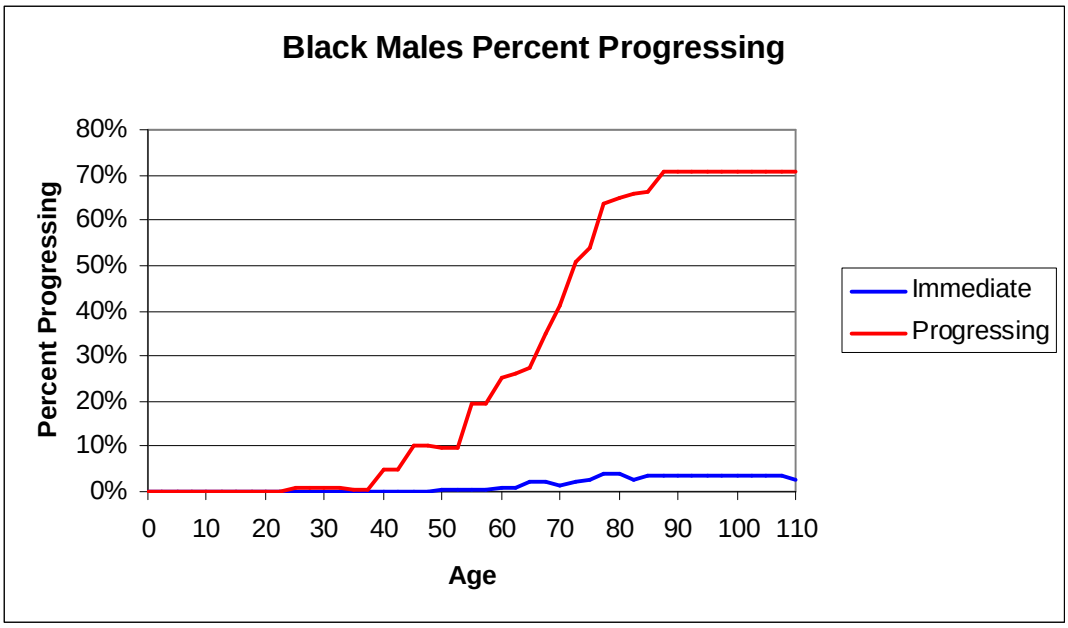
41. Yorke-Smith, N. and C. Gervet. Data Uncertainty in Constraint Programming: A Non-Probabilitistic Approach. 2001. American Association for Artificial Intelligence.

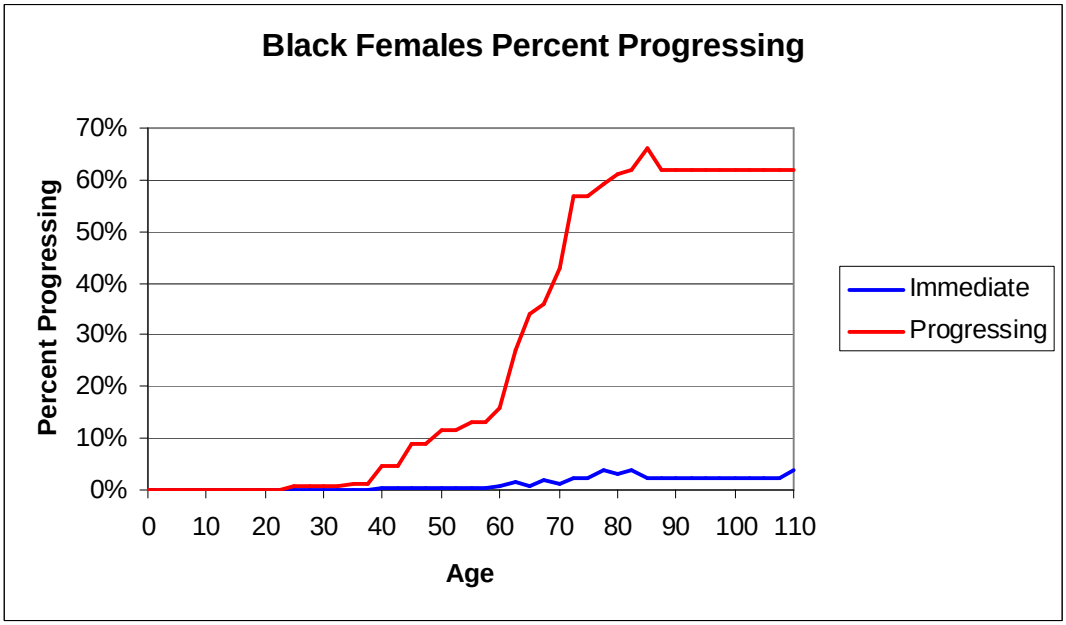
Appendices

Appendix A: Percent of Adenomas Progressing and Immediate

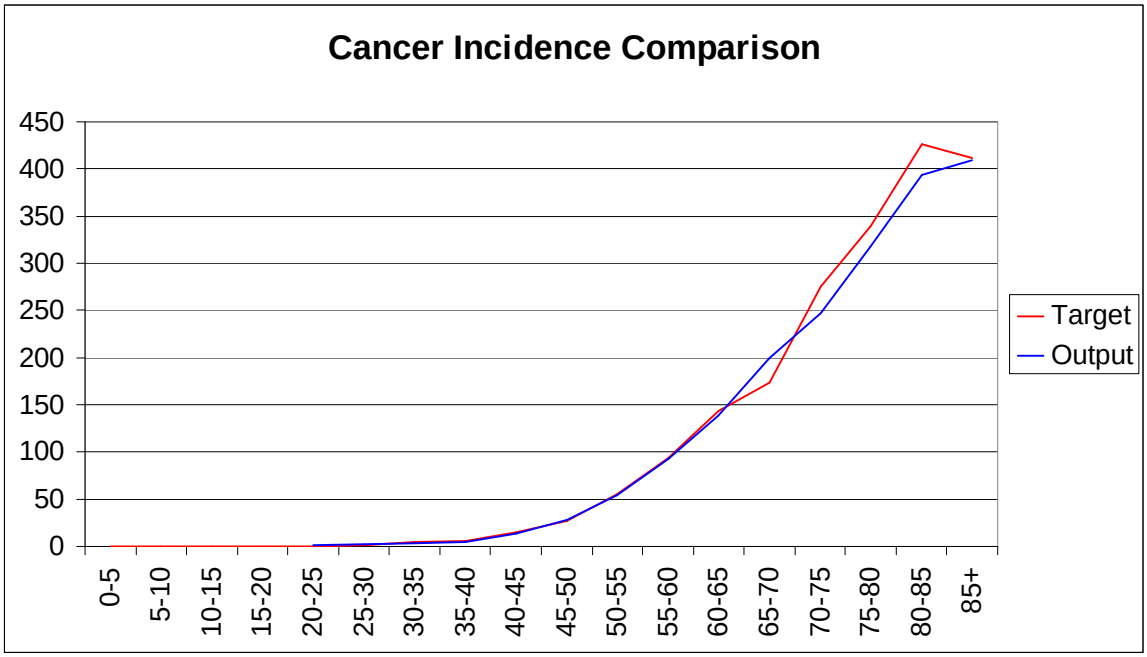


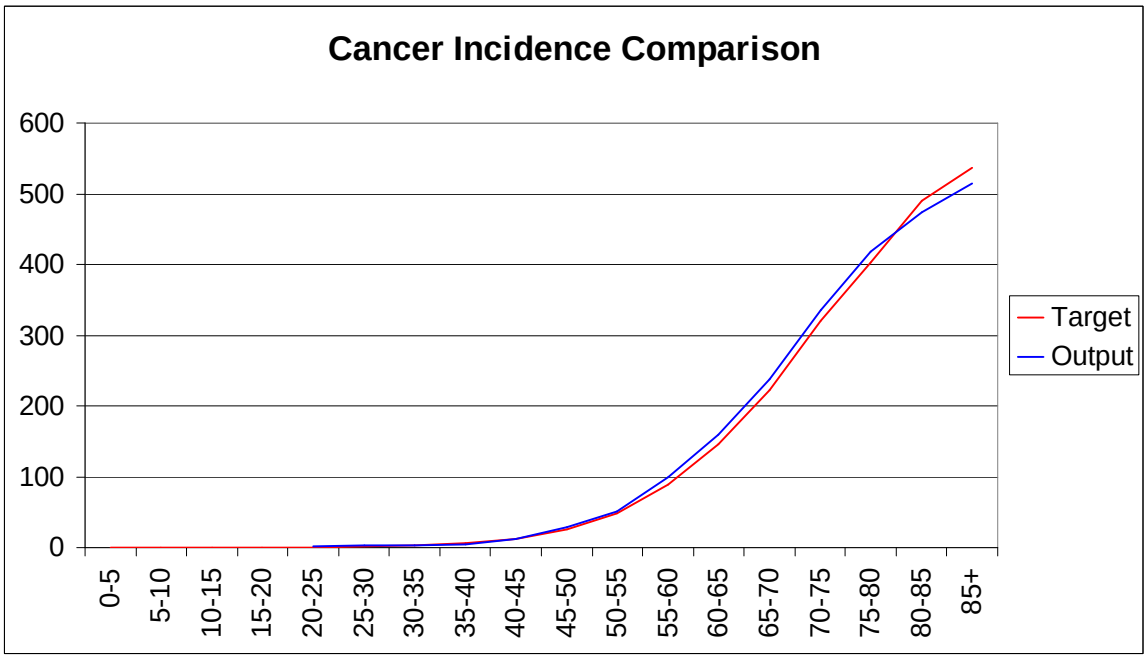



Black Males Percent Progressing
Percent Progressing
80%
70%
60%
50%
40%
30%
20%
10%
0%
0
10
20
30
40
50
60
70
80
90
100
110
Age
Immediate
Progressing

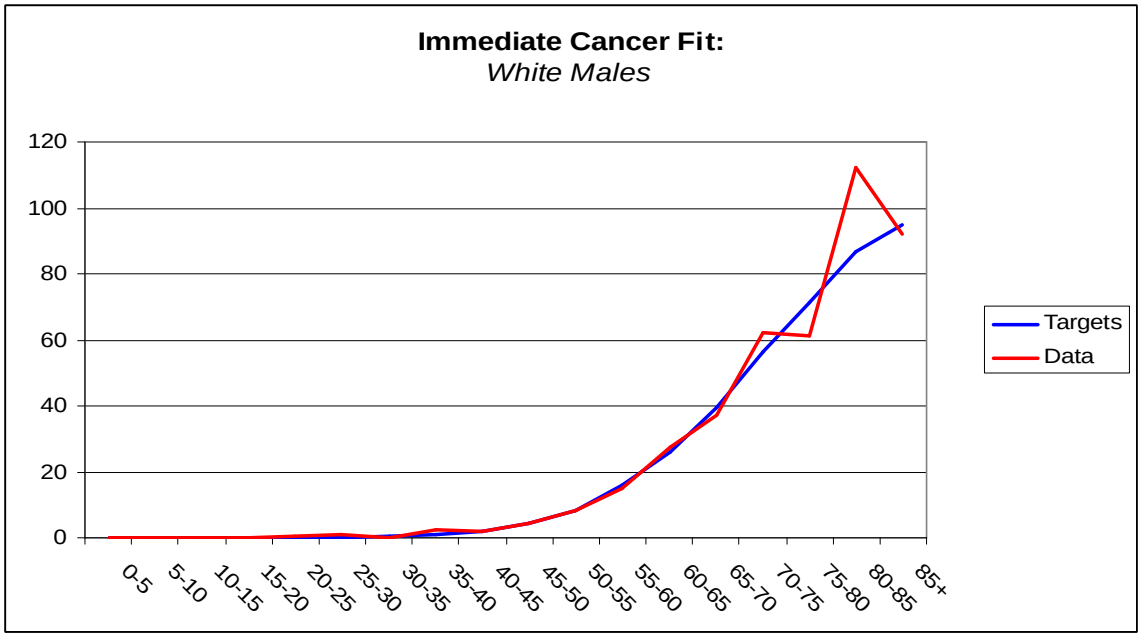

Black Females Percent Progressing
Percent Progressing
70%
60%
50%
40%
30%
20%
10%
0%
0
10
20
30
40
50
60
70
80
90
100
110
Age
Immediate
Progressing

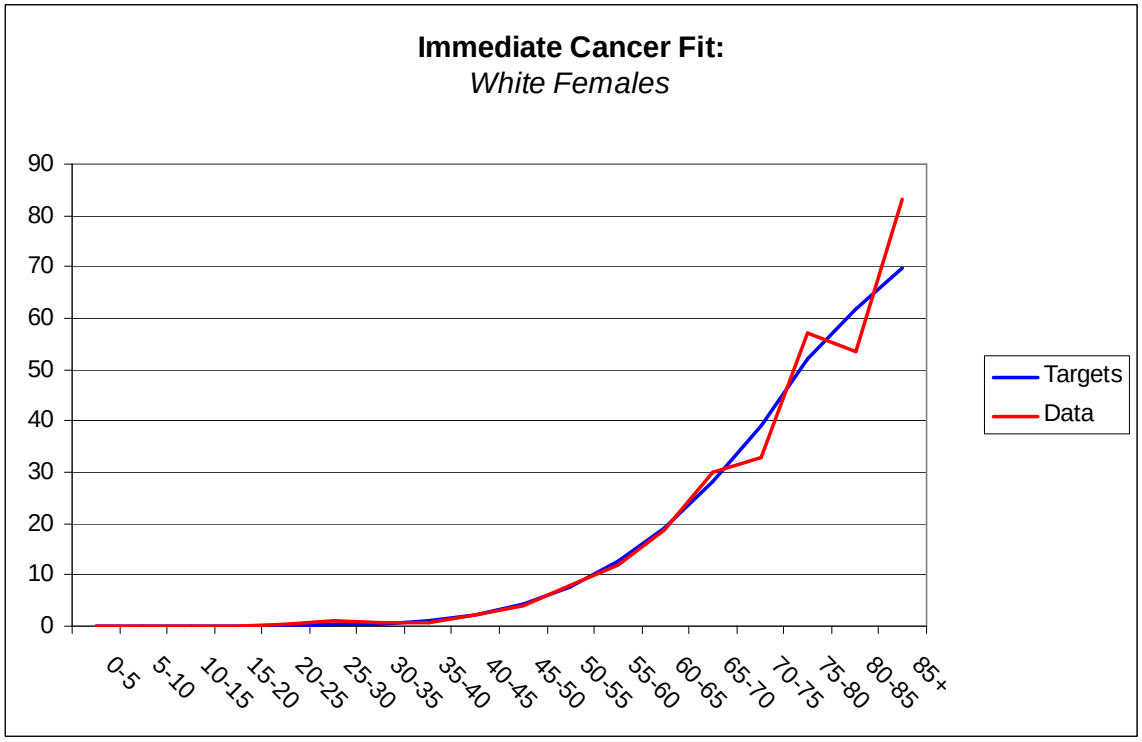
Appendix B: Progressive Adenoma Cancer Fit

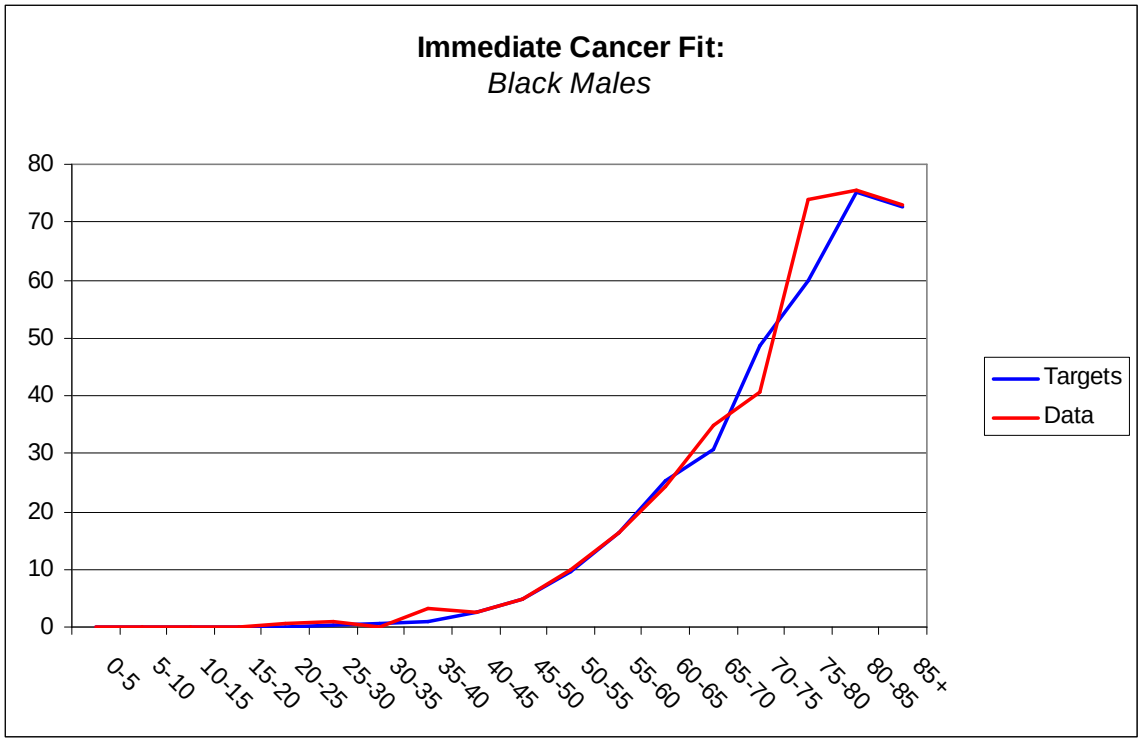


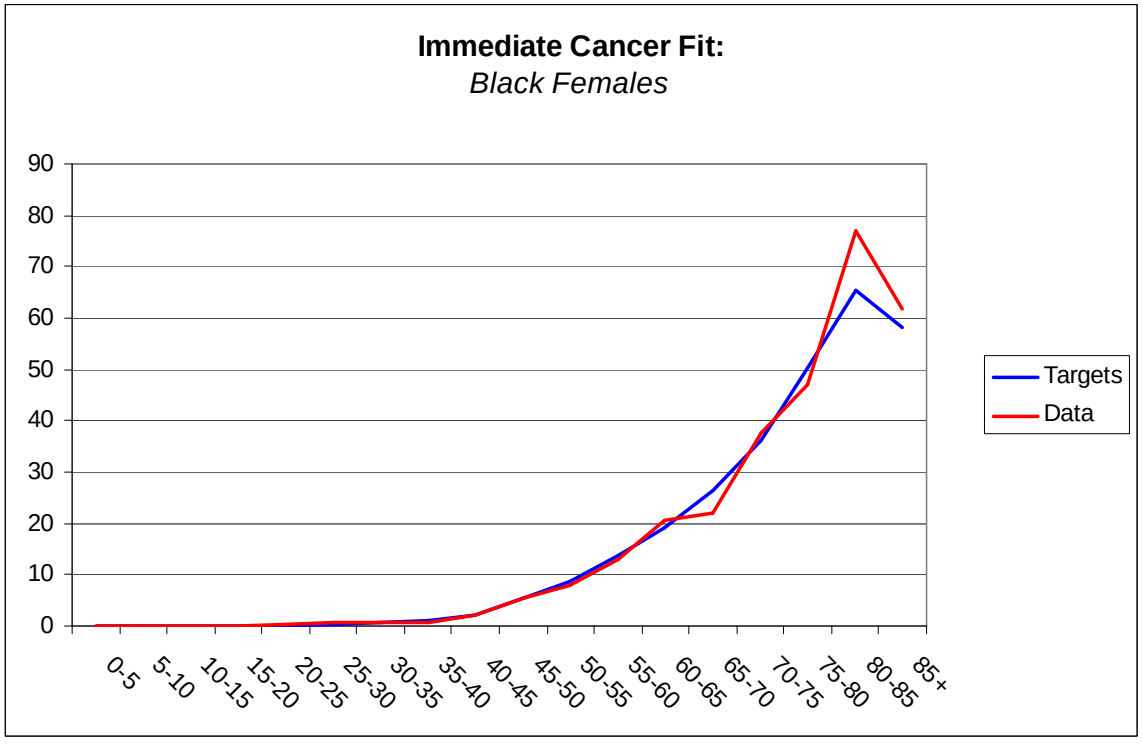


Appendix C: Initial Immediate Cancer Fit

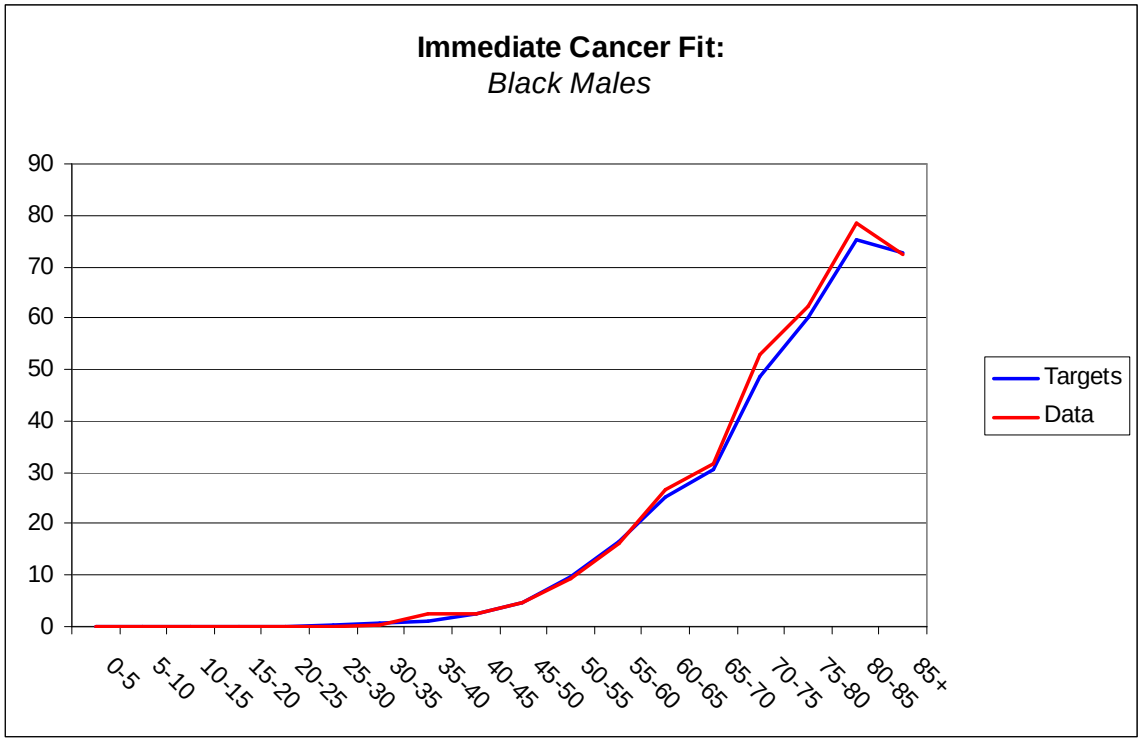


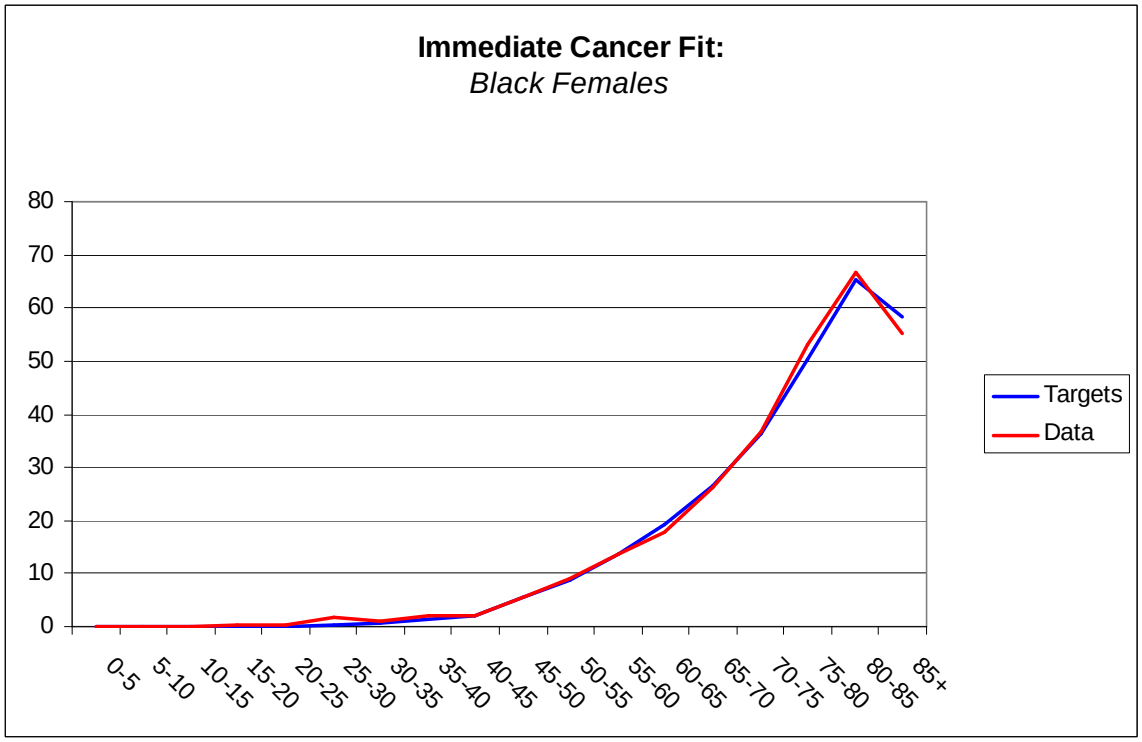


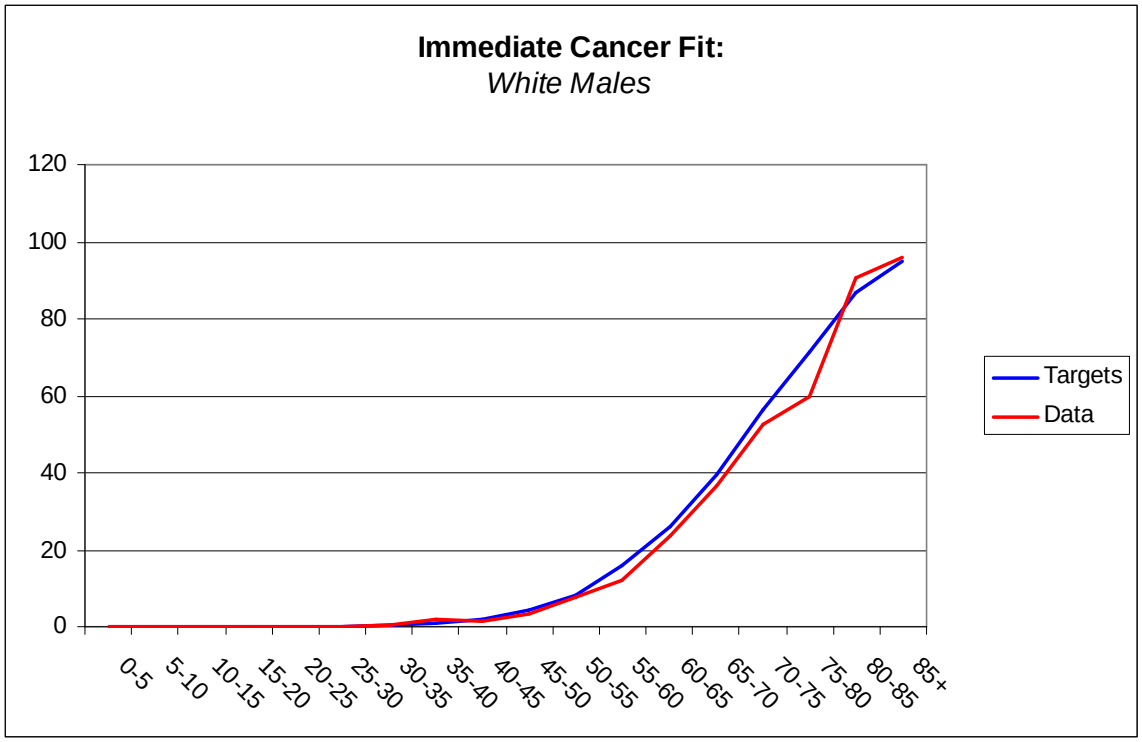

Immediate Cancer Fit:
Black Males
80
70
60
50
40
30
20
10
0
0-5
5-10
10-15
15-20
20-25
25-30
30-35
35-40
40-45
45-50
50-55
55-60
60-65
65-70
70-75
75-80
80-85
85+
Targets
Data

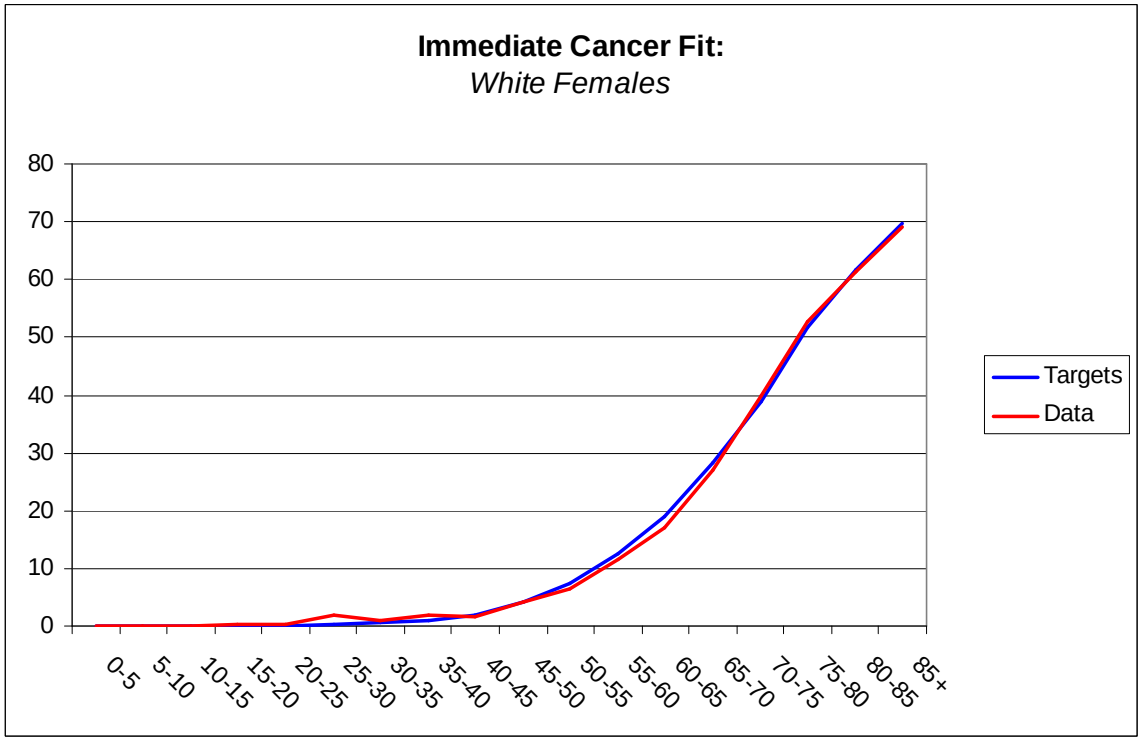

Immediate Cancer Fit:
Black Females
90
80
70
60
50
40
30
20
10
0
0-5
5-10
10-15
15-20
20-25
25-30
30-35
35-40
40-45
45-50
50-55
55-60
60-65
65-70
70-75
75-80
80-85
85+
Targets
Data

Appendix D: Final Immediate Cancer Fit






Immediate Cancer Fit:
White Males
120
100
80
60
40
20
0
0-5
5-10
10-15
15-20
20-25
25-30
30-35
35-40
40-45
45-50
50-55
55-60
60-65
65-70
70-75
75-80
80-85
85+
Targets
Data

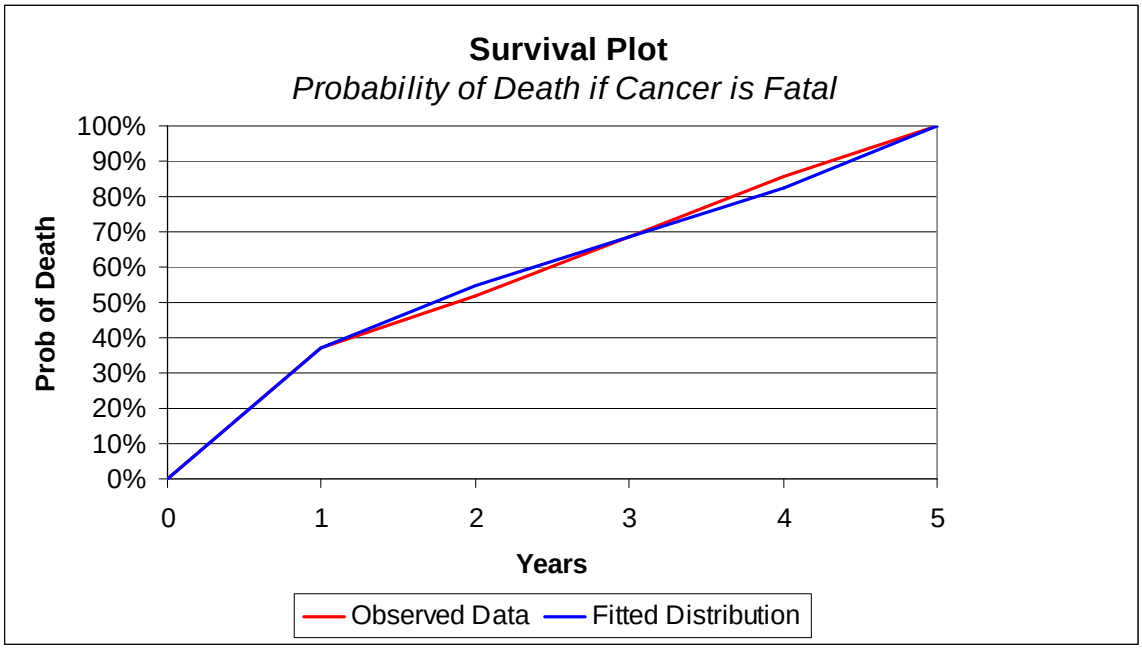

Immediate Cancer Fit:
White Females
80
70
60
50
40
30
20
10
0
0-5
5-10
10-15
15-20
20-25
25-30
30-35
35-40
40-45
45-50
50-55
55-60
60-65
65-70
70-75
75-80
80-85
85+
Targets
Data

Appendix E: Percent progressing values used to derive initial immediate cancer fit

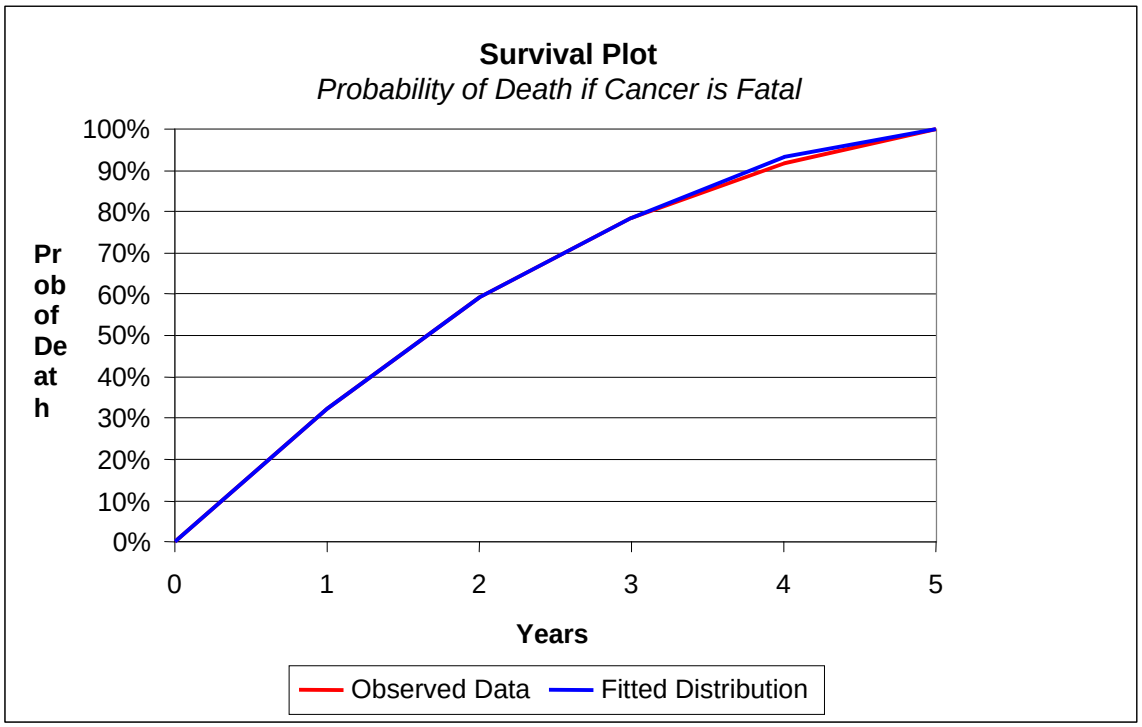
Age	WM	WF	BM	BF
0	0%	0%	0%	0%
25	1%	1%	1%	1%
35	1%	1%	1%	1%
40	4%	5%	5%	5%
45	9%	8%	10%	9%
50	9%	13%	10%	12%
55	20%	16%	20%	13%
60	31%	18%	25%	16%
62.5	32%	27%	24%	27%
65	37%	34%	27%	34%
67.5	45%	35%	35%	25%
70	51%	43%	30%	43%
72.5	66%	57%	51%	57%
75	75%	57%	49%	57%
77.5	81%	59%	64%	59%
80	83%	55%	62%	55%
82.5	61%	62%	71%	62%
85	63%	66%	67%	66%
110	63%	62%	71%	62%

Appendix F: Survival Distributions by Stage

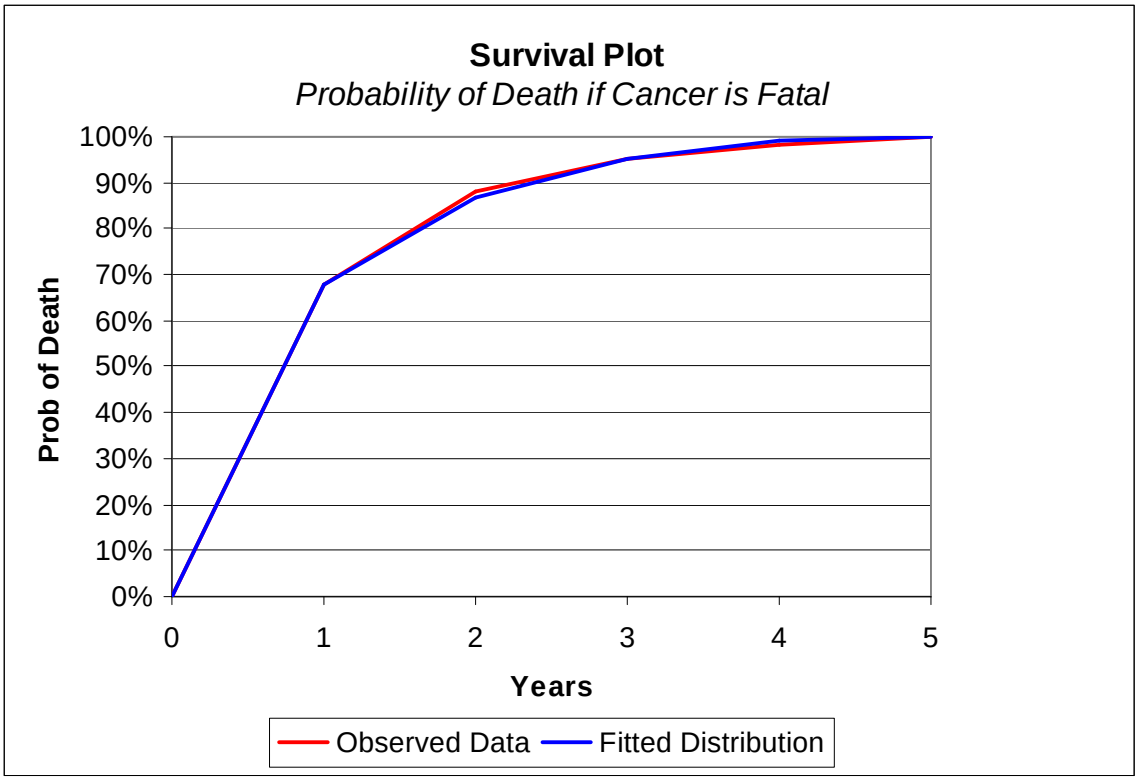
Local Cancer



Regional Cancer



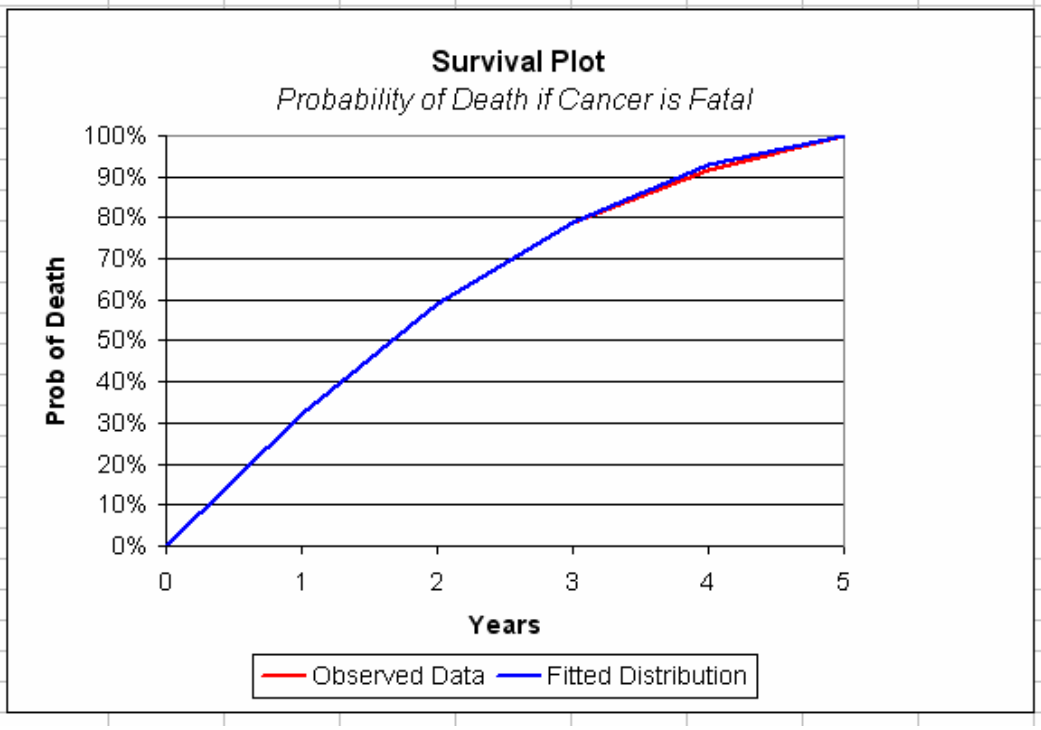
Distant Cancer



Appendix G: Cancer Survival Fitting Spreadsheet

Year	Data	Modified D	Johnson S	Error
0	0	0	0	NA
1	0.1269	0.3228	0.3228	0.00%
2	0.2324	0.5912	0.5903	0.15%
3	0.3095	0.7873	0.7873	0.00%
4	0.3607	0.9176	0.9313	1.49%
5	0.3931	1	1	N/A
			Overall	0.41%

ohnson SB Paramante	
Min	0
Max	5
Gamma (a	0.5127
Delta (a2)	0.7016



Appendix H: Cancer Fitting Spreadsheet

	A	B	C	D	E	F	G	H	I	J	K	L	M	
1		**With History**												
2		Population	180000											
3		Age	Can. Inc	Ad Inc	First Ad	Deaths	polyps dead	first dead	Immediate	canc inc	% with	ad inc	Imm Inc	Pop
4	0	0	0	0	0	0	0	0	0	0.0	0.0%	0.0%	0.0%	
5	0-5	0	0	0	0	5226	0	0	0	0.00	0.00%	0.00%	0.00	
6	5-10	5	0	0	0	425	0	0	0	0.00	0.00%	0.00%	0.00	
7	10-15	10	0	0	0	460	0	0	0	0.00	0.00%	0.00%	0.00	
8	15-20	15	1	13653	12590	1070	45	43	0	0.58	7.24%	7.85%	0.00	
9	20-25	20	10	13931	11137	1342	186	167	2	5.79	13.70%	15.94%	1.16	
10	25-30	25	11	14491	10049	1259	248	206	5	6.41	19.57%	24.40%	2.92	
11	30-35	30	12	15322	9447	1375	452	329	1	7.05	25.15%	33.44%	0.59	
12	35-40	35	22	16317	8846	2081	856	613	8	13.03	30.40%	43.11%	4.74	
13	40-45	40	46	19424	9286	3223	1666	1128	10	27.58	35.98%	54.78%	6.00	
14	45-50	45	96	24177	10049	5490	3544	2203	13	58.70	42.14%	69.62%	7.95	
15	50-55	50	197	26403	9263	8833	6973	4024	54	124.64	48.07%	86.51%	34.17	
16	55-60	55	340	25425	7664	12533	12397	6443	39	227.86	53.36%	103.99%	26.14	
17	60-65	60	504	23006	5836	18019	20910	10147	105	368.74	57.81%	121.29%	76.82	
18	65-70	65	670	19755	4468	22338	29834	13407	91	564.62	61.80%	138.73%	76.69	
19	70-75	70	679	15353	2910	25136	38034	16040	133	704.90	65.23%	155.87%	138.07	
20	75-80	75	712	10525	1820	25921	43820	17620	81	1000.14	68.29%	172.27%	113.78	
21	80-85	80	486	9415	1419	22344	42203	15764	85	1073.58	71.60%	194.91%	187.77	
22	85-90	85	329	6203	730	14986	32814	11233	29	1435.11	75.81%	227.84%	126.50	
23	90-95	90	101	1807	170	6321	16672	5014	3	1272.20	79.57%	267.35%	37.79	
24	95-100	95	24	320	29	1345	4027	1101	1	1483.31	82.32%	301.17%	61.80	
25	100-105	100	3	67	5	212	682	182	0	1098.90	86.45%	334.43%	0.00	
26	105-110	105	0	5	0	61	236	54	0	0.00	88.52%	386.89%	0.00	
27	Total		4243	255599	105718	180000	255599	105718	660					
28	**85+**									1993.46	76.70%	237.43%	143.95	
29														

Cell Definitions

Column A – Age category for the population

Column B – Starting age of the population group. This is the start point for the copy from the fitting spreadsheet

Column C – Cancer incidence that become symptomatic from the simulation for that age group

Column D – Adenoma incidence from the simulation output

Column E – The number of people who get an adenoma for the first time in that age group

Column F – The number of people who died after making it to that age group

Column G – The number of polyps that were there on the people who died that period

Column H – The number of people with adenomas who died that age group

Column I – The number of immediate cancers that become symptomatic in that age group

Column J – This column is the first calculated column in the sheet. The formula for this cell at the age group of 10 to 15 (J7) is C7/(C$2-Sum(F$4:F6))*100000/5. This turns the cancer incidence observed by the simulation into a cancer incidence rate of cases/100,000.

Column K – This column is the percent of living people with adenomas. The formula is (Sum(E$5:E7)-SUM(H$4:H6))/(C$2-SUM(F$4:F6)). What this formula does is find the number of people with adenomas at this age and divides by the total living population.

Column L – This column is the adenomas per 100 people. The formula is (Sum(D$5:D7)-SUM(G$4:G6))/(C$2-Sum(F$4:F6)). What this formula does is find the number of living adenomas in this age and divides by the total living population.

Column N – This column is the first calculated column in the sheet. The formula for this cell at the age group of 10 to 15 (N7) is I7/(C$2-Sum(F$4:F6))*100000/5. This turns the immediate cancer incidence observed by the simulation into a immediate cancer incidence rate of cases/100,000.

These columns are then compiled in a separate sheet that calculates the weighted average of each of these based upon family history and without family history (weighting of 0.19 for with family history). The outputs of the simulation are then compared to the target values using an average percent error.