ECOLE POLYTECHNIQUE FEDERALE DE LAUSANNE
SCHOOL OF LIFE SCIENCES

Massachusetts Institute of Technology

Master's project in Bioengineering and Biotechnology

Analysis of familial risk for prostate, colon and female breast cancer in Sweden

Done By

Paulo Refinetti

Carried out in the laboratory of Epidemiology at Massachusetts Institute of Technology Under the supervision of Prof. William G. Thilly, Head of the Lab.

Under the direction of Prof. Stephan Morgenthaler
In the laboratory of Applied statistics, EPFL

LAUSANNE, EPFL 2011

Table of Contents

Summary

Using the Swedish Familial Cancer Database, organized and studied by the Hemminki group since 1997, the hypotheses that breast, prostate and colorectal cancers were affected by familial risks in Sweden (1958-2008) are tested. The null hypotheses were rejected for the three cancer types up to the age interval 70-74 yrs for which probative data are available.

Familial risks, represented as age-specific mortality rate ratios, are found to be age invariant: 1.9±0.3 (95% CI) for sons or daughters ages 45 to 74 of parents dying of colorectal cancer, 2.0 ± 0.2 for sons ages 50 to 74 of fathers dying of prostate cancer. The ratio of the risks for daughters of mothers dying of breast cancer ages 35 to 74, suggests a different behavior between pre-menopause and post-menopause ages. Can be seen as constant at 2.3± 0.3 for all observed age groups. The data could also be interpreted as showing a different behavior between the ages before and after menopause. The ratio could be decreasing linearly before menopause and be constant after menopause

One potential risk parameter is eliminated as familial because this parameter is identical for the general and familial populations: growth rate of preneoplastic lesions. Initiation or promotion oncomutation rates in stem cells could account for the familial risks in accord with wide variations of mutation rates observed in developing colon and lungs. Population parameters such as the fraction of persons in whom initiated

stem cells would continue to grow after maturity could also have caused the observed familial risks, with oncomutation rates. For familial colorectal cancer, the two population parameters considered are of little effect, permitting the conclusion that familial colorectal cancer risk lies solely in oncomutation rates.

Studies of immigrants have demonstrated that fetal/juvenile environmental factors are major determinants of organ-specific adult cancer risk. From these considerations emerges the hypothesis that the prime determinants of colorectal and other cancer risks lie in the fetal/juvenile oncomutation rates. This hypothesis may be tested directly by measuring and comparing somatic mutations at defined genetic loci (a.) in parents and their adult children (b.) among all adults with and without cancers in a specified organ.

Introduction

In 1896 Roentgen declaimed: "Kampf den Krebs! Krebs ist heilbar." Because he interpreted shrinkage of tumors after x-irradiation as evidence that means could be found to cure this dread disease. But in 2009, the American Senator Arlen Specter observed: "The lack of progress in cancer therapy in my lifetime has been scandalous."

Specter's criticism is supported by the cancer mortality data of the United States recorded from 1895-2008 (the data can be viewed at http://mortalityanalysis.mit.edu) and similar observations in Asian and European countries in the 20th century. The U.S. age-specific adult mortality rates for female breast (Figure1), male prostate (Figure 2) and male and female colorectal cancers (Figures 3 and 4) are nearly constant. Such improvements as have been made are ascribable to early detection and surgery, e.g., cervical and colorectal cancers, or public health efforts as in anti-cigarette campaigns that decreased lung cancer incidence. There are no explanations for other significant decreases such as in the mortality rates for solid tumors of juveniles and young adults beginning circa 1940. See female breast cancer, Figure 1.

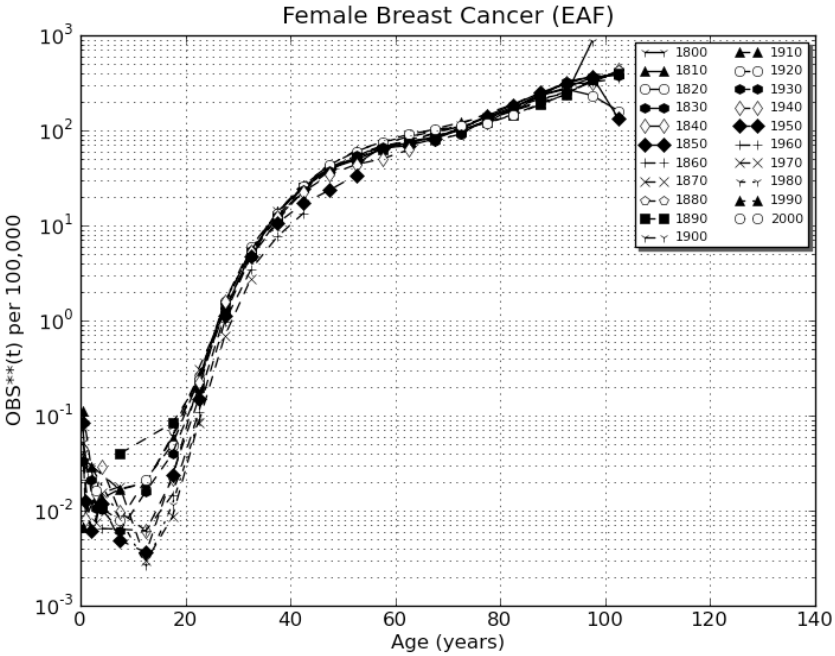

Figure 1: Mortality from breast cancer in all reporting U.S. counties for European American females 1895-2008. **[http://mortalityanalysis.mit.edu]**. Data recorded by the U.S. Census Bureau 1895-1935, the U.S. Public Health Service, 1936-2008.

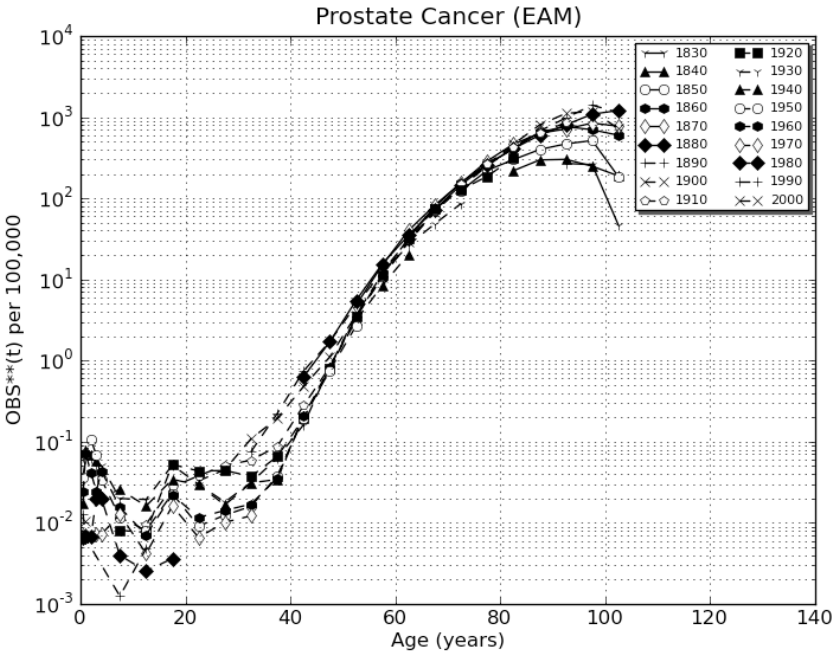

Figure 2: Mortality from prostate cancer in all reporting U.S. counties for European American males 1895-2008. **[http://mortalityanalysis.mit.edu]**. Data recorded by the U.S. Census Bureau 1895-1935, the U.S. Public Health Service, 1936-2008.

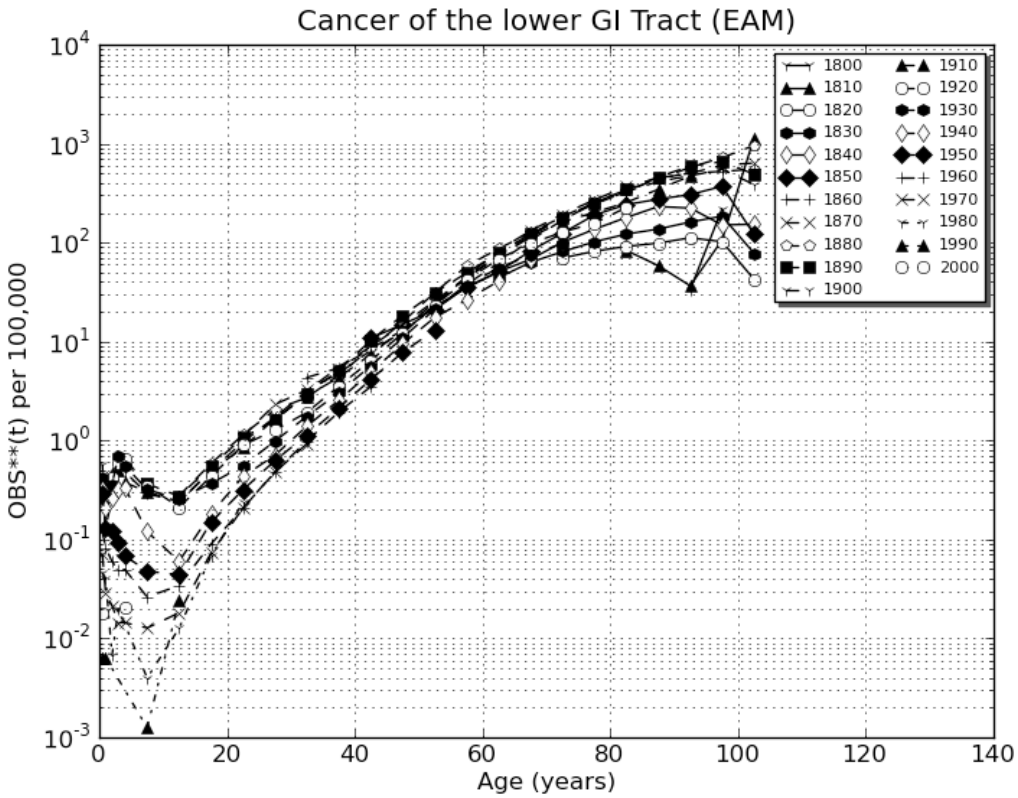

Figure 3: Mortality from colorectal (lower digestive tract) cancers in all reporting U.S. counties for European American males 1895-2008. **[http://mortalityanalysis.mit.edu]**.

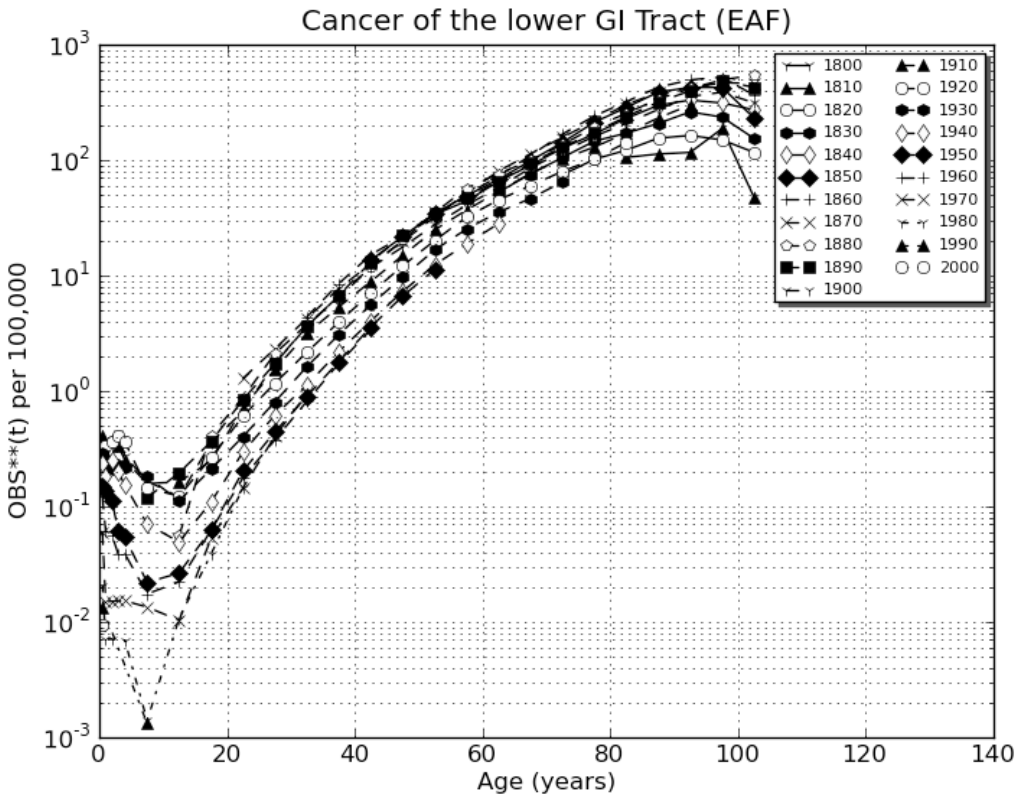

Figure 4: Mortality from colorectal (lower digestive tract) cancer in all reporting U.S. counties for European American females 1895-2008. **[http://mortalityanalysis.mit.edu]**.

Senator Specter might also have aimed his criticism of scientists in the related area of cancer causation. In 1971, the oncologist/chemist Charles Heidelberger pointed out that "We now know what causes cancer: chemicals, radiation, viruses or something else. " Today other than the risk of lung cancer from cigarette use and skin cancer from exposure to sunlight there are no explanations for the causation of the most common forms of cancers including colorectal, prostate and breast. For less common forms of cancers viral etiology has been demonstrated as in papilloma virus and cervical cancer, Epstein-Barr virus and Burkitt's lymphoma. (http://mortalityanalysis.mit.edu).

Note should be made that Dr. Herrero-Jimenez created the extraordinary database for disease mortality records in the United States from which Figures 1-4 are copied. The data are provided for many diseases back to 1895 organized in ways that permit examination of historical changes at a level unavailable in any other large human population. This database has been publically available since 1996 and is

presently found at http://mortalityanalysis.mit.edu. This thesis would not have been possible without his contribution.

Carcinogenesis:

The general "two-stage" cascade hypothesis and recent modifications.

Cancer has been studied at the population, tissue, cell, genetic and molecular levels. From these studies a coherent model of carcinogenesis expressible as a "two-stage" cascade of rare events in the stem cell lineage of organogenesis and preneoplasia has evolved (NORDLING 1953; PLANT 1955; ARMITAGE & DOLL 1957; ARMITAGE 1953; Moolgavkar & A G Knudson 1981; A G Knudson 1971; Cairns 1975; Herrero-Jimenez et al. 2000). The general picture as of 2008 was that inactivation by mutation of about two tumor suppressor genes in a tissue stem cell lineage were required to "initiate" some form of "cell at risk," such as a stem cell. The initiated cell continues to grow slowly and to produce more initiated stem cells, thus forming a preneoplastic lesion such as an adenomatous polyp of the colon. One or more rare events such as mutations are then believed to "promote" an initiated preneoplastic stem cell to become a founder neoplastic stem cell for a neoplastic lesion that would grow, usually metastasize, and kill the patient.

This developing model has been adjusted in response to the discovery that mutations in adult tissues arose solely in the stem cells of developing tissue: tumor initiating mutations appear to be limited to the stem cells of the fetal/juvenile period(Sudo et al. 2008)(Kini et al. 2011 as an Appendix).

Nature of human stem cells.

In a surprising advance in cancer cell biology, the stem cells of organogenesis, preneoplasia and neoplasia have been found to be an entirely unexpected life form, the metakaryote, the discussion of which is beyond the scope of this thesis but the reality of which must influence all future research (Gostjeva et al. 2006; Gostjeva et al. 2009; Sudo et al. 2008; Zheng et al. 2006). In brief, stem cells of organ

development in plants and animals are non-eukaryotic cells that use a double stranded RNA/DNA genomic replicative intermediate, separate the sister dsRNA/DNA genomes without the use of mitosis, then degrade the RNA strand replacing it with a DNA strand employing at least two "by-pass" DNA polymerases with high intrinsic mutation rates. These cells are highly resistant to killing by x-rays (and radiomimetic drugs) and these characteristics explain in large part the failed hopes of Roentgen and others who tried to cure cancer by the use of treatments that killed eukaryotic but not "metakaryotic" cancer stem cells.

Tumor suppressor genes in which required initiating oncomutations occur.

Rare familial forms of common cancers killing juvenile or young adults have permitted oncogeneticists to identify a few forms of autosomal cancer tumor suppressor genes in which mutations required for tumor initiation occur. In these cases the afflicted individual inherits one inactivated tumor suppressor allele under circumstances wherein the inactivation of both parental copies appears to be sufficient for tumor "initiation": colorectal cancer (APC) (Fearnhead & Wilding 2002), kidney cancer (VHL),(Kaelin 2007) nervous system cancers (NFI, NFII) (Brooks 2004), retinoblastoma and osteosarcoma (RB)(Yamasaki 2004), basal cell skin carcinoma (PTCH)(Leiter & Garbe 2008). To the author's knowledge, no measurements of somatic mutation rates of these genes have yet been reported for the tissues at risk.

No genes are known in which required promoting oncomutations occur.

No genes have been found to be the sites of mutations of "promotion". Mutations in the murine/avian oncogenes such as RAS and the gene TP53 were originally mistaken for promotional genes because these mutations are frequently found in human tumors. However, mutations in these genes are distributed in sectors of human tumors and are not found in all tumor cells of a cancer as would be required for a required oncomutation creating a monoclonal tumor. (A hypothesis has been advanced that mutations in these genes make the mutant cells responsive to hypertrophy by autocrine factors released by tumor cells.)

Genes that modulate the rate of mutation in metakaryotic stem cells.

Other rare familial forms of cancers involve inheritance of inactivated alleles in the putative pathways for environmentally induced mutation and thus oncomutation of fetal/juvenile stem cells. Examples include the several genes in which inheritance of inactivated alleles from both parents results in xeroderma pigmentosum, in which it appears that the stem cells of skin epithelium become hypersensitive to mutation by sunlight (Cleaver 2000).

Genes that carry mutations that modulate the growth potential of initiated stem cells.

From the example of cigarette smoking and lung cancer causation comes the teaching that a carcinogen need not cause mutations in tumor suppressor genes or oncogenes. In fact, a carcinogen may permit/stimulate the growth of pre-existing initiated preneoplastic stem cells that do not grow in the absence of the carcinogenic agent(Sudo et al. 2008).

It is possible that persons may be exposed to an environmental agent or inherit a mutation that places them at risk of a particular form of cancer relative to unexposed/unmutated persons. Such conditions could very well be familial in nature.

Genes that carry mutations that modulate susceptibility to competing forms of mortality that share risks with the observed form of cancer.

Demonstrable sharing of risk between breast cancer and ovarian mortality supports the possibility that an environmental exposure or inherited condition could alter the degree of shared risk and that persons so affected would have higher or lower expectations of death by the observed cancer type relative to the general population. Such conditions could also be familial in nature.

Nature of genetic changes that can inactivate tumor suppressor genes or activate putative oncogenes.

Tumor suppressor gene inactivation could occur by point mutations or larger chromosomal deletions, rearrangements involving reciprocal or non-reciprocal recombination or changes in the epigenetic “punctuation” of the genome. Such

changes could be driven by spontaneous processes, exposure to exogenous chemicals or radiation or by viruses.

The change(s) required for tumor promotion remain unknown. Calculations for colorectal cancer risk (Kini et al., 2011, in Appendix) indicate that the requirement of a single genetic event would occur at the rate of a typical gene inactivation process, such as loss of an active *APC* colorectal tumor suppressor allele. However, the small set of amino acid substitutions that activate known oncogenes would be expected to occur at rates approximately 1/100 of gene inactivation rates raising the possibility that oncogene activation is not a part of colorectal cancer promotion (Sudo et al. 2008) .

The continuing search for genes that carry risk for common forms of cancer.

Modern biology has so far failed to uncover genes that carry risk for common cancers, vascular diseases, diabetes and a host of nonlethal conditions. This failure may in large part be attributed to the misapplication of the methods of linkage disequilibria. These methods have been successfully applied to several thousand rare diseases in small families because risk in each family was generally represented by a single gene and allele. However for common diseases in which risk would be expected to be carried by multiple alleles in any gene and possibly by multiple mutations in several different genes the methods of linkage disequilibrium are inapplicable (Morgenthaler & Thilly 2007).

A general approach to studying the genetics of persons with and without common cancers was derived to use advanced technology for scanning all known human genes to discover which, if any, carried inherited mutations that affected cancer risk (Tomita-Mitchell et al. 1998). This same technology, cycling temperature denaturing capillary electrophoresis (CDCE) is also applicable to the enumeration of point mutations as a function of DNA sequence in human blood samples and organs (Ekstrom et al. 2008). These technological advances and demonstrations are cited here, as they will be the basis of the suggested future research exploring familial risks of cancer.

Quantitative cascade models.

Despite the many unknowns in cancer development and etiology, the last ~60 years have seen development of quantitative cascade models. These models describe the occurrence of the necessary tumor initiation events (mutations), tumor promotion events and their physiological sequelae, growth of preneoplastic lesions and growth of lethal neoplasias, respectively (Moolgavkar & A G Knudson 1981; Herrero-Jimenez et al. 2000)(Alfred G Knudson 2001)(Meza et al. 2005). In light of recent discoveries about stem cells and limitations of initiation mutations to fetal/juvenile stem cells an adjusted model of carcinogenesis developed principally at MIT and EPFL groups led by Professors W. Thilly and S. Morgenthaler in which Dr. Pablo Herrero-Jimenez and later Mr. Lohith Kini both at MIT should be cited for seminal contributions.(Herrero-Jimenez et al. 2000)(Kini et al. Appendix I)

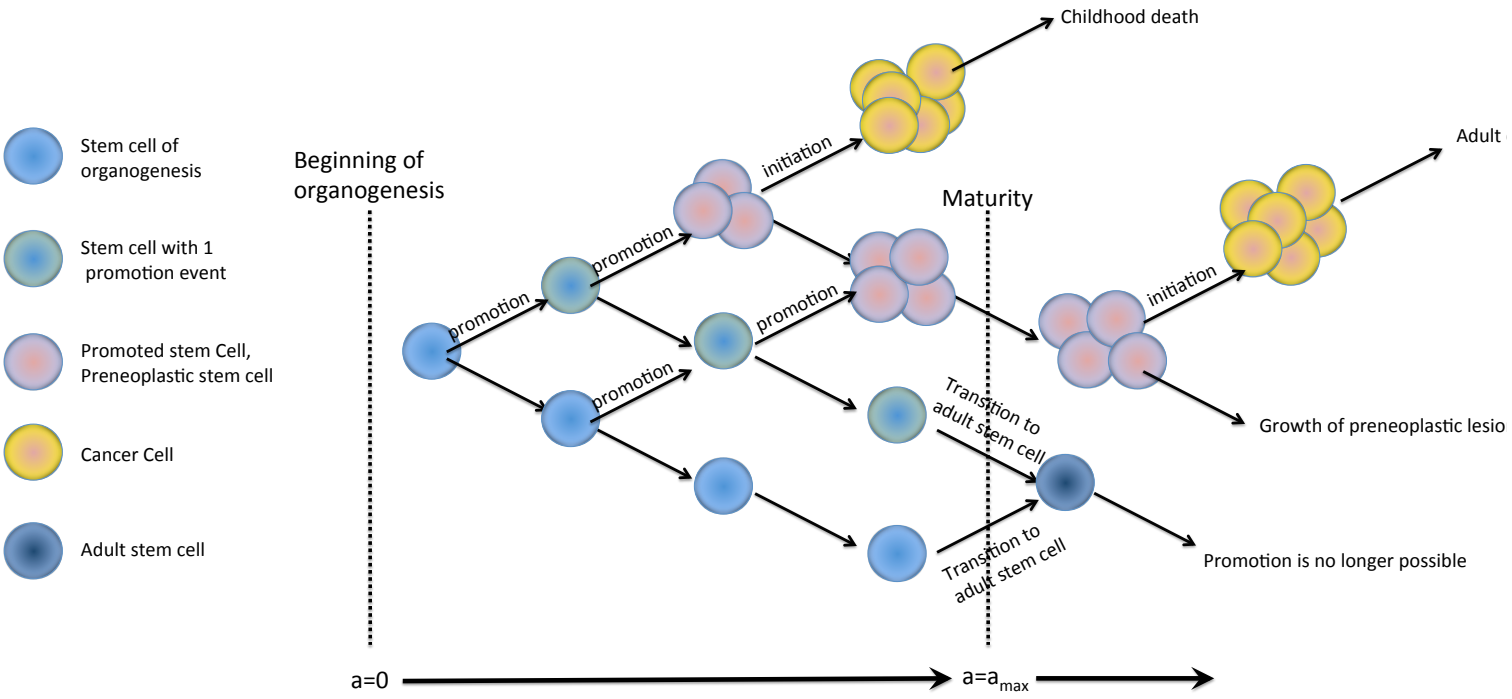

Figure 5: General idea behind the model of 2-stage carcinogenesis. Stem cells of organogenesis start dividing at the beginning of organogenesis and keep on dividing until maturity is reached. After maturity, stem cells of organogenesis become adult stem cells that can no longer undergo initiation. There are a_{max} divisions from the beginning of organogenesis until maturity of the organ. Stem cells that undergo 2 promotion events become preneoplastic stem cells that give rise to preneoplastic lesions. These preneoplastic lesions can be initiated both before and after maturity. The initiation of a preneoplastic stem cell gives rise to a tumor that ultimately kills the patient.

Their general model presently posits "n" required initiating events with rates per organogenic stem cell doubling of $R_{i,}$ and "m" promotion events with rates per preneoplastic stem cell doubling of R_A.

Organogenic stem cells are formally modeled as a binomially expanding population increasing from 1 to 2^{a}_{max} stem cells at the end of the juvenile period (maturity).
The growth rate per year of preneoplastic stem cells in mature organs is represented as the parameter "μ".

Mortality rate for a particular form of cancer, for the period from maturity to extreme old age, at age "t", can be described using $V_{OBS}(t)$: the average expected sum of lethal events in an individual in any year "t" for t > 16.5 yrs.

$$V_{OBS}(t) = nR_i^n \int_{a=0}^{a_{max}} a^{n-1} 2^a \frac{d(1 - e^{-mR_A^m (g-a)^{m-1} 2^{g-a}})}{d(g-a)}$$ Equation 1

Where "g" is the age of the individual in terms of total stem cell doublings since the earliest fetal period, g = a = 0, to the age of individuals observed for cancer death rate, where $g = \mu (t - 19) + a_{max}$.

In this treatment it is assumed that initiated organogenic stem cells grow after maturity as preneoplastic lesions and no other common form of mortality shares the same risk factors as the form of cancer being observed. In the case that such preneoplastic lesions grow and can undergo promotion in only a fraction of the population, that fraction at risk of promotion is designated as "F_{prom}". If there is a competing form of mortality sharing the risk factors with the observed cancer, the fraction of all deaths among the persons dying from these risk factors in an age interval is represented as the constant "f". Using these population risk factors the function $V_{OBS}(t)$ can be transformed into the expected number of lethal events at age "t" given that "F_{prom}" and/or "f" are less than one.

$$r_{OBS}(t) = \frac{F_{prom}\left(1 - e^{-V_{OBS}(t)}\right)}{-\frac{1}{f}\int_0^t V_{OBS}(a)\,da}$$ Equation 2

This form of cancer cascade model is an approximation ignores the probability that there may be multiple independent pathways of initiation and/or promotion but does provide a platform in which each of the biological variables, initiation mutation rates, promotion event rates and preneoplastic growth rates may be considered along with population variables in terms of exposure to environmental agents and/or inherited conditions of risk. (Details and derivation are in Kini et al., 2011 in preparation, APPENDIX I)

Applied to comparison of the age-specific rates of a specified form of cancer, the observed age-specific mortality rates of any two defined populations, OBS_1 (t) and OBS_2(t) can be defined over intervals 15-19 through 100-104 as the RISK $RATIO_{1,2}$ (t) = OBS_1(t) / OBS_2(t). In this thesis comparison will be between a "familial" population, OBS_1 (t), defined as the children of parents who have died of a particular form of cancer in Sweden,1958-2008, and the general population, OBS_2(t) comprised of all persons born in Sweden, 1933-2008.

Changing cancer risks among immigrant populations

Population studies distribution of common cancers has been used to try to identify the major determinants of cancer risk. In such studies, cancer incidence (i.e diagnosis) and/or mortality are compared between different populations. For example, comparing the general population and people with an affected siblings (especially fraternal and identical twins) or parents (A Altieri & K Hemminki 2007; Goldin et al. 2005; Kari Hemminki & Czene 2002; Kari Hemminki & Xinjun Li 2004a; Lichtenstein et al. 2000; Couto & K Hemminki 2007).

One of the earliest population studies revealed that daughters of Japanese immigrants to the United States born and raised in America had the high breast cancer mortality of other North Americans while young Japanese women emigrating to America showed the low breast cancer mortality of the native Japanese population in Japan (BUELL & DUNN 1965; Thomas & Karagas 1987). The data established a basis for believing that some unknown environmental factor was acting in the

fetal/juvenile period to define risk for breast, and by analogy, other cancers later in life.

More recently the Hemminki group at DKFZ in Heidelberg has extended these observations to the children of Swedish immigrants from countries with different patterns of common adult cancers. These studies have found for gastric cancer (Mousavi et al. 2011) that the country of fetal/juvenile development defines the adult risk. Similar results were shown for other types of cancer as well (K Hemminki 2002) (Kari Hemminki et al. 2002). Curiously, the effects appear to be organ specific, a key point in thinking about cancer etiology. The mechanism of increased or decreased cancer risk in immigrant children as adults has not been defined but as in the case of breast cancer in Japanese-American female children a clear case for an undefined environmental factor has been made.

These immigrant population studies are equally important for revealing that immigration of mature individuals does not change the pattern of age-specific risk among cancer sites. These data alone support a conclusion that environmental factors that differ among countries do not effect the apparent growth rate of preneoplastic lesions, μ, the rate of an event(s), R_A, required for promotion of a preneoplastic to a neoplastic stem cell, the fraction of persons in whom preneoplastic lesions of a particular kind grow, F_{prom}, or the degree of competition among causes of death that share risks with the observed form of cancer.

Cancer risks among spouses with long cohabitation in Sweden

The Hemminki group has also studied the coincidence of cancer rates among Swedish couples with at least thirty years of cohabitation and found that there is no evidence of shared spousal risk indicating that within Swedish society the adult environment does not contribute to risk via the same set of parameters found in adult immigrants to Sweden. It appears from these different forms of population studies in Sweden that environmental risk factors act only in the fetal/juvenile period.

Population studies of common, late-onset cancers reveal both environmental and inherited forms of risk

The idea that certain cancer types run in families was noted in antiquity but remains a matter of contention in modern epidemiology. In 1915 after exhaustive study of available data for the United States the actuarial epidemiologist Hoffman was unable to conclude that familial cancer rates differed from those in the general population and concluded: "With regards to heredity and family history, some additional observations reemphasize earlier conclusions that the available evidence in this respect is in the negative" (Hoffman 1915). On reflection we can now see that at that time the numbers of families with two generations living past 65 years of age, when cancer rates are highest, was small and that for many major forms of cancer the age-specific rates were increasing with each succeeding birth cohort from ~1850 through 1900. The general impression of familial cancer risks would have been derived from those few families of longer-lived individuals dying of the same form of cancers and the even fewer families in whom deaths of young adults spanned several generations, e.g. familial early-onset cancers.

Since 1997 the group headed by Professor Kari Hemminki, first at the Karolinska Institute and later at the German Center for Cancer Research, has been organizing and analyzing the general and familial cancer incidence and mortality data available in Sweden since 1958 with the purpose of discovering if a familial risk exists for the most common cancers (Kari Hemminki & Bowang Chen 2004c; K Hemminki et al. 2007a; Ji & Kari Hemminki 2006; K Hemminki & X Li 2004b; Lorenzo Bermejo & Kari Hemminki 2005; Kari Hemminki & Bowang Chen 2005b; Kari Hemminki & Xinjun Li 2004c; Kari Hemminki et al. 2007b; Bermejo & Kari Hemminki 2005; Kari Hemminki & Bowang Chen 2004b; K Hemminki & B Chen 2004d; Kari Hemminki & Xinjun Li 2004a; Kari Hemminki & Czene 2002; Kari Hemminki & Bowang Chen 2004a; Kari Hemminki & Bowang Chen 2005a; Andrea Altieri et al. 2005; Hiripi et al. 2009). When the fraction of persons in the general population diagnosed with a cancer in maturity, 30-75 yrs, in the general population was compared to that fraction of persons in whom either parent had been diagnosed with the identical form of cancer, the ratio (standard incidence rate, SIR.) was found to be significantly greater than unity for all cancers. In this case the parents and

children were predominately of Swedish genetic heritage and had been raised in a relatively homogeneous society. The increased familial risk could thus be attributed to inherited traits, environmental factors or combinations of both. Computations of the standard incidence rate (SIR) between the population with a familial history of cancer and the general population was found to be significantly greater than 1 for all common cancers. This means there is an underlying familiar factor which increases the risk. Such a factor could be inherited, environmental or a combination of both. It is by the kind permission and supervision of Professor Hemminki that the current study was made possible.

Problem of historical changes in age specific cancer mortality

A problem arises in comparing cancer mortality rates among persons born at different periods of history. Inspection of increases in prostate cancer among European American males born 1840-49 to those born 1920-29 shows up to a five-fold increase among age of death intervals 60-64 through 95-99. A direct comparison of fathers' and sons' prostate cancer mortality rates for these birth cohorts would be performed under a condition in which "fathers" would perforce have lower age-specific mortality rates than "sons". Hemminki, based on prostate cancer incidence data 1958-2008, reported about a two-fold familial prostate cancer risk. The interaction of two phenomena, familial risk and changing historical age-specific risk could pose a difficult analytical problem.

Fortunately, for the cancers examined herein, the age-specific mortality rates 50-54 through 70-74 years were essentially invariant in the Swedish population 1958-2008 obviating the problem presented if mortality rates were changing.

Problem of geographical change

Cancer rates at specific sites differ between various countries. The existence of variations in cancer rates within a country would introduce a bias in the study of

familial risk. A familial risk could be due to offspring having a higher chance of living in the same region as the one in which their parents grew up.

A study by Vatland has shown that there are no variation in cancer rates in the United States other than the one expected by chance (Vatland 2001). In this study, the author compared the distribution of cancer rates between the communities of the 6 largest states in the US. The conclusion was that all variations could be explained by chance alone and therefore there is no environmental factor affecting cancer rates varying between the different communities in the US.

A study of such completeness has not been performed in Sweden. Data from the family cancer database does not allow such a complete study to be performed. However, regional cancer rates and mortality data can be found on the Swedish Health and Welfare Statistical Databases. This database is freely accessible online (http://192.137.163.40/epcfs/FisFrameSet.asp?FHStart=ja&W=1440&H=900). The inspection of the data does not reveal any major variations between the regions for each age-group. Combining the information from Vatland's work and the Swedish database supports the hypothesis that there is no internal variation in cancer rates within Sweden.

DATA: Definitions and treatment.

Mortality vs. incidence data.

A problem has arisen in recent years as new methods of detecting what are believed to be early forms of cancer have been reported as "cancers". Both breast and prostate cancer diagnoses have been affected by this phenomenon such that the number of false positive diagnoses has risen to half or more than half of all diagnoses. For this reason the data analyzed in this thesis is restricted to diagnosis at the time of death.

This does not mean that the diagnosis at death is accepted here as wholly accurate. It means only that of two imperfect forms of data the one with the lesser degree of bias has been chosen. The rationale of this choice is presented below using Nordcan, a database incorporating the cancer registries of all Nordic countries, publicly available at http://www-dep.iarc.fr/NORDCAN/ (Engholm et al. 2010).

The data for incidence and mortality of breast cancer (Figures 6 and 7, of prostate cancer (Figures 8 and 9) and of colorectal cancer in males (Figures 10 and 11) and females (Figures 12 and 13) permit comparison of these two means of expressing the population experience of cancer.

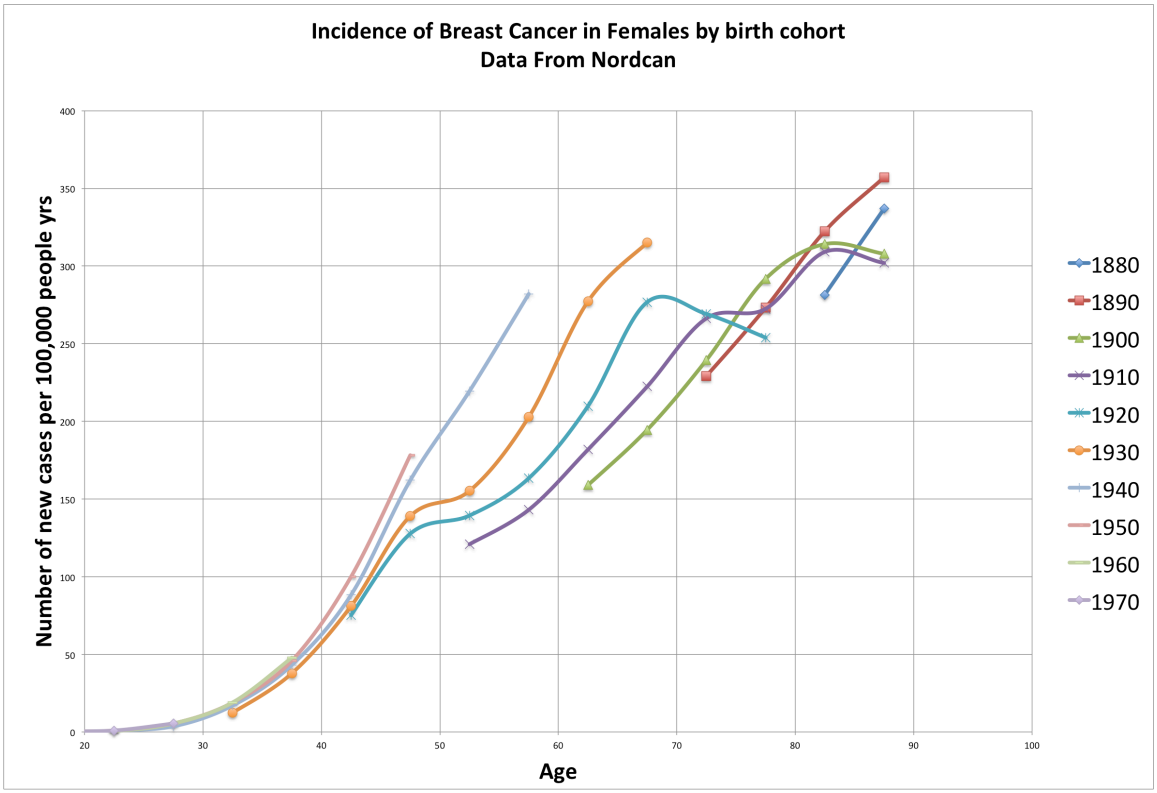

Figure 6: Age-specific incidence rates of breast cancer in Sweden for each birth cohort for which at least 2 five-year age intervals were available.

Breast cancer incidence (Figure 6) was recorded as rising significantly in birth decade cohorts from 1880-89 through 1940-49. The recorded incidence in women 55-59 years of age rose from ~140 x 10^{-5} in the 1910-19 cohort to ~280 x 10^{-5} in the 1940-49 cohort. In contradistinction, mortality (Figure 7) changes are small or undetectable in these same birth decades. The recorded mortality in women 55-59 years of age was ~60 x 10^{-5} in the 1910-19 cohort and also ~60 x 10^{-5} in the 1940-49 cohort. Reference to the American mortality data [http://mortalityanalysis.mit.edu] shows a near constant breast cancer mortality rate for adult women born in the late 19th through the 20th centuries. The near constant mortality rates suggest small progress in treatment of this cancer and the increasing incidence suggests a general rise in false positive diagnoses as physicians sought to begin treatments of breast cancer with minimum evidence that lesions detected were pre-cancerous or

cancerous. It is clear that familial breast cancer risk would better be estimated using mortality data.

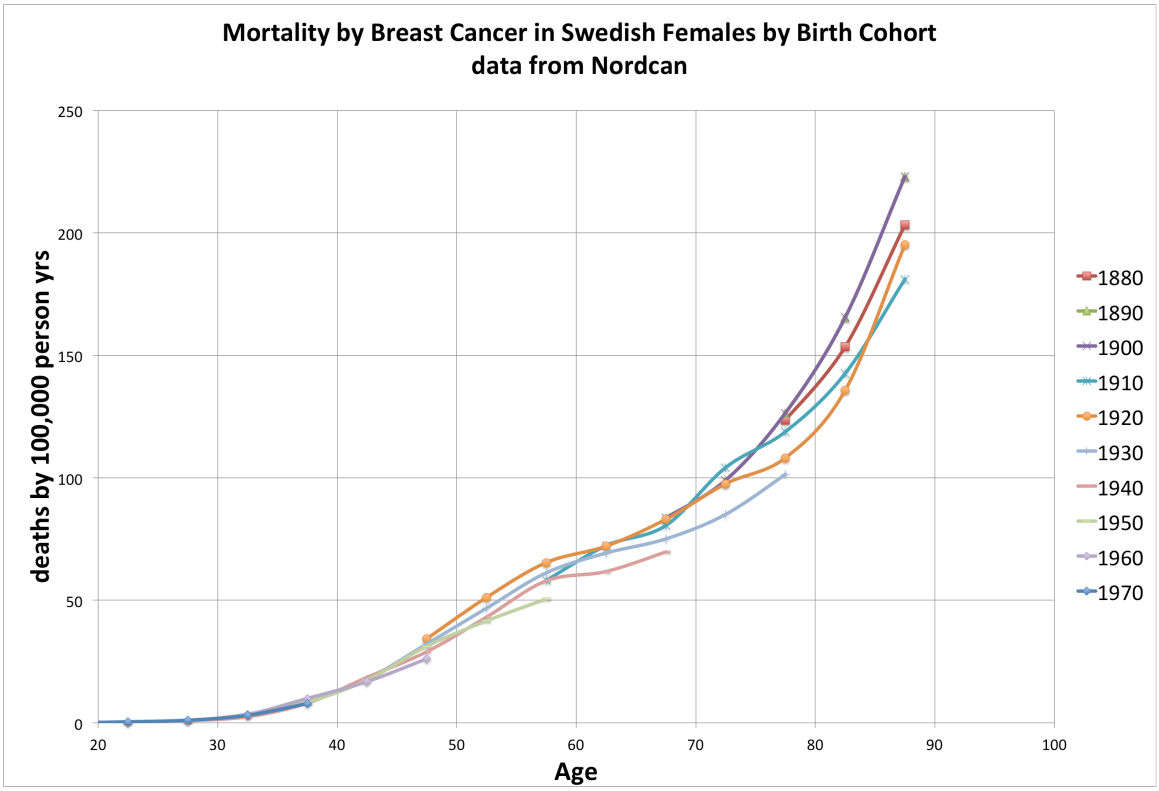

Figure 7: Age-specific mortality rates of breast cancer in Sweden for each birth cohort for which at least 2 five-year age intervals were available.

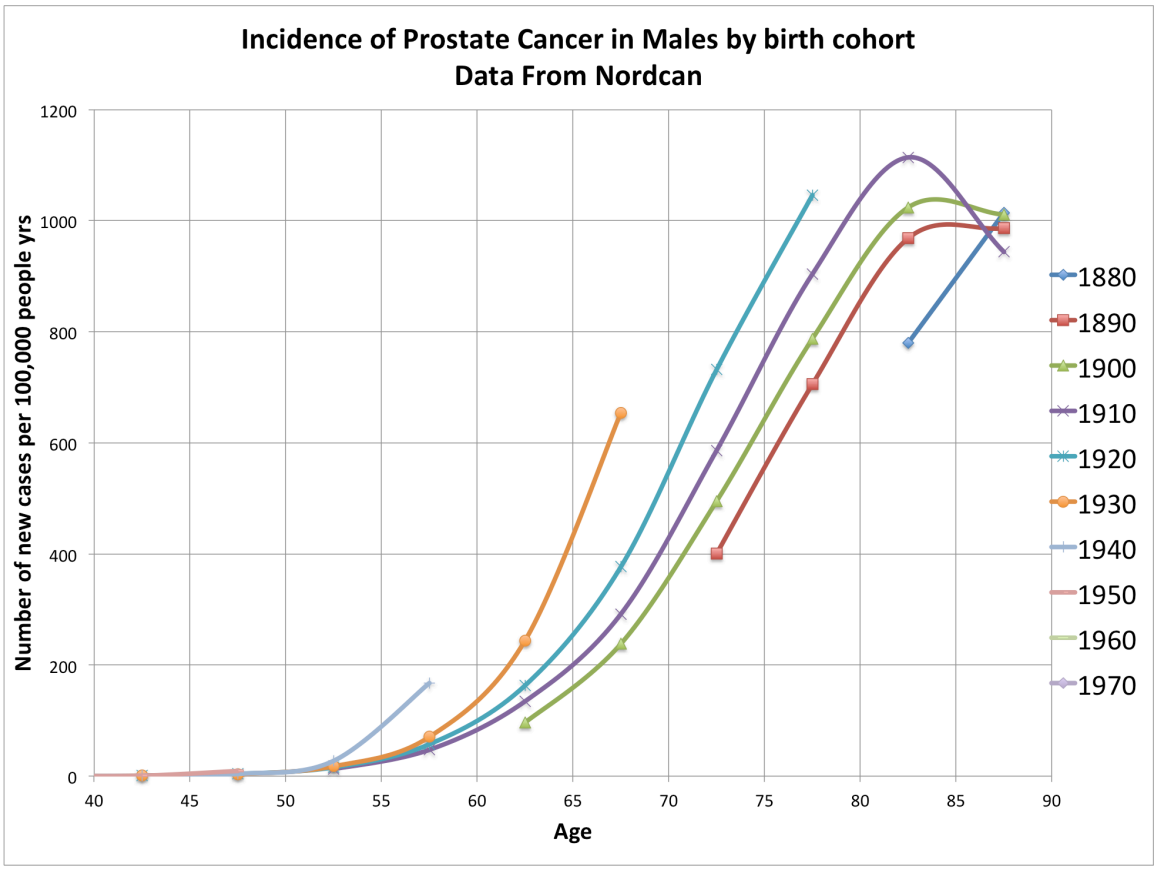

Figure 8Age-specific incidence rates of prostate cancer in Sweden for each birth cohort for which at least 2 five-year age intervals were available.

Prostate cancer incidence (Figure 8) was recorded as rising significantly in birth decade cohorts from 1880-89 through 1940-49. The recorded incidence in men 60-64 years of age rose from ~100 x 10^{-5} in the 1900-09 cohort to ~220 x 10^{-5} in the 1930-39 cohort. In contradistinction, mortality (Figure 9) changes are small or undetectable in these same birth decades. The recorded incidence in men 60-64 years of age was ~40 x 10^{-5} in the 1900-09 cohort and also ~40 x 10^{-5} in the 1930-39 cohort. Reference to the American mortality data [http://mortalityanalysis.mit.edu] shows a near constant prostate cancer mortality rate for adult men born in the 20th century. The near constant mortality rates suggest small progress in treatment of this cancer and the increasing incidence suggests a general rise in false positive diagnoses as physicians sought to begin treatments of prostate cancer with minimum evidence that lesions detected were pre-cancerous or cancerous. It seems clear that familial prostate cancer risk would better be studied using mortality data.

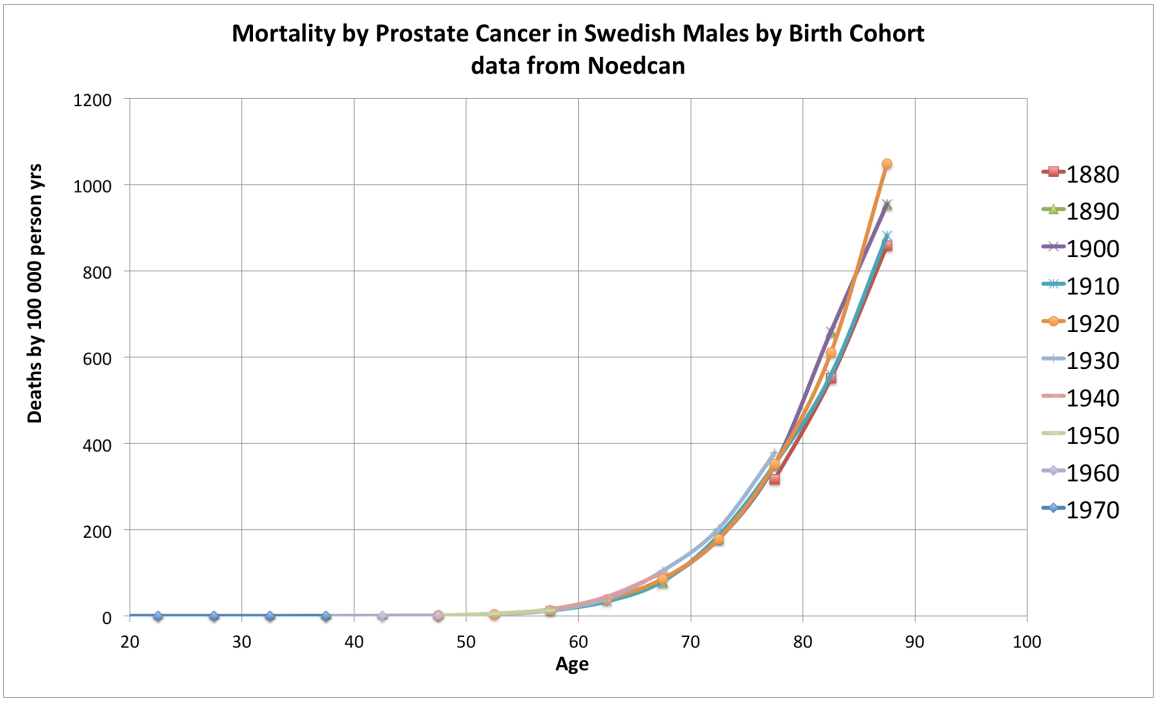

Figure 9: Age-specific mortality rates of prostate cancer in Sweden for each birth cohort for which at least 2 five-year age intervals were available.

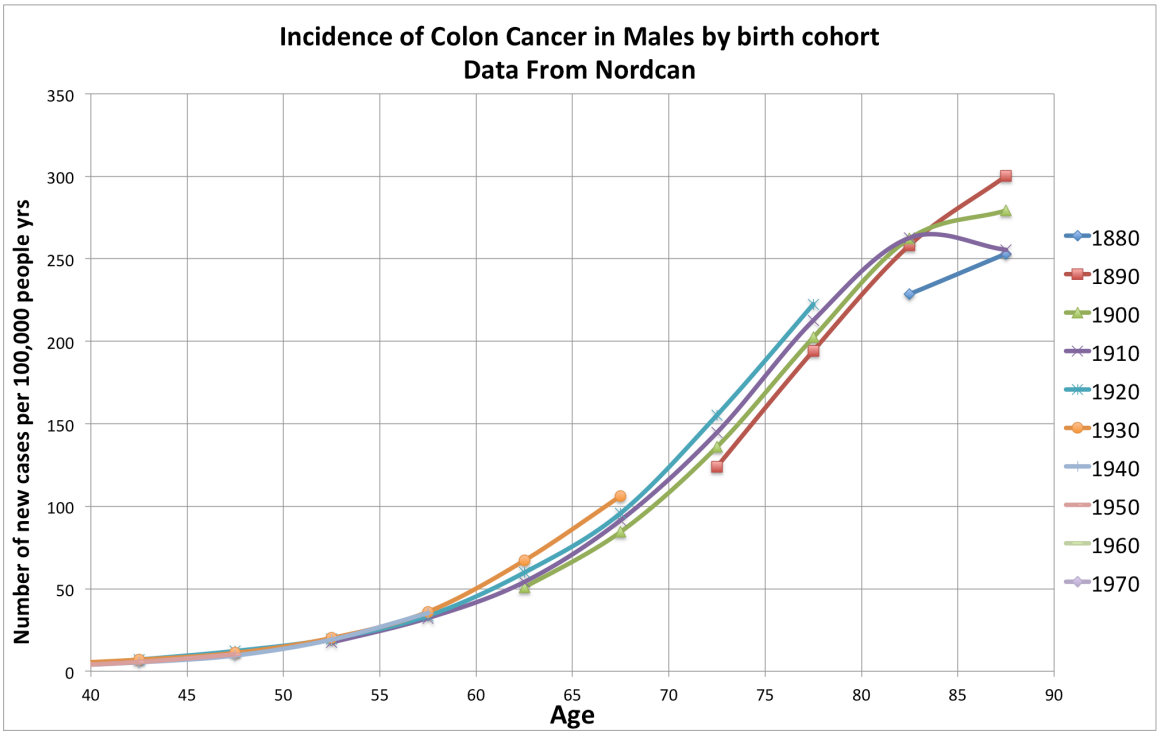

Figure 10. Age-specific incidence rates of colon cancer in Swedish males for each birth cohort for which at least 2 five-year age intervals were available.

Colon cancer incidence in males (Figure 10) was recorded as rising significantly in birth decade cohorts from 1880-89 through 1930-39 but to a lesser degree than recorded for prostate cancer. The recorded incidence in men 60-64 years of age rose from ~50 x 10^{-5} in the 1900-09 cohort to ~70 x 10^{-5} in the 1930-39 cohort. In contradistinction, mortality (Figure 11) changes are small or undetectable in these same birth decades up to age intervals 70-74. The recorded incidence in men 60-64 years of age was ~40 x 10^{-5} in the 1900-09 cohort and also ~40 x 10^{-5} in the 1930-39 cohort. The near constant mortality rates suggest small progress in treatment of this cancer and the increasing incidence suggests a general rise in false positive diagnoses as physicians sought to begin treatments of prostate cancer with minimum evidence that lesions detected were pre-cancerous or cancerous. It seems clear that familial male colon cancer risk would better be studied using mortality data.

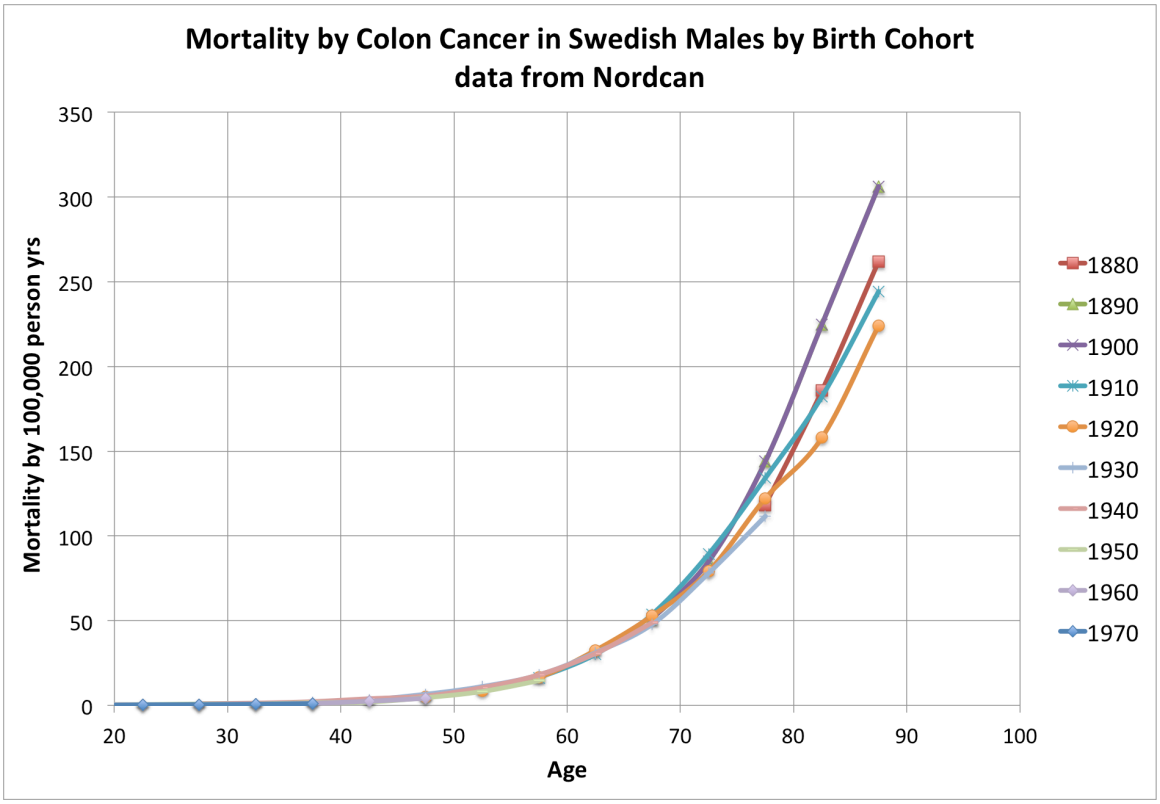

Figure 11. Age-specific mortality rates of colon cancer in Swedish males for each birth cohort for which at least 2 five-year age intervals were available.

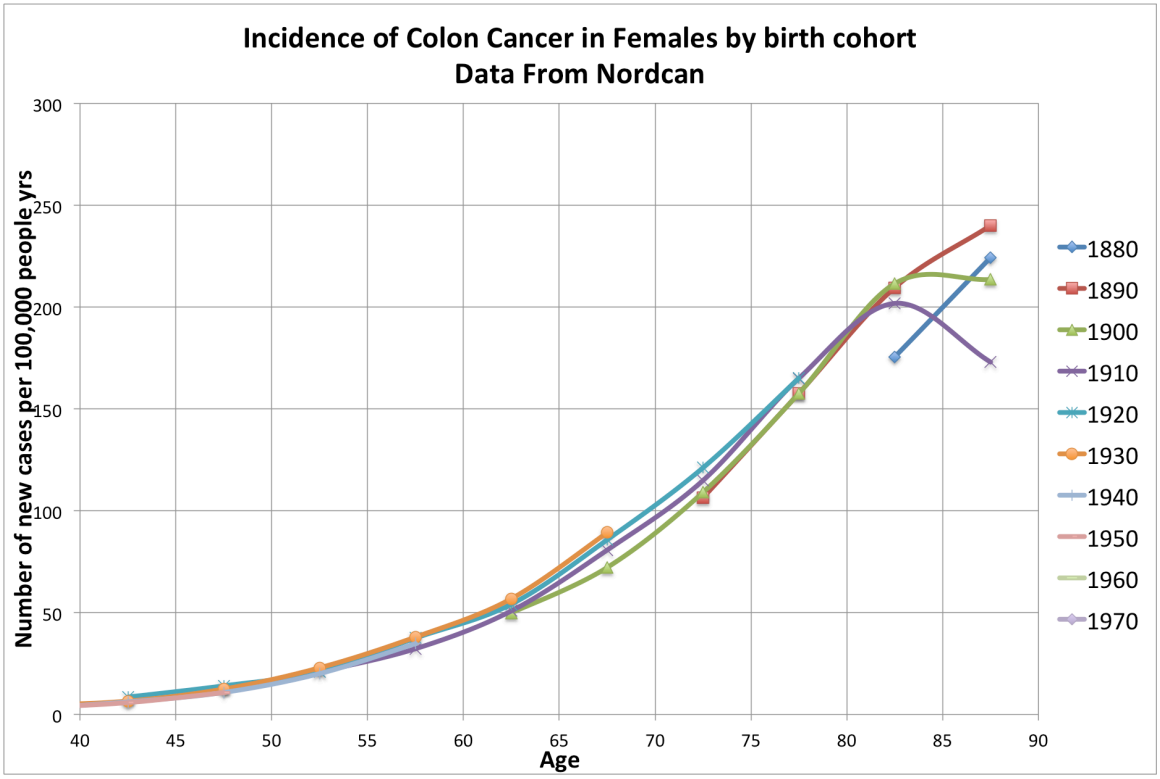

Figure 12. Age-specific incidence rates of colon cancer in Swedish females for each birth cohort for which at least 2 five-year age intervals were available.

Colon cancer incidence in females (Figure 12) was recorded as rising slightly in birth decade cohorts from 1900-09 through 1930-39. This rise is much smaller than the one observed for breast cancer. The recorded incidence in women 65-69 years of age rose from ~70 x 10^{-5} in the 1900-09 cohort to ~90 x 10^{-5} in the 1930-39 cohort. Mortality rates (Figure 13) changes are small or undetectable in these same birth decades up to age intervals 65-69 but show evidence of decrease in older age-intervals in successive birth decade cohorts. The recorded incidence in women 65-69 years of age was ~40 x 10^{-5} in the 1900-09 cohort and also ~40 x 10^{-5} in the 1930-39 cohort. As was the case for male colon cancer it seems clear that familial female colon cancer risk would better be studied using mortality data. That remains true even if, unlike breast and prostate cancers, colon cancer mortality data of females contain evidence that medical intervention has reduced mortality rates in males and females in Sweden in the late 20th century.

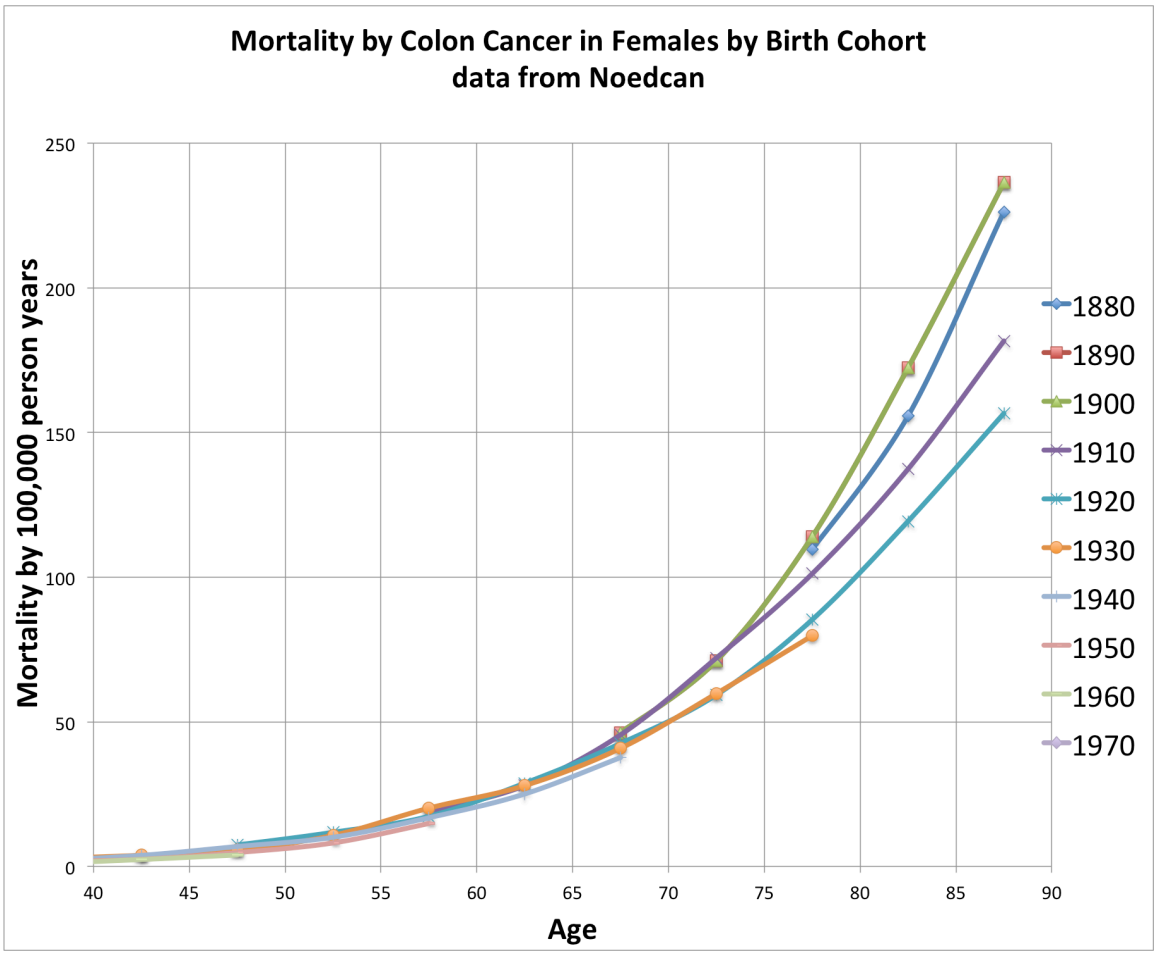

Figure 13. Age-specific mortality rates of colon cancer in Swedish females for each birth cohort for which at least 2 five-year age intervals were available.

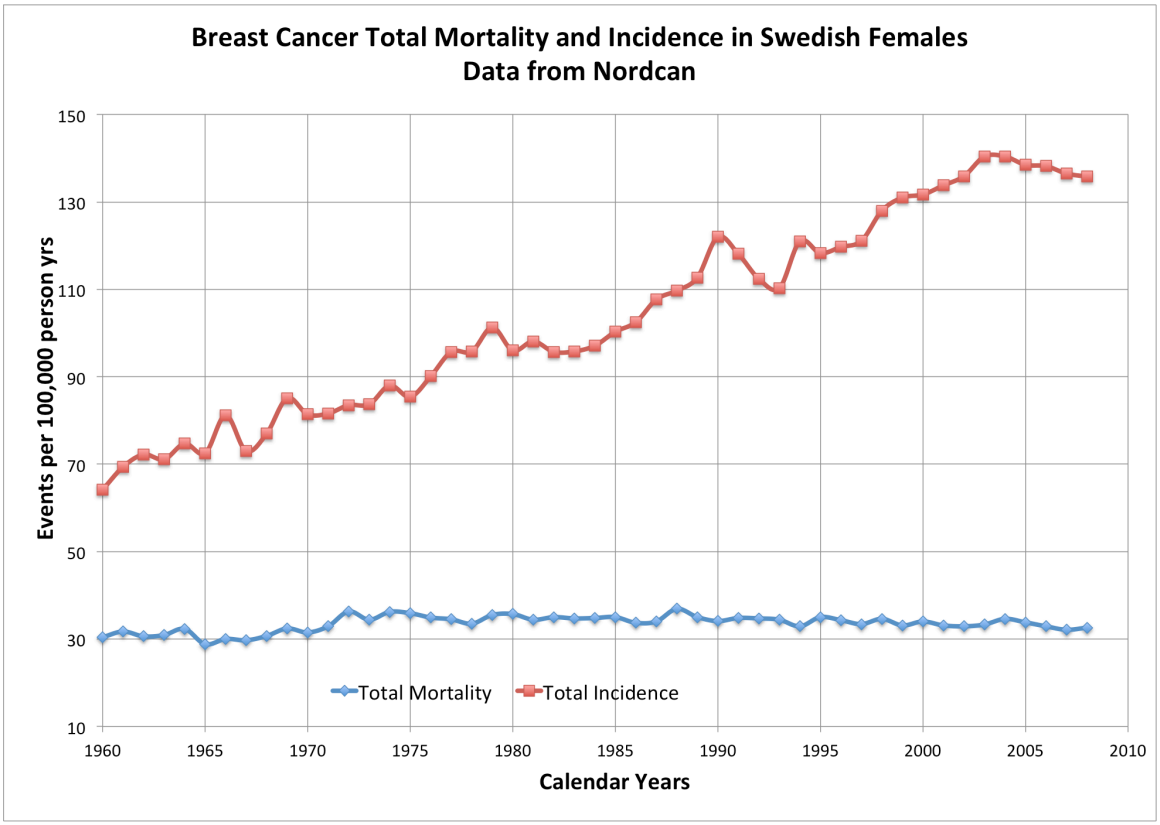

Figure 14: Incidence and mortality for breast cancer in Sweden from 1960 to 2008 averaged over all age intervals.

The basic choice to be made between incidence and mortality is illustrated by Figure 14 which shows the incidence and mortality for breast cancer in Sweden, 1958-2008. It is simply not credible to imagine that a constant rise in cancer incidence was matched each year by an offsetting increase in therapeutic efficacy. The use of mortality data has, however, the drawback of not taking into account improvements in medical practices as has been observed for colon cancer. Recognizing this limitation, especially in undercounting the persons actually with a form of cancer because they do not die from it, mortality data are used throughout the remainder of this thesis.

Source and presentation

In 1997 the group led by prof. Hemminki organized the family cancer database (FCD) in Sweden. They linked the administrative family registry of all Sweden since 1932 with the Swedish Cancer Registry (K Hemminki et al. 2001). The family registry is based on its assignment of identification numbers to all residents beginning in 1933. In 1958, the Swedish government instituted a clinical record of cancer diagnoses within its health care system and simultaneously registered the primary cause of death for each resident. Therefore, the cancer status of every individual that died after 1958 is recorded. The group of prof. Hemminki had the insight of linking the 2 databases and organizing the data for research purpose. The result is the FCD. This is a dataset unique in its kind, for both its completeness and size. In Appendix II are provided the raw data for population size and deaths from the three forms of cancer treated herein such that the number of deaths recorded 1958-2008 for each five-year age interval are made available for inspection.

Within these data, those individuals born 1933-2008 serve as the general cohort of children. Each recorded death has been matched to the death records of parents to define the familial cohort consisting of the children who died of a specific form of cancer with at least one parent who died of the same form of cancer in the same 1958-2008 period. Both the general and the familial population are from the same generation and contain the same age intervals, from 0-4 through 70-74 yrs.

The total number of deaths recorded for each death age group is as follows:

Age 2Group	Number of primary deaths by breast cancer: general cohort	Number of primary deaths by breast cancer: familial cohort
2.5	0	0
7.5	0	0
12.5	1	1
17.5	2	0
22.5	1	0

27.5	6	2
32.5	13	2
37.5	49	10
42.5	132	20
47.5	347	41
52.5	571	46
57.5	842	69
62.5	759	35
67.5	539	30
72.5	252	17
SUM	3442	258

Table 1: Total number of primary deaths by breast cancer in females during each 5 year interval. The familial cohort is the set of women whose mother died of breast cancer.

Age Group	Number of primary deaths by prostate cancer: general cohort	Number of primary deaths by prostate cancer: familial cohort
2.5	0	0
7.5	0	0
12.5	0	0
17.5	0	0
22.5	0	0
27.5	0	0
32.5	0	0
37.5	0	0
42.5	2	0
47.5	20	3
52.5	134	13
57.5	399	37
62.5	766	65
67.5	930	71

72.5	625	47
Total	2876	236

Table 2: Total number of primary deaths by prostate cancer during each 5 year interval. The familial cohort is the set of men whose father died of prostate cancer.

Age Group	Number of primary deaths by colon cancer in males, general cohort	Number of primary deaths by colon cancer in males, familial cohort	Number of primary deaths by colon cancer in females, general cohort	Number of primary deaths by colon cancer in females, familial cohort
2.5	0	0	0	0
7.5	0	0	0	0
12.5	0	0	0	0
17.5	1	0	0	0

22.5	2	0	0	0
27.5	7	0	3	2
32.5	3	0	2	0
37.5	24	0	18	1
42.5	52	2	38	5
47.5	95	10	104	7
52.5	219	14	211	16
57.5	382	23	379	24
62.5	538	29	476	19
67.5	496	30	417	24
72.5	271	11	221	8
SUM	2090	119	1869	106

Table 3: Total number of primary deaths by colon cancer during each 5 years interval. The familial population is the set of men or women with at least one parent dead of colon cancer.

These mortality rates are defined for each cohort and age-interval as the number of persons dying within the interval by the population size within that interval. Thus the recorded mortality rate for cohort and age interval in statistical parlance is the maximum likelihood estimator (MLE) for cancer risk. It is obvious by inspection of Tables 1-3 that

The low mortality rates early in life prevent meaningful comparisons of general and familial cohorts. We have set the minimum number of recorded deaths in an interval at 9 before considering the data probative. In Figures 15-18 display all non-zero values in the Swedish records. However the calculations used to define relative familial to general cohort age-specific risk only considers an age group if nine (9) or more deaths are recorded in the age interval in the familial cohort.

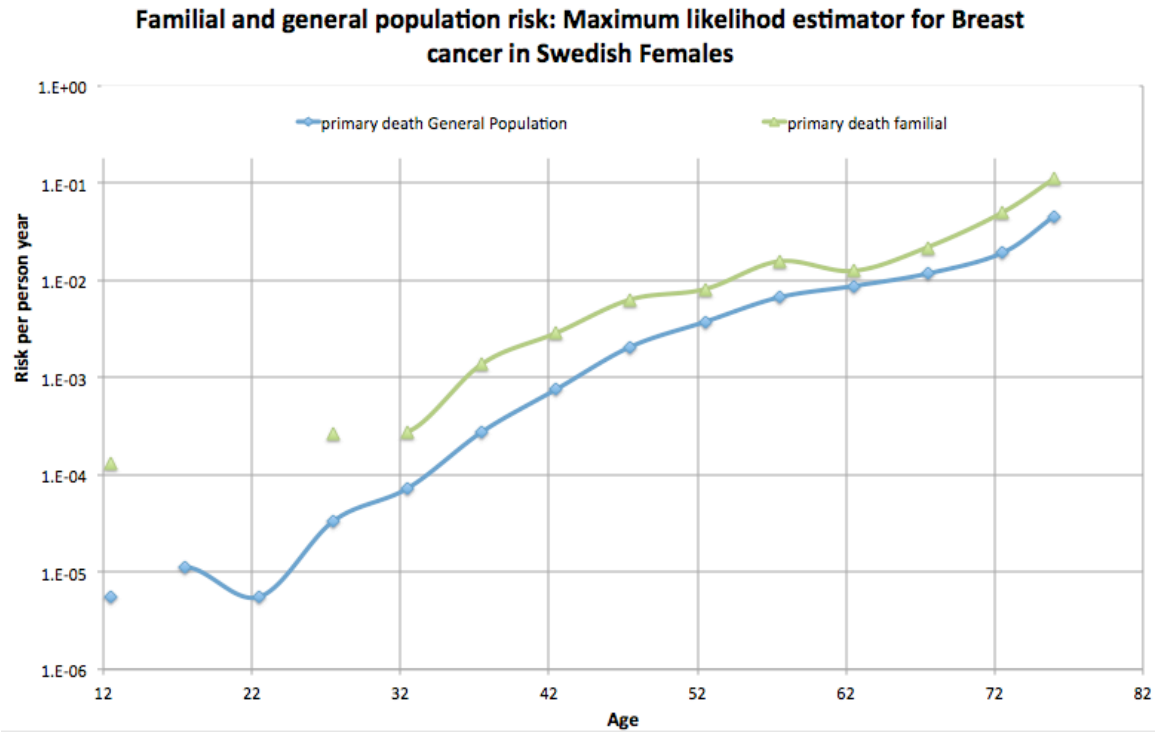

Figure 15: The MLE of risk for breast cancer for both the familial and general cohorts. The probative data points for which there are 9 or more cases are from 35-39 to 70-74

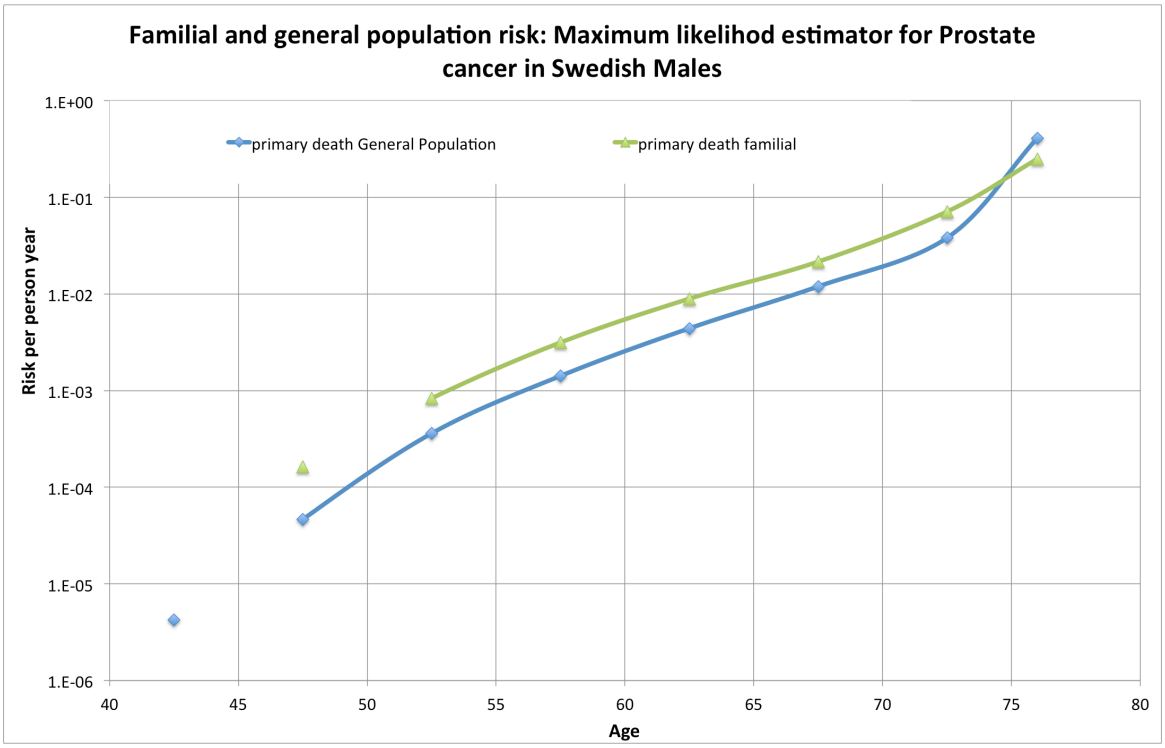

Figure 16: The MLE of risk for prostate cancer for both the familial and general cohorts. The probative data points for which there are 9 or more cases are from 50-54 to 70-74

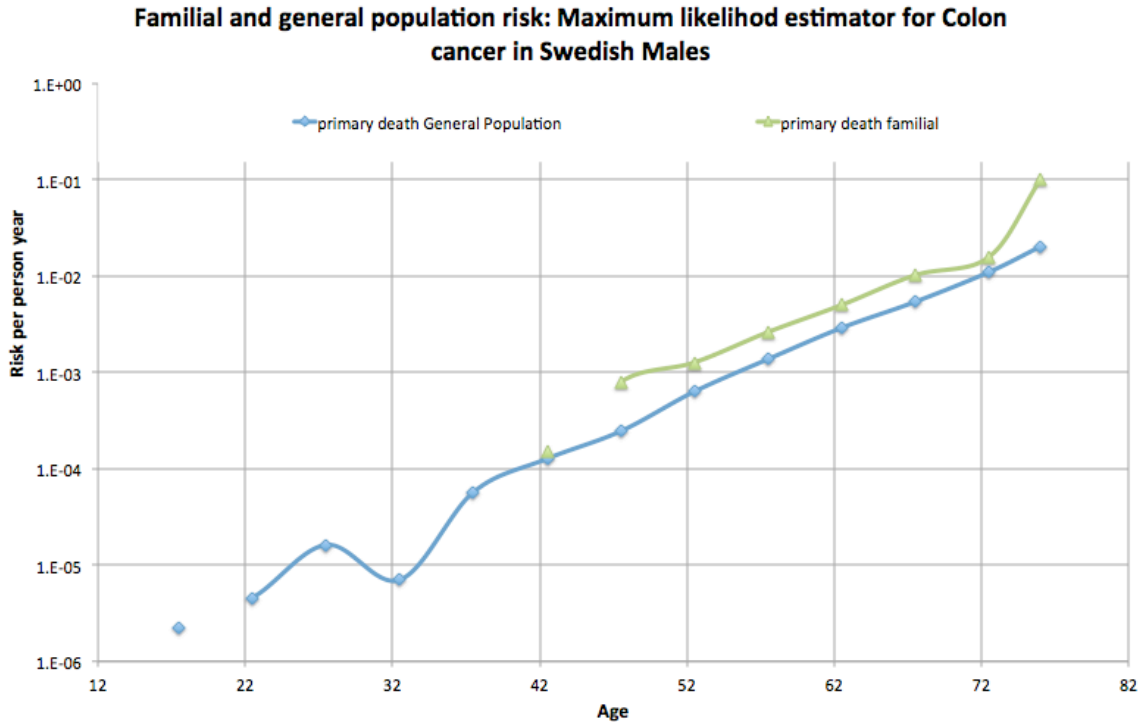

Figure 17: The MLE of risk for colon in males cancer for both the familial and general cohorts. The probative data points for which there are 9 or more cases are from 45-49 to 70-74

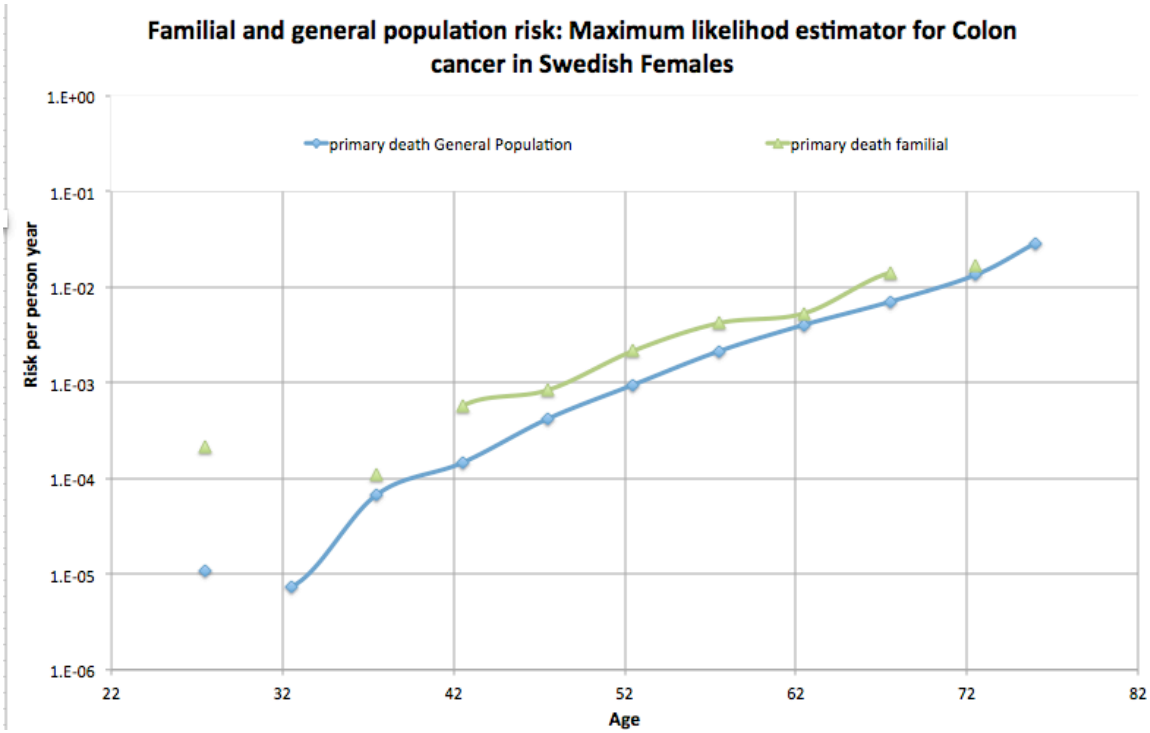

Figure 18: The MLE of risk for colon in females cancer for both the familial and general cohorts. The probative data points for which there are 9 or more cases are from 50-54 to 65-69

Figure 16Figure 18 show that the familial age- and cancer-specific mortality rates are consistently greater than in the general cohorts. In the next section the degree of statistical confidence in this conclusion is determined for each cancer and age interval for which probative data were available. It also appears that the ratio of familial to general cohort mortality rates are similar over the age-intervals examined. This perception is also analyzed blow.

Statistical Methods

Definition of risk

The definition of cancer mortality risk in statistical terms is the chances of success (death) given a trial (one year of human life). This risk is probability that includes 0, 1 and all intervening values.

The risk of dying of cancer is defined as $H \in [0,1]$ (H stands for hazard. R for risk is not used in order not to create confusion with R_i and R_A, the mutation rates).

As a person either dies or does not die of the cancer studied in any year of observation, the probability of the number of deaths among many persons can be expressed as a Binomial distribution.

Binomial distribution: $P_k(k|H)=\binom{N}{k}H^k(1\text{-}H)^{N-k}$

Where H is the risk, N is the number of trials, k the number of successes (deaths) and P_k denotes the probability density function of k.

It is important to note, that the binomial distribution assumes that the risk (probability of a success given one trial) is the same for all trials. If we have the number of trials and the number of successes (deaths), it is possible to calculate the probability distribution of the risk for a given outcome.

According to Bayes theorem: $P_H(H|k)=\frac{P_k(k|H)P_H(H)}{\int P_H(k|H)P_H(H)\,dH}$

This equation has two unknowns. This means that in order to solve it, we need to make some hypothesis about $P_H(H)$. $P_H(H)$ is the *a priori* distribution of the risk, or the distribution of the risk, before any trial is done. A good hypothesis is to assume it is uniform between zero and one. As said before, the risk is only defined between zero and one, and having no real reason why some risks should be more likely than others, it is reasonable to assume it is constant.

Hypothesis: $P_H(H)=1\ \forall\ H\in[0,1]$

$$P_H H(H|k)=\frac{P_k(k|H)}{\int P_k(k|H)\,dH}=\frac{H^k(1\text{-}H)^{N-k}}{\int_0^1 H^k(1\text{-}H)^{N-k}\,dH}$$

Thus :

With this last equation it is now possible to calculate the probability distribution (PDF) of the risk, for each set of trials and successes. Although no analytical solution will be proposed for the integral in the denominator, it's numerical calculation is trivial.

The use of age-specific risk

Having defined risk, it is now possible to define cancer risk. In that purpose, we must first define what is meant by a trial. The chosen definition of a trial is a person year. A person year is one person living one year. The number of successes is then defined as the number of persons dying of cancer within that person year (as explained above mortality will be used instead of incidence). Thus risk is defined as the risk of dying of a specific type of cancer during one year. Defined as such, the maximum likelihood estimator (MLE) for the cancer risk is:

MLE of cancer risk: $$H = \frac{\text{Number\_of\_cases}}{\text{Person\_years}}$$

It is important to note that the risk for cancer changes during the lifetime depending on the age of the individual. The binomial approximation assumes the risk to be constant over all trials. Therefore, for the risk has to be defined as age-specific. The smaller the age-interval, the greater the validity of the assumption of constant risk. However, if the age-intervals are made to be too small, the number of cases within each will be to small to be studied. Taking into account this dilemma, 5 year age groups are chosen. The risk will be assumed constant over 5 year periods. This remains an approximation. However, reducing the width of the age-groups would reduce too much the numbers available, and thus increase the variance of the risk within an individual age-group.

Testing the Null Hypothesis

Let H_t and H'_t be the general population risk and the familial population risk respectively, for each age group t, $t \in [1, 2 \ldots n]$. These are defined by N_t, k_t, the number of person years within each age interval and the number of cancer cases for the general population. Respectively, we have N'_t and k'_t for the familial population.

We now define a parameter $\gamma_t = H'_t - H_t$

From probability theory, we have that $P_{\gamma_t}(\gamma_t) = P_{H'_t}(H'_t) * P_{-H_t}(-H_t)$ (where the asterisk means convolution and P_γ is the PDF of the parameter γ).

As for every t: $N_t >> N'_t$ and $k_t >> k'_t$ we can say that the PDF of H_t P_{Ht} is narrow compared to the PDF of H'(t), $P_{H'(t)}$. This follows from the observation that the general population is much larger than the population with a familial risk for a specific cancer.

We can therefore assume: $P_{-H_t} \approx \delta_{-H_t}$ (where δ_x is the Dirac function, the neutral element of the convolution shifted by x. δ_x Is an infinitely narrow function with an area of 1)

This means we can write $P_{\gamma_t}(\gamma_t) = P_{H'_t}(H'_t - H_t)$ (Nothing more than the PDF of H'_t shifted by H_t).

Having the PDF of γ for each age group, allows us to test if the risk is higher at each age group. However, we would be interested in knowing if the whole risk is higher in general for the familial population. With that goal, we define:

$$\gamma_T = \sum_{i=1}^{n} \gamma_i$$

Using the same property as before:

$P_{\gamma_T}(\gamma_T) = P_{\gamma_1} * P_{\gamma_2} * ... * P_{\gamma_n} = P_{H_1}(H'-H) * P_{H_2}(H_2'-H_2) * ... * P_{H_n}(H_n'-H_n)$, this means that the PDF of the sum of the risk differences is equal to the convolution of all the PDF of the familial risks shifted by the general population risk.

Having a parameter with known PDF which takes into account the differences in all the age specific risks, we can formulate the null and alternate hypothesis as follows:

H_0: The cancer risk for the familial population is the same as the one for the general population thus $\gamma_T = 0$.

H_1: The cancer risk for the familial population is greater than the risk for the general population thus $\gamma_T > 0$.

As we can calculate the PDF of γ_T, we can write: $P(\gamma_T>0)=\int_0^\infty P_{\gamma_T}(\gamma_T)d\gamma_T$. By definition, the P-value $=1\text{-}P(\gamma_T>0)=P(\gamma_T<0)$.

In the end, for every cancer studied, by looking at the P-value and the PDF of γ_T, it is possible to decide whether or not the evidence is sufficient to reject the null hypothesis.

A second way of testing the null hypothesis will also be performed. This test, assumes that the number of deaths by cancer in a specific age interval to be a random variable following a Poisson distribution. This assumption is in part justified by the fact that we have a certain number of events happening in a definite time period. It is also assumed that for a high enough number of deaths, the Poisson distribution and the Normal distribution are very similar. It is further assumed that the number of person years is a constant with no variance in relation to the number of cancer deaths for each person year. This seems reasonable since the number of person years is so much larger than the number of cancer cases.

Under these assumptions, we can say that the risk follows a normal distribution with a variance equal to the mean. If we further assume that the mean is the observed risk, then we can say that:

$k\sim N(k,\sqrt{k})$ and $k'\sim N(k',\sqrt{k'})$

Consequently $H\sim N(H,\frac{\sqrt{k}}{N})$ and $H'\sim N(H',\frac{\sqrt{k'}}{N'})$

At this point we ask what is the distance between R and R'. The distance is expressed in quantiles, or standard deviations apart. This process will be done for each age group where the number of cases is considered sufficient, that means larger or equal to 10 cases. It can be seen that N is far larger than N'. This means that the standard deviation of R will be much smaller than the standard deviation of R'

For each of these cases we will calculate the distance in quantiles as:

$$Q=\frac{H'-H}{\frac{\sqrt{k'}}{N'}}$$

The null hypothesis will be considered rejected if all the tested age intervals for a single cancer will have a distance between the two risks of three standard deviations.

Calculating the constant ratio between the risks

From Figure 15Figure 18, it can be observed that the ratio between the familial risk and the risk for the general population could be constant. This section will show how the ratio, if constant, can be calculated.

For every age interval, the probability distribution of the familial risk H' has been calculated. R is assumed to be constant relative to H' and will thus not be considered as a random variable. If the ratio is constant:

For each t: $H'_t = aH_t$

Since R_t is a constant, it is reasonable to say that:

$$P(a=A)=\prod^{t} P_{H_t'}(AH_t)$$

For every α:

meaning that the probability of α taking a given value A is the product over all t, of the probabilities of observing $H'_t=AH_t$.

This calculation gives an estimate of the probability of α taking a certain value. The probability distribution of α can be obtained by normalizing the probability function:

$$P_a(a)=\frac{\prod^{t} P_{H_t'}(aH_t)}{\int P(a=A)\,dA}$$

The α chosen will be the expectation of α: E(α).

This method allows for a 3rd test of the null hypothesis. The probability that $\alpha \leq 1$ is the P-value of the null hypothesis under the assumption that the ratio is constant.

Results

Test of the null hypothsis

In this study, the null hypothesis has been tested for the following cancers: breast, prostate, and colon in males and females. The null hypothesis is tested in two different ways.

For every age group, in each of the four cases, an individual P-value has been calculated, to know if $\gamma_t>0$. The result is the following table:

Age group	Prostate	Breast	Colon Male	Colon Female
2.5	6.20E-01	2.63E-01	4.41E-01	3.15E-01
7.5	6.19E-01	2.62E-01	4.40E-01	3.14E-01
12.5	6.17E-01	1.97E-02	4.40E-01	3.14E-01
17.5	6.11E-01	3.63E-01	4.38E-01	3.13E-01
22.5	6.03E-01	2.58E-01	4.34E-01	3.11E-01
27.5	5.94E-01	1.55E-02	5.70E-01	1.13E-02
32.5	5.85E-01	4.81E-02	4.26E-01	3.06E-01
37.5	5.73E-01	2.25E-05	7.44E-01	1.93E-01
42.5	5.54E-01	3.39E-07	3.21E-01	4.33E-03
47.5	4.16E-02	4.73E-10	7.27E-04	3.73E-02
52.5	4.07E-03	1.00E-06	8.85E-03	1.52E-03
57.5	7.88E-06	1.53E-10	2.10E-03	1.01E-03
62.5	1.25E-07	1.69E-02	2.50E-03	9.89E-02
67.5	1.76E-06	5.73E-04	5.23E-04	6.66E-04
72.5	2.25E-05	1.44E-04	8.97E-02	2.07E-01

Table 4: For each age group, we calculated the P-value of the null hypothesis: γ(t) = 0. Each $P-value=\int_0^\infty P_{\gamma(t)}(\gamma)d\gamma$. Every age group represents a five-year interval. For example: the age group 37.5 represents the age interval from 35 to 39 years of age.

For each distribution, the P-value has been calculated.

	Prostate	Breast	Colon Male	Colon Female
P-value	1.30E-10	0	2.03E-09	1.60E-04
Total Cases	236	258	119	103

Table 5: The results from the testing of the null hypothesis. The P-value refer to the results from the first type of test. Pval= $\int_0^\infty \gamma_T P(\gamma_T) d\gamma_T$

These results show that the null hypothesis is rejected for all four cases. In the case of breast cancer, the P-value is so low it cannot be calculated within the floating comma approximation of MATLAB™.

We can now show the results of the second test of the null hypothesis. In this test, only age intervals that are considered probative are considered, i.e. having at least nine (9) deaths recorded in the familial cohorts for any five-year age interval. The results show distance in quantiles between the risk of the familial and general population.

Age Group	Prostate	Breast	Colon Males	Colon Females
37.5		2.53		
42.5		3.30		
47.5		4.31	2.20	
52.5	2.04	3.66	1.88	2.23
57.5	3.34	4.76	2.29	2.41
62.5	4.09	1.79	2.27	1.06
67.5	3.76	2.55	2.58	2.46
72.5	3.19	2.54	1.01	
Average	3.12	3.18	2.04	2.04

Table 6: The results from the second test of the null hypothesis. Each value here represents the distance in quantiles between the familial risk and the risk for the general population. Only age groups with 9 or more familial cases are here considered. The value of each distance

$$Q = \frac{R' - R}{\frac{\sqrt{k'}}{N'}}$$

.

A distance Q =2 means that there is a 95% chance that the 2 values are different. All probative age intervals for prostate cancer have Q values greater than 2. However, one age interval for breast cancer and three among male and female colon cancers have values of Q < 2. Therefore according to this test, it is not possible to reject the null hypothesis for all probative age groups. This is the reason why the first test was performed. It allows for the combination of evidence across all age groups. The results of which are presented in Table 5.

Ratio between the risks

Having rejected the null hypothesis, the ratio between the familial and general population can be calculated. For each age groups with probative data, the ratio was calculated along with it's 95% confidence interval. The results are as follows:

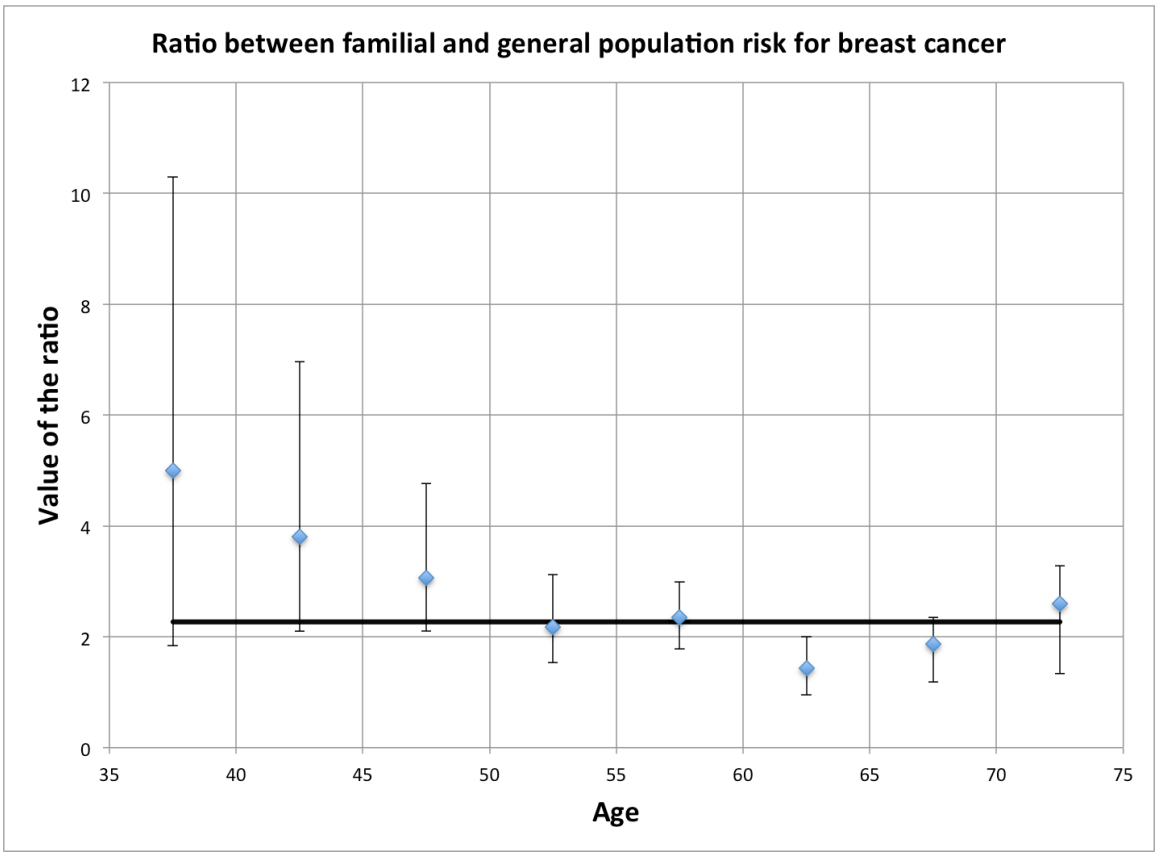

Figure 19: The ratio between the familial and general population risk for breast cancer as a function of age. Showing 2 standard deviation of the ratio (95% confidence interval). The black line indicates the calculated value of the ratio if constant, $\alpha = 2.27$.

It would be possible to claim that the ratio remains constant over all ages for breast cancer. However, the ratio seems to decrease linearly with age during the first 4 age groups (35-55 years). The interval showed are 95% CI. Thus, such a result could be explained by chance alone and the claim made, that the ratio remains constant during the observation period.

Other interpretations of the evolutions of the ratio between the familial and general population risk for breast cancer are also possible. Figure 19 could be broken into 2 separate parts, one relating to women before menopause and the other to women after menopause. If the 2 parts are studied separately, the section after menopause (approximately over 50) would have a constant ratio of roughly 2, whereas the part before menopause would show a downward trend of the ratio.

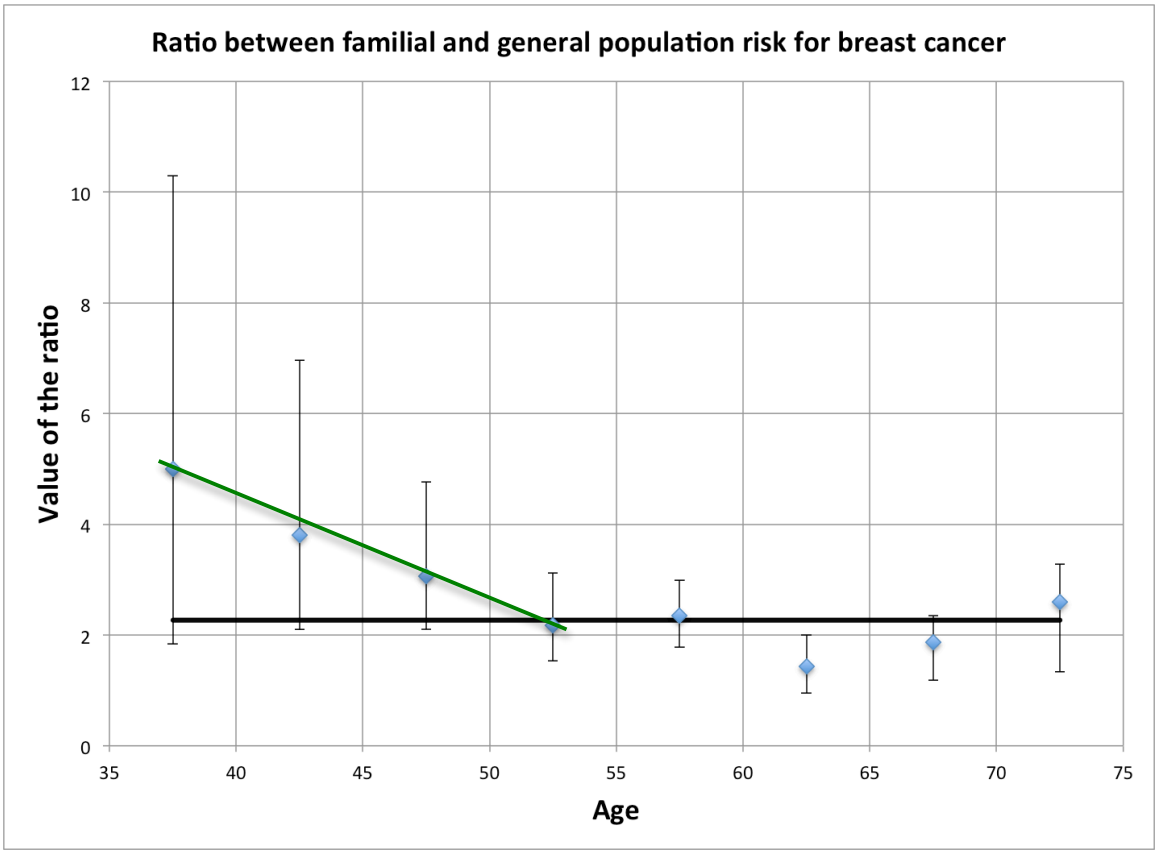

Figure 20: The same as Figure 19. The first 4 age groups seem to indicate a linear decrease in the ratio during the period 35-55 years of age. This decrease is illustrated by a hypothetical regression line (the green line) for the first 4 age groups. This difference in ratio could be due to early familial breast cancer such as BRCA1.

Figure 20 shows that the first 4 age groups of breast cancer are aligned on a downward slope. The ratios of the first 3 age groups have very large standard deviations. As a consequence, the observed downward slope could simply be due to chance. However, the possibility that it is driven by some biological phenomenon cannot be ruled.

The decrease in ratio could be due to the small fraction of women that die of early familial breast cancer. Such forms have been identified i.e. BRCA1/BRCA2 {Dite:2003de}. The high early ratio would be due to a small with a very high risk for early breast cancer. As age increases, the risk of the rest of the population increases and the fraction of women with a higher risk decreases as they die. The simultaneous decrease in the fraction of women at high risk and increased risk of the rest of the

population causes the ratio to stabilize between ages 50 and 55. Other explanations are also possible.

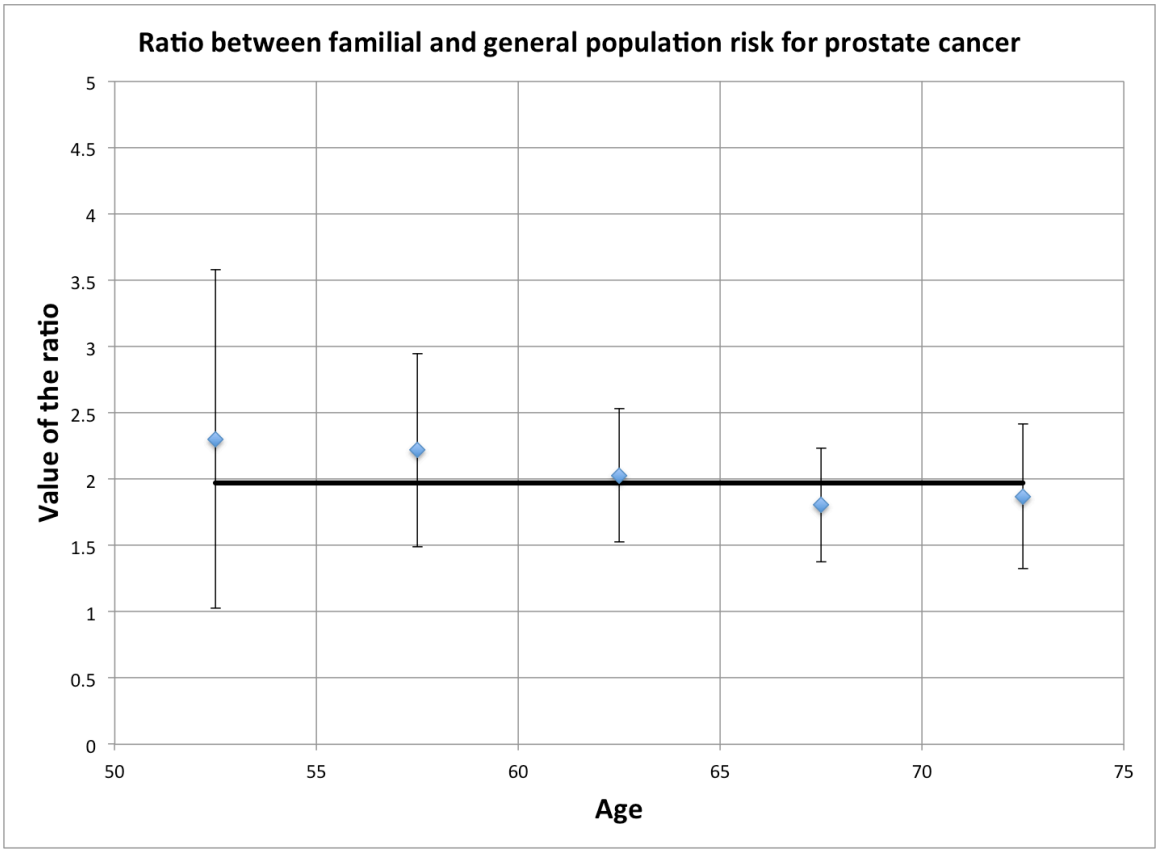

Figure 21: The ratio between the familial and general population risk for prostate cancer as a function of age. Showing 2 standard deviation of the ratio (95% confidence interval). The black line indicates the calculated value of the ratio if constant, $\alpha = 1.97$.

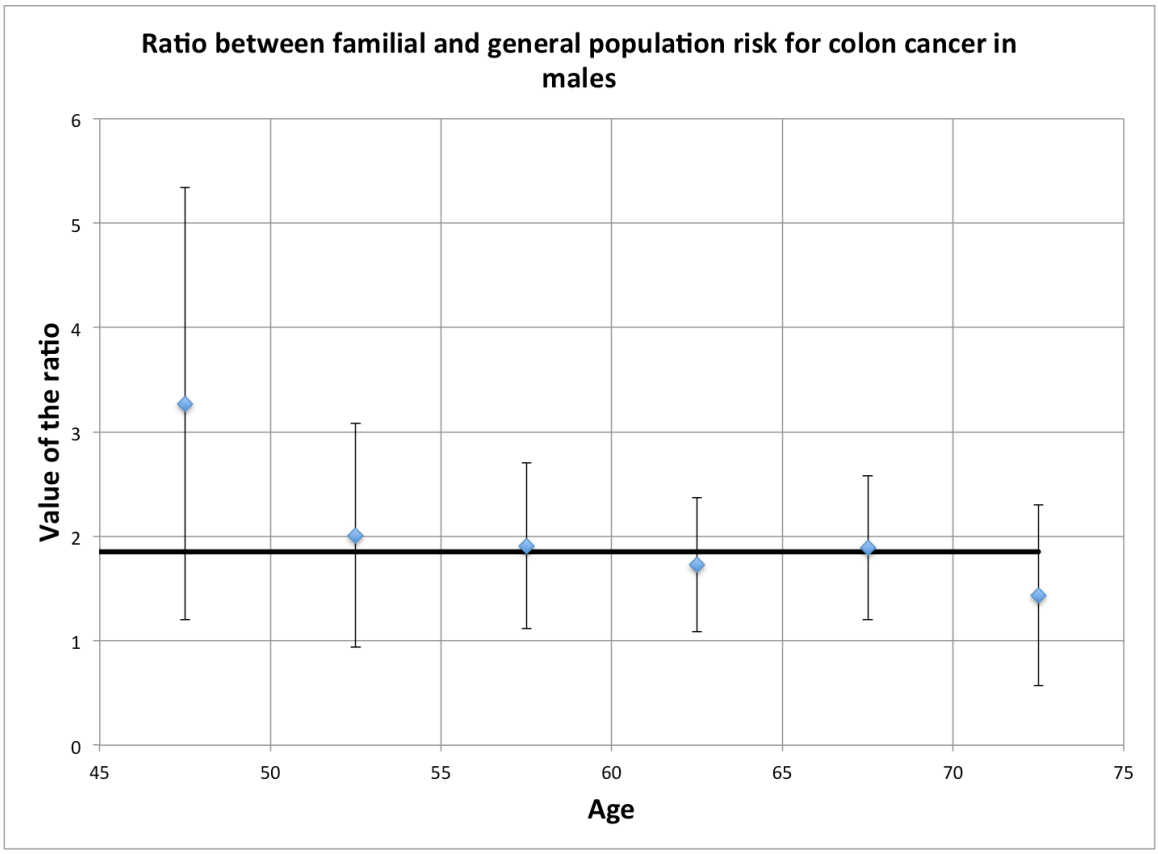

Figure 22: The ratio between the familial and general population risk for colon cancer in males cancer as a function of age. Showing 2 standard deviation of the ratio (95% confidence interval). The black line indicates the calculated value of the ratio if constant, $\alpha = 1.85$.

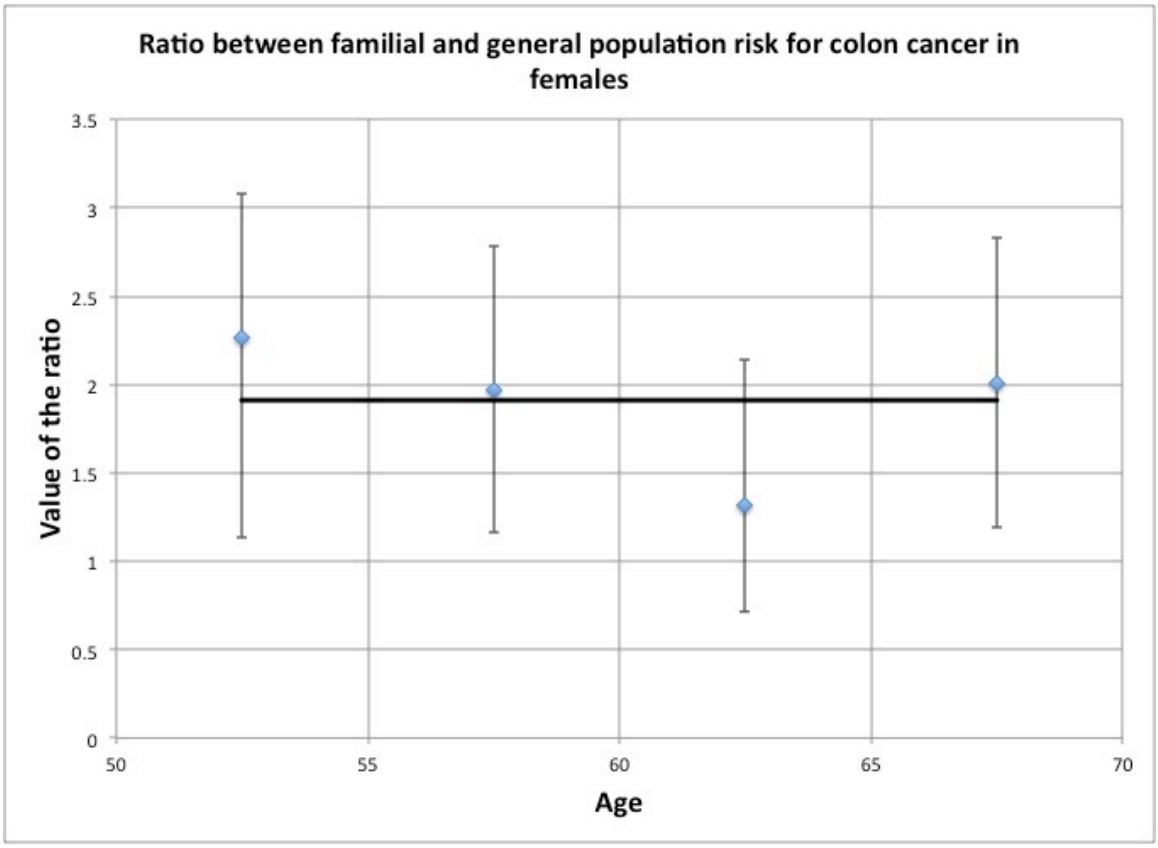

Figure 23: The ratio between the familial and general population risk for colon cancer in females as a function of age. Showing 2 standard deviation of the ratio (95% confidence interval). The black line indicates the calculated value of the ratio if constant ,α =1.91.

Figure 21Figure 23 suggest that there is no statistical difference between the ratios of any 2 age groups within a cancer. The only exception is the comparison between age 37.5 and 62.5 for breast cancer. Among over 20 age groups, it is expected by chance alone that one would be different when testing with 95% probability. It is therefore possible to conclude that the ratios are constant for all age groups for which probative data is available.

Using the regression method described in the method section, we calculate the expected ratio between the familial and general population risk α, in all 4 cases.

	Prostate	Breast	Colon Males	Colon Females

Expected slope: α	1.97	2.27	1.85	1.91
P(α≤1)	1.53E-21	3.14E-33	8.86E-10	4.97E-09
95% confidence Interval	[1.74 2.23]	[2.02 2.57]	[1.54 2.19]	[1.56 2.23]

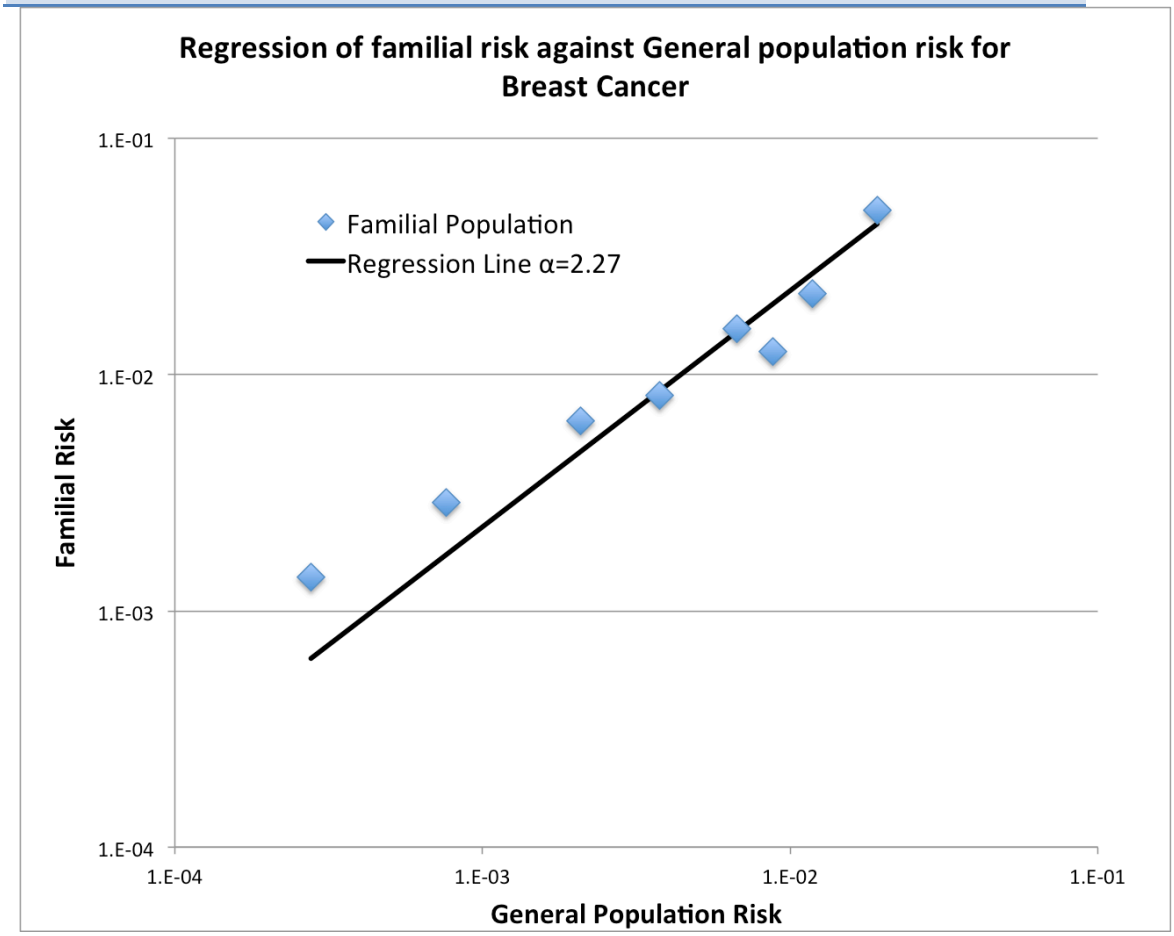

Figure 24: The familial risk as a function of the general population risk for breast cancer. The regression was calculated mentioned in the methods. It is the representation of the ratio between familial and general population if assumed constant over all age groups. The value of the calculated ratio is α. Since the axis are logarithmic, the value of the slope seen is 1 and is shifted by log(α)

These results can be presented in the form or a plot: familial risk as a function of general population risk for all age groups:

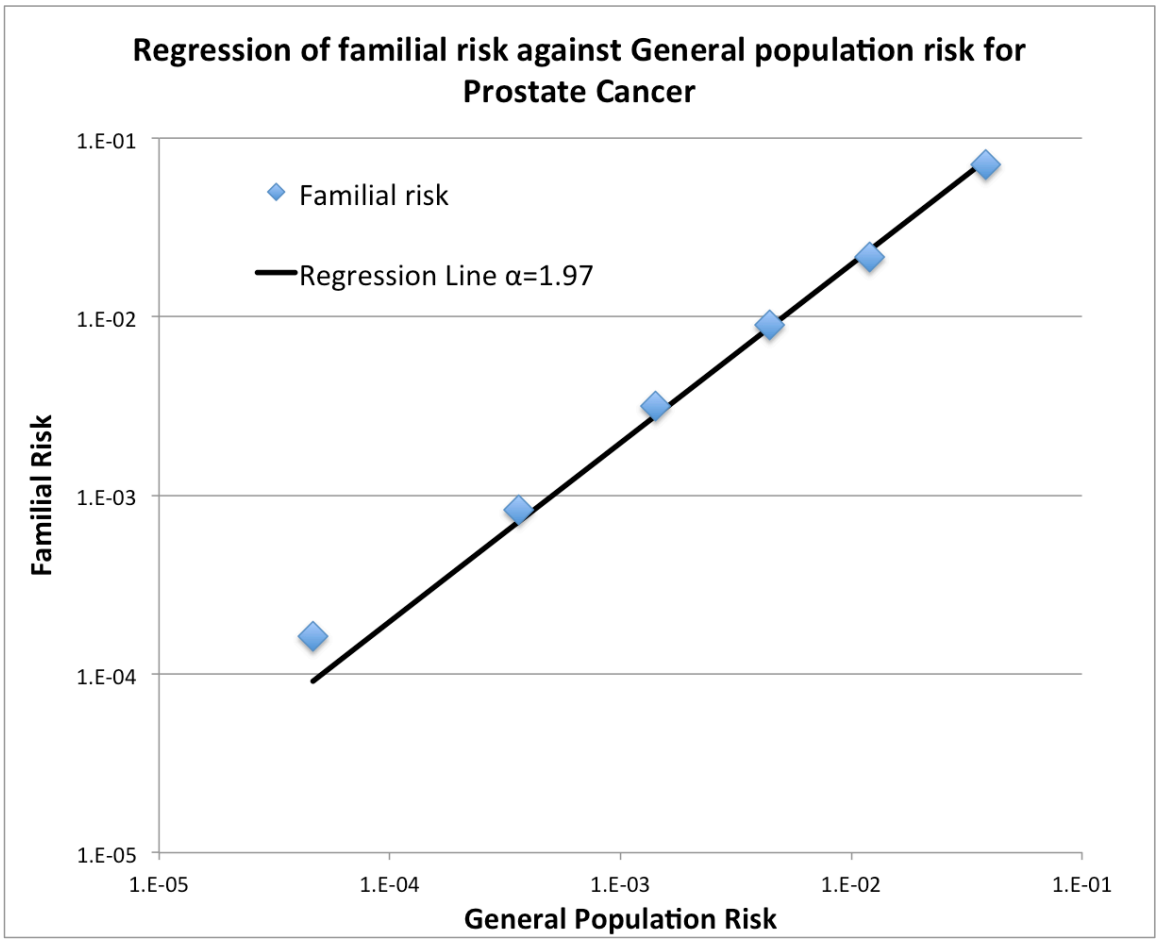

Figure 25: The familial risk as a function of the general population risk for prostate cancer. The regression was calculated mentioned in the methods. It is the representation of the ratio between familial and general population if assumed constant over all age groups. The value of the calculated ratio is α. Since the axis are logarithmic, the value of the slope seen is 1 and is shifted by log(α)

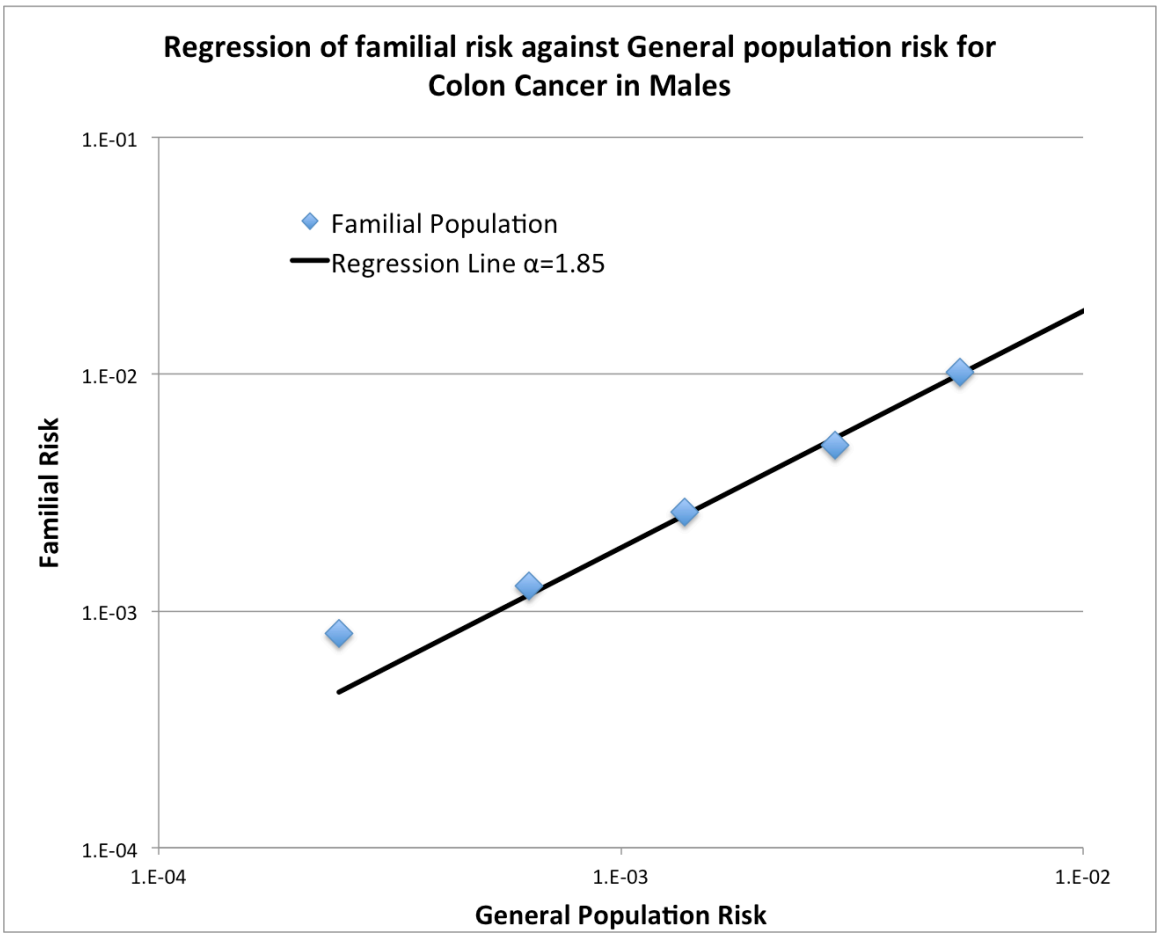

Figure 26: The familial risk as a function of the general population risk for colon cancer in males. The regression was calculated mentioned in the methods. It is the representation of the ratio between familial and general population if assumed constant over all age groups. The value of the calculated ratio is α. Since the axis are logarithmic, the value of the slope seen is 1 and is shifted by log(α)

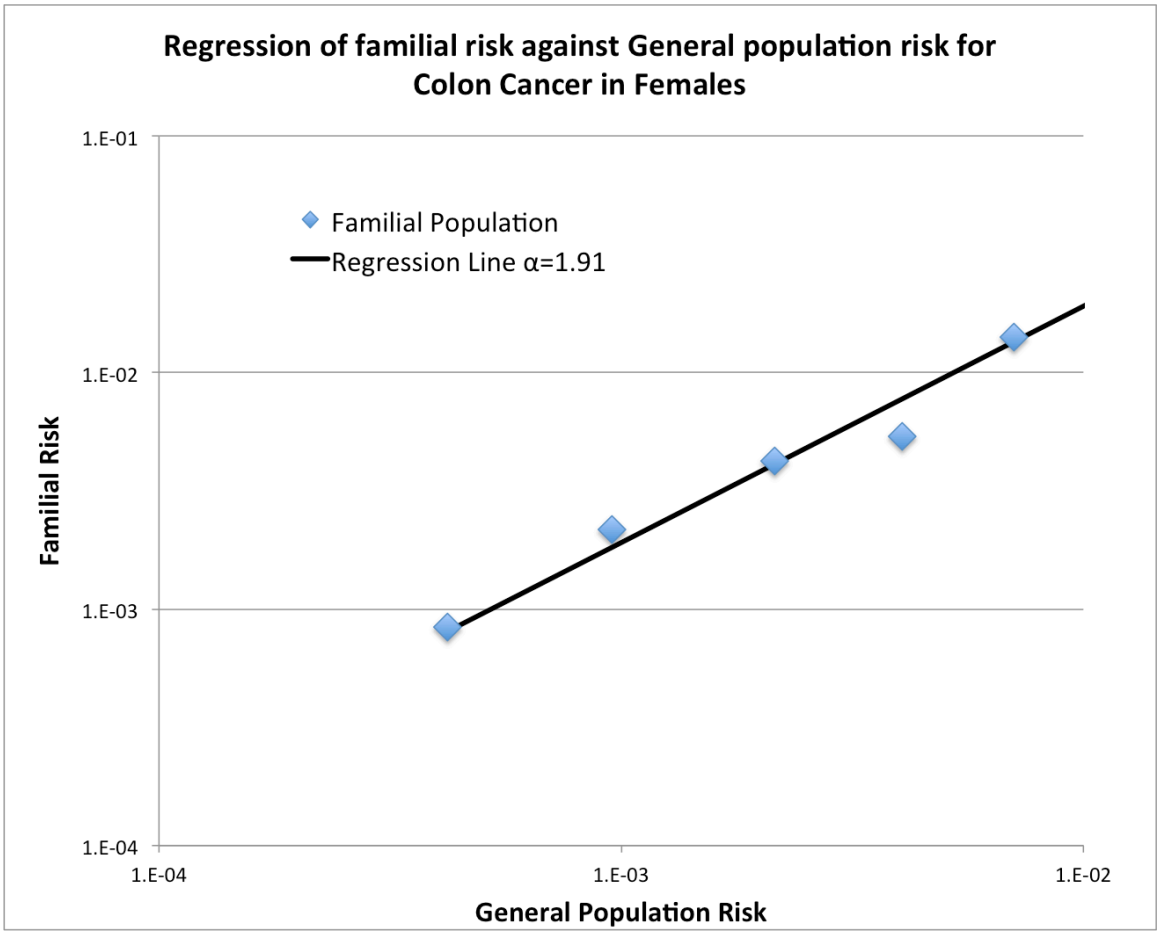

Figure 27: The familial risk as a function of the general population risk for colon cancer in females cancer. The regression was calculated mentioned in the methods. It is the representation of the ratio between familial and general population if assumed constant over all age groups. The value of the calculated ratio is α. Since the axis are logarithmic, the value of the slope seen is 1 and is shifted by log(α).

For prostate, male colon and female colon cancer, the data indicates a constant ratio. The case for breast cancer could be argued both ways. The simplest explanation would be that the ratio remains constant over the observed period. Earlier in this section is described another possible explanation.

Inspection of Figure 21,Figure 22Figure 23 suggests that the ratio remains constant over all age groups with probative data for prostate and colon cancer in both males and females. There are not 2 age groups that can be said to be significantly different at 95% confidence under a pairwise comparison.

Biologically based model

The model used in this investigation is a two-stage carcinogenesis model published in Kini et al. This model uses biological parameters to describe age-specific mortality by cancer.

As postulated by Armitage and Doll (ARMITAGE & DOLL 1957)cancers arise in two stages: initiation and promotion.

Initiation was defined as the process by which an undefined but numerically constant population of cells at risk experienced rare events at a constant annual rate. The number of specific oncogenic events required to transform a cell at risk into a founder "initiated" cell of a preneoplastic lesion was defined as "n". Promoted stem cells give rise to preneoplastic lesions that continue to divide at the same rate as fetal/ juvenile stem cells. This doubling occurs at an annual rate μ, which is kept constant even after maturity. The number of promotion events needed (n) is thought to be 2.

As quoted from Kini et al:

"Familial heterozygosity for the *APC* gene is fully penetrant; all heterozygotes display multiple adenomas and, if untreated, adenocarcinomas (Kwak & Chung 2007). This indicates that for most colorectal tumors the number of required initiating mutations "n" ≥ 2.

Values of the geometrical mean of initiation mutations assuming values of n other than 2 are thus clearly discordant with *APC* and *OAT* colonic mutation rate estimates. These facts, derived from clinical genetic observations in both inherited and sporadic forms of colorectal cancers are wholly consistent with the conclusion that n = 2 and inconsistent with values of $n \neq 2$."

The number of specific oncogenic events required to transform a cell at risk into a founder "promoted " cell of a neoplastic lesion was defined as "m". Kini et al. shows that best fits are obtained with m=1 when n=2.

Kini et al. (2011) have incorporated in a model for colorectal cancer the findings that promotion has to happen in the fetal/juvenile period. This model stipulates that the promotion of a stem cell give rise to a preneoplastic lesion that grows with the fetal/juvenile organ up to maturity. It then continues to increase at a growth rate close to that of juvenile tissues (Herrero-Jimenez et al., 2000). In each

growing preneoplastic colony a cell can undergo promotion at any point after initiation. Only people that have become promoted during the fetal/juvenile period can undergo initiation.

The resulting incorporating these findings to calculate age-specific mortality (OBS) is:

$$OBS(g)=\frac{F_{prom}\left(1-e^{-2\ln(2)R_i^2R_A2^g\sum_{a=0}^{a_{max}}ae^{-R_A2^{g-a}}}\right)}{F_{prom}+(1-F_{prom})e^{-\frac{1}{f}2\ln(2)R_i^2R_A\sum_{k=0}^{g}2^k\sum_{a=0}^{a_{max}}ae^{-R_A2^{g-a}}}} \quad \text{Equation 3}$$

In this equation, time is expressed in g, the number of stem cell doubling. If we wish to convert the time in years, we apply the following relation:

$$t=\frac{g-a_{max}}{\mu}+19 \text{ thus: } g=\mu t+a_{max}-19 \quad \text{Equation 4}$$

The parameters meaning is as follows:

R_i: is the rate of an initiation mutation for every stem cell division

R_A: is the rate of the promotion mutation for every stem cell division

F_{prom}: The fraction of the population in which an initiated stem cell will grow to form a preneoplastic lesion.

f: "Represent the fraction of persons that die of the observed cause among the set of mortal diseases with shared risks and synchronous changes in death rates" (Quoted from Kini et al 2011, in appendix).

a_{max}: the number of stem cell divisions in an organ when an individual reaches maturity.

μ: The division rate of stem cells in an individual's preneoplastic lesion.

The model can also be presented in the same form as is done in the introduction.

$$OBS(g)=\frac{F_{prom}\left(1-e^{-V_{OBS}(g)}\right)}{F_{prom}+(1-F)e^{-\frac{1}{f}\int_0^g V_{OBS}(a)da}} \qquad \text{Equation 2}$$

with:

$$V_{OBS}(g) = 2\ln(2)R_i^{\,2}R_A 2^g \sum_{a=0}^{a_{max}} a e^{-R_A 2^{g-a}} \qquad \text{Equation 1}$$

N.B. The value and meaning of V_{OBS} has already been discussed in the introduction.

The model assumes that the number of promotional events happening at age g follows a Poisson distribution. As a consequence, the probability of observing no events at age g is: $e^{-V_{OBS}(g)}$. Thus the probability of observing at least one event is:

P(at least one event at age g) = $1-e^{-V_{OBS}(g)}$.

As F_{prom} is the fraction of people in which a neoplastic lesion grows, the upper term represents the probability of having a promotional event in the general population of people alive at age g.

If the whole population is at risk, i.e $F_{prom}=1$, the lower term =1. Therefore:

$OBS(g)=1-e^{-V_{OBS}(g)}$

In the case $F_{prom} \neq 1$, the population at risk is dying faster than the general population. This would need to be accounted for. The fraction of people alive is the fraction of people at risk, minus the fraction of people that died of the modeled disease and related concurrent causes.

The assumption that the number of lethal events in an interval [0,g] follows a Poisson distribution is made. It follows from the previous assumption, where the number of lethal events happening during an interval [g-1,g], is Poisson distributed. From the definition of $V_{OBS}(g)$:

$\int_0^a V_{OBS}(a)da$ = The expected number of lethal events happening in an individual up to age g.

Using the same argument as the upper term,

$e^{-\frac{1}{f}\int_0^g V_{OBS}(a)da}$ =The probability of not having died of the studied disease or any concurrent form of death up to age g.

The probability of being alive at time g is:

$F_{prom}+e^{-\frac{1}{f}\int_0^g V_{OBS}(a)da}-F_{prom}e^{-\frac{1}{f}\int_0^g V_{OBS}(a)da}$ =The fraction of people at risk + The fraction of the people that did not die – Those that have died of the studied disease or any concurrent form of death.

This equation can be factorized to: $F_{prom}+(1-F_{prom})e^{-\frac{1}{f}\int_0^g V_{OBS}(a)da}$, the lower term of the general equation.

Cancer is though to arise through a variety of possible pathways. The observed mortality is the sum of the mortality caused by the all the different pathways. These distinct routes leading to cancer could each produce a different age-specific mortality curve in the population where they occur. However, not knowing the nature of these individual pathways, the only possibility is to consider the average mortality for the general population. It follows that this observed mortality is the average mortality for the different groups of the population in which cancer occurs through a different pathway.

Having defined the general equation, we will now formalize the ratio between mortality for two populations: the familial population and the general population. Without loss of generality, we can write 2 separate equations, one for the age specific risk in the general population and one for age specific risk in the familial population. The only difference will be the values the parameters can take.
For the general population we have:

$$OBS_0(g) = \frac{F_0\left(1-e^{-2\ln(2)R_{i0}{}^2R_{A0}2^g\sum_{a=0}^{a_{max0}}ae^{-R_{A0}2^{g-a}}}\right)}{F_0+(1-F_0)e^{-\frac{1}{f_0}2\ln(2)R_{i0}{}^2R_{A0}\sum_{k=0}^{g}2^k\sum_{a=0}^{a_{max0}}ae^{-R_{A0}2^{g-a}}}}$$

and for the familial population:

$$OBS_1(g) = \frac{F_1\left(1 - e^{-2\ln(2)R_{i1}{}^2R_{A1}2^g\sum_{a=0}^{a_{max1}} ae^{-R_{A1}2^{g-a}}}\right)}{F_1 + (1-F_1)e^{-\frac{1}{f_1}2\ln(2)R_{i1}{}^2R_{A1}\sum_{k=0}^{g}2^k\sum_{a=0}^{a_{max1}} ae^{-R_{A1}2^{g-a}}}}$$

The ratio between the age specific risk for the general population and familial population is thought to remain constant over the ages. The ratio can be expressed as:

$$\frac{OBS_1(t)}{OBS_0(t)} = \frac{F_1\left(1 - e^{-2\ln(2)R_{i1}{}^2R_{A1}2^g\sum_{a=0}^{a_{max1}} ae^{-R_{A1}2^{m_1t+a_{max1}-a-19}}}\right)}{F_1 + (1-F_1)e^{-\frac{1}{f_1}2\ln(2)R_{i1}{}^2R_{A1}\sum_{k=0}^{g}2^k\sum_{a=0}^{a_{max1}} ae^{-R_{A1}2^{m_1t+a_{max1}-a-19}}}} \Bigg/ \frac{F_0\left(1 - e^{-2\ln(2)R_{i0}{}^2R_{A0}2^g\sum_{a=0}^{a_{max0}} ae^{-R_{A0}2^{m_0t+a_{max0}-a-19}}}\right)}{F_0 + (1-F_0)e^{-\frac{1}{f_0}2\ln(2)R_{i0}{}^2R_{A0}\sum_{k=0}^{g}2^k\sum_{a=0}^{a_{max0}} ae^{-R_{A0}2^{m_0t+a_{max0}-a-19}}}}$$

This remains constant for each 5 years age interval between 30 and 74 years of age, as shown by analysis of the data.

If F_0 equals1, it follows that F_1 also equals 1. In this situation, the equation simplifies to:

$$\frac{OBS_1(t)}{OBS_0(t)} = \frac{1 - e^{-2\ln(2)R_{i1}{}^2 R_{A1} 2^g \sum_{a=0}^{a_{max1}} a e^{-R_{A1} 2^{m_1 t + a_{max1} - a - 19}}}}{1 - e^{-2\ln(2)R_{i0}{}^2 R_{A0} 2^g \sum_{a=0}^{a_{max0}} a e^{-R_{A0} 2^{m_0 t + a_{max0} - a - 19}}}}$$

In all individuals that are heterozygous for the APC gene, APC -/- polyps appear to grow (Alfred G Knudson 2001; Kwak & Chung 2007). Therefore F=1. The similarity between the colon and prostate age-specific mortality curve suggest that the assumption should be maintained. For a more complete description of the assumption that F=1, please refer the reader to Kini et al.

The equation is highly non-linear. Therefore, no attempt to analyze the parameters analytically will be made.

Application of model to data to discover parameters that may be affected by familial risk. (limitation to case of n=2, m=1)

The first set is to calculate values of the parameters that accurately fit the mortality data of prostate cancer and colon cancer in males. Only male cancers will be analyzed in this section.

The fitting process will give an estimate of the values the parameters could take in the general mortality by a specific form of cancer. The different parameters will be modified one by one until a curve that portrays an increase similar to the one observed in the familial population is observed. This process will give an estimate of the variation of each parameter needed to explain the observed increase.

Note that the values in the model are population averages. People vary in size, and therefore the size of the organs varies as well. The mutation rate can also vary from person to person by a range of nearly 10-fold (Sudo et al. 2008). The values used in the model are therefore averages of a heterogeneous population. A difference

in the value of a parameter between the familial and general population, therefore reflects a change in the distribution.

The effects of the variables in the models are not independent of each other. For example, for a large a_{max} , a variation in R_A will have more impact than if a_{max} is small.

Preneoplastic growth rate: μ and fraction of deaths among persons at risk from cancer at observed site: f

The effect of a variation in f or μ is shown in Figure 28 and Figure 29, for prostate and colon cancer respectively.

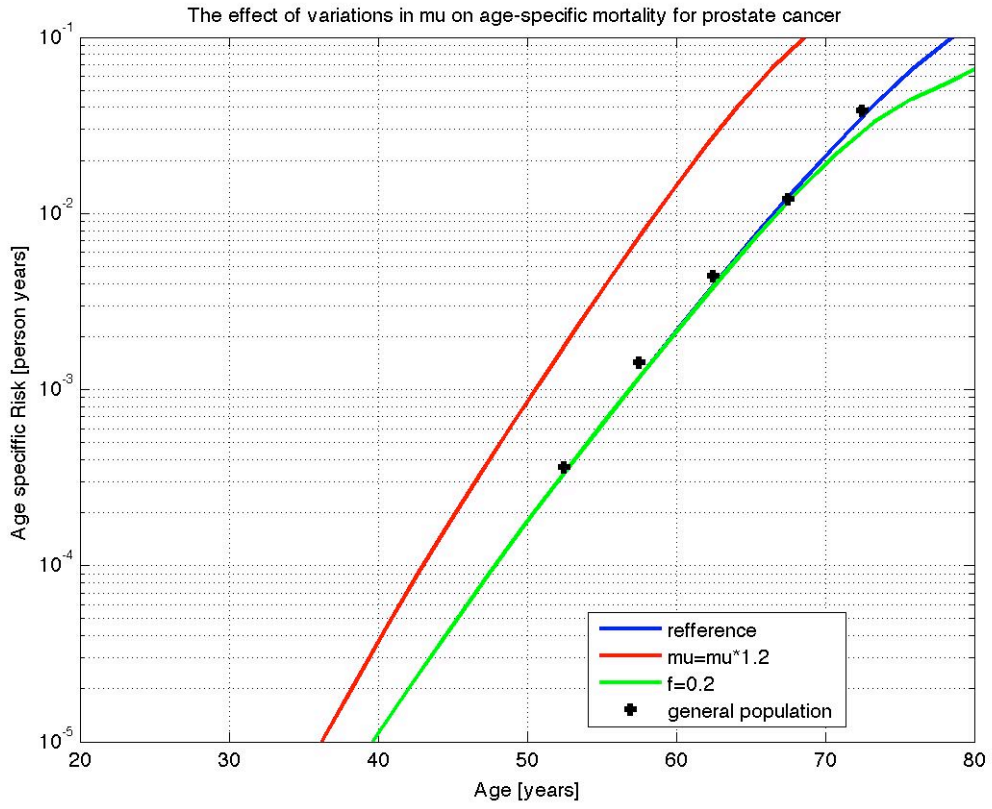

Figure 28: The effect of variations in μ (mu) and f on the age specific mortality by prostate cancer. We can clearly see that a variation in f or μ would not keep the ratio between the original curve and the varied cure in the observed age-range.

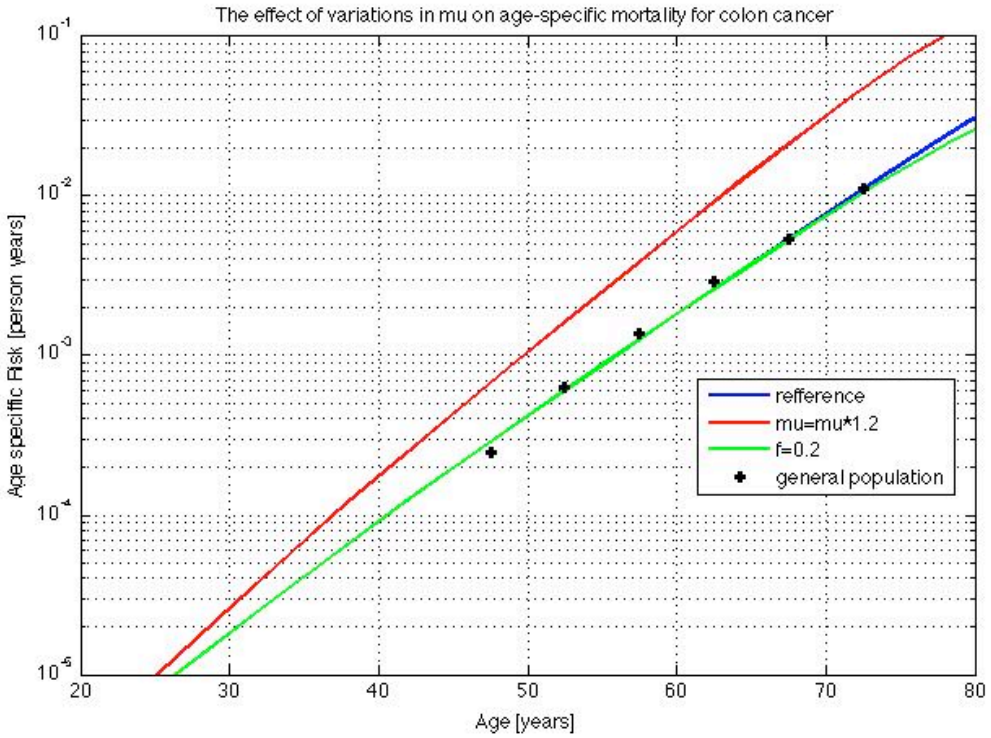

Figure 29: The effect of variations in μ (mu) and f on the age specific mortality by male colon cancer. We can clearly see that a variation in f or μ would not keep the ratio between the original curve and the varied cure in the observed age-range.

The graphs show that a variation in μ or f would result in a change in the observed slope of the curve. This results in the ratio changing during the observation period. Consequently the conclusion that difference between the familial population and the general population is a difference in a_{max}, R_i, R_A, F_{prom} or a combination of these factors can be reached. However, due to the complexity of the problem they will be analyzed separately.

Study of familial APC concluded that everyone is at risk for colon cancer. Thus F is assumed equal to 1. This assumption is extended to prostate cancer based on the similarities of the mortality curves.

Figure 30 and Figure 31 show the curves resulting from the variation of a_{max}, R_i, and R_A independently for prostate and colon cancer. The reference curve is the one that best fits the mortality for the general population. The variations of the parameters are aimed to fit the mortality of the familial population.

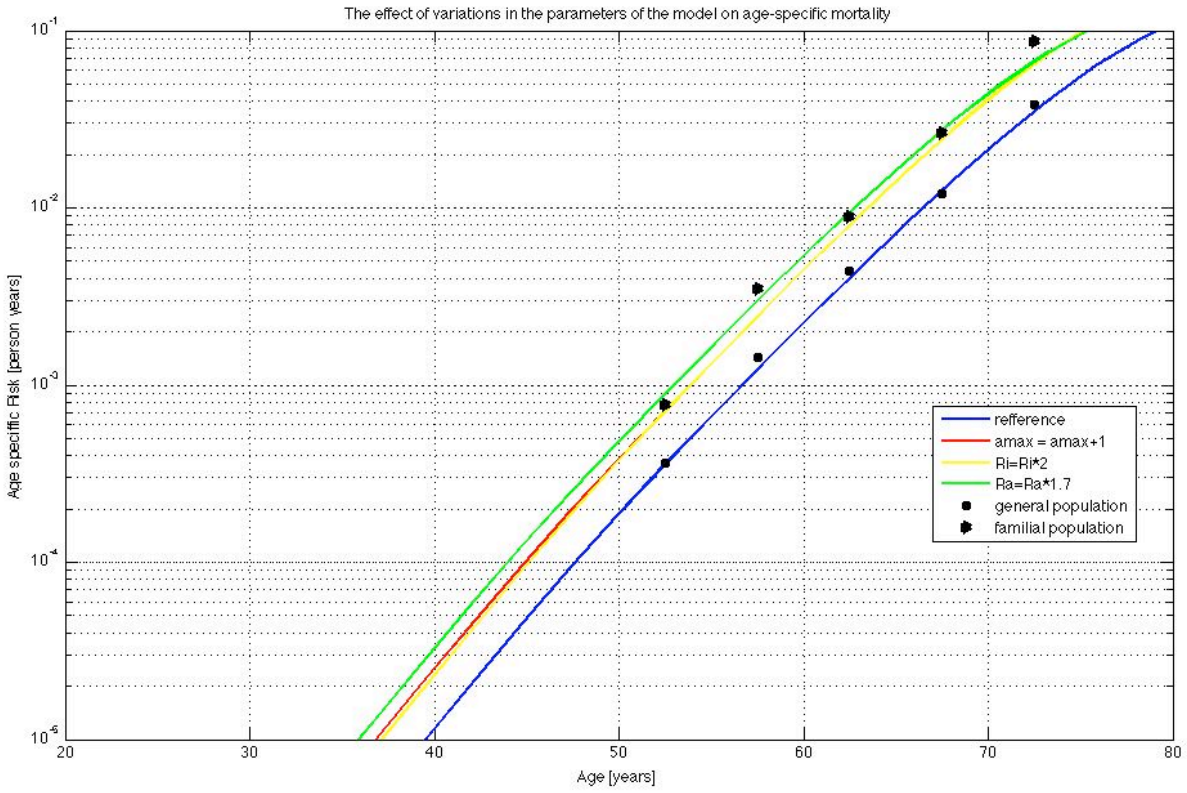

Figure 30: The effect of the variation of a_{max}, R_i and R_A, on the age specific mortality for prostate cancer. The reference is the curve calculated with the parameters supposed to fit the general population mortality. To obtain the other 3 curves, we have varied the corresponding parameter while keeping the other constant until it was close enough to the mortality of the familial population. In order to fit the familial population, R_i is multiplied by 2, R_A is multiplied by 1.7 and 1 is added to a_{max}.

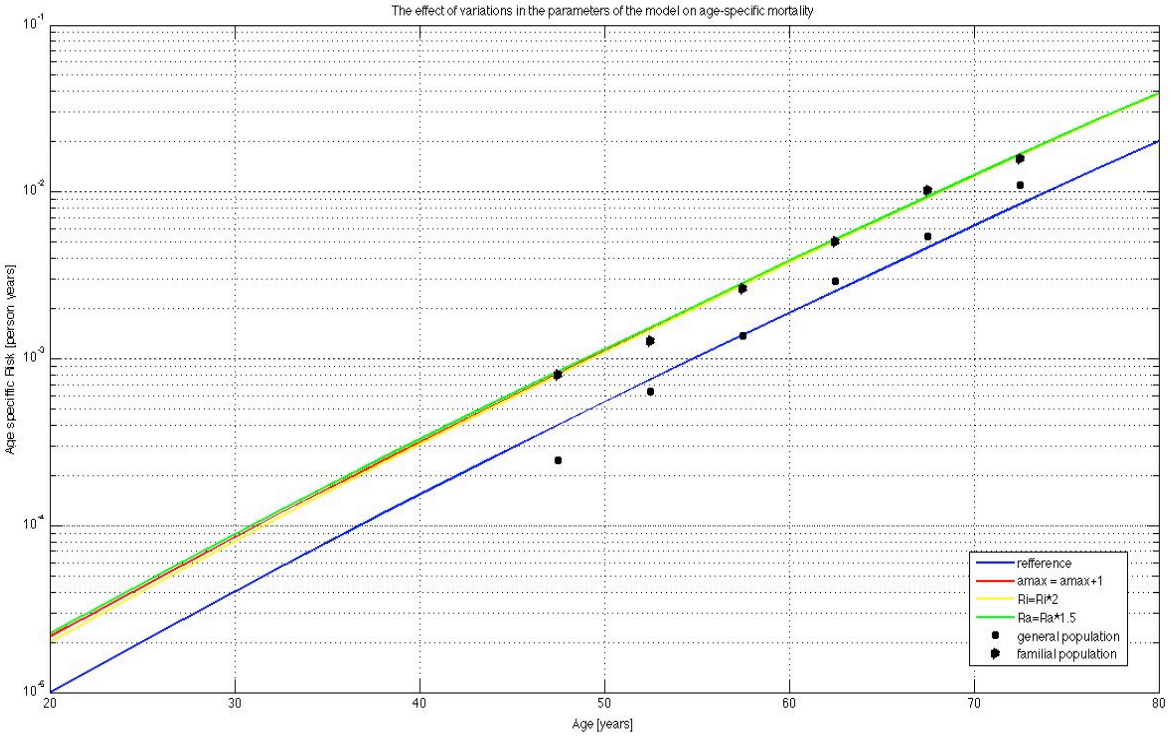

Figure 31: The effect of the variation of a_{max}, R_i and R_A, on the age specific mortality for colon cancer in males. The reference is the curve calculated with the parameters supposed to fit the general population mortality. To obtain the other 3 curves, we have varied the corresponding parameter while keeping the other constant until it was close enough to the mortality of the familial population. In order to fit the familial population, R_i is multiplied by 2, R_A is multiplied by 1.5 and 1 is added to a_{max}.

Number of stem cell doubling until maturity (related to organ size): a_{max}

a_{max} represents the number of stem cell doubling in the organ studied. For the case of the colon, the number of stem cells is thought to be equal to the number of crypts since each crypt has a stem cell at the bottom.

The graphs show that for both colon and prostate, a_{max} would need to increase by 1 in order to fit the risk of the familial population. An increase of 1 means a doubling of the number of stem cells in the organ and therefore a doubling of the size of the organ.

A doubling in the number of stem cells of the colon corresponds to a doubling in the length. For a_{max} to explain familial risk, the familial population must have a colon twice as long as the general population. This increase in colon length must be shared between parents and children. A link between the length of the colon in parents and the length of the colon in children is very plausible. However, it is not

plausible that the population whose parents died of colon cancer have a colon twice as large as the general population.

Based on the results of the model, the conclusion that a variation in colon size could not alone explain the familial risk for cancer is reached.
Similar arguments can be made for prostate cancer. In both cases, a doubling of the organ is needed to double the risk of cancer.

Rate of initiation mutations: R_i

Figure 30 and Figure 31 show that R_i is 1.8 and 2 times larger in the familial population for prostate and colon cancer respectively. The difference between 1.8 and 2 can be considered as not significant. Therefore the variation needed for R_i to explain familial risk is roughly 2-fold.

These numbers reflect an average of the population under study. The population is known to be heterogeneous for mutation rates (Sudo et al.). This heterogeneity is due to inheritable factors, environmental factors or both.

According to the model, the average R_i among the people that died of cancer is higher. For R_i to be the underlying factor of familial risk, the combination of shared environment and inheritable factors must account for a doubling in R_i among the offspring.

Since the distribution of mutation rates in the population is unknown, no attempt will be made to calculate the correlation between parent R_i and offspring R_i.

These findings suggest R_i is possibly the underlying factor behind familial risk.

Rate of promotion mutation: R_A

Figure 30 and Figure 31 show that R_A needs to be 1.5 times larger for the prostate and 1.4 times larger for the colon cancer to describe familial risk. This difference can be considered as small.

It is not known if R_i and R_A are different. It is well possible that $R_A=R_i$. In which case they would vary together. This, however, would not change the problem. In this case, $R_i=R_A$ would 1.7 times higher in the familial population as in the general population.

These findings suggest R_A is possibly the underlying factor behind familial risk.

Fraction at risk of promotion given initiation : F_{prom}

The fraction of the people at risk is believed to be equal to 1. A change in F_{prom} would decrease the mortality rate. This change in age-specific mortality has a constant ratio to the original curve in the ages between 20 and 75.

The mortality of the general population was fitted under the assumption that $F_{prom}=1$. Any change in F_{prom} will therefore reduce the calculated mortality. As a result, no attempt will be made to fit the familial population when changing the value of F_{prom}. Figure 32 Figure 33 show the prostate and colon cancer curves and the modified curve with F_{prom} =0.5.

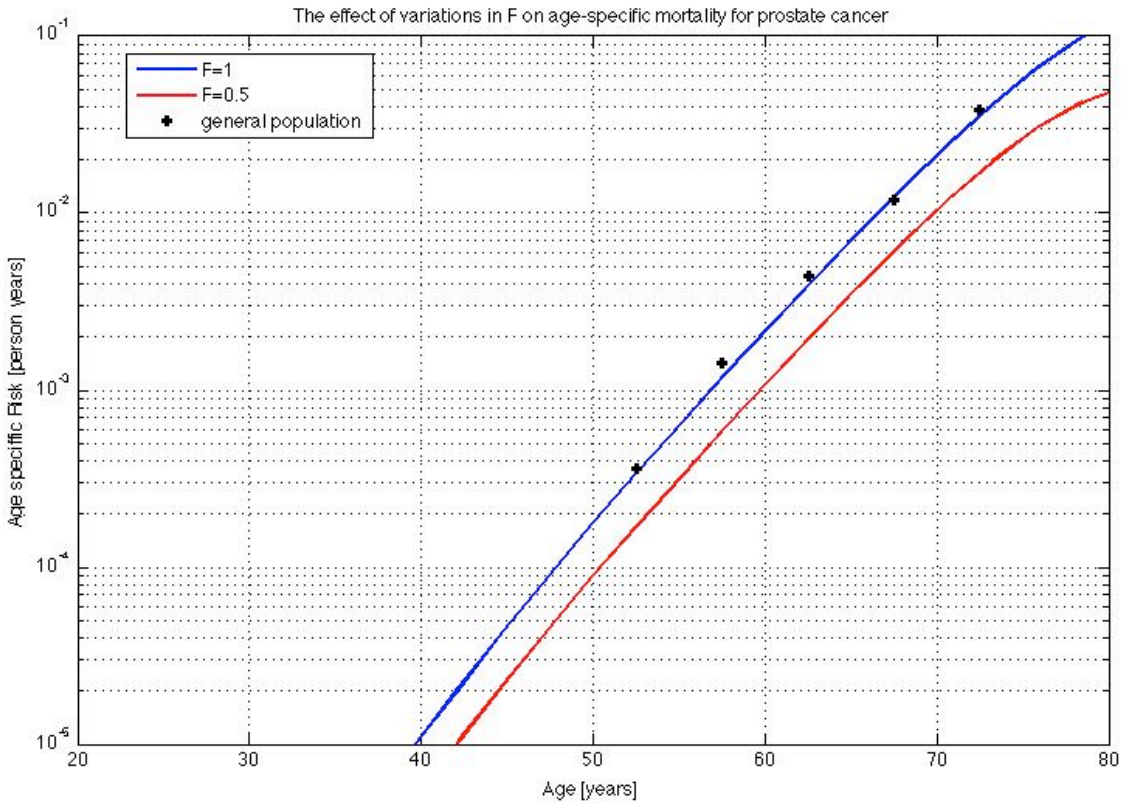

Figure 32: The effects of varying F_{prom} on the age specific mortality by prostate cancer. Age specific mortality was calculated using the same parameters as in Figure 30. The other curve was calculated using again the same parameter but by using an $F_{prom}=0.5$. Since we believe that $F_{prom}=1$, the curve calculated using $F_{prom}=0.5$ does not fit the familial population and is even smaller than the general population mortality. This graph is informative as it tells us how a variation of F_{prom} will impact age-specific mortality.

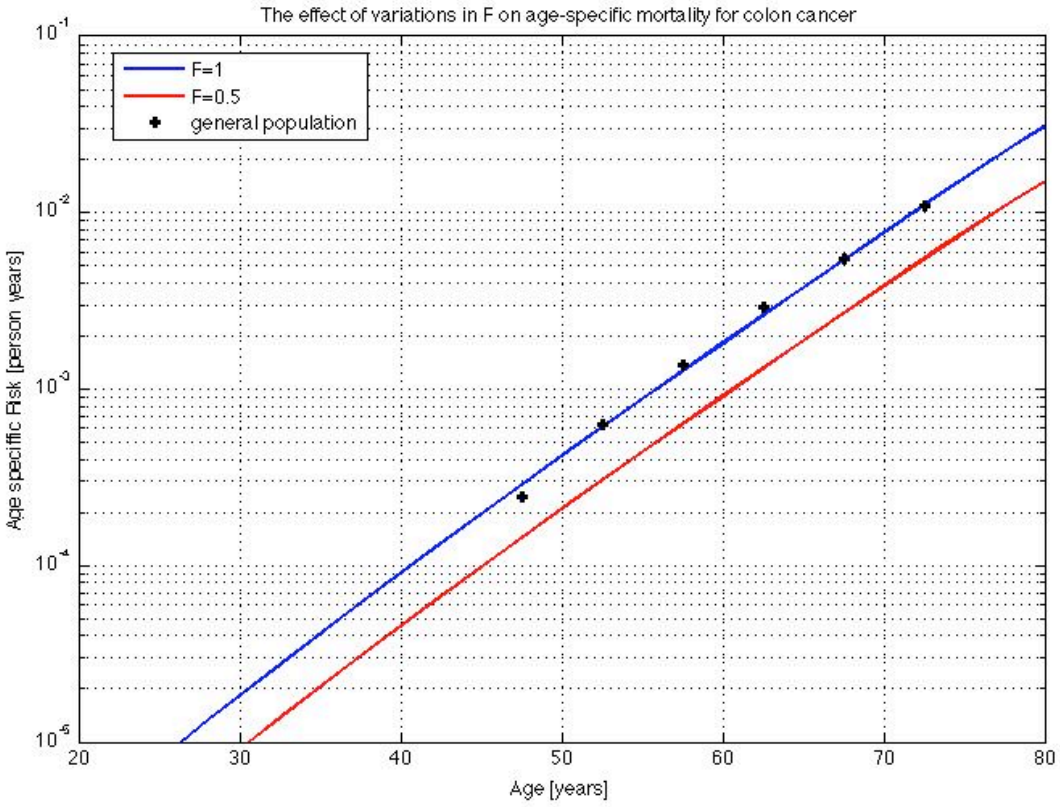

Figure 33: The effects of varying F_{prom} on the age specific mortality by colon cancer. Age specific mortality was calculated using the same parameters as in Figure 31. The other curve was calculated using again the same parameter but by using an F_{prom}=0.5. Since we believe that F_{prom}=1, the curve calculated using F_{prom}=0.5 does not fit the familial population and is even smaller than the general population mortality. This graph is informative as it tells us how a variation of F_{prom} will impact age-specific mortality

This means that if F_{prom} is different from 1 and is the underlying cause of familial risk, it would need to be smaller than 0.5 in the general population. This follows from the fact that F_{prom} is a fraction and is therefore always smaller or equal to 1.

If F_{prom} is to explain familial risk it would have to be very small in the general population (ie. Much smaller that 0.5). Otherwise, the link between a parent being at risk and an offspring being at risk would be too strong.

For example, if F_{prom} =0.5: If a parent dies of cancer he is necessarily at risk. Since $F_{prom1}=2*F_{prom0}$ it follows that $F_{prom1}=1$. This means that if a parent is at risk, the offspring are necessarily at risk as well. A small value of F_{prom} is not compatible with the observed mortality rates by prostate and colon cancer in the general population.

Furthermore, it would contradict the observations made in the study of APC mutated families.

These results therefore suggest that the fraction of people at risk is not the underlying factor of a familial risk for cancer.

Discussion

Strengths and limitations of the statistical tests

The difference between risks is tested. Risk is assumed constant within every interval studied. No correction is applied to account for the resolution (i.e. the approximation that mortality remains constant in 5 years intervals). However risk is changing within every studied interval. In some cases it doubles between the beginning and the end of the 5 years interval. Taking smaller intervals would have reduced the numbers too much. These findings need to be taken into account when interpreting the result of the test.

Difference in risk was calculated for every age group. The final test verifies whether the sum of all differences is greater than zero. What does this mean when the risk is changing by four orders of magnitude during lifetime? It means that the difference in risks is also changing by a similar order of magnitude. Consequently, the later, and larger risks, will contribute much more to the final sum than the earlier, and smaller, ones. On one hand, this is a problem. However, the later, and larger risks, have much larger numbers that the earlier ones. As a consequence, the way the total difference in risk is tested, gives more importance to the age groups with larger numbers because they are the ones with larger risks. Some age groups can have no cases, as happens for all the age groups before 40 years of age in prostate cancer. The difference in risk is thus centered around zero for those ages. The number of person years in those early groups are very high. Consequently, the distribution will be extremely narrow and approach a Dirac function, the neutral element of the convolution. These age groups will therefore only contribute in a minor way to the final PDF of γ_T, and thus to the test result. Even so, the early age groups, if they are the same and with a risk equal to zero, will still contribute by slightly broadening of

the final distribution of γ_T. It is impossible to know if this is the optimal way of doing or not. The only thing is to keep these facts in mind when interpreting the results.

The second test uses a simpler approach. It assumes that the number of cases follows a Poisson distribution in all age groups. It also assumes that the Poisson distribution is approximated by a normal distribution with a variance equal to its mean. These assumptions are only valid if there are sufficient numbers of familial cases in an age group. The limit is set to 9 cases, a number that is reasonably large to make the second hypothesis valid. This test can therefore only test some of the data points. It is also to be noted that it provides no way of combining evidence. This means that the test result needs to be interpreted for each age group individually. This visualization allows a better understanding of the significance of each age group. In fact, from the first method it is hard to judge if there are a small amount of age groups dominating the general effect.

The impossibility of testing the null hypothesis globally raises some questions. For example, we can only test four age groups in the case of female colon cancer. Of these four age groups, three have a distance in quantiles between the means, Q, greater than 2 and around 2.5. However, one of them is around 1. What can be concluded? With a distance Q of 1 it is impossible to reject the null hypothesis. A distance of 2.5 is enough to reject it. In the example of colon cancer in females, the null hypothesis is rejected in 3 out of 4 age groups. Using this method alone, it is impossible to draw a conclusion as to whether or not we should reject the global null hypothesis that the overall familial risk is greater than the general population risk. However, supported with the 1st method, it offers strong evidence to refute the null hypothesis.

Calculating constant ratio

Figure 21Figure 23, indicate that within each case, the ratio does not vary significantly between the age groups. There is today no method to measure absolute evidence in favor or against a linear regression. Therefore the claim that the ratio is constant can only come from the pairwise comparison between the risk ratios of each age group within a cancer. Once it is observed that there is no statistical difference

between the ratios, the total ratio over all age groups was calculated using a form of linear regression. This method provides a way of estimating the constant ratio between the familial and general population risk for cancer. It also provides another test of the null hypothesis. It is to be noted, that the parameter $P(\alpha < 1)$ tests the null hypothesis under a very specific assumption: That the ratio between the two risks is constant for all age groups. It is a test that alone would not be sufficient, since one could always contradict the hypothesis that the ratio is constant, even when it looks constant. However, it is a further support for the previous tests done.

Identified parameters

The results indicate that if a parameter is to be the source of a familial risk for cancer, it is most likely to be mutation rates. The model used considers a situation where initiation and promotion can have different rates. Modeling using only one mutation rate for both initiation and promotion is also possible.

To support the conclusion is the fact that mutation rate is know to vary at least 7-fold in the lung, and 6-fold in the colon. This means that only a weak link in mutation rates between parents and offspring would cause an increase in cancer risk in the offspring in parents that died of cancer.

Since the distribution of mutation rates is unknown in the general population, the correlation between mutation rate in parents and offspring is impossible with the current data.

The model used is valid within the assumptions described in Kini et al. It is a way in which one can interrogate the influence of various biological parameters on the age-specific mortality by cancer. Therefore the drawing of conclusions needs to be very careful. Consequently, the conclusions that have been drawn from it are valid as long as it is assumed that cancer is the result of dividing cells undergoing random mutations that can lead to the disease.

For a complete discussion of the model used, please refer to Kini et al. (attached as an appendix).

Future studies

The results presented here clearly establish the existence of a familial risk. Thus, the offspring of affected parents are at higher risk for cancer than the general population, for all age groups. It is also shown that the ratio between the risks of the familial population to the general population is roughly 2. This value seems to be true for all studied cases. Using a model for carcinogenesis, it can be shown that a change in the size of the organ, the fraction of people at risk or an increase in promotion or initiation mutation rates, would increase the age-specific risk by a constant factor for all the observed age groups. Any combination of these changes would also keep the ratio constant. From the analysis of the changes, it can be concluded that the mutation rates are the most likely factors to explain a large portion of the familial risk. A future study in the area of familial risk for cancer, should focus on measuring mutation rates in the general population and establishing its distribution. Such a study must also investigate the link between the mutation rates in the parents and offspring. Furthermore, this study should establish the distribution of mutation rates among cancer patients. Last, the study should measure the mutation rate distribution in the offspring of parents that died of cancer. If this distribution turns out to be different, it will be possible to determine how large is the contribution of familial mutation rates to the familial risk of cancer. This measurement would also help validate the model for carcinogenesis used in this project.

The study described in Sudo et al.(Sudo et al. 2008) describes a method that can be used to measure mutation rates in human organs. Unfortunately it is an extremely tedious method requiring large amount of organ tissue. Furthermore, there is no reason for which one should believe that mutation rates are the same throughout the human body. As a consequence, the study would have to be performed for every organ of interest. Thus a different method would have to be developed to study mutation rates in people.

In Sudo et al, biopsies of different sizes were excised from human lungs. These biopsies were then analyzed using constant denaturing gel electrophoresis for known genes (CDGE). CDGE is a method that allows for the quantification of mutant DNA strands in a DNA pool. CDGE allows the quantitative measurement of

DNA mutations in a large DNA pool with resolution up to one mutant copy in 10^4 normal copies [reference]. Furthermore, using large thermo-cyclers, it has a throughput of 96 DNA pools per hour. The use of CDGE to calculate mutation rates in an organ is described in Sudo et al. Biopsies were taken at various positions in a lung. From each biopsy, DNA was extracted, and amplified. CDGE was used to quantify the mutant fraction in those pools for various gene fractions. With the mutant fraction and the biopsy size, the size of the mutant colony size was calculated. Knowing the size of the colony gives information about which time in the fetal/juvenile development the mutation occurred. Form there, the authors calculated the mutation rates.

A similar method to the one used by Sudo et al. could be used for calculating mutation rates in any organ. However, this would require many biopsies being taken from various organs in the same persons, and then repeating the experiment for many people. This would be far too cumbersome to do on a large scale. Furthermore, it would not be possible on live people. This calls for another method of obtaining DNA from organs. One such possibility is the isolation of macrophages from the blood. Macrophages phagocytize dead cells from all organs. Thus, a macrophage could be treated in a similar way as a biopsy was in Sudo et al. Running a CDGE analysis on a macrophage containing DNA from a specific organ, would give the mutant fraction in the region the macrophage was located previous to being in the blood circulation.

The challenges involved will be discussed individually and can be stated as follows:

1. Efficiently isolating the macrophages from a blood test:

Macrophages represent about 1 in 10^5 cells in circulation in the blood. Since there are about 5 10^6 cells/µl in blood, there are about 5 10^4 macrophages/ml of blood.

2. Defining which organ the DNA in the macrophage came from.

To calculate organ-specific mutation rates, the macrophages need to be sorted according to the organ of provenience. This could be achieved through the identification of organ specific mRNA or proteins identified and measured by specific a fluorescent antibodies. Since the macrophages phagocytized cells in the organ, it

still contains some mRNA original from the dead cell. The use of fluorescent probes followed by fluorescent activated cell sorting (FACS) sorting should separate the macrophages. It is still unclear whether this method will provide the accuracy and throughput required.

3. Estimating the size of the mutant colony from the fraction of mutated DNA in the macrophage:

CDGE allows the quantification of mutant fractions. However, what is needed to measure the mutation rate is the size of various mutated colonies. A link must be established between the mutant fraction in the DNA inside a macrophage and the size of the mutant colony on the site from where the macrophage comes.

4. Defining for which genes will be used to calculate the mutation rates:

CDGE can only be used for one DNA fragment at a time. Furthermore, the use of CDGE requires the a priori knowledge the sequence that is being analyzed. Therefore, the set of gene fragments that will be analyzed will have to be determined a priori. Mutation rates for different DNA locus can be different [reference missing]. Thus the gene fragments could introduce a source of bias in the measurement.

5. Increasing the throughput of CDGE:

Using a standard 96 well thermo-cycler, it is possible to analyze 96 DNA pools in 1 hour. Each pool can contain up to 10^4 different DNA strands for a single mutant to be detected. If one needs to analyze 100 sectors per organ, 5 organs, 100 gene fragments per person, the number of pools to analyze would become 50 000 per person. This amounts to roughly 500 hours of CDGE per person. It is yet unclear how many people the study should involve. However, it is clear the throughput of CDGE will have to be increased. The easiest way could simply be to increase the number of capillaries in one thermo-cycler.

The study would be possible provided solutions are found to the here-above stated challenges.

Conclusion

The difference in risk between the familial and general population is found to be significant for prostate, breast and colon cancer. Three methods were used to reject the null hypothesis. The first used the definition of risk as the parameter of a binomial distribution and calculated the probability distribution of the risk using an assumption on the a priori distribution of the risk. The second method assumed the number of cases to be Poisson distributed and as a consequence so is the risk. The third method is an extension of the first method . The calculated probability distribution of the risk is used to calculate the ratio between the familial and general population risk. The probability that the ratio is greater than 1 is a test of the null hypothesis under the assumption that the ratio remains constant throughout the age groups. Of these three methods, the first and second are independent. Thus the null hypothesis is tested and rejected by at least 2 independent means.

The ratio between the familial and the general population risk is calculated. The calculated ratio is observed to remain constant. The behaviors or the ratio between the risks of breast cancer for the earlier age groups could indicate a change in the ration in the ages before menopause. Even so, the evidence is not sufficient to reject the possibility that the ratio remains constant for breast cancer as well.

The data shows that pairwise comparison between the ratios taken for various age groups within a cancer cannot establish that they are significantly different. Thus a method of regression is used to calculate the ratio across all age groups. This total ratio is found to be approximately 2. No significant difference was found between the values of the ratios of prostate and colon cancer. There is one pairwise comparison in breast cancer that would indicate a significant difference. It is not possible to conclude that this difference is not due to chance alone. Thus the general conclusion is that the ratio between familial cancer risk and general population risk remains constant over all observed ages.

The observation of constant ratio is not compatible with a difference in preneoplastic growth rate between he familial and general population (μ). It is not compatible either with a variation in the fraction of people dying of diseases with shared risk (f) between both populations.

Mathematical modeling using the 2-stage model shows that to explain familial risk, the number of stem cells in the average organ should double between the familial

and general population. These findings restricted the explanation of familial risk to the fraction of people at risk and the mutation rates.

Studies of rare early onset familial colon cancers (due to a mutation in the APC gene) suggest that the fraction of people at risk is 1 for colon cancer. Therefore leaving mutation rates as the only explanatory factor for familial cancer risk.

Finally, a method to measure mutation rates in the general population is outlined. Carrying out the described experiment would measure the distribution of mutation rates in a population. The comparison of the mutation rates distribution in the familial and general population would confirm or reject hypothesis here proposed and give insights into the process of carcinogenesis.

Citations

Altieri, A & Hemminki, K, 2007. Number of siblings and the risk of solid tumours: a nation-wide study. *British journal of cancer*, 96(11), pp.1755–1759.

Altieri, Andrea, Bermejo, J.L. & Hemminki, K., 2005. Familial risk for non-Hodgkin lymphoma and other lymphoproliferative malignancies by histopathologic subtype: the Swedish Family-Cancer Database. *Blood*, 106(2), pp.668–672.

ARMITAGE, P., 1953. A note on the time-homogeneous birth process. *Journal of the Royal Statistical Society. Series B (….*

ARMITAGE, P. & DOLL, R., 1957. A two-stage theory of carcinogenesis in relation to the age distribution of human cancer. *British journal of cancer*, 11(2), pp.161–169.

Bermejo, J.L. & Hemminki, Kari, 2005. Familial lung cancer and aggregation of smoking habits: a simulation of the effect of shared environmental factors on the familial risk of cancer. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology*, 14(7), pp.1738–1740.

Brooks, D.G., 2004. The neurofibromatoses: hereditary predisposition to multiple peripheral nerve tumors. *Neurosurgery clinics of North America*, 15(2), pp.145–155.

BUELL, P. & DUNN, J.E., 1965. CANCER MORTALITY AMONG JAPANESE ISSEI AND NISEI OF CALIFORNIA. *Cancer*, 18, pp.656–664.

Cairns, J., 1975. Mutation selection and the natural history of cancer. *Nature*,

255(5505), pp.197–200.

Cleaver, J., 2000. ScienceDirect - Journal of Dermatological Science : Common pathways for ultraviolet skin carcinogenesis in the repair and replication defective groups of xeroderma pigmentosum. *Journal of dermatological science*.

Couto, E. & Hemminki, K, 2007. Estimates of heritable and environmental components of familial breast cancer using family history information. *British journal of cancer*, 96(11), pp.1740–1742.

Ekstrom, P.O. et al., 2008. Analysis of mutational spectra by denaturing capillary electrophoresis. *Nature Protocols*, 3(7), pp.1153–1166.

Engholm, G. et al., 2010. NORDCAN--a Nordic tool for cancer information, planning, quality control and research. *Acta oncologica (Stockholm, Sweden)*, 49(5), pp.725–736.

Fearnhead, N. & Wilding, J., 2002. Genetics of colorectal cancer: hereditary aspects and overview of colorectal tumorigenesis. *British medical bulletin*.

Goldin, L.R. et al., 2005. Familial aggregation and heterogeneity of non-Hodgkin lymphoma in population-based samples. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology*, 14(10), pp.2402–2406.

Gostjeva, E. et al., 2006. Bell-shaped nuclei dividing by symmetrical and asymmetrical nuclear fission have qualities of stem cells in human colonic embryogenesis and carcinogenesis. *Cancer genetics and cytogenetics*, 164(1), pp.16–24.

Gostjeva, E. et al., 2009. Metakaryotic stem cell lineages in organogenesis of humans and other metazoans. *Organogenesis*.

Hemminki, K, 2002. Cancer risks in second-generation immigrants to Sweden - Hemminki - 2002 - International Journal of Cancer - Wiley Online Library. *International Journal of Cancer*.

Hemminki, Kari & Chen, Bowang, 2004a. Familial association of colorectal adenocarcinoma with cancers at other sites. *European journal of cancer (Oxford, England : 1990)*, 40(16), pp.2480–2487.

Hemminki, Kari & Chen, Bowang, 2005a. Familial association of prostate cancer with other cancers in the Swedish Family-Cancer Database. *The Prostate*, 65(2), pp.188–194.

Hemminki, Kari & Chen, Bowang, 2004b. Familial risk for colon and rectal cancers. *International journal of cancer Journal international du cancer*, 111(5), pp.809–810.

Hemminki, Kari & Chen, Bowang, 2004c. Familial risk for colorectal cancers are

mainly due to heritable causes. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology*, 13(7), pp.1253–1256.

Hemminki, K & Chen, B, 2004d. Familial risk for colorectal cancers are mainly due to heritable causes. *Cancer Epidemiology Biomarkers & Prevention*, 13(7), p.1253.

Hemminki, Kari & Chen, Bowang, 2005b. Familial risks for colorectal cancer show evidence on recessive inheritance. *International journal of cancer Journal international du cancer*, 115(5), pp.835–838.

Hemminki, Kari & Czene, K., 2002. Attributable risks of familial cancer from the Family-Cancer Database. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology*, 11(12), pp.1638–1644.

Hemminki, Kari & Li, Xinjun, 2004a. Age-specific familial risks for renal cell carcinoma with evidence on recessive heritable effects. *Kidney international*, 65(6), pp.2298–2302.

Hemminki, K & Li, X, 2004b. Familial risk in testicular cancer as a clue to a heritable and environmental aetiology. *British journal of cancer*, 90(9), pp.1765–1770.

Hemminki, Kari & Li, Xinjun, 2004c. Familial risks of cancer as a guide to gene identification and mode of inheritance. *International journal of cancer Journal international du cancer*, 110(2), pp.291–294.

Hemminki, Kari, Li, X. & Czene, K., 2002. Cancer risks in first-generation immigrants to Sweden. *International Journal of Cancer*, 99(2), pp.218–228.

Hemminki, K, Sundquist, J. & Ji, J., 2007a. Familial risk for gastric carcinoma: an updated study from Sweden. *British journal of cancer*, 96(8), pp.1272–1277.

Hemminki, Kari, Ji, J. & Försti, A., 2007b. Risks for familial and contralateral breast cancer interact multiplicatively and cause a high risk. *Cancer research*, 67(3), pp.868–870.

Hemminki, K et al., 2001. The nation-wide Swedish family-cancer database--updated structure and familial rates. *Acta oncologica (Stockholm, Sweden)*, 40(6), pp.772–777.

Herrero-Jimenez, P. et al., 2000. Population risk and physiological rate parameters for colon cancer. The union of an explicit model for carcinogenesis with the public health records of the United States. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 447(1), pp.73–116.

Hiripi, E. et al., 2009. Familial association of pancreatic cancer with other malignancies in Swedish families. *British journal of cancer*, 101(10), pp.1792–1797.

Hoffman, F.L., 1915. *The mortality from cancer throughout the world*,

Ji, J. & Hemminki, Kari, 2006. Familial risk for histology-specific bone cancers: an updated study in Sweden. *European journal of cancer (Oxford, England : 1990)*, 42(14), pp.2343–2349.

Kaelin, W.G., 2007. The von Hippel-Lindau Tumor Suppressor Protein and Clear Cell Renal Carcinoma. *Clinical Cancer Research*, 13(2), pp.680s–684s.

Knudson, A G, 1971. Mutation and cancer: statistical study of retinoblastoma. *Proceedings of the National Academy of Sciences of the United States of America*, 68(4), pp.820–823.

Knudson, Alfred G, 2001. Two genetic hits (more or less) to cancer. *Nature Reviews Cancer*, 1(2), pp.157–162.

Kwak, E.L. & Chung, D.C., 2007. Hereditary colorectal cancer syndromes: an overview. *Clinical Colorectal Cancer*, 6(5), pp.340–344.

Leiter, U. & Garbe, C., 2008. *Advances in Experimental Medicine and Biology* J. Reichrath, ed. New York, NY: Springer New York.

Lichtenstein, P. et al., 2000. Environmental and heritable factors in the causation of cancer--analyses of cohorts of twins from Sweden, Denmark, and Finland. *The New England journal of medicine*, 343(2), pp.78–85.

Lorenzo Bermejo, J. & Hemminki, Kari, 2005. Familial risk of cancer shortly after diagnosis of the first familial tumor. *Journal of the National Cancer Institute*, 97(21), pp.1575–1579.

Meza, R., LUEBECK, E. & MOOLGAVKAR, S., 2005. Gestational mutations and carcinogenesis. *Mathematical biosciences*, 197(2), pp.188–210.

Moolgavkar, S.H. & Knudson, A G, 1981. Mutation and cancer: a model for human carcinogenesis. *Journal of the National Cancer Institute*, 66(6), pp.1037–1052.

Morgenthaler, S. & Thilly, W., 2007. A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: A cohort allelic sums test (CAST). *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 615(1-2), pp.28–56.

Mousavi, S.M., Sundquist, J. & Hemminki, K., 2011. Does immigration play a role in the risk of gastric cancer by site and by histological type? A study of first-generation immigrants in Sweden. *Gastric cancer : official journal of the International Gastric Cancer Association and the Japanese Gastric Cancer Association*.

NORDLING, C.O., 1953. A new theory on cancer-inducing mechanism. *British journal of cancer*, 7(1), pp.68–72.

PLANT, R.K., 1955. Genital tuberculosis. *Western journal of surgery, obstetrics, and*

gynecology, 63(2), pp.81–87.

Sudo, H. et al., 2008. Fetal-juvenile origins of point mutations in the adult human tracheal-bronchial epithelium: absence of detectable effects of age, gender or smoking status. *Mutation research*, 646(1-2), pp.25–40.

Thomas, D.B. & Karagas, M.R., 1987. Cancer in first and second generation Americans. *Cancer research*, 47(21), pp.5771–5776.

Tomita-Mitchell, A. et al., 1998. Single nucleotide polymorphism spectra in newborns and centenarians: identification of genes coding for rise of mortal disease. *Gene*, 223(1-2), pp.381–391.

Vatland, J.A., 2001. Analysis of community cancer mortality rates.

Yamasaki, L., 2004. *Cancer Treatment and Research* D. A. Frank, ed. Boston: Kluwer Academic Publishers.

Zheng, W. et al., 2006. Origins of human mitochondrial point mutations as DNA polymerase gamma-mediated errors. *Mutation research*, 599(1-2), pp.11–20.

Aknowledgments

This master thesis would not have been possible without the help and contribution of various people. I would therefore like to thank those that have helped me during this year of research:

Prof William G. Thilly, for his supervision at MIT, continuous support, invaluable contributions and tough scientific discipline imposed on the research.

Prof. Stephan Morgenthaler, for creating the connection to Prof. Thilly, meticulous supervision from EPFL, counseling in moments of distress and mathematical teachings.

Prof. Kari Hemminki, for hosting me in Heidelberg, allowing me to access the Swedish family cancer database and helping me to understand the data used.

Dr. Per O. Ekstrom, for hosting me in Oslo and teaching me to perform high throughput DNA mutation spectroscopy using CDCE.

Appendix

Data for mortality of the general and familial population

Follows the raw data of mortality that were used in the analysis of familial risk. For each cancer will be presented the number of deaths for the familial and general population within each year for each age. Both the familial and general population number represents the number of deaths in the offspring population (i.e. the offspring of recorded parents).

On the horizontal axis are the calendar years and on the vertical axis the age of death.

Female Breast cancer, general population:

	1964	1976	1978	1979	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008
13		1																															
15				1																													
17									1																								
21					1																												
28													1							1													1
29													1								1								1				
30																								1									
31																					2												
32	1						1																					1					
33																1		2															
34															1			1			1	1											

35																1					2	1	1	1									
36																1	1								1	2		1	1				
37																	1	2	2			1	2		1			2					
38							1	2																1									
39											1			1	1	1	3	2			3		2	1	1	1		1	1	2			
40																1	1	1			1	1		1					2	1			
41								1					1		2	1	4		1	5	3		3	2	1		1	1		1			
42							1							3			4	5	1	1	1	2	3		1	2			1		1		
43								1				1			1	2	4	4	1	4	2			3	3	1	2	1			1	2	1
44													1			3	5	5	4	2	2	2	3	1	3	1	3					1	
45					1	1	2									4	6	6	9	2	4	2		1	1	1	1	3		1			3
46			1							1					1	4	4	10	4	2	2	1	2	2	3	1	3	4		1		2	1
47						1					1	1	1		2	5	11	1	9	8	7	5	5	5	5	5	2	5	2	2	3	1	1
48									1	1						5	7	11	4	2	4	2	6	5	3	2		2	1	1	4	1	
49							1			1		1		2	2	6	10	8	6	10	11	6	3	9	7	2	3	3	5	3		1	
50																6	13	9	6	6	7	9	4	6	5	3	6	5	2	1	3		
51													1		1	8	8	7	10	10	11	5	4	9	3	6	4	5	2	5	2	1	1
52																5	12	6	4	11	18	12	7	6	6	8	6	5	5	2	1	2	3
53													1			7	8	13	8	9	7	10	7	4	7	6	3	8	5	5		1	2
54														1		7	13	17	10	3	6	15	13	14	10	8	7	5	5	5	3	4	1
55																4	11	10	9	8	11	6	15	15	9	12	16	3	5	10	7	1	5
56																4	11	9	9	6	15	8	6	12	20	12	6	9	10	10	6	6	4
57																6	12	11	8	10	6	9	11	11	11	14	8	6	12	9	6	4	4
58																5	12	17	9	4	11	13	9	10	15	9	14	13	14	7	7	6	7
59																1	7	8	12	15	11	12	10	15	9	16	15	10	10	8	8	9	6

60																		9	7	6	17	14	13	10	11	11	9	10	5	15	9	13	13	15
61																			3	2	11	8	12	10	9	12	13	10	9	10	14	8	10	12
62																				3	4	10	13	16	7	15	7	6	14	11	17	22	11	5
63																					1	12	7	11	6	11	10	8	7	11	12	11	8	9
64																						2	9	5	5	20	6	13	17	12	9	17	8	11
65																							3	9	6	7	9	14	6	11	10	14	13	16
66																									10	2	13	11	12	18	14	12	14	10
67																									2	6	7	5	16	11	16	10	21	16
68																										1	8	11	21	12	9	14	10	15
69																											2	12	6	20	13	10	14	17
70																												3	10	12	13	17	8	18
71																													2	9	10	14	13	6
72																														4	8	21	13	13
73																															4	13	10	7
74																																5	6	13
75																																	6	8
76																																		5

Female breast cancer, familial poulation

	1976	1983	1987	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008
13	1																					
28																						1
29										1												
32																	1					
34							1															

35										1												
36															1		1					
37											1	1					1					
39															1			1	2			
40							1															
41									1	1		2		1					1			
42						1	2					1										
43		1	1		1				1				1									
44						1	1						1			1						
45					1			2														
46				1	2	1			1		1			1	1							
47						1		1	1	1	1	1	1		1	1	1		1	1	1	
48							1				1			1	1				1			
49							2		1	1			2	1			1	4				
50								1		1		1		1			1			1		
51						3	2		1			1	1				1				1	
52									1	1	1							2	1	1		
53						2		1		1	1		1			1		1	1			1
54						2	1				2	1	1	1	2					2		1
55						1	1	2				2	3	2	1	3		1				
56							2			2			1	1	2				1		1	
57									1	1		1			3	3		2	2	1	2	1
58							2	2	1			1	1	3		1	1	2		1		1
59							1		1		1	1				2	1	1			2	
60										1	1			2		2		1	1		1	1

61									2	1		1				1			1		1	
62									1						1	2	1		1		3	
63												1		2					1			2
64													1				1					1
65															1			2		1		
66													1		2			2	2	1	1	1
67														1		1		1	1		2	2
68																2	1				1	
69																	1		2	1		
70																			1	1		1
71																		1	1	2	2	
72																				3		2
73																					1	1
74																						1
75																					1	

Male prostate cancer, general population

	1988	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008
41					1				1						1	1				
43																		1		
44			1																	
45					1	2			1											
46		1																		

47					1									2	1					
48					3					1					1			1		
49					1		2	1			1	1	2	2	2					
50	1				1		3	2	6	2	2	3		1						
51			4	1	4	3	2	2	2	2	3	2		2	2	1	1		1	1
52			1	3	1	2	2	2		1	2	2	2		3		1		4	
53		1	1	4	7		1	3	3	4	1	3		2	1		6	2	3	1
54			2	2	3	2	2	3	7	4	3	3	1	2	1	4		2	2	4
55		1	1	1	4	4	3	4	3	2	4	2	3	7	2	6	5		2	1
56		2	1	6	5	4	4	3	2	3	5	5	3	3	8	4	2	4	5	7
57			2	2	5	4	4	3	2	5	8	3	7	7	10	8	7	4	5	2
58			3	3	6	3	4	3	7	3	5	3	10	13	6	7	7	5	8	4
59				1	4	6	4	7	7	7	13	7	6	10	9	5	13	7	12	8
60					5	7	3	10	14	9	10	7	6	13	13	13	9	13	10	13
61					2	8	8	9	7	7	11	7	13	11	21	11	16	16	12	9
62							7	7	8	7	9	17	13	10	16	15	11	8	14	12
63							2	8	11	12	9	20	12	11	12	7	12	18	15	23
64								3	8	16	11	15	18	13	12	19	14	28	22	14
65									1	6	11	13	20	14	23	20	21	14	15	20
66										2	7	21	20	18	25	21	26	15	10	26
67											2	10	19	23	17	22	24	21	22	21
68												2	15	13	20	32	34	21	24	23
69													8	16	18	33	33	32	32	26
70														3	11	22	17	20	33	32
71															10	18	27	31	26	26

72																12	31	26	35	43
73																	12	17	33	29
74																		10	26	36
75																			12	27
76																				13

Male prostate cancer, familial population

	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008
41				1															
43																	1		
45					1														
48				1															
49										1		1							
50								2											
51		1		1	1	1				1			1						
52					1						1					1			
53			1					1	1	1									
54					1	1	1	1										1	1
55	1			1	2		1					1		1				1	
56			1			1				1	1	1		3				2	
57			1	1							1	2	1	2	1		1		
58								1				1				1			1
59					1				1	1				2		2		1	
60				1	1	1	1	1	1	1	2		1	2		3	3	1	2

61						1	1		1		2	1	1	2	1	2		1	1
62						1	2	1			2		1	3		1	1	2	2
63									1	1	1	2	1	2		1	1	1	4
64							1				3		1	1	3	3	3	1	
65										1	1	1	2	5		1	1	1	3
66												2	1	2		4	4		3
67										1			1		2	1	3		
68												2	4	2	1	3	3	3	3
69														2	3	3	3	2	1
70														2		1	2	3	
71														2	3	3	2	6	1
72																3	1	1	3
73																		2	5
74																		1	5
75																			1
76																			1

Male colon cancer, general population

	1961	1963	1967	1971	1973	1974	1975	1978	1979	1980	1981	1982	1983	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008
19																			1																		
20																											1										
21	1																																				
26			1								1																										
28			1																													1					

29		1			1							1																										
33														1																								
34						1														1																	
35				1														1												1							
37																1					1																
38							1			1	1												1	1										1				
39					1				1		1				1							1		2	1		1		1			1						
40					1			1					1										1	1					1	1							
41									1							1	1		1	2	1	1	1		2	1		1									
42																	1		1	2	1						1				1			1			
43																	1				4	1	1	1			2	1				1					
44									1				1		1	1					1	1		1	1	2		1				1					
45										1							1	1	2	1	1		1			1	1	1			1						
46													1								3	2	1	3	1	1	1		1	2		1		1				
47																	1			1	2	1	1	1	2		1	1		1	1			1			
48														1				1		3	3	1	3	1	1		2	3	1	1	1				1			
49														1		1				1	2	2	2	2	3	1		2		3	3	1	2		2			
50															1						5	1	2	2	3	1	1		1	2	2	1			2	1		
51															1		1				2	3	2	1	4	4		4	2	4	4	2	2		2	1		
52																					4	2	1	2	1	3	4		3	3	4	2	1		2		1
53																	1			3	6	3	2	6	1	3	2	2	4	4	3	4	2	1	1	1	1
54																	1		1	4	9	1	2		4	5	6	7	7	2	3	4	4	4	3	4	1
55																				3	2	7	4	4	4	3	4	3	3	5	3	7	3	3	4		
56																				1	6	5	2	9		3	2	5	4	2	4	5	2	3	6	2	1
57																				2	5	4	5	4	3	4	4	11	7	3	5	1	3	4	5	4	6

58																				1	7	5	2	3	4	3	3	9	5	7	4	7	8	3	7	8	4
59																					1	5	2	4	7	5	5	5	3	5	12	6	9	7	4	5	3
60																					2	2	6	8	5	5	6	3	10	7	6	3	7	7	8	2	8
61																							4	5	5	5	8	7	3	11	9	7	6	9	9	5	11
62																							1	5	7	4	12	15	6	5	7	7	12	13	10	7	10
63																								3	3	5	12	9	14	12	13	4	13	7	11	2	10
64																										3	3	7	11	15	9	5	11	6	5	13	12
65																										2	6	9	5	13	13	12	10	13	10	9	15
66																											1	2	7	7	15	9	11	16	13	7	10
67																													11	8	8	17	11	9	17	10	8
68																													1	4	11	13	15	9	13	17	17
69																														3	6	9	15	10	14	15	10
70																															2	8	13	7	16	12	21
71																																1	14	14	9	15	21
72																																	5	9	5	18	14
73																																		5	13	11	15
74																																			4	6	13
75																																				2	11
76																																					2

Male colon cancer, familial population

	1989	1990	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008
40				1															
42													1						

45	1	1																	
46					1				1										
48				1			1						1						
49						1				1		1							
50																	1		
51										1		1							
52			1									1							
53			2		1						1								
54					1			1	1	2									
55			1											1					
56													1				1		
57				1				1		1			1			1	1		1
58											1					1		1	
59						2	1	1		1	1		1			1		1	
60										1		1						1	
61								1					1			1			
62						1		1		1		1			2		3	1	3
63									2		1				1				1
64												1	1		1	1		1	
65									1		1		2	1	1		1	2	
66												1	1			1		1	1
67															3		2		
68														3	1	1			3
69														1	1				1
70																1		2	

71															1			1		
72															1					
73																1	1	1		
74																	1		1	
76																			1	

Female colon cancer, general population

	1964	1974	1976	1977	1978	1979	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008
26										1																									
28									1																										
29	1																																		
31																																			1
32																		1																	
35			1																																
36		1													2																1				
37		1			1													1				1	1												
38										1				1							1	1													
39										1		1										1	1												
40				1						1									1	1			1				1				1				
41						2																		1		1									
42																1		1		1	1	1	1	1	2				1	1					
43							1											1	1							1	1	1							
44								1										1	2		2					1				2	1				
45																		1	1	2	1	1	1	2		2	3			1		1			

46													3			1		3	4		1		2	1		3	1		1			1	2	1	
47									1				1			1				1	3	1			1	1							1		
48														1		1		3	2		2	4	1	2	1	1	2				2	1			1
49													1		2	2	1	1	4	1	2	3	3			2	1	3		1	2				
50												1				1		1	4	7	1	1	4	1	2	1	1		1			1	1	3	1
51											1							3	4	2	5	4	5	6	1	3	2	2	1	1	1			1	
52																		3	2	2	2	2	2	4	4		5	2		1	4			1	1
53																1			4	1		4	5	3	7	3		4	3	2	1	1	2	1	
54															1		1	3	6	3	2	6	5	3	6	2	1	5	4	2	1	1	5	1	2
55																		2	4	9	5	3	6	2	5	4	4	6	9	2	2		2	1	3
56																	2	3	8	7	1	9	6	3	6	2	4	2	4	3	6	6	3	3	2
57																		2	4	6	5	5	6	4	6	3	7	6	3	2	4	7	3	1	7
58																		1	1	5	2	3	4	4	5	4	5	3	7	4	4	3	7	7	5
59																			1	2		3	7	4	6	7	5	5	4	7	9	3	2	7	3
60																			1	4	6	6	8	14	3	3	4	6	5	6	4	8	5	5	7
61																				1	1	5	5	5	4	9	7	5	11	6	2	8	5	4	5
62																					1	5	5	7	9	8	14	6	7	11	11	7	6	10	5
63																						1	2	12	6	5	6	7	5	11	7	6	13	10	9
64																							1	3	6	6	6	8	12	10	7	3	8	8	8
65																								2	6	6	9	10	5	8	8	7	8	12	12
66																									1	4	8	9	8	10	6	11	9	7	10
67																										2	6	11	8	6	6	9	10	12	13
68																											2	6	6	7	10	10	12	10	8
69																												2	7	11	10	15	11	17	14
70																													3	6	6	9	16	17	19

71																														4	10	9	8	8	12
72																															2	7	7	13	13
73																																2	3	16	11
74																																	4	7	9
75																																		2	11
76																																			2

Female colon cancer, familial population

	1964	1982	1989	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008
28		1																			
29	1																				
36																	1				
40					1																
41										1											
42							1	1													
44												1									
45										1											
47							1				1										
48					1																
49									2				1								
50					1													1			
51									1	2				1							
52						1							1								
53			1			1												1			

54								1	1		1			1	1						
55								1	2					2	1						1
56																1				2	
57				1	1					1			1	2	1			1		1	
58															2				1		
59													1					1			
60															1						
61																	1				
62															2		2	1	1		1
63										3			2					1		1	
64															1				1	1	
65															2			2		1	1
66											1	2	1			1		1		1	
67													1						1		
68														1	1		2				
69															1	1	1	1	1		
70																1					
71																	1		1		
72																	1	1			
73																			1	1	1

Kini et al. 2011

This work has been submitted for publication to the British journal of cancer. Reference to it is made during the thesis; it is therefore attached in its complete form

Fetal/Juvenile Tumor Initiation in Human Colorectal Cancer

Lohith G. Kini[1], Pablo Herrero-Jimenez[2], Jayodita Sanghvi[3], Efren Gutierrez, Jr.[4], David Hensle[5], John Kogel[5], Rebecca Kusko[6], Karl Rexer[7], Ray Kurzweil[8], Elena V. Gostjeva[1], Stephan Morgenthaler[9] and William G. Thilly[1*]

[1]Department of Biological Engineering, Massachusetts Institute of Technology, 16 - 771, 77 Massachusetts Ave., Cambridge, MA 02139; [2]SLC Ontario, 690 Dorval Drive, Suite 200, Oakville, Ontario, L6K 3W7, Canada; [3]Department of Bioengineering, Stanford University, 232 Aryshire Farm Lane #103-A, Stanford, CA 94305, [4]Massachusetts General Hospital for Children, 55 Fruit Street, Boston, MA 02115, [5]30 W Oak Street, Apt. 22B, Chicago, IL 60610, [5]211 East Ohio Street, Apt. 1122, Chicago, IL 60611, [6]Boston University Medical Center, 715 Albany Street, Boston, MA 02118, [7]Rexer Analytics, 30 Vine Street, Winchester, MA 01890, [8]Kurzweil Technologies, Inc.15 Walnut Street, Wellesley, MA 02481, [9]Department of Mathematics, Ecole Polytechnique Federale Lausanne, EPFL MA -103, 1015 Lausanne, Switzerland

Keywords: tumor initiation | cancer models | fetal/juvenile mutation | metakaryotic stem cells

Classification: BIOLOGICAL SCIENCES: Medical Genetics, Hematology and Oncology

Abbreviations: APC, adenomatous polyposis coli; FAPC, familial adenomatous polyposis coli; OAT, sialylmucin O -acetyl transferase

ABSTRACT

Age-specific colorectal cancer rates increase sub-exponentially from maturity, reach a maximum, then decline in old age. The "two-stage" model of Armitage and Doll related adult cancer rates to exponential growth rates of preneoplastic colonies and the number and rates of mutations necessary to begin preneoplasia (initiation) and then neoplasia (promotion). Initiation mutations were assumed to occur throughout adult life in a population homogeneous for risks but the model did not account for the maximum, then declining rates observed. This discordance has been ascribed to diagnostic errors in old age, stratification of population risks and/or limitation of initiation mutations to the fetal/juvenile period. Now, however, nuclear and mitochondrial mutations in adult lungs have been found distributed as Luria-Delbruck expansions that indicate restriction of initiation mutations to stem cells of the fetal/juvenile period, specifically to newly-discovered mutator/hypermutable *metakaryotic* stem cells. A computer program, CancerFit v5.0, has been developed to test the fetal/juvenile initiation hypothesis. Best fits with lifetime incidence data are obtained with (a.) *APC* gene mutation rates derived from observations in FAPC patients, (b.) two, and only two, initiation events and (c.) preneoplastic stem cell doubling rates approximating that of juvenile growth. Concordance is further improved by population risk stratification consistent with variation in adenoma numbers observed among FAPC patients. Public health implications: conditions affecting stem cell fetal/juvenile oncomutation rates should potentiate adult cancer rates. They may account for shifts of organ-specific cancer rates among immigrants and exponentially decreasing untreatable pediatric tumor rates during the past seventy years.

INTRODUCTION

The rates of most forms of cancer decline from the first year of life to a minimum in mid- juvenile years, increase sub-exponentially into old age then decline in extreme old age [1, 2]. Armitage and Doll [3] sought to reconcile the increasing rate of cancers with adult age in terms of new understandings of genetics and genetic change of the 1950s. They posited that any single cell at risk in a static adult cell population could be "initiated" by "n" required genetic events. An initiated cell founded an exponentially growing preneoplastic colony in which any cell at risk was "promoted" to a tumor-founding cell by "m" independent events. Their original "two-stage" model of initiation and promotion assumed that initiation mutations occurred throughout life in adult tissues only and that the population was homogeneous with regard to oncomutation and preneoplastic growth rates. Their model did not predict the observed maximum cancer rate of old age. (A brief history of quantitative cancer models is offered in Supporting Information.)

More than fifty years after their pioneering postulates, the number, nature and origins of oncomutations that *initiate* the slow growth of preneoplastic lesions and later *promote* a preneoplastic cell to the rapid growth of neoplastic lesions remain obscure for most common cancers, e.g. lung, breast, prostate, pancreas. Clinical observations have shown that preneoplastic lesions of several organs continue to grow after maturity and give rise to lethal neoplasias through extreme old age. Promotional genetic changes, if any, must therefore occur throughout life in preneoplastic colonies. But genetic or epigenetic events have not yet been found that fulfill the logical requirements for a rare human promotion mutation: each tumor cell of the tumor and its derived metastases must carry the rare promoting genetic/epigenetic event(s).

However, for a few common cancer types, e.g. colorectal, kidney, nervous system, and skin, autosomal tumor suppressor genes have been identified both copies of which must be genetically inactivated for initiation to occur [1]. We chose one of these, colorectal cancer, for this study.

Familial heterozygosity for the *APC* gene is fully penetrant; all heterozygotes display multiple adenomas and, if untreated, adenocarcinomas [2]. This indicates that for most colorectal tumors the number of required initiating mutations "n" ≥ 2. The fact that colorectal preneoplastic or neoplastic colonies can grow in all persons in which

they arise indicates that the population is not stratified relative to growth of preneoplastic or neoplastic colonies. But the range of hundreds to thousands of adenomas observed in colons among individuals inheriting an inactivated *APC* allele suggests population stratification with regard to APC somatic initiation rates [1, 2].

Based on evidence drawn from epidemiology, clinical histopathology and somatic genetics we and others have postulated that colorectal cancer initiating mutations in the *APC* gene occur only in the fetal/juvenile period [4, 5, 6] specifically in a mutator/hypermutable metakaryotic stem cell lineage [7, 8, 9, 10] and that the majority of point mutations inactivating this gene are attributable to copying errors of DNA polymerase β on undamaged DNA [11]. Here we provide (a.) newly organized data on mortality from lower gastrointestinal tract cancer, (b.) a model of carcinogenesis in which initiation is limited to the fetal juvenile period, (c.) a suitable computer program and (d.) tests of the concordance of predictions of the model with observed cancer rates.

Mortality Data Base: historical age-specific cancer rates.

U.S. cancer mortality numbers and populations recorded 1900-2006 by the U.S. Census Bureau (1900-1935) and the U.S. Public Health Service (1936-2006) have been matched and organized with regard to gender, ethnicity, calendar interval of birth, "**h**", (ten years: 1800-09, 1810-19, …), calendar year interval of death, "**y**" (five years, 1900-04, 1905-09,…), and age at death interval, "**t**" (five years, 0-4, ….100-104)) [11, http://mortalityanalysis.mit.edu] (Supporting Information Table S1). These data allow computation of raw age-specific lifetime mortality rates, OBS(h,t), as the number of deaths by the observed cause divided by the number of persons alive at the beginning of the one-year interval "**t**". Thus OBS(h,t) is an approximation of the conditional probability that a person would have died of the observed cause given that he or she was still alive. However, cancer models predict incidence rates, INC(h,t), as a calculated approximation, CAL(h,t), of conditional rates of deaths absent covariant factors such as competing forms of death or the effect of medical intervention in the age/time interval observed. Transforming observed raw mortality rates, OBS(h,t), to estimates of incidence rates, INC(h,t), requires correction for several sources of bias. In extreme old age (t = 100-104) death rates approach ~0.3 per year and must have reduced the number of deaths by the observed cause. Correction for this bias consists

of determining the total raw mortality rate for each five year age interval, TOT(h, t), and defining the coincidence-corrected mortality rate at the third or middle year, OBS*(h,t) as OBS(h,t) /[1-TOT(h,t) + OBS(h,t)]. Accounting for historically improving five-year survival rates, SUR(h,t), is also required for some cancers such as colorectal cancers [12, 13, 14]. The expected incidence rate, INC(h, t), adjusted for these considerations is:

INC(h,t) = OBS(h,t)/([1-SUR(h,t)][1-TOT(h,t) + OBS(h,t)]). **Equation 1**

Diagnostic errors at death may also be expected and these would vary among cancer types, age at death, historical year of reporting etc. so that INC(h,t) as defined here is an approximation and its uncertainties must be considered in comparing predictions of models, CAL(h,t), to incidence represented by INC(h,t).

Lower gastrointestinal tract cancer mortality in the U.S., 1900-2006.

Deaths from cancer of the lower gastrointestinal tract (present ICD9 codes: 152, 153, 154) are predominantly deaths from colorectal cancer [http://mortalityanalysis.mit.edu]. Figure 1a uses a semi-log plot to illustrate the transformations from raw mortality rates to incidence rates: from OBS(h=1890-99,t) to OBS*(h=1890-99,t), then to INC(h=1890-99, t). Also shown is the result of the same transforms of the succeeding decade, INC(h= 1900-09,t), which yielded similar estimates of for all values of "**t**". (Supporting Information Table S2)

In Figure 1b the OBS*(h,t) vs. "t" is presented on a semi-log scale for each of successive birth cohort intervals, "h", from 1800-09 through 2000-06. These data demonstrate that colorectal cancer death rates in older adults (t > 65) rose throughout the birth cohorts of the 19th century reaching an approximately stable maximum in and after the birth cohort of 1880-89. In Figure 1c these same data are plotted so that changes OBS*(h,y) as a function of age of death interval, "t" may be seen as functions of historical year of death interval "y". This form of presentation shows a significant decrease in older adult death rates ascribable in whole or part to advances in medical practice after 1950 that increasingly effect each successive birth cohort (Supporting Information Table S2, Figure 1) [http://mortalityanalysis.mit.edu, 12, 13].

In Figure 1d the data are plotted as OBS*(y, t) vs. "y" to illustrate the historical changes within each age of death interval, "t" as a function of historical year of death

interval, "y". This form of presentation reveals that colorectal cancer mortality rates in children and young adults have decreased exponentially from ~ 1940 through 2006.

Algebraic elements of the two-stage model.

Limitation of initiation mutations to the fetal/juvenile stem cell doublings.

Growth of normal fetal/juvenile stem cells is here modeled as a series of "**a**" net binomial doublings (a = 0, 1, 2, …, a_{max}) in which "**n**" required initiation mutations, i, j, …n, occur in any order at constant mutation rates R_i, R_j, …,R_n per doubling [13, 14]. The number of newly initiated stem cells in doubling period "**a**" is n ($\prod_n R_i$) $a^{(n-1)}$ 2^a. In the fetal/juvenile model organogenic stem cells are posited to reach maturity represented by "$\mathbf{a_{max}}$" doublings with high constant mutation rates and to undergo metamorphosis to maintenance stem cells with no net additional net cell growth and much lower mutation rates [9].

Assuming each of the ~10^7 adult colonic crypts to be represented at juvenile/adult metamorphosis by a single metakaryotic stem cell, the number of net doublings at maturity, a_{max}, is about 23.25, i.e., $10^7 \sim 2^{23.25}$ [13,14]. The metakaryotic mutator/hypermutable stem cell lineage of human organ anlagen appears to begin in gestational week 4-5 with creation of two metakaryotic stem cells from symmetrical amitosis of a single precursor embryonic mitotic stem cell at a = 0 [10]. At birth, a colon contains ~2^{20} colonic crypts each containing a basal metakaryotic stem cell [10]; thus at birth, a ~ 20, at maturity, a = a_{max} ~ 23.25.

Promotion mutations during preneoplastic stem cell doublings.

After initiation in any fetal/juvenile doubling "**a**" growth of preneoplastic stem cells as a colony is modeled as a series of "**g - a**" binomial doublings (g-a = 0, 1, 2, …) in which "**m**" required promotion mutations (A, B, …m, occur at constant mutation rates R_A, R_B, …,R_m per doubling) [13,14]. The expected number of newly initiated stem cells in preneoplastic doubling period "**g-a**" is $\mathbf{m\ (\prod_m R_A)\ (g\text{-}a)^{(m-1)}\ 2^{(g-a)}}$**.** Under these assumptions the number of organogenic doublings "**a**" at initiation and the number of preneoplastic doublings "**g-a**" after initiation sum to "**g**" which is a very useful continuous variable because it describes the age of humans in terms of continuous stem cell doublings through fetal/juvenile and then preneoplastic growth. In each organogenic doubling interval "**a**" new preneoplastic colonies are created (initiated) and these colonies grow until promotion and subsequent death remove

them. The extinction of preneoplastic colonies at "**a**" and at "**g - a**" is driven by the supra-exponential term **exp[-m ($\prod_m$ R_A $(g-a)^{(m-1)}$ $2^{(g-a)}$]).**

If all persons have the same numbers and rates of "**n**" required initiation and "**m**" required promotion oncomutations and all initiated cells grow at the same average rate as preneoplastic stem cells (homogeneous risk) the expected number of promotional events at the binomial doubling age interval "g", V(g) may be represented as:

$$V(g) = n (\prod_n R_i) \int(a=0, a=a_{max}) a^{(n-1)} 2^a d(1-\exp[-m \prod_m R_A (g-a)^{(m-1)} 2^{(g-a)}]) /d(g-a)\ da$$

$$\approx n (\prod_n R_i) \sum(a=0, a=a_{max}) a^{(n-1)} 2^a d(1-\exp[-m (\prod_m R_A (g-a)^{(m-1)} 2^{(g-a)}]) /d(g-a)$$

Equation 2

This process is illustrated in Figure 2 in which the contribution to promotion at age "g" from initiation at each organogenic doubling "a" is shown to rise and fall with "g-a". The sum of these terms from initiations in all organogenic doubling intervals "a" approximates well the observed lifetime incidence rate of many cancer types including colorectal cancer: it increases sub-exponentially, reaches a maximum in old age and declines appreciably in extreme old age. The earliest initiations of fetal organogenesis drive the tumor incidence rate of juveniles and young adults, the initiations of adolescent organogenesis drive the tumor incidence rate in extreme old age.

In addition to homogeneous risk it is here assumed *pro tempore* that there are no synchronously competing forms of death with shared risks for colorectal cancer and that all events of initiation and promotion are randomly assorted among persons as in a Poisson distribution. Under these conditions the expected number of newly promoted lesions through the end of any doubling interval "g", CAL(g), is:

$$CAL(g) = (1-e^{-V(g)})$$ **Equation 3**

*Age of death, "**t**", and doubling age of promotion, "**g**".*

Cancer mortality data corrected for coincident deaths within the year of death, OBS*(h,t) and its derived estimate of incidence, INC(h,t) are calculated in five year

age-of-death intervals 5-9 ,…., 100-104 years such that deaths in any age interval are plotted at the mid interval , i.e. deaths within the interval 50-54 are plotted at 52.5 years. CAL(h, g) is, however, approximated as the instantaneous rate of promotion at the *end* of each stem cell doubling interval "**g**".

To account for the difference between adult age at promotion and death we adopt Armitage and Doll's [3] estimate of 2.5 yr, which is in accord with estimates of net tumor growth rates that would produce a gross untreated tumor mass in excess of three kilograms [12, 13]. Death at age t = 72.5 is thus attributed to promotion at age t = 70.

The relationship between human age at death in years, "**t**", and stem cell doubling age at promotion, "**g**" is then defined if there is a constant average adult preneoplastic stem cell annual doubling rate, "**μ**". Given the age of maturity for males as 16.5 yr [13, 14] at $g = a_{max}$ and an average interval of 2.5 yrs between promotion and death absent medical intervention:

$$g = \mu\,(t - 16.5 - 2.5) + a_{max} = \mu\,(t - 19) + a_{max} \qquad \textbf{Equation 5}$$

Stratification of risks in the population.

We have previously represented the fraction of the population in whom all of the potential conditions necessary for cancer death are present as "F" imagining that a person either is or is not at risk for a necessary oncogenic process. The corresponding fraction in which one or more necessary condition is absent has been represented as (1-F) [12, 13, 14]. Stratification need not, however, be an "all or none" phenomenon. Children's grow to different sizes, which may create stratification with regard to the maximum number of stem cells at risk of initiation $2^{a\ max}$. Children grow at different rates and so may the preneoplastic lesions in different persons. Stratification with regard to mutation rates in fetal/juvenile expansion has been noted for both mitochondrial and nuclear genes [9]. In progress is an effort to incorporate stratification with regard to initiation and promotion oncomutation rates and the growth rates of preneoplasias. The use of "F" in this present report serves as first approximation in stratification accounting insofar as it values of (1-F) > 0 could represent in part a fraction of persons that are not initiated greater than expected by a Poisson distribution of initiations among all persons with a single set of initiating

mutation rates. Equation 4 rewritten to account for stratification in this way creates the model:

$CAL(g) = F(1-e^{-V(g)})/ [F + (1-F)\, e^{\int V(\gamma)\, d\gamma}]$ evaluated from $\gamma = 0$ to g. **Equation 6.**

Competing synchronous forms of mortality.

Epidemiological observations have also demonstrated that forms of cancer may share environmental or inherited risk factors with another, e.g. breast and ovarian cancers, in which the death rates increase synchronously with age [http://mortalityanalysis.mit.edu]. The term "f" has been introduced to represent the fraction of persons that die of the observed cause among the set of mortal diseases with shared risks and synchronous changes in death rates [12, 13, 14]. Equation 6 rewritten to account for both stratification and a hypothetical synchronous competing form of mortality with shared risk factors with the observed disease in this way creates the model:

$CAL(g) = F(1-e^{-V(g)})/ [F + (1-F)\, e^{\int 1/f\, V(\gamma)\, d\gamma}]$ evaluated from $\gamma = 0$ to g. **Equation 7.**

RESULTS

Application of the model to age-specific colorectal cancer incidence in a specific cohort. To test the null hypothesis that the function CAL(h,t) calculated from stipulated parameters did not differ significantly from the observed age-specific incidence function INC(h,t), a means was required to assess the variation expected by chance among independent trials. As an approximation of this variation we compared age-specific incidence rates of lower G.I. tract cancer observed for European-American males born 1890-99 to the rates for those born 1880-89 and 1900-1909 (Figure 1a) as shown in Supporting Information Table 2. This comparison comprises any real changes in colorectal cancer incidence during the twenty year historical interval (expected to be negligible by inspection of Figure 1a) sampling error arising from the smaller numbers of recorded death in early and late adult life, and biases arising from historically changing diagnostic accuracy and estimates of five-year survival rates. This comparison yielded a GOF(h,t) of 0.043 that serves here as a best available approximation of the minimum GOF(h,t) that may be expected between any

model and the age-specific incidence data. This is akin to an average standard deviation of +/- 5 % for the estimate for each of eighteen five-year age of death intervals. Not included in this estimate are any age-specific ascertainment biases such as may be expected if the accuracy of cause of death diagnoses declines in extreme old age relative to younger ages. Were an underestimate of 30% postulated for the four oldest age of death intervals then the GOF(h,t) of an accurate model would be about 0.08. Given these boundaries it seems appropriate to reject the null hypothesis if GOF(h,t) exceeds 0.16 for any set of model conditions as this accounts to a reasonable degree for both uncertainty in the accuracy of the incidence data and the errors associated with sampling.

First, the best fits of CAL(h=1890-99, 15< t <104) were calculated for the twenty-five combinations of n = 1-5 and m = 1-5 under the parsimonious conditions of homogeneous risk and no synchronous mortal diseases sharing risk factors with colorectal cancer. Values of $(\prod R_i)^{1/n}$and $(\prod R_A)^{1/m}$ were permitted to range from 10^{-9} to 10^{0} and the range of μ was set at 0.1 to 0.3. The complete matrix of results is provided in Supporting Information as Table S3. The value of 0.052 was the minimum GOF(h,t) observed for n = 4, m =2. For n = 2, an appropriate biological value if only the loss of both copies of the *APC* gene were necessary and sufficient for initiation in most colorectal cancers, and m = 1, a default assumption absent any evidence about mutations required for promotion, the GOF(h,t) was 0.085. It should be noted that discordance was greatest at t >75 yr where underestimation of colorectal cancer as a cause of death by as much as 30% has been suspected in extreme old age [13-16].

Second, we observed the best fits of CAL(h,t) under the additional assumption of inhomogeneous risk, i.e., the parameter "F" representing a hypothetical fraction of the population at risk was allowed to range from 0 to 1. The values of GOF(h, t) ranged from 0.035 (n = 3, m = 4) to 0.064 (n = 3, m = 5). The value of 0.043 was the minimum GOF(h,t) observed for n = 2, m = 1. Further discussion of "F' and the complete matrix of these results is provided in Supporting Information as Table S4.

Thirdly, we considered the possibility of both population inhomogeneity and a competing synchronous mortal disease having genetic and/or environmental risks shared with colorectal cancer, i.e., the parameter "f" representing this possibility was

allowed to range from 0 to 1 as discussed in Supporting Information. This assumption did not, however, further reduce the values of GOF(h,t).

Figure 3 depicts the degree of concordance of the two trial conditions: F = 1, f = 1 (population homogeneity, no synchronous competing risk) and F < 1, f =1 (population inhomogeneity, no synchronous competing risk)) with adult lifetime incidence data for lower G.I. tract cancer in European American males born 1890-99 INC(h,t). The model presented herein incorporating the observation of limitation of mutation to the fetal/juvenile period but without the assumption of risk stratification fits the data well except at extreme old age when it predicts incidence rates as much as 30% higher than recorded. Assuming accuracy of the data in old age, when the model accounts both for the *fact* of fetal/juvenile mutation limitation [17] and the *possibility* of risk stratification in the population (F <1) fits are sensibly improved.

DISCUSSION

We find that the observed lifetime adult age-specific incidence of colorectal cancer in European-American males born 1890-99 or 1900-09 are wholly consistent with initiation limited to the stem cells of the fetal/juvenile period. We further find that when *APC* initiation mutation rates derived from observations in FAPC patients are employed the model is consistent with observation only if n = 2 initiation mutations are employed.

Fetal/juvenile mutation.

Early models posited that a constant number of adult cells were at continuous risk of acquisition of a required set of oncomutations that resulted in lethal tumors [18-20]. Platt's questions about selective growth advantage of early oncomutants [21] stimulated the creation of the two-stage model of Armitage and Doll in which initiation was restricted to a constant number of adult cells [3]. An alternate theory, fetal/juvenile or, more narrowly, gestational initiation was based on Luria's demonstration that in a bacterial culture growing from a small number of cells early mutations give rise to large mutant colonies and late mutations to small ones [22]. This concept was clearly applicable to the stem cells of a growing organ. By maturity a stem cell initiated early in development would create a large preneoplastic colony, an initiation in the last juvenile doubling would create a single initiated stem cell. Early cancers would arise from early developmental initiation, cancers of extreme old

age would arise from the last juvenile initiation events [4, 5, 6, 23]. In the algebraic model provided here each successive doubling of stem cells of the fetal/juvenile period provides initiated colonies. Deaths from colonies initiated in any specific doubling interval are distributed over subsequent periods of life (Figure 2). Biological data supporting limitation of mutations to the fetal/juvenile period were scant but suggestive. New colonic adenomatous polyps are more rarely detected after age sixty [24] whereas if tumor initiation continued throughout life, a continuing increase in polyps would be expected [7]. The first direct evidence of a juvenile limitation of a human parenchymal tissue mutation was offered by Brash and Pontén who reported that increases in mutant *TP53* colonies in human skin were restricted to the juvenile years and subsequent solar exposure in adults increased mutant colony size but not number [25, 26]. More recently, the distribution of five nuclear and seventeen mitochondrial point mutations in adult human lung epithelium was found to match a simple Luria-Delbruck expansion of mutant stem cells for ten stem cell doublings prior to maturity without further increase in adult life [9, 25].

Mutator/hypermutable metakaryotic stem cells of organogenesis and carcinogenesis.

The observed high fetal/juvenile mutation rates have been associated with amitotic, non-eukaryotic cells that arise in the 4^{th}-5^{th} week of gestation that appear to serve as the stem cells of human organogenesis and carcinogenesis [7 – 10]. These large "metakaryotic" cells with bell shaped nuclei found in human fetal/juvenile organs, preneoplastic lesions and neoplasias increase by symmetric amitotic fission and have been observed to produce all parenchymal, subsequently mitotic, cell forms by asymmetrical mitotic fissions in tissues derived from all three primordial germ layers [10]. Metakaryotes display a bizarre mode of DNA segregation that occurs prior to, or concomitant with, DNA replication in sister cells [10] that appears to involve DNA copying by DNA polymerase β insofar as about half of cancer-initiating *APC* point mutations sampled are attributable to errors of DNA polymerase β copying undamaged DNA in vitro [11].

Environmental cancer risks during the fetal/juvenile period.

Agreement between the origins of adult somatic mutations in mutator/hypermutable fetal/juvenile human stem cells and age-specific cancer rates offers an explanation of

epidemiological associations between fetal and early childhood exposure to known mutagenic stimuli, particularly sunlight [26, 27].

Generational changes of age and organ-specific cancer rates in immigrant populations towards those of the new country of residence may also be thought of in terms of fetal/juvenile initiation mutations. Adult immigrants would have experienced tumor initiation in the country of origin while early *in utero* immigrants would experience the tumor initiations specifying risk after age 20 in the new country. In both older and younger immigrants the environment of the new country would define promotional stresses that might act by inducing promotional oncomutations and/or by selecting through growth stimulation conditionally initiated stem cells acquired in the fetal/juvenile period [28].

Similarly, the marked decrease in death rates from lower G.I. tract (Fig. 1d.) and other incurable cancers in children and young adults may be attributable to a decline in fetal/juvenile initiation rates that began circa 1940 and continued through 2006 [http://mortalityanalysis.mit.edu].

Save for sunlight exposure of juveniles, however, there is no direct evidence that environmental mutagens cause mutations in humans [29]. “Spontaneous” mutation caused by simple DNA polymerase mis-incorporation errors or errors following DNA damage by endogenous processes may account for all fetal/juvenile mutations save for juvenile mutation by sunlight[11, 30]. Environmental mutagens may have been or still may be indirectly responsible for some of these oncomutations

The assumption of population homogeneity.

If it is posited that the population is in some way inhomogeneous with regard to risk, $F < 1$, the fit to the recorded INC(h,t) is sensibly improved (Figure 3). One form of risk stratification is represented by the ~tenfold range in colonic adenomatous polyp numbers observed among individuals of FAPC families [2] that indicate a range of fetal/juvenile *APC* mutation rates. A similar variation of numbers of mutant crypts displaying loss of the second allele of sialylmucin O-acetyl transferase (*OAT)* has been observed in adult colon [31]. The discrepancies between the two-stage model adjusted for fetal/juvenile initiation and the colorectal cancer incidence data may thus lie principally in the assumption of a homogeneous oncomutation rate that assumes a Poisson distribution of initiating events among persons in each organogenic doubling period. The wide and approximately rectangular distribution of estimated lung

fetal/juvenile mutation rates suggests that the fraction of persons without any initiation events by maturity is significantly greater than predicted by the assumption of mutation rate homogeneity [9]. This error is compensated in part by positing a value of F <1 (Figure 3).

The assumption of synchronous competing forms of death.

Assuming a competing synchronous form of death with shared risk(s) with colorectal cancer does not further increase goodness of fit under best-fit conditions (Figure 3, Supporting Information Table S3) in accord with epidemiological studies of familial colorectal cancer [17]. This possibility must however be considered for cancers such as breast and ovarian cancers in which synchronous age-specific death rates display shared familial risk.

Number and rates of required initiating mutations.

Patients heterozygous for the *APC* gene are reported to have some "hundreds to thousands" of adenomatous polyps in adulthood presumably resulting from a single mutation inactivating the single active inherited *APC* allele [1, 2]. The last stem cell doubling would be expected to account for half of the total polyp number, the penultimate doubling for one quarter and so forth. With an estimated geometric mean of ~1000 polyps among 10^7 total colonic crypts we may therefore estimate that the rate of loss/inactivation of the normal parental *APC* allele occurs at ~$(1000/2)/10^7 = 5 \times 10^{-5}$ events per stem cell doubling. The similarly estimated geometrical mean rate of loss of the second allele of sialylmucin O-acetyl transferase (*OAT)* is also ~5×10^{-5} [31].

The term $((\prod R_i)^{1/n})$ represents the geometrical mean of "n" initiating mutations. Estimates of this parameter derived from best fits of the model given the assumption of population stratification for n = 2 and m = 1-5 were 3.5, 2.3, 2.8, 2.8 and 4.3×10^{-5} respectively (Figure 3) [12] agreeing well with estimates from clinical enumeration of mutant polyps or crypts [1, 2]. For n = 1, m =1, F <1, the rate estimate of $((\prod R_i)^{1/n})$ is about 5.3×10^{-8}, for n = 3, m = 1, F < 1 about 2.8×10^{-4}. Values of the geometrical mean of initiation mutations assuming values of n other than 2 are thus clearly discordant with *APC* and *OAT* colonic mutation rate estimates. These facts, derived from clinical genetic observations in both inherited and sporadic forms of colorectal cancers are wholly consistent with the conclusion that n = 2 and inconsistent with values of $n \neq 2$.

Estimation of preneoplastic growth rate, μ.

Using the best fit stipulated values for n = 2, F < 1 and any value of m = 1-5, the preneoplastic growth rate "μ" was estimated to be ~0.18. This is close to the juvenile growth rate of mass in males, 0.158, and females, 0.167 [12-14].

Numbers and rates of required promotion mutations.

It does not yet seem possible to estimate "m" or the related geometrical mean of promotion mutation rates, $((\prod R_A)^{1/m})$. Reasons for this limitation are discussed in Supporting Information.

SUMMARY

Quantitative analysis of age-specific colorectal cancer incidence rates indicate that they are wholly consistent with the hypothesis that initiation mutations occur only in the fetal/juvenile period. When matched to fetal/juvenile *APC* mutation rates derived from observation of polyp number in FAPC patients the number of initiation mutations is consistent with two and only two rare initiation events. Population risk stratification for initiation mutation rates is indicated by a wide distribution of mutant numbers in adult colons and accounting for said stratification improves the concordance with age-specific colorectal cancer rates.

ACKNOWLEDGMENTS

We gratefully acknowledge the contributions of the late Professor Lars Ehrenberg (Wallenberg Laboratory and the University of Stockholm) and those of Dr. Suresh Moolgavkar (Fred Hutchinson Cancer Research Center, Seattle, WA.). Jonathan Jackson and Aaron Fernandes contributed to the initial design of CancerFit. Five of us (LK, SM, RK, EVG, WGT) were supported in part by a research contract to M.I.T. from United Therapeutics, Inc., Silver Spring, MD for the study of human carcinogenesis.

REFERENCES

1. Knudson AG. (2001) Two genetic hits (more or less) to cancer. Nat Rev Cancer 1:157-162.
2. Kwak EL and Chung DC. (2007) Hereditary colon cancer syndromes: an overview. Clin. Colorectal Cancer 6:340-344.
3. Armitage P, Doll R. (1957) A two-stage theory of carcinogenesis in relation to the age distribution of human cancer, Br. J. Cancer 92:161–169.

4. Thilly WG. (1988) Looking ahead: Algebraic thinking about genetics, cell kinetics and cancer. *In:* H Bartsch, K Hemminiki, and IK O'Neill, eds., *Methods for Detecting DNA Damaging Agents in Humans: Applications in Cancer Epidemiology and Prevention*, IARC Monograph No. 89, IARC Scientif. Publ., Lyon, pp. 486-92.
5. Moolgavkar SH. (1988) Biologically motivated two-stage model for cancer risk assessment Toxicol. Lett. 43: 139-150
6. Thilly WG. (1991). What actually causes cancer? Technology Review, Mar. / Apr. 48-54.
7. Gostjeva EV, Thilly WG. (2005) Stem cell stages and the origins of colon cancer: a multi-disciplinary perspective. Stem Cell Reviews, 1:243-252.
8. Gostjeva EV, Zukerberg L, Chung D, Thilly WG. (2006) Bell-shaped nuclei dividing by symmetrical and asymmetrical nuclear fission have qualities of stem cells in human colonic embryogenesis and carcinogenesis. Cancer Gen. Cytogen. 164:16-24.
9. Sudo H, Li-Sucholeiki X-C, Marcelino LA, Gruhl AN, Herrero-Jimenez P, Zarbl H, Willey JC, Furth EE, Morgenthaler S, Coller HA, Ekstrom PO, Kurzweil R, Gostjeva EV, Thilly WG. (2008) Fetal-juvenile origins of point mutations in the adult tracheal-bronchial epithelium: absence of detectable effects of age, gender or smoking status. Mutat. Res. 646:25-40.
10. Gostjeva EV, Koledova V, Tomita-Mitchell A, Mitchell M, Goetsch MA, Varmuza S, Fomina JN, Darroudi F, Thilly WG. (2009) Metakaryotic stem cell lineages in organogenesis of humans and other metazoans. Organogenesis 5: 109-118.
11. Muniappan BP, Thilly WG. (2002) The DNA polymerase beta replication error spectrum in the adenomatous polyposis coli gene contains human colon tumor mutational hotspots. Cancer Res. 62:3271-3275.
12. Herrero-Jimenez P. Determination of the historical changes in primary and secondary risk factors for cancer using U.S. public health records. (2001) MIT Sc.D. Thesis.
13. Herrero-Jimenez P, Thilly G, Southam PJ, Tomita-Mitchell A, Morgenthaler S, Furth EE, Thilly WG. (1998) Mutation, cell kinetics, and subpopulations at risk for colon cancer in the United States. Mutation Research, 400:553-578.
14. Herrero-Jimenez P, Tomita-Mitchell A, Furth EE, Morgenthaler S, and Thilly WG. (2000) Population risk and physiological parameters for colon cancer. The union of an explicit model for carcinogenesis with the public health records of the United States. Mutation Research, 447:73-116.
15. Berge T, Lundberg S. Cancer in Malmo 1958-1969. An autopsy study. Acta Pathol Microbiol Scand Suppl 1977 260:1-235
16. Linstrom P, Janzon L, Sternby NH. Declining autopsy rate in Sweden: a study of causes and consequences in Malmo, Sweden. J Intern Med (1997) 242: 157-165
17. Hemminki K, Granstrom C, Chen B. (2005) The Swedish family-cancer database: update, application to colorectal cancer and clinical relevance. Hered. Cancer Clin. Pract. 3:7-18.
18. Holloman JH, Fisher JC. (1950) Nucleation and growth of cell colonies Science 111: 489-491
19. Nordling CO. (1953) A new theory on the cancer-inducing mechanism, Br. J. Cancer 7: 68–72.

20. Armitage P, Doll R. (1954) The age distribution of cancer and a multi-stage theory of carcinogenesis, Br. J. Cancer 81:1–12.
21. Platt R. (1955) Clonal Aging and Cancer. The Lancet. 265: 867.
22. Luria SE, Delbruck M. (1943) Mutations of bacteria from virus sensitivity to virus resistance. Genetics 28: 491-511.
23. Meza R, Luebeck EG, Moolgavkar SH. (2005) Gestational mutations and carcinogenesis. Math Biosci 197:188-210
24. Atkin W, Rogers P, Cardwell C, Cook C, Cuzick J, Wardle J, Edwards R. (2004) Wide variation in adenoma detection rates at screening flexible sigmoidoscopy. Gastroenterology 126, 1247-56.
25. Coller HA, Khrapko K, Herrero-Jimenez P, Vatland JA, Li-Sucholeiki X-C, Thilly WG. (2005) Clustering of mutant mitochondrial copies suggests stem cells are common in human bronchial epithelium. Mutat. Res. 578:256-271.
26. Brash DE, Rudolph JA, Simon JA, Lin A, McKenna Gj, Baden HP, Halperin AJ, Ponten J. (1991) A role for sunlight in skin cancer: UV-induced p53 mutations in squamous cell carcinoma. Proc Natl Acad Sci USA 88:10124-10128
27. Brash DE, Ponten J. (1998) Skin precancer. Cancer Surv. 32:69-113.
28. Hemminki K, Ji J, Brandt A, Mousavi SM, Sundquist J .(2010) The Swedish Family-Cancer Database 2009: prospects for histology-specific and immigrant studies. Int J Cancer 126: 2259-2267.
29. Thilly WG. Have environmental mutagens caused oncomutations in people? (2003)Nature Genetics 34:255-259.
30. Zheng W, Khrapko K, Coller HA, Thilly WG, Copeland WC.(2006) Origins of human mitochondrial mutations as DNA polymerase gamma-mediated errors. Mutat. Res. 599: 11-20.
31. Campbell F, Williams GT, Appleton MA, Dixon MF, Harris M, Williams ED. (1996) Post-irradiation somatic mutation and clonal stabilisation time in the human colon. Gut 39: 569–573.

FIGURE LEGENDS

Figure 1. Lower gastrointestinal cancer in European American males, calendar years of death, 1890 to 2006 . 1a. Transformations linking raw mortality rates, OBS(h,t), with estimates of incidence rates, INC(h,t). Shown are the raw mortality rate function, OBS(h,t), for h = 1890-99 corrected for coincident forms of death which becomes the function OBS*(h,t) which, in turn, corrected for historically increasing survival rate, becomes the incidence rate function, INC(h,t). Also shown is INC(h=1900-09, t). **1b.** OBS*(h, t), coincidence corrected death rates for birth cohort intervals, h = 1800-09, … , 2000-06, vs. age intervals of death, “t” = 0.5, 1, 2, 3, 4, 5-9, …, 100-104. **1c.** OBS*(h,y), coincidence corrected death rates for birth decade intervals, h vs. calendar year of death intervals , “y”=1900-1904, …, 2000-2004, 2005-2006. **1d.** OBS*(t,y), coincidence corrected death rates for intervals of age at death “t” vs. calendar year of death intervals, “y”.

Figure 2. V(a -> g). Expected number of lethal events/person at lifetime stem cell division "g", arising from initiation at fetal/juvenile division "a". Here it is assumed that g = a = 0 at the embryonic turning point ~4-5 weeks of gestation, g = a = 20 at birth (t = 0) and g = a = a_{max}= 23.25 at maturity (t =17.5). In adult males "g" increases with age according to the relationship g = μ(t - 19) + 23.25. Note that earliest initiations presage fetal, juvenile and young adult deaths. Late initiations account for deaths in extreme old age. Note also that in this example the SUM of all values V(a->g) at stem cell division age "g", increases sub-exponentially from g = 23 to a maximum at g = 37 and then declines rapidly. In this example g = 37 represents ~ 105 yrs of age. [Stipulated parameter values for this example: n = 2, m = 1, μ= 0.16, $(R_iR_j)^{1/2}$ = 2.2 x 10^{-5} and R_A = 4.4 x 10^{-5} with homogeneous risk (F = 1) and absence of competing synchronous forms of death with risk factors shared by colorectal cancer (f=1)].

Figure 3. Comparisons of observation INC(h,t) to CAL(h,t) for the two-stage cancer model with varying assumptions. n = 2, m = 1 is stipulated for all examples. (a.) CAL(h,t), homogeneous risk, no synchronous competing form of death (parsimonious model, F = 1, f = 1). GOF(h,t) = 0.085. (b.) CAL(h,t), heterogeneous risk (F = 0.43, no synchronous competing form of death (f = 1)) . GOF(h,t) = 0.043. For CAL(h,t), heterogeneous risk (F < 1), synchronous competing forms of death (f < 1). GOF(h,t) = 0.041. See Supporting Information (Table S4) for definitions and ranges of "F" and "f" employed.