TEST FOR DECREASE IN AGE AT DIAGNOSIS OF LYNCH SYNDROME OVER SUCCESSIVE GENERATIONS

YANJING HE

Test for Decrease in Age at Diagnosis of Lynch Syndrome over Successive Generations

by

©Yanjing He

A thesis submitted to the School of Graduate Studies
in partial fulfillment of the requirement for the Degree of
Master of Science in Statistics

Department of Mathematics and Statistics
Memorial University of Newfoundland

May, 2011

St. John's Newfoundland and Labrador Canada

Abstract

The study of age at onset anticipation and parent-of-origin effects on age at onset in Lynch Syndrome (LS) are of interest to both clinical medicine and research. Although several studies have suggested the presence of age at onset anticipation and parent-of-origin effects on age at onset of LS, the question remains as to whether this evidence reflects ascertainment bias rather than the phenomenon under study. The aim of this thesis is to assess decrease in age at diagnosis of LS over successive generations as well as parent-of-origin effects on age at diagnosis of LS based on the data provided by the Colon Cancer Family Registry. We first demonstrate that the variable age at diagnosis in the sample is right truncated by the closing date of the study and, as a result, the variable age at diagnosis is a biased sample of the target populations. To assess decrease in age at diagnosis of the disease over successive generations, we use the symmetry test proposed by Tsai et al. (2005) which accounts for the bias caused by the right truncation of both the parent's and child's ages at diagnosis. To test parent-of-origin effect, we examine and improve the method used by Lindor et al. (2010). Based on our preliminary analysis, we did not find sufficient statistical evidence from this sample to claim that there exists a parent-of-origin effect on age at diagnosis of LS relating to either the gender of the parent or the gender of the offspring after accounting for the sampling bias. The results given by the symmetry test suggest that there exists a decrease in age at diagnosis of LS over successive generations. This result should be free of the sampling bias caused by the right truncation. What remains uncertain is whether true genetic anticipation contributes to the decrease in age at diagnosis over successive generations observed in this disease.

Contents

List of Tables

List of Figures

Acknowledgements

I would like to thank my supervisor Dr. J Concepción Loredo-Osti and co-supervisor Dr. Michael Woods for all their support and guidance through all the stages of my master program, especially for providing me the opportunity to work on this interesting data. Their efforts were fundamental for the completion of this thesis.

I sincerely thank readers Dr. Kenneth Morgan and Dr. Hong Wang for their valuable comments and constructive suggestions. I also would like to express my thanks to Dr. Yunqi Ji and Ms. Mary Fujiwara for their help with my thesis proofreading and editing. Their effort have improved the thesis a lot. Their influence on the thesis can be seen throughout. I appreciate their hard and valuable work.

Department of Mathematics and Statistics has provided the support I have needed to produce and complete my thesis. I would like to thank all the people working in the department.

Chapter 1

Introduction

1.1 Lynch Syndrome and Age at Onset Anticipation

Lynch Syndrome (LS) is an autosomal, dominantly inherited, colorectal cancer predisposition syndrome, exhibiting high penetrance (80% - 85%) and accounting for 2% - 10% of the total colorectal cancer burden [Lynch and de la Chapelle (1999)]. LS patients typically develop colorectal cancer (CRC) at an early age (mean age 45 years). In addition to colorectal cancer, the tumor spectrum includes cancers of the endometrium, stomach, small bowel, ovary, ureter/renal pelvis, brain, hepatobiliary tract, and skin. LS is caused by germline mutations in the *DNA mismatch repair* (MMR) genes *MSH2*, *MLH1*, *MSH6*, and *PMS2*, with *MLH1* and *MSH2* accounting for more than 90% of all germline mutations identified [Westphalen et al. (2005)].

Genetic anticipation is a term that refers to a tendency for the onset of a genetic disease to occur at progressively earlier ages or with progressively greater severity in successive generations. That is, if the offspring of patients develop the disease, they will tend to do so at an earlier age than their parents. Whether anticipation

actually exists for any disease has been a controversial subject for some time. However, for some diseases, genetic anticipation is a well-recognized clinical feature with a completely characterized molecular mechanism, but LS is not one of these diseases [Gruber et al. (2009)], although genetic anticipation (that is, earlier age at onset of colorectal cancer in offspring) has been postulated to occur in LS. It remains debatable whether successive generations are truly affected at earlier ages than their ancestors and/or whether the severity of the disease is more pronounced. Thus far, only limited and controversial data are available on this issue, ranging from single case reports to a few systematic investigations in LS families [Menko et al. (1993), Rodriguez-Bigas et al. (1996); Tsai et al. (1997); Vasen et al. (1994); and Nilbert et al. (2009)].

Another related topic is parent-of-origin effects on age at onset of LS. A parent-of-origin effect is the phenomenon where even though parental alleles may segregate in a Mendelian fashion, their expression and, consequently, their effect on the trait under study (parent-of-origin effect on age at onset of LS) depends on whether the allele was inherited from mother or father. Parent-of-origin effects relate to the gender of the parent, and in autosomes, are not expected to be associated with the gender of the offspring. Lindor et al. (2010) report that their study of parent-child pairs in which both parent and child were affected by nonsyndromic colorectal cancer showed that the affected offspring of affected fathers were younger on average than offspring of affected mothers (53.7 vs. 55.8 years; $p = 0.0003$). When the data was divided into sons and daughters, the difference was driven by younger age at diagnosis in daughters of affected fathers compared to sons (52.3 years vs. 55.1 years; $p = 0.0004$). That is, an earlier age at diagnosis of colorectal cancer in female offspring appeared to be more pronounced when the disease allele was transmitted from the father compared to transmission through the mother. These findings seem to suggest some genetic factors are associated with the X-chromosome rather than with parent-of-origin effects.

The study of age at onset anticipation, that is, a tendency for the onset of a genetic disease to occur at progressively earlier ages, and parent-of-origin effects are of interest both in the clinical and research settings. If confirmed, these results may have important implications for genetic counseling and clinical management of LS families [Westphalen et al. (2005)]. Individuals with LS are at a high lifetime risk of colorectal cancer. Appropriate surveillance to detect tumours before they progress to late stage colorectal cancer is very important for the welfare of the patient and for cost effectiveness in the health care system. The existence of age at onset anticipation of LS underscores the need to initiate surveillance programs at a young age. "It should also stimulate research into the genetic mechanisms that determine age at onset and whether the genetic instability that characterizes LS can be linked to anticipation" [Nilbert et al. (2009)].

1.2 Statistical Issues on Assessment of Age at Onset Anticipation and the Related Statistical Methods

Apparent anticipation is a problem arising in the statistical assessment of age at diagnosis (onset) anticipation. Apparent changes in age at diagnosis (onset) of disease between generations, which could suggest genetic anticipation (a true biological occurrence), may simply be a statistical artifact of inadequate statistical analysis based on ages at diagnosis in cohorts that have not been followed for a sufficiently long time [Picco et al. (2001)]. Although several studies have suggested the presence of anticipation in LS, the question remains as to whether this evidence reflects ascertainment bias rather than genetic anticipation.

Statistically, decreased age at onset over successive generations could result from inappropriate sampling of family data, that is, sampling bias. Sampling bias is defined as a sampling anomaly that causes some members of the population to be less likely to be included than others. It results in a departure from random sampling of a population causing that all participants are not equally balanced or objectively represented in the sample. If this is not accounted for, results can be erroneously attributed to the phenomenon under study rather than to the method of sampling. It is also referred to as ascertainment bias. This type of bias involves systematic selection of families in which both parental and offspring generations are affected. Two sources of sampling bias may be particularly relevant to the study of the age at onset anticipation. The first source is due to inadequate follow-up time, where persons who have not completed the risk period for a disease are included with those who have. When parents who have passed through most of the period of risk for the disease are compared with children who have not yet completed the risk period, children who are unaffected at the time of analysis but go on to manifest the disease at later ages are not included in the calculation of average age at onset. Consequently, some late-onset cases in younger generations may be missed at the time of ascertainment, which can mimic anticipation. This makes the average age of disease onset in children appear younger than it would be if this group were followed for a longer time, which in turn produces a false impression of genetic anticipation. The second source of sampling bias is fertility bias, where cases with an earlier onset are less likely to have children, thus reducing the probability of finding parent-offspring pairs where the parent shows earlier onset. The resulting data may appear to reflect decreased age at onset over successive generations. In summary, apparent anticipation can be caused by sampling bias. For a sample subjected to the above mentioned sampling bias, appropriate statistical analysis methods should be employed to take the bias into account.

In addition to sampling bias, the effect of secular trends, for example, increased smoking rate, changing dietary habits, or changing quality of health care, can also produce

spurious evidence of age at diagnosis (onset) anticipation.

Finally, though not per se a source of spurious evidence of anticipation, both family-clustered structured data, that is, intra-familial correlation due to shared genotype and/or environment and information in censored observations representing undiagnosed family members can distort the results of an analysis, and therefore must be taken into account when analyzing data.

In summary, when investigating anticipation, a distinction between biological and statistically artificial anticipation must be made. It is possible that apparent genetic anticipation can be explained by sampling bias without invoking any additional genetic influences. Thus, claims for genetic anticipation must be based on methods that properly take into account study design and the duration of observation for all individuals in the study.

Several statistical methods are commonly used to assess age at diagnosis (onset) anticipation. Each method targets specific issues with the data.

If one confines an analysis to affected individuals only, standard statistical methods include the paired t-test for age at onset of affected parent-offspring pairs and non-parametric ANOVA of age at onset of all affected on the predictor generation.

To account for differences in the length of follow-up "at risk" duration between generations, one can introduce information on age at interview to develop a test which incorporates right truncation of the age at onset by assuming that the ages at onset of affected parent-child pairs may be modeled as being right truncated by the age at interview. Huang and Vieland (1997) use this approach and consider that the age at onset of affected parent-child pairs may be modeled as being right truncated by the age at interview and that the age at onset random variable follows a bivariate normal

distribution. They used maximum likelihood methods for truncated data to perform the test. Rabinowitz and Yang (1999) proposed a nonparametric approach for right truncated age at onset data, and their tests represent generalizations of the sign test and the Wilcoxon rank-sum test. Tsai et al. (2005) proposed a simple generalized paired t-test and a Wilcoxon signed-rank test to adjust for the bias caused by the right truncation of both the parent's and child's ages at onset. Alternatively, one can prolong the follow-up period to identify multiple generations with comparable years of follow-up for comparison or use methods that properly account for the duration of observation in all individuals being studied.

To handle the effect of secular trends, the paired t-test may be applied to different birth cohorts.

To deal with family-clustered structured data, i.e., intra-familial correlation due to shared genotype and/or environment, a random effects model that is an extension of the generalized paired t-test has been proposed by Tsai et al. (2005).

If one extends an analysis to include unaffected individuals as right censored data, the log-rank test in univariate survival analysis and semi-parametric survival analysis (Cox proportional hazards model) can be employed to test and assess the difference in age at disease onset between two generations. To handle family-clustered structured data, several methods have been developed, for example, the non-parametric paired test by Hsu et al. (2000), which is a generalization of the log-rank test, Cox proportional hazards models with the so-called robust sandwich estimate of the covariance matrix by Binder (1992) and Haynatzki et al. (2007), and the gamma frailty model by Haynatzki et al. (2007) and Klein (1992). To handle birth cohort effects/secular trends, the effect of birth cohorts/secular trends can be incorporated into the survival analysis as in Daugherty et al. (2005).

1.3 Objectives of the Thesis

Although the objective may be to study age at onset, this variable cannot be measured under the present study design. Instead age at diagnosis is used because it is generally recorded and is considered to be an acceptable proxy for age at onset.

The first objective of this work is to assess the decrease in age at diagnosis of LS over successive generations. The second objective is to evaluate parent-of-origin effects on age at diagnosis of LS over successive generations based on the data provided by the Colon Cancer Family Registry. The third objective is to reveal how apparent changes in age at diagnosis of the disease between generations, which could suggest genetic anticipation, can be an artefact of inadequate analysis based on age at diagnosis in cohorts that have not been followed for a sufficiently long time.

The present study, rather than theoretically developing advanced statistical methods for testing for anticipation, will examine and apply existing methods that are currently used for this purpose to the sample provided by the Colon Cancer Family Registry.

The organization of the thesis is as follows. In Chapter 2, a description of the data set provided by the Colon Cancer Family Registry is given. This is followed by an exploration of the data. The exploration demonstrates that the data is subject to the sampling bias due to the different durations of follow-up time in different generations. In Chapter 3, an analysis of parent-of-origin effects on age at diagnosis of LS over successive generations is presented. In Chapter 4, an analysis of age at diagnosis anticipation of LS over successive generations is presented. Finally, in Chapter 5, we conclude and briefly present possible future work.

Chapter 2

Description and Exploration of the Data Set

2.1 Description of the Data Set

2.1.1 The source of the data

The data was provided by the Colon Cancer Family Registry [http://epi.grants. cancer.gov/ CFR/]. Families were ascertained through the Colon Cancer Family Registry from both population-based and clinic-based sources. Details about the data can be found in Subsection 2.1.2 and Section 2.2.

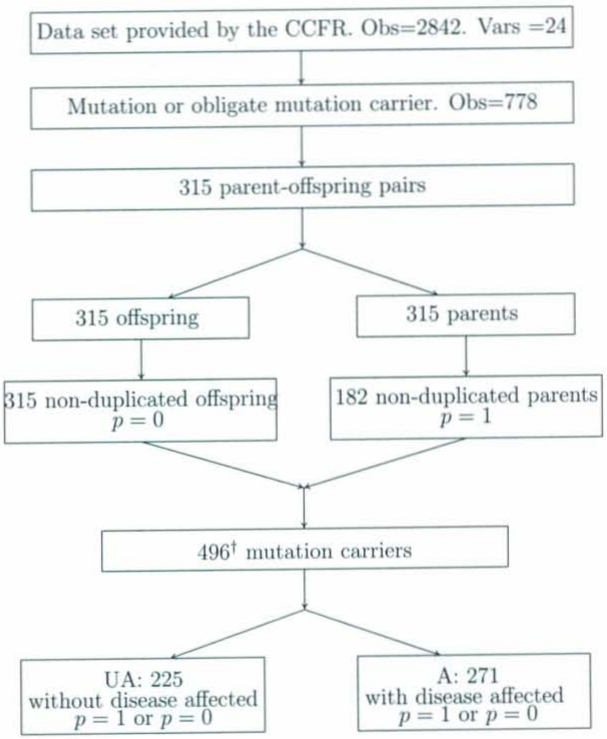
2.1.2 Variables

Variable name	Definition	Values
Centre.no		
FamilyID		
FSRC	1=population-based, 2=clinic-based	
PersonID		
PROBAND.FLAG	1=proband	
MotherID		
FatherID		
Sex		
GENE	mismatch repair (MMR) gene types	*MLH1*, *MSH2*, *MSH6*, *PMS2*
MUT.STATUS	indicates whether the person was identified as MMR mutant. c=carrier, n=non-carrier	{Missing, c, n}
ObligateCarrier	c=obligate carrier	{Missing, c}
mut.descrip	Description of the mutation	
DOB	date of birth: dd/mm/yyyy	
Age.Death.or.Last.Known.Age		
colorectal.CA	Colorectal cancer diagnosis (1=yes, 0=no)	
age.at.colorectalCA	age at first diagnosis of primary colorectal cancer	
endometrial.CA	1=yes, 0=no	
age.at.endometrialCA	age at first diagnosis of primary endometrial cancer	
Prostate.CA	1=yes, 0=no	
age.at.ProstateCA	age at first diagnosis of primary prostate cancer	
otherlynch	Other Lynch cancer (1=yes, 0=no)	
age.at.otherlynch	age at first diagnosis of Lynch cancer	
nonlynch	1=yes, 0=no	
age.at.nonlynch	age at first diagnosis of non Lynch cancer	
Note: Non Lynch = all cancers except colorectal, endometrial, prostate, and other Lynch cancers.		
Note: Other Lynch = stomach, small intestine, kidney, ureter, brain, and ovary cancer.		

Table 2.1: Explanation of variables

2.2 Data Preparation

The data sets A and UA which are relevant to our analysis in later chapters can be obtained from the original raw data set provided by the Colon Cancer Family Registry as follows.



Note: †Age at diagnosis was not available for some individuals, so these individuals were not included in the study. Since some parent-offspring pairs share the same parent, there are only 182 unique parents among 315 parents.

2.3 Exploration of the Data

Figure 2.1 is based on data set A described in the last subsection. Parents are represented by "1", offspring by "0"; "time" represents age at diagnosis, and "DOB" represents date of birth.

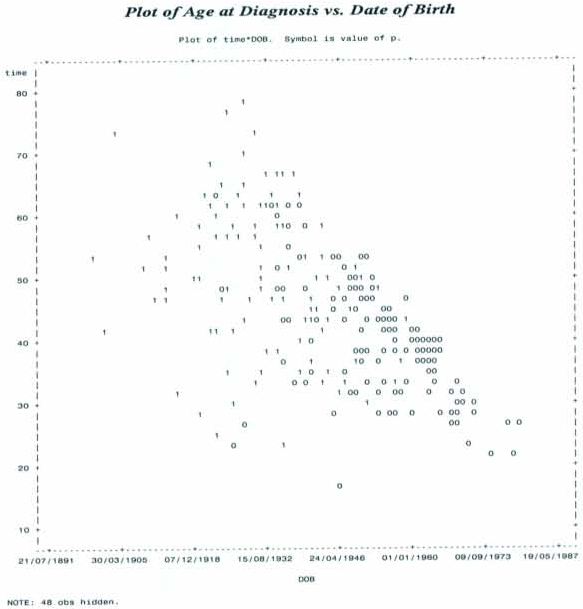


Figure 2.1: Plot of age at diagnosis vs. date of birth

2.3.1 Findings and discussions

From Figure 2.1, we observe that parents ($p = 1$) tend to have later age at diagnosis than offspring ($p = 0$). The average age at diagnosis for parents is approximately 50

years old, while the average age at diagnosis for offspring is approximately 40 years old. We also observed the following :

1. For the disease affected patients born before 1930, thus having completed the risk period for the disease by the year 2009, age at diagnosis mainly lies between 30 and 70, which provides a reference for the risk period for the disease.

2. For the disease affected patients born in and after 1930, the ages at diagnosis are mainly larger than 30 but less than the age given by the formula

$$\text{AGER} \equiv 2009 - \text{their birth year.} \tag{2.1}$$

For example, a disease affected patient born in 1960 has an age at diagnosis less than 49 whereas an affected patient born in 1970 has an age at diagnosis less than 39. This phenomenon is due to the fact that among the mutation carriers born after 1930, who have not completed the risk period for the disease by the year 2009, only early-onset patients, whose age at diagnosis lies within the observational window (0, AGER), are observed and thus are included in data set A. Late-onset patients, whose age at diagnosis are later than AGER, are not observed due to having not been followed up as long as the previous generation and thus are not included in data set A. That is, the variable age at diagnosis is right truncated by the closing date of the study being the year 2009. By definition, right truncation of survival data occurs when only those individuals whose event time lies within a certain observational window (0, YR) are observed. An individual whose event time is not in this interval is not observed and no information on this subject is available to the investigator. This is in contrast to censoring where there is at least partial information on each subject. Because we are only aware of individuals with event times in the observation window, the inference for truncated data is restricted to conditional estimation. Since the age at diagnosis distribution in parents and children are right truncated relative to the distribution of the target population under study, only individuals with age at diagnosis prior to their current age are eligible for

inclusion. As a result, the observed variable age at diagnosis is a biased sample of the target population in which individuals with early age at diagnosis are over-represented relative to others in the target population. The obtained sample actually represents a population other than the target one.

3. Most parents (p=1) were born before 1946 (which also can be seen from Figure 2.2) and thus have a relatively long follow-up time. Therefore, the data are a better representation of the parental generation. However, most offspring (p=0) were born after 1946 and thus have a shorter follow-up time, the data are a poor representation of the offspring generation with an over-representation of individuals with an early age at diagnosis cases. In short, since children are younger than parents, the truncation effect is more pronounced in the children than in the parents.

Since we are only aware of patients with age at diagnosis less than AGER due to the right truncation, the inference for truncated data should be restricted to conditional estimation. Otherwise, the inference based on the observed decrease in age at diagnosis over successive generations from the sample would be valid only to the populations that the sample actually represents but not to the target population. Failing to take this into account, age at diagnosis anticipation could be erroneously claimed. Since individuals with early age at diagnosis in the sample are over-represented, especially for the offspring, relative to the target population, an apparent age at diagnosis anticipation could be due to underestimating the age at diagnosis for offspring in the target population, which could be erroneously attributed to anticipation rather than to the method of sampling.

Figure 2.1 demonstrates that the variable age at diagnosis is right truncated. Top panel represents the offspring; bottom panel represents the parents.

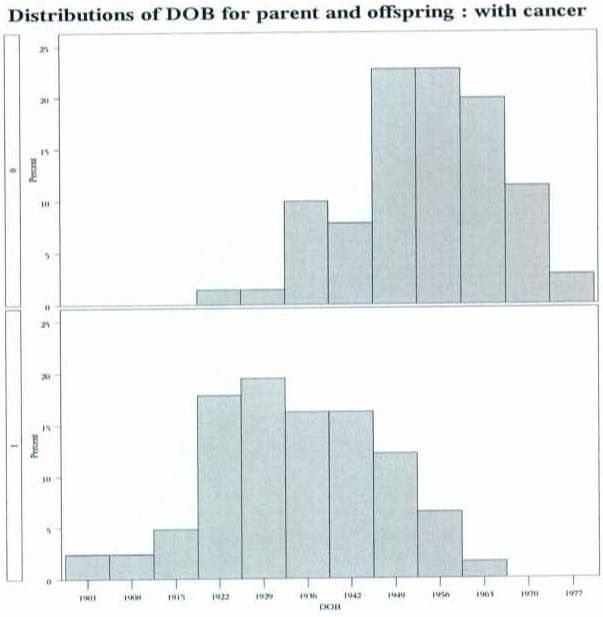


Figure 2.2: The distribution of DOB: affected individuals

Chapter 3

Analysis: Parent-of-Origin Effects on Age at Diagnosis of LS

In this chapter, we seek statistical evidences for a parent-of-origin effect on age at diagnosis of LS by studying decrease in age at diagnosis over successive generations for a parent-offspring pair in which both the parent and child were affected by the disease. Specifically, we seek statistical evidence that the gender of the disease-allele transmitting-parent influences age at diagnosis of the disease in offspring and that the disease-allele transmitting-parent's influence on age at diagnosis of the disease in offspring depends upon the gender of the offspring.

3.1 Methods

To assess parent-of-origin effects on age at diagnosis of LS, we can formulate the problem as follows:

For an affected parent-offspring pair, let

$$dage = \text{age at diagnosis for the parent} - \text{age at diagnosis for the offspring.}$$

dage is a random variable and the mean of *dage* can be written as

$$\text{mean}(dage) = \mu + \alpha \cdot g_o + \beta \cdot g_p + \gamma \cdot g_o \cdot g_p + f(g_o \cdot x,\ g_p \cdot x) + x$$

where g_o and g_p are the gender of offspring and the gender of parent, respectively, and x is an amount due to the sampling bias. From Section 2.3, we know that the sampling bias, thus x, is associated with an individual's date of birth, especially, the offspring's date of birth. If we code the nominal variables g_0 and g_p as follows:

$$g_o = \begin{cases} 0 & \text{for son;} \\ 1 & \text{for daughter.} \end{cases} \quad \text{and} \quad g_p = \begin{cases} 0 & \text{for father;} \\ 1 & \text{for mother.} \end{cases}$$

then testing whether phenotypic effect of the parental alleles on daughters depends on their paternal or maternal origin is equivalent to testing $\beta + \gamma = 0$. Testing whether phenotypic effect of the parental allele on sons depends on their paternal or maternal origin is equivalent to testing $\beta = 0$. Testing whether mother's alleles have a different phenotypic effect on offspring in a gender-specific manner is equivalent to testing $\alpha + \gamma = 0$. Testing whether father's alleles have a different phenotypic effect on offspring in a gender-specific manner is equivalent to testing $\alpha = 0$. Under the assumption that $f(g_o \cdot x,\ g_p \cdot x) \equiv 0$, that is, there exists no interaction between g_o, g_p and x, the tests mentioned above are equivalent to two-sample tests given below, as used in Lindor et al. (2010).

1. Test 1 tests whether phenotypic effect of parental alleles on offspring depends on their paternal or maternal origin. Here, the phenotypic effects on offspring are related to the gender of the parent. Two samples are involved in the two-sample test related to Test 1.

   Sample 1 is

   $$\{\ dage_i \equiv (\text{age}_{m(i)} - \text{age}_{o(i)}) \quad \text{for } i\text{th mother-offspring pair} \mid i = 1, \cdots, 58\},$$

   and sample 2 is

   $$\{\ dage_i \equiv (\text{age}_{f(i)} - \text{age}_{o(i)}) \quad \text{for } i\text{th father-offspring pair} \mid i = 1, \cdots, 40\},$$

where age_m is the mother's age at diagnosis, age_f is the father's age at diagnosis, age_o is the offspring's age at diagnosis, and *dage* is the decrease in age at diagnosis over successive generations.
Statistical test: two-sample t-test

$$\mathcal{H}_0: \quad \text{mean (sample 1)} = \text{mean (sample 2)}$$

vs.

$$\mathcal{H}_\alpha: \quad \text{mean (sample 1)} \neq \text{mean (sample 2)}$$

for a large sample size or for a small size sample if *dage* follows a normal distribution; and the two-sample Wilcoxon signed-rank test

$$\mathcal{H}_0: \quad \text{median (sample 1)} = \text{median (sample 2)}$$

vs.

$$\mathcal{H}_\alpha: \quad \text{median (sample 1)} \neq \text{median (sample 2)}$$

for a small size sample if *dage* does not follow a normal distribution.

2. Test 2 tests whether parental alleles' phenotypic effect on daughter depends on their paternal or maternal origin. The phenotypic effects on female offspring are related to the gender of the parent. Two samples are involved in the two-sample test related to Test 2.
Sample 1 is

$$\{\ dage_i \equiv (\text{age}_{m(i)} - \text{age}_{d(i)}) \quad \text{for } i\text{th mother-daughter pair} \mid i = 1, \cdots, 33\},$$

and sample 2 is

$$\{\ dage_i \equiv (\text{age}_{f(i)} - \text{age}_{d(i)}) \quad \text{for } i\text{th father-daughter pair} \mid i = 1, \cdots, 18\},$$

where age_d is the daughter's age at diagnosis.
Statistical test: two sample t-test

$$\mathcal{H}_0: \quad \text{mean (sample 1)} = \text{mean (sample 2)}$$

vs.

$$\mathcal{H}_\alpha : \quad \text{mean (sample 1)} \neq \text{mean (sample 2)},$$

and for the two-sample Wilcoxon signed-rank test

$$\mathcal{H}_0 : \quad \text{median (sample 1)} = \text{median (sample 2)}$$

vs.

$$\mathcal{H}_\alpha : \quad \text{median (sample 1)} \neq \text{median (sample 2)}.$$

3. Test 3 tests whether phenotypic effect of parental alleles on sons depends on their paternal or maternal origin. The phenotypic effects on male offspring are related to the gender of the parent. Two samples are involved in the two-sample test related to Test 3.
   Sample 1 is

$$\{\ dage_i \equiv (\text{age}_{m(i)} - \text{age}_{s(i)}) \quad \text{for } i\text{th mother-son pair} \mid i = 1, \cdots, 25\},$$

   and sample 2 is

$$\{\ dage_i \equiv (\text{age}_{f(i)} - \text{age}_{s(i)}) \quad \text{for } i\text{th father-son pair} \mid i = 1, \cdots, 22\},$$

   where age_s is the son's age at diagnosis.
   Statistical test: two-sample t-test

$$\mathcal{H}_0 : \quad \text{mean (sample 1)} = \text{mean (sample 2)}$$

   vs.

$$\mathcal{H}_\alpha : \quad \text{mean (sample 1)} \neq \text{mean (sample 2)},$$

   and for the two-sample Wilcoxon signed-rank test

$$\mathcal{H}_0 : \quad \text{median (sample 1)} = \text{median (sample 2)}$$

   vs.

$$\mathcal{H}_\alpha : \quad \text{median (sample 1)} \neq \text{median (sample 2)}.$$

4. Test 4 tests whether the mother's alleles have a different phenotypic effect on offspring in a gender-specific manner. The phenotypic effects on offspring relate to the gender of the offspring. Two samples are involved in the two-sample test related to Test 4.
   Sample 1 is

   $$\{\ dage_i \equiv (\text{age}_{m(i)} - \text{age}_{d(i)}) \quad \text{for } i\text{th mother-daughter pair} \mid i = 1, \cdots, 33\},$$

   and sample 2 is

   $$\{\ dage_i \equiv (\text{age}_{m(i)} - \text{age}_{s(i)}) \quad \text{for } i\text{th mother-son pair} \mid i = 1, \cdots, 25\}.$$

   Statistical test: two-sample t-test

   $$\mathcal{H}_0: \quad \text{mean (sample 1)} = \text{mean (sample 2)}$$

   vs.

   $$\mathcal{H}_\alpha: \quad \text{mean (sample 1)} \neq \text{mean (sample 2)},$$

   and for the two-sample Wilcoxon signed-rank test

   $$\mathcal{H}_0: \quad \text{median (sample 1)} = \text{median (sample 2)}$$

   vs.

   $$\mathcal{H}_\alpha: \quad \text{median (sample 1)} \neq \text{median (sample 2)}.$$

5. Test 5 tests whether father's alleles have a different phenotypic effect on offspring in a gender-specific manner. The phenotypic effects on offspring relate to the gender of the parent. Two samples are involved in the two-sample test related to Test 5.
   Sample 1 is

   $$\{\ dage_i \equiv (\text{age}_{f(i)} - \text{age}_{d(i)}) \quad \text{for } i\text{th father-daughter pair} \mid i = 1, \cdots, 18\},$$

   and sample 2 is

   $$\{\ dage_i \equiv (\text{age}_{f(i)} - \text{age}_{s(i)}) \quad \text{for } i\text{th father-son pair} \mid i = 1, \cdots, 22\}.$$

Statistical test: two-sample t-test

$$\mathcal{H}_0: \quad \text{mean (sample 1)} = \text{mean (sample 2)}$$

vs.

$$\mathcal{H}_\alpha: \quad \text{mean (sample 1)} > \text{mean (sample 2)},$$

and for the two-sample Wilcoxon signed-rank test

$$\mathcal{H}_0: \quad \text{median (sample 1)} = \text{median (sample 2)}$$

vs.

$$\mathcal{H}_\alpha: \quad \text{median (sample 1)} > \text{median (sample 2)}.$$

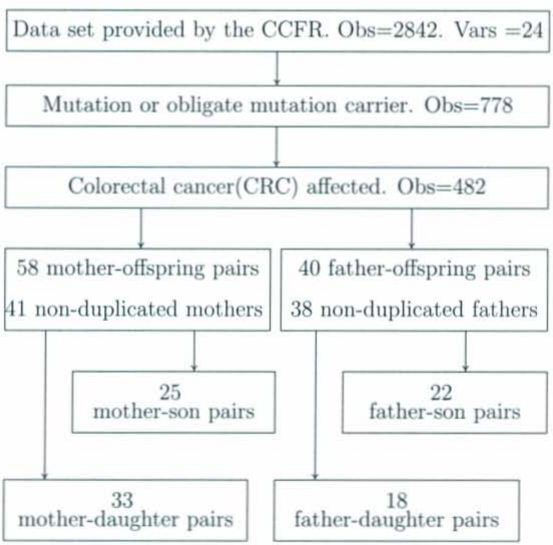
Remark: when we test the effects listed in Test 1 to Test 5 by comparing means and medians of the two independent samples (sample 1 and sample 2), though each sample is subject to the sampling bias, the comparison would not be biased by the sampling bias if the assumption $f(g_o \cdot x,\ g_p \cdot x) \equiv 0$ holds. Since this condition does not necessarily hold, the method should be used with caution. If the assumption $f(g_o \cdot x,\ g_p \cdot x) \equiv 0$ does not hold, an alternative discussion or method is required.

3.2 Data Preparation

All individuals considered in this analysis satisfy the following:

- Mutation carrier (MUT STATUS is c) or obligate mutation carrier (Obligate Carrier is c)
- Affected with colorectal cancer
- Population-based (FSRC is 1) or clinic-based (FSRC is 2)

The data sets relevant to Test 1 to Test 5 in Section 3.1 can be obtained from the raw data set provided by the Colon Cancer Family Registry as follows.



Note: Age at diagnosis was not available for some individuals, so these individuals were not included in the study.

3.3 Results of Two-Sample Tests

A summary of the results of the statistical analysis, based on the two-sample tests as used in Lindor et al. (2010), is given in Table 3.1. We also present some summary statistics for i) unique parental age at diagnosis (mothers vs. fathers); ii) offspring age at diagnosis (mothers vs. fathers); iii) female offspring age at diagnosis (mothers vs. fathers); and iv) male offspring age at diagnosis (mothers vs. fathers).

Parent's observed mean age at diagnosis (mother = 52.88 and father = 48.21), parent's observed median age at diagnosis (mother = 52.00 and father = 49.50), as well as the range of parent's age at diagnosis (mother: 25-80 and father: 24-70) can provide us with a rough idea about age at diagnosis of LS. Since the parent's age at diagnosis is right truncated, though much less severe than for the offspring, the real figures should be at least of this magnitude. These figures can provide us with a reference of the extent of the sampling bias in the offspring data. The results also show that mother's age at diagnosis is later on average than the father's. We will come back to this topic later.

For both mother and father, their offspring's age at diagnosis is much earlier (by about 13 years) on average than theirs. These phenomena are at least partially due to the sampling bias.

The age at diagnosis for the offspring of the fathers was earlier on average than those for the offspring of the mothers (observed mean for fathers, 36, vs. observed mean for mothers, 39, years old). When divided into sons and daughters, this difference was being driven by the earlier age at diagnosis for daughters (observed mean 33.5 and median 35) of affected fathers compared to the age at diagnosis for sons (observed mean 38 and median 39) of affected fathers. The age at diagnosis for daughters of affected fathers is also earlier on average than both the age at diagnosis for daughters of affected mothers (observed mean 40 and median 39) and the age at diagnosis for

Table 3.1: Statistical analysis of the age at diagnosis

	Affected Mother	Affected Father	Mother vs. Father
Unique Parent	N(m) = 41 Mean(m) = 52.88 Median(m) = 52 Range(m) = 25 - 80	N(f) = 38 Mean(f) = 48.21 Median(f) = 49.50 Range(f) = 24 - 70	
All offspring	N(o) = 58 Mean(o) = 39.66 Median(o) = 39 Range(o) = 21 - 63	N(o) = 40 Mean(o) = 35.85 Median(o) = 36 Range(o) = 22 - 50	Test 1 58 pairs vs. 40 pairs H_0: mean(m) = mean(f) H_a: mean(m) $\neq$ mean(f)
Age(m/f) - Age(o)	mean = 14.33 Std.Dev = 13.6 median= 13.5 Range = -28 - 44	mean = 12.45 Std.Dev = 13.86 median= 14 Range = -16 - 39	t.test † p-value = 0.5 t.test ‡ p-value = 0.5 w.test¶ : p-value=0.6
Female offspring	N(d) = 33 Mean(d) = 39.88 Median(d) = 39 Range(d) = 24 - 63	N(d) = 18 Mean(d) = 33.5 Median(d) = 35 Range(d) = 22 - 45	Test 2 33 pairs vs. 18 pairs H_0: mean(m) = mean(f) H_a: mean(m) $\neq$ mean(f)
Age(m/f) - Age(d)	**mean = 14.39** Std.Dev = 12.97 **median = 13** Range = -7 - 39	**mean = 17.22** Std.Dev = 10.95 **median= 18** Range = -2 - 39	t.test† p-value = 0.44 t.test‡ p-value = 0.41 w.test¶ : p-value=0.42
Male offspring	N(s) = 25 Mean(s) = 39.36 Median(s) = 39 Range(s) = 21 - 62	N(s) = 22 Mean(s) = 37.77 Median(s) = 39 Range(s) = 27 - 50	Test 3 25 pairs vs. 22 pairs H_0: mean(m) = mean(f) H_a: mean(m) $\neq$ mean(f)
Age(m/f) - Age(s)	**mean = 14.24** Std.Dev = 14.69 **median = 14** Range = -28 - 44	**mean = 8.55** Std.Dev = 14.98 **median = 8.5** Range = -16 - 39	t.test† p-value = 0.2 t.test‡ p-value = 0.2 w.test¶ : p-value=0.17
Female offspring vs. Male offspring	Test 4 33 pairs vs. 25 pairs H_0: mean(d) = mean(s) H_a: mean(d) $\neq$ mean(s) t.test† p-value=0.97 t.test‡ p-value=0.97 w.test¶ : p-value=0.9	Test 5 18 pairs vs. 22 pairs H_0: mean(d) = mean(s) H_a: mean(d) > mean(s) t.test† p-value=**0.024** t.test‡ p-value=**0.02** w.test¶ : p-value=**0.04**	

Note: †Two-sample t test with variance equal.
Note: ‡Two-sample t test without variance equal.
Note: ¶Wilcoxon signed-rank test for the corresponding test regarding medians.

sons of affected mothers (observed mean 39 and median 39).

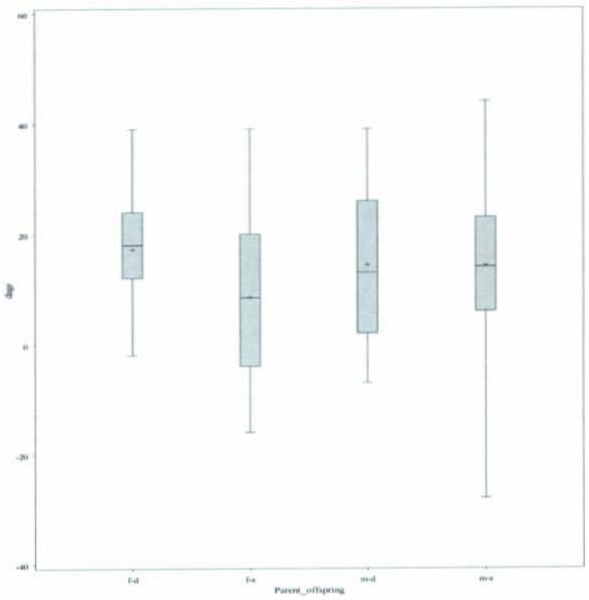


Figure 3.1: Comparison of *dage* among father-daughter (f-d) group, father-son (f-s) group, mother-daughter (m-d) group and mother-son (m-s) group

Both the t-test and the Wilcoxon signed-rank test show that the mean and median decrease in age at diagnosis over successive generations for the father-daughter group is significantly larger than for the father-son group ($p < 0.05$, Test 5) while the mean and median decreases in age at diagnosis over successive generations are not significantly different either between the father-daughter group and the mother-daughter group (Test 2), between the father-son group and the mother-son group (Test 3), or between the mother-daughter group and the mother-son group (Test 4) with p-value

0.4, 0.2, and 0.9, respectively. Please see Figure 3.1 and Table 3.1. These results could further suggest that decrease in age at diagnosis over successive generations depends not only on the gender of the parent but also on the gender of his/her offspring if the assumption $f(g_o \cdot x,\ g_p \cdot x) \equiv 0$ holds. However, imprinting, the most common form of parent-of-origin effects, relates to the gender of the parent, but not the gender of the offspring, that is, imprinting effects are not expected to be affected by the gender of the offspring. Thus, it remains to be seen whether this result is merely an artifact of inadequate statistical analysis.

3.4 Discussion

In this section, we will look closer at the results presented in the previous section, especially the results relating to Test 5. Recall that for those born after 1935, the later an individual was born the more severe was the right truncation. Thus earlier age at diagnosis contributes to the mean and median age at diagnosis (see Figure 2.1 and the summary statistics for the patient's age at diagnosis in Table 3.1).

Regarding the observation that father's age at diagnosis is earlier on average than mother's (48.21 vs. 52.88 for mean and 49.5 vs. 52 for median), Figure 3.2 shows that compared to mother's group, more individuals were born after 1935 for father's group, thus more cases of early age at diagnosis are included in the father's group. Therefore, the father's earlier age at diagnosis is, to a certain extent, a result of the sampling bias.

Regarding the observation that age at diagnosis for the father's offspring is earlier on average than for the mother's offspring (35.85 vs. 39.66 for mean and 36 vs. 39 for median), Figure 3.3 shows that compared to the DOB distribution of the mother's offspring, the DOB distribution of the father's offspring is weighted towards a later DOB thus more cases of early age at diagnosis are included in the fathers offspring

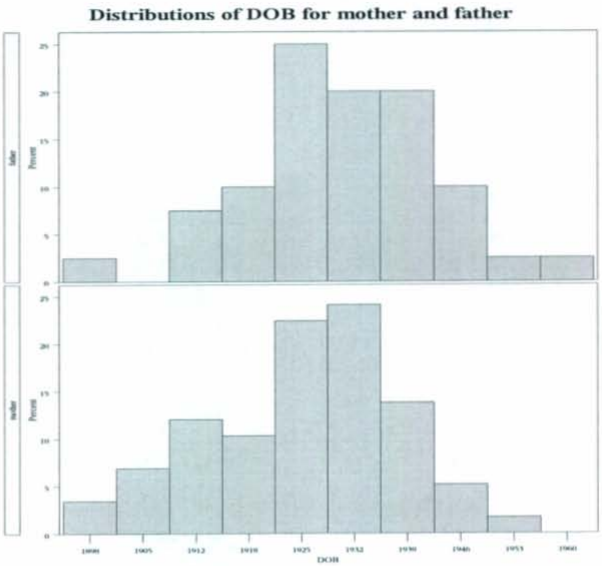


Figure 3.2: The distributions of date of birth (DOB): fathers vs mothers

data set. Therefore, sampling bias is a factor that causes an observed earlier age at diagnosis for the father's offspring than for the mother's offspring. Figure 3.3 also shows that compared to the three other *DOB* distributions of father's sons, mother's sons, and mother's daughters, the *DOB* distribution of the father's daughters is weighted towards a later *DOB*. Therefore, the sampling bias can explain, at least to a certain extent, the observed earlier age at diagnosis on average for the father's daughter than for the father's son, the mother's son, and the mother's daughter (33.50 vs. 37.77, 39.96 and 39.88 for the mean and 35 vs. 39, 39, and 39 for the median).

Regarding Test 5, recall that the results obtained from the two-sample test (two observed *dage* samples) in Table 3.1 are valid for Test 1 to Test 5 in section 3.1 only

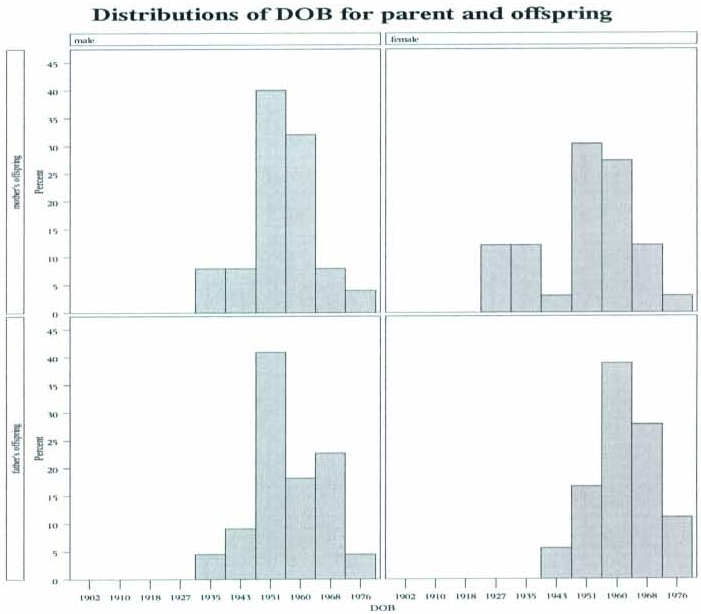


Figure 3.3: The distributions of date of birth (DOB): parents vs. offspring

under the assumption that there exists no interaction between g_p, g_o and x, or equivalently date of birth, which relates to the sampling bias. However, Figure 3.3 suggests that there exists interaction between g_p, g_o and date of birth (that is, x) for the data under investigation. Therefore, it may be questionable for the results given by the two-sample tests (two observed *dage* samples) to be valid for Test 1 to Test 5, thus an alternative analysis is required.

Figure 3.4 provides insight into the two samples involved in Test 5. The figure clearly shows that there is an interaction between g_o and DOB of father's offsprings, and therefore x (for the definition of x, please see page 16). More affected sons (of affected

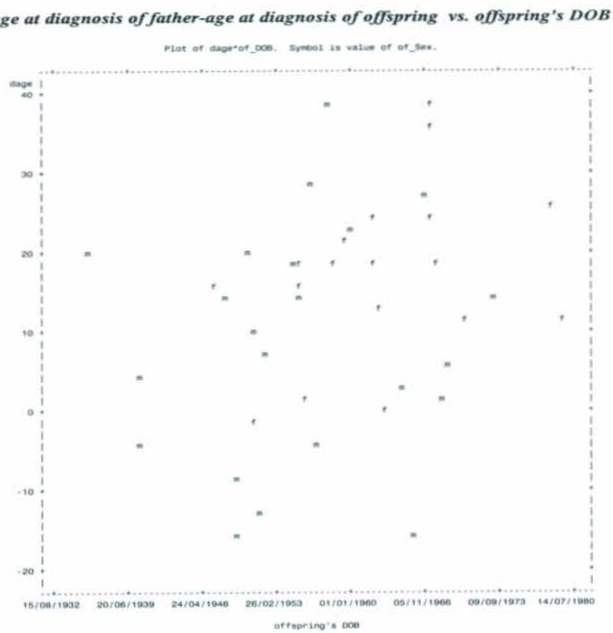


Figure 3.4: *dage* for a father-offspring pair vs. the offspring's date of birth (*DOB*). "m": father-son pair. "f": father-daughter pair.

fathers) were born before 1953 and were less subjected to the sampling bias, thus contributing smaller values of *dage* to the mean and median of sample 2, while more affected daughters (of affected fathers) were born after 1953 and were more subjected to sampling bias, thus contributing larger values of *dage* to the mean and median of sample 1. In both the t-test and Wilcoxon signed-rank test, only one dimensional information, *dage*, for two samples is used, while the second dimensional information, the sampling bias related to *dage*, is ignored. As a result, the t-test and Wilcoxon signed-rank test lead to the conclusion that the mean and median decrease in age

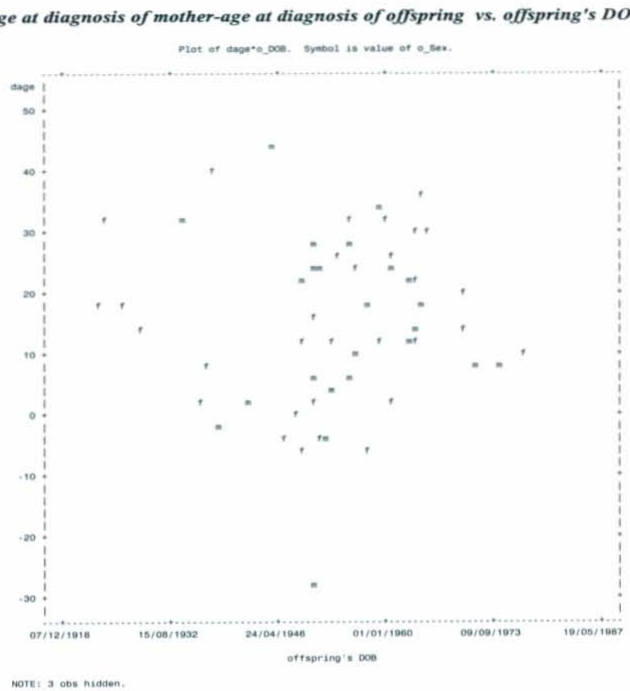


Figure 3.5: *dage* for a mother-offspring pair vs. the offspring's date of birth (DOB). "m": mother-son pair. "f": mother-daughter pair.

at diagnosis over successive generations for the father-daughter group is significantly larger than for the father-son group. After accounting for the sampling bias, it can be seen from

$$\mu + \alpha \cdot g_o + \beta \cdot g_p + \gamma \cdot g_o \cdot g_p = \text{mean}(dage) - f(g_o \cdot x,\ g_p \cdot x) - x$$

that the difference in

$$\text{mean}(dage) - f(g_o \cdot x,\ g_p \cdot x) - x,$$

and thus the difference in

$$\mu + \alpha \cdot g_o + \beta \cdot g_p + \gamma \cdot g_o \cdot g_p, \quad \text{or equivalently,} \quad \alpha \cdot g_o$$

between the father-daughter group and the father-son group is less than the difference in *dage* between the two groups. Thus, the apparent parent-of-origin effects related to the gender of the offspring due to failing to account for the bias partially or completely disappears. Combining this observation with the results given in Table 3.1 and Figure 3.1, we can infer that the difference in $\alpha \cdot g_o$ between the two groups may not be significant any more or equivalently, α is not significantly different from zero. Recall that testing whether father's alleles have a different phenotypic effect on offspring in a gender-specific manner is equivalent to testing $\alpha = 0$ (page 16). Alternatively, this inference can also be drawn from a direct inspection of Figure 3.4. This figure shows that for pairs whose offspring were born in the same year thus whose *dage* subjected to similar sampling bias, the two distributions of *dage* corresponding to father-daughter pairs and father-son pairs do not systematically separate from each other. This observation implies that there is insufficient evidence from this sample to claim that father's alleles have a different phenotypic effect on offspring in a gender-specific manner. Of course, the analysis presented here is a qualitative one.

Regarding Test 4, Figure 3.5 shows that the interaction between g_o and DOB of the mother's offspring is not significant for offspring who were born after 1946. Since the assumption $f(g_o \cdot x,\ g_p \cdot x) \equiv 0$ holds for this subset, the results of the t-test and Wilcoxon signed-rank test are not affected by the sampling bias. Though there exists an interaction between g_o and DOB of the mother's offspring for offspring who were born before 1946, the corresponding *dage* have a similar distribution as *dage* for either mother-son pairs or mother-daughter pairs after 1946 as a whole, suggesting the inclusion of the former (before 1946) into the latter (after 1946) will not affect

the mean and median of the latter. Therefore, the results of two-sample t-test and Wilcoxon signed-rank test in Table 3.1 for Test 4 are not distorted by the sampling bias and are valid. Alternatively, this inference can also be drawn from a direct inspection of Figure 3.5. This figure shows that for pairs whose offspring were born in the same year thus whose *dage* subjected to similar sampling bias, the two distributions of *dage* corresponding to mother-daughter pairs and mother-son pairs do not systematically separate from each other. This observation is consistent with the results given by the two-sample t-test and the Wilcoxon signed-rank test in Table 3.1.

Having gained insight into the data, we next use an appropriate quantitative method to adapt to the circumstances.

3.5 An Alternative Analysis - Regression Analysis

To get a quantitative inference about the effect of gender of either the parent or the offspring on *dage*, we use a regression method to repeat the above analysis. We chose a model which appropriately fit the data according to R-Square (adj-R-Square) and fit diagnostics for inference.

Corresponding to the two-sample t-test for Test 2 (3), we regress *dage* on gender of the parent and test β in

$$dage = \alpha + \beta \cdot gender_p + \epsilon \tag{3.1}$$

to test for significant difference from zero for parent-daughter (son) pairs. For Test 4 (5), we regress *dage* on gender of offspring and test β in

$$dage = \alpha + \beta \cdot gender_o + \epsilon \tag{3.2}$$

to test for significant difference from zero for mother (father)-offspring pairs. The inferences from Models 3.1 and 3.2 which are fit to the data used in the two-sample tests without excluding influential observations, are the same as the corresponding

ones obtained by the two-sample tests in Table 3.1 (for parameter estimates and results from model fitting, please see Figures A.1-A.8).

To account for the interaction between gender of offspring, gender of parents and date of birth, we regress *dage* on gender of the parent or gender of the offspring. adjusting for *DOB* of the parent and *DOB* of the offspring. For Test 2(3),we test β in

$$dage = \alpha + \beta \cdot gender_p + \gamma \cdot pdob + \delta \cdot odob + \epsilon \tag{3.3}$$

for significant difference from zero for parent-daughter (son) pairs. For Test 4(5), we test β in

$$dage = \alpha + \beta \cdot gender_o + \gamma \cdot pdob + \delta \cdot odob + \epsilon \tag{3.4}$$

for significant difference from zero for mother (father)-offspring pairs.

We find that Models 3.1 and 3.2 do not fit the data well, and therefore the inferences based on these models are not reliable. By dropping influential observations from the data, we obtain Models 3.3 and 3.4 which fit the data appropriately according to R-Square (adj-R-Square) and fit diagnostics (please see Figures A.9-A.16). For more about identifying and handling influential observations, please see Bowerman et al. (1993) Chapter 5. The inferences based on the best fit Models 3.3 and 3.4 are listed in Table 3.2. We find that after adjusting to *DOB* of the parent and of the offspring, the effect of gender, either of the parent or of the offspring, on *dage* is not significant at a p-value 0.05. The signs of estimates of δ and γ are consistent with our expectation.

3.6 Conclusions

In conclusion, based on the analysis presented in this chapter, we found no enough statistical evidence from this sample to claim the existence of parent-of-origin effects on age at diagnosis of the disease related to either the gender of the parent or the

Table 3.2: Inferences based on the best fit models

Test	Parameter	Estimate	Standard Error	t value	Pr $>\|t\|$
Test 2	δ (*odob*)	1.183	0.420	2.82	0.0075
	γ (*pdob*)	-0.422	0.381	-1.11	0.2746
	β (mother)	**0.184**	3.493	0.05	**0.9582**
Test 3	δ (*odob*)	1.514	0.455	3.33	0.0020
	γ (*pdob*)	-1.137	0.400	-2.84	0.0080
	β (father)	**-6.369**	3.414	-1.87	**0.0710**
Test 4	δ (*odob*)	1.573	0.388	4.06	0.0002
	γ (*pdob*)	-0.793	0.364	-2.18	0.0355
	β (son)	**-0.420**	2.938	-0.14	**0.8869**
Test 5	δ (*odob*)	0.773	0.436	1.77	0.0864
	γ (*odob*)	-0.193	0.355	-0.54	0.5914
	β(son)	**-6,524**	4.503	-1.45	**0.1577**

gender of the offspring. This may further suggest that there exists no parent-of-origin effects on age at onset of LS.

Our analysis also shows that in order to account for the sampling bias caused by the right truncation, one should include into the analysis the date of birth of the offspring and of the parents which relates to the magnitude of the effect of right truncation on *dage*. Otherwise, wrong conclusions may be drawn.

Chapter 4

Analysis: Age at Diagnosis Anticipation

We are interested in testing for decrease in age at diagnosis of LS over successive generations. As we have seen from Figure 2.1 in Chapter 2, the data under study is subject to sampling bias, that is, the parental generation has passed through most of the risk period for the disease while the offspring generation has not yet completed the risk period. As a result, offspring who are unaffected at the time of analysis but go on to manifest the disease at later ages are not included in this sample. In this situation, standard statistical methods, such as the paired t-test, are inappropriate because these methods do not adequately adjust for sampling bias. Here, we will use a test method proposed by Tsai et al. (2005). They model the age at diagnosis of affected parent-child pairs as being right-truncated by age at interview and formulate the problem in terms of symmetry tests. They propose a simple generalized paired t-test and a Wilcoxon signed-rank test to adjust for the bias caused by the right truncation of both the parent's and child's age at diagnosis. For the advantage of Tsai's method over other methods used in the circumstance of the age at diagnosis of affected parent-child pairs being right-truncated by the age at interview, please see the paper by Tsai et al. (2005).

4.1 The Symmetry Tests

4.1.1 Methods

In this subsection, we give a summary of the symmetry tests proposed by Tsai et al. (2005), which will be used in this section, for the reader's convenience.

The symmetry tests are a simple generalized paired t-test (for large sample sizes) and the Wilcoxon signed-rank test (for small sample sizes) to adjust for the bias caused by the right truncation of both the parent's and child's age at diagnosis. Also, Tsai et al. (2005) extends the generalized paired t-test to a random effects model that enables analysis of correlated data from nuclear families and could be further extended to larger family structures. This approach circumvents some of the sampling bias that plagues anticipation testing, specifically, the sampling bias caused by the right truncation of both parental and child's age at diagnosis. However, some power is lost because some of the parent-child pairs are discarded, but simplicity and lack of bias are gained.

Recall that for right-truncated data, only individuals for whom the event has occurred by a given date are included in the study. The main impact on the analysis when data are truncated is that the investigator must use a conditional distribution in constructing the likelihood.

For a given affected parent-child pair, let C_p and C_c denote the age at interview of the parent and child, respectively, and let T_p and T_c denote their respective age at diagnosis. Assume the ordered pairs (T_p, T_c) and (C_p, C_c) are independent, given that both parent and child became affected before beginning of the study's enrollment period. Here, "parent-child pair" always denotes a pair in which both the parent and the child are affected. Let n denote the number of parent-child pairs included in the study and let the quadruple $(T_{pi}, T_{ci}, C_{pi}, C_{ci})$, $i = 1, \ldots, n$, denote the ages at

diagnosis and ages at interview of the ith parent-child pair. For a parent-child pair to be included in the study, the diagnosis age must be lower than the interview age for each person. If this is not true, then the diagnosis age is truncated by the interview age. Therefore, it is reasonably assumed that the (T_{pi}, T_{ci}, C_{pi}, C_{ci}) quadruples are independent and identically distributed, according to the conditional probability density function of (T_p, T_c, C_p, C_c), given that $T_p \leq C_p$ and that $T_c \leq C_c$. Moreover, this conditional probability density function is given by

$$\frac{f(t_p, t_c)g(c_p, c_c)}{P(T_p \leq C_p, T_c \leq C_c)} I(T_p \leq C_p, T_c \leq C_c)$$

where I(A) equals 1 if A is true and 0 otherwise, and f and g are the probability density functions of (T_p, T_c) and (C_p, C_c), respectively. Let F and G denote the cumulative distribution functions corresponding to f and g, respectively. The goal is to use the observed data (T_{pi}, T_{ci}, C_{pi}, C_{ci}), $i = 1, \ldots, n$, to test the null hypothesis that f is symmetric, that is, that the age at diagnosis of parent and child are exchangeable. If f is symmetric (equivalent to F is symmetric), then there is no age at diagnosis anticipation. If f is not symmetric, then age at diagnosis anticipation provides one reasonable explanation, though not the only one. If the onset ages T_{pi} and T_{ci} are subject to the same truncation effect, then one can compare them directly because they have an identical bias effect.

A parent-child pair, (T_{pi}, C_{pi}) and (T_{ci}, C_{ci}), is called comparable if $max(T_{pi}, T_{ci}) \leq min(C_{pi}, C_{ci})$. That is, the child's age at diagnosis must be lower than the parent's age at interview, and the parent's age at diagnosis must be lower than the child's age at interview. The first condition is usually met in typical studies, but the second condition is not met by all parent-child pairs and necessitates discarding some pairs. Tsai et al. (2005) define an indicator δ_i as equaling 1 when that condition is met, and 0 otherwise.

Let $n^* = \sum_{i=1}^{n} \delta_i$ denote the number of comparable parent-child pairs in the observed data, and let ($T^*_{pi}, T^*_{ci}, C^*_{pi} C^*_{ci}$), $i = 1, \ldots, n^*$, denote the observed quadruple of

comparable parent-child pairs. Note $n^* \leq n$, and $max(T^*_{pi}, T^*_{ci}) \leq min(C^*_{pi}, C^*_{ci})$ for $i = 1, \ldots, n^*$.

Let $F^*(s,t)$ denote the marginal cumulative distribution function of (T^*_{pi}, T^*_{ci}). Under the assumption that pairs (T_p, T_c) and (C_p, C_c) are independent, it is easily established

$$\begin{aligned} F^*(s,t) &= Pr(T_p \leq s, T_c \leq t \mid \delta = 1) \\ &= \frac{\int_{-\infty}^{s}\int_{-\infty}^{t} f(x,y)G^*(x \vee y)dydx}{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)G^*(x \vee y)dydx} \end{aligned}$$

where $a \vee b = max(a,b)$ and

$$G^*(t) \equiv Pr(C_p \geq t, C_c \geq t) = 1 - G(t,\infty) - G(\infty,t) + G(t,t)$$

The independence assumption is essential in deriving the above equality. If (T_p, T_c) and (C_p, C_c) are dependent, then the density f cannot be nonparametrically identified. It follows from the above equality that F^* is symmetric if and only if F is symmetric.

To test symmetry of F, Tsai et al. (2005) now test for symmetry of F^*. Then they use known standard statistics, such as the paired t-test statistic and/or the Wilcoxon signed-rank test statistic, to the data (T^*_{pi}, T^*_{ci}), $i = 1, \ldots, n$ for testing the symmetry of F^*.

The Paired t-test

It is known that the paired t-test statistic $t = \bar{d}/\sqrt{var(\bar{d})}$, where

$$\bar{d} = \frac{\sum_{i=1}^{n^*}(T^*_{pi} - T^*_{ci})}{n^*} = \frac{\sum_{i=1}^{n}(T_{pi} - T_{ci})\delta_i}{\sum_{i=1}^{n}\delta_i}$$

and

$$var(\bar{d}) = \frac{1}{n^*}\frac{\sum_{i=1}^{n}(T_{pi} - T_{ci} - \bar{d})^2\delta_i}{n^* - 1}$$

Even when the original bivariate distribution of (T_p, T_c) is bivariate normal, the statistic t does not follow a t distribution. However, it can be approximated by a

normal distribution for large n^*.

Wilcoxon Signed-Rank Test

The Wilcoxon signed-rank test is a nonparametric alternative to the paired t-test. This test assumes that there is information in the magnitude of the differences between paired observations, as well as the signs. Take the paired observations, calculate the differences, and rank them from smallest to largest by absolute value. Add all the ranks associated with positive differences, giving the T_+ statistic. Finally, the p-value associated with this statistic is found from an appropriate table.

The Wilcoxon signed-rank test is used when there are two nominal variables and one measurement variable. One of the nominal variables has only two values, such as "before" and "after", and the other nominal variable often represents individuals. This is the non parametric analog to the paired t-test, and should be used if the distribution of differences between pairs is not normally distributed.

The null hypothesis : The null hypothesis is that the median difference between pairs of observations is zero. Note that this is different from the null hypothesis of the paired t-test, which is that the mean difference between pairs is zero, or the null hypothesis of the sign test, which is that the numbers of differences in each direction are equal.

The test statistic : The absolute value of the differences between observations are ranked from smallest to largest, with the smallest difference getting a rank of 1, then next larger difference getting a rank of 2, etc. Ties are given average ranks. The ranks of all differences in one direction are summed, and the ranks of all differences in the other direction are summed. The smaller of these two sums is the test statistic,W.

The sign test is less efficient than the Wilcoxon signed-rank test when the underlying distribution is symmetric. The Wilcoxon rank-sum test is also less efficient than the Wilcoxon signed-rank test when data are paired.

Tsai et al. (2005) also provide a random effects model, an extension of the generalized paired t-test. The random effects model represents a way to handle multiple affected parent-child pairs within the same family, and to adjust for their mutual dependence. One can apply it only to the comparable pairs in the data set.

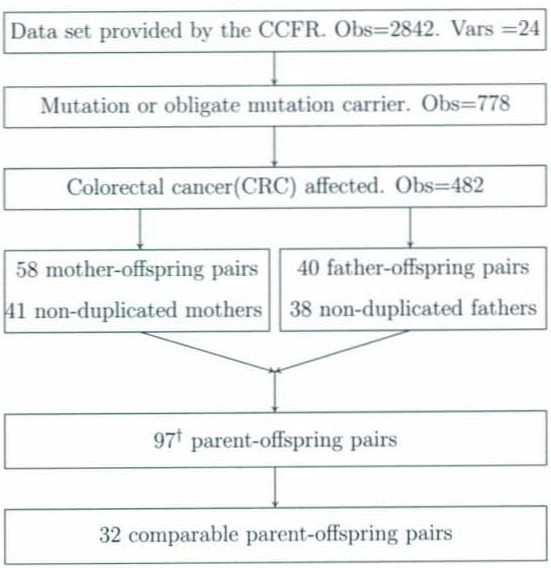
Let (T^*_{pij}, T^*_{cij}) denote the jth comparable parent-child pair in the ith pedigree and let $d^*_{ij} = T^*_{pij} - T^*_{cij}$. It is clear that d^*_{ij} and d^*_{ik} are not independent for $j \neq k$. A random effects model can be used to analyze the d^*_{ij}. The model assumes that correlation arises among measures d^*_{ij} in comparable parent-child pairs within a pedigree i. Specifically, they assume d^*_{ij} follows

$$d^*_{ij} = \alpha + \beta_i + \varepsilon_{ij}$$

where the β_i's are independent and identically distributed random variables with mean zero and variance σ^2_β, the ϵ_{ij}'s are random errors with mean zero and variance σ^2_ε. They further assume that β_i and ϵ_{ij} are independent. That is, there are unobserved factors represented by the β_i that are common to all d^*_{ij} for a given pedigree j but that vary across pedigrees. Note that this approach is equivalent to testing $\alpha = 0$.

4.1.2 Data preparation

To get the relevant data set (comparable parent-offspring pairs) from the original data set, we screen the data set as follows.



Note: †*Age.Death.or.Last.Known.Age* was not available for some individuals, so these also were not included in the analysis.

Getting Comparable Parent-Child Pairs

As mentioned before we know that a parent-child pair, (T_{pi}, C_{pi}) and (T_{ci}, C_{ci}), is comparable if

$$max(T_{pi}, T_{ci}) \leq min(C_{pi}, C_{ci})$$

Here C_{pi} is the age at interview of a parent, which is *p.Age.Death.or.Last.Known.Age*, C_{ci} is the age at interview of child, which is *o.Age.Death.or.Last.Known.Age* in the

current study, T_{pi} is the age at diagnosis of a parent, which is *p.age.at.colorectalCA*, and T_{ci} is the age at diagnosis of a child, which is *o.age.at.colorectalCA*. The ages at interview, that is, *Age.Death.or.Last.Known.Age*, are less than AGER (For AGER, please see definition 2.1 on page 12).

We use the following method to obtain comparable parent-child pairs in R package.

Getting Comparable Parent-Child Pairs

```
AMF$min <- rep(0,97)
for (i in 1:97)
{AMF$min[i] <- min(AMF$o_Age_Death.or.Last.Known.Age[i],
                   AMF$p_Age_Death.or.Last.Known.Age[i])}
amf<-subset(AMF,o_age_at_colorectalCA<min+1 & p_age_at_colorectalCA<min+1)
```

4.1.3 Results

The paired data : (T^*_{pi}, T^*_{ci}), $\quad i = 1, \ldots, 32$.
The correlation of T^*_{pi} and T^*_{ci} : $\quad corr(T^*_{pi}, T^*_{ci}) = 0.534$.

Table 4.1: Summary of $(T^*_{pi} - T^*_{ci})$

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Obs
-16.000	-2.500	2.000	2.688	8.000	20.000	32

For this comparable parent-child pairs sample, the offspring's age at diagnosis is earlier on average than their parent's age at diagnosis with an observed mean decrease of the paired age amounting to 2.69 years and an observed median decrease of the

Table 4.2: The paired t-test - one-sample t-test: one side

$\mathcal{H}_0$: true mean of $(T^*_{pi} - T^*_{ci}) = 0$
$\mathcal{H}_\alpha$: true mean of $(T^*_{pi} - T^*_{ci}) > 0$
$t = 1.8371$, $df = 31$, p-value $= 0.0379$
95% mean confidence interval: $(0.207, +\infty)$

Table 4.3: The paired t-test - one-sample t-test: two sides

$\mathcal{H}_0$: true mean of $(T^*_{pi} - T^*_{ci}) = 0$
$\mathcal{H}_\alpha$: true mean of $(T^*_{pi} - T^*_{ci}) \neq 0$
$t = 1.8371$, $df = 31$, p-value $= 0.0758$
95% mean confidence interval: $(-0.296, 5.671)$

Table 4.4: Wilcoxon signed-rank test with continuity correction: one side

$\mathcal{H}_0$: true location of $(T^*_{pi} - T^*_{ci}) = 0$
$\mathcal{H}_\alpha$: true location of $(T^*_{pi} - T^*_{ci}) > 0$
$V = 318$, p-value $= 0.0399$

Table 4.5: Wilcoxon signed-rank test with continuity correction: two sides

$\mathcal{H}_0$: true location of $(T^*_{pi} - T^*_{ci}) = 0$
$\mathcal{H}_\alpha$: true location of $(T^*_{pi} - T^*_{ci}) \neq 0$
$V = 318$, p-value $= 0.0798$

Table 4.6: Bootstrap test for mean = 0

$\mathcal{H}_0$: true mean of $(T^*_{ci} - T^*_{pi}) = 0$
$\mathcal{H}_\alpha$: true mean of $(T^*_{ci} - T^*_{pi}) > 0$
p-value $= 0.03408$
95% mean confidence interval: $(0.281, +\infty)$

Table 4.7: Bootstrap test for median = 0

$\mathcal{H}_0$: true median of $(T^*_{ci} - T^*_{pi}) = 0$
$\mathcal{H}_\alpha$: true median of $(T^*_{ci} - T^*_{pi}) > 0$
p-value = 0.02311
95% median confidence interval: (1.000, $+\infty$)

paired age amounting to 2 years, which is much less than the corresponding ones without adjusting for the sampling bias in Table 3.1. The test for the null hypothesis $\mathcal{H}_0$: true mean of $(T^*_{pi} - T^*_{ci}) = 0$ against the alternative hypothesis $\mathcal{H}_\alpha$: true mean of $(T^*_{ci} - T^*_{pi}) > 0$ yields a p-value of 0.0379 by the paired t-test and 0.03408 by the bootstrap test. The test for the null hypothesis $\mathcal{H}_0$: true median of $(T^*_{pi} - T^*_{ci}) = 0$ against the alternative hypothesis $\mathcal{H}_\alpha$: true median of $(T^*_{ci} - T^*_{pi}) > 0$ yields a p-value of 0.0399 by the Wilcoxon signed-rank test and 0.02311 by the bootstrap test. These test results imply that there is evidence in the data to suggest the existence of anticipation for age at diagnosis in LS.

We could extend the generalized paired t-test to a random effects model that enables analysis of correlated data from nuclear families. However, due to the limited data for comparable parent-child pairs and the observation from the following tables, we do not analyze the data further with the random effects model in this thesis.

The distribution of Family ID of the comparable parent-child pairs

110004550	110006763	110008508	125000010	125000146	125001235	130151007	130151008	130151010
1	1	1	1	1	2	1	2	2
130151021	130151064	130350052	130350067	130350110	130350166	130350224	130350281	130454009
2	1	1	2	1	1	1	1	1
130656093	130757006	130757046	150550229	150550641	150550744	150570273	150570476	
1	1	1	1	2	1	1	1	

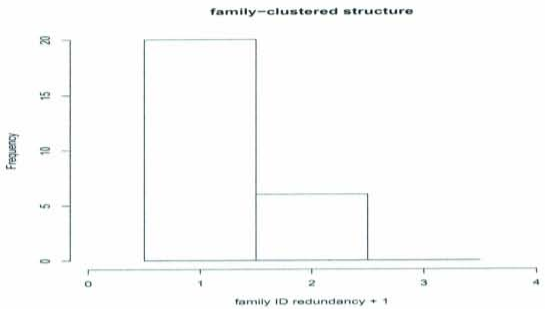


Figure 4.1: Family ID frequency distribution. Note: the two individuals of a parent-offspring pair have same family ID. Each Family ID is counted once for each parent-offspring pair, so the count of each Family ID = family ID redundancy + 1

We can see from the distribution of family ID of the comparable parent-child pairs, the number of multiple affected parent-child pairs within the same family are few and family clustering is not serious.

4.1.4 Conclusion

Based on the symmetry test, there is evidence in the data suggesting the existence of anticipation for age at diagnosis in LS.

4.2 Survival Analysis

In the above analysis, we only used the comparable parent-offspring pairs data. However, the comparable parent-offspring pairs only account for a small part of the available data, thus the available information is not efficiently utilized. Can we use more available data at hand to get a more reliable result?

It is tempting to analyze age at diagnosis data with survival analysis. This way, we can conduct an analysis based on more available data by including the mutation carriers, the mutation carriers who had not been affected or diagnosed with the disease by the time they were censored, and therefore hope to get a more accurate result.

Among variables available in the data set on hand, the only candidate for censoring time is the variable *Age.Death.or.Last.Known.Age*. At this point, let *Age.Death.or.Last.Known.Age* be censoring time for those mutation carriers not affected with LS.

4.2.1 Exploration of the data set

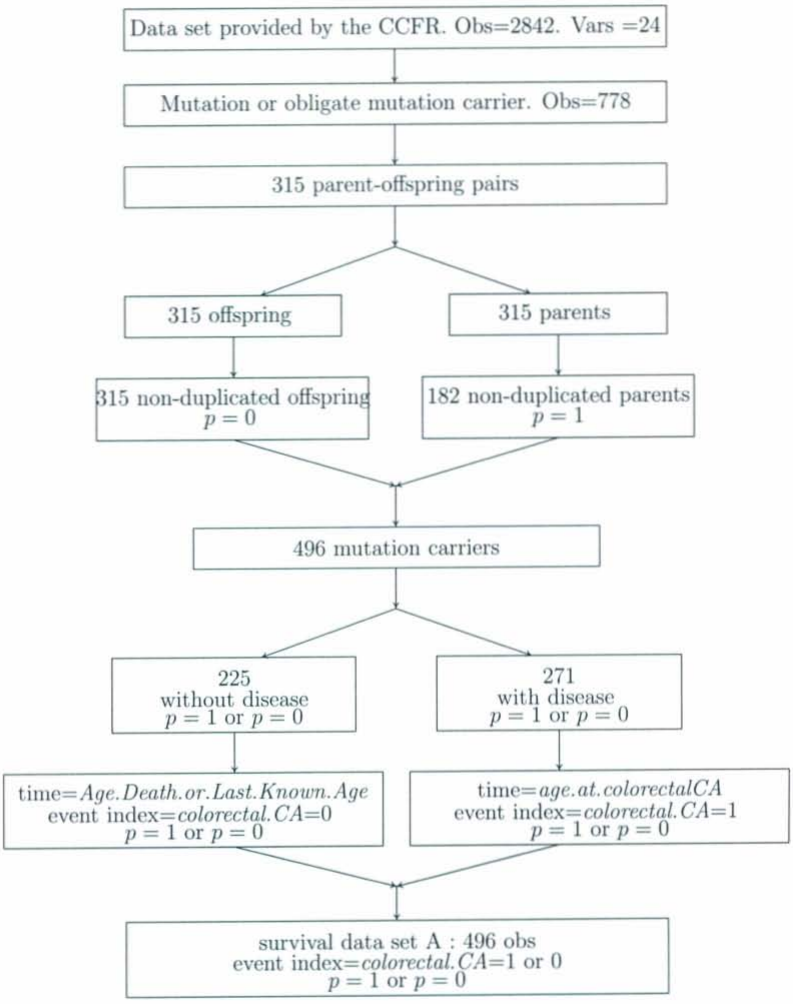
Survival time data preparation

We constructed survival time as follows. For the disease-affected mutation carrier (*colorectal.CA*=1), we set the time = age at diagnosis = *age.at.colorectalCA*, and the event *index* = 1. For the disease-unaffected mutation carrier (*colorectal.CA*=0), we set the time = age at censoring = *Age.Death.or.Last.Known.Age* and the event *index* = 0.

```
MUT_0923_pair_merge$time <- rep(0, dim(MUT_0923_pair_merge)[1])

MUT_0923_pair_merge$time[MUT_0923_pair_merge$colorectal.CA==1] <-
MUT_0923_pair_merge$age.at.colorectalCA[MUT_0923_pair_merge$colorectal.CA==1]
MUT_0923_pair_merge$time[MUT_0923_pair_merge$colorectal.CA==0] <-
MUT_0923_pair_merge$Age.Death.or.Last.Known.Age[MUT_0923_pair_merge$colorectal.CA==0]

MUT_0923_pair_merge_new <- subset(MUT_0923_pair_merge, time<999)
```


Data set provided by the CCFR. Obs=2842. Vars =24
Mutation or obligate mutation carrier. Obs=778
315 parent-offspring pairs
315 offspring
315 parents
315 non-duplicated offspring
$p = 0$
182 non-duplicated parents
$p = 1$
496 mutation carriers
225
without disease
$p = 1$ or $p = 0$
271
with disease
$p = 1$ or $p = 0$
time=*Age.Death.or.Last.Known.Age*
event index=*colorectal.CA*=0
$p = 1$ or $p = 0$
time=*age.at.colorectalCA*
event index=*colorectal.CA*=1
$p = 1$ or $p = 0$
survival data set A : 496 obs
event index=*colorectal.CA*=1 or 0
$p = 1$ or $p = 0$

The exhibition of the relation between age at diagnosis/age at censoring and date of birth

Figures 4.2 and 4.3 reveal the relation between age at diagnosis and date of birth and the relation between age at censoring and date of birth, respectively.

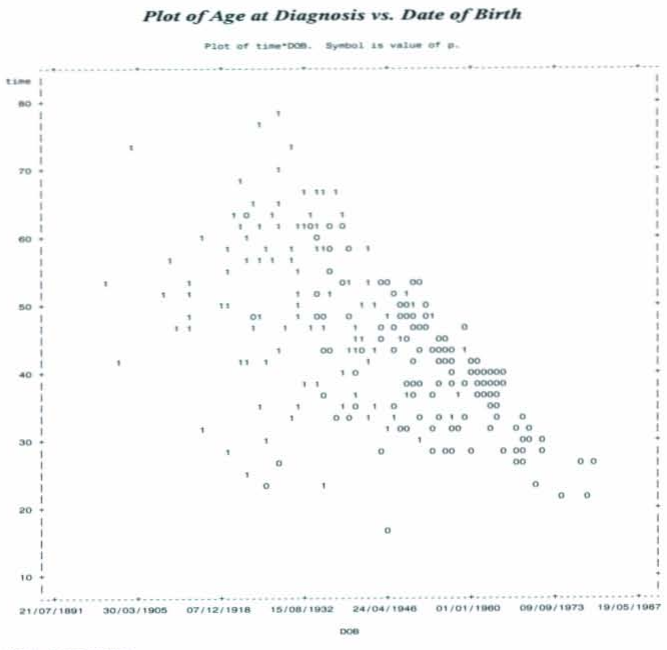


Figure 4.2: Plot of age at diagnosis vs. date of birth. "1" represents parent and "0" represents offspring

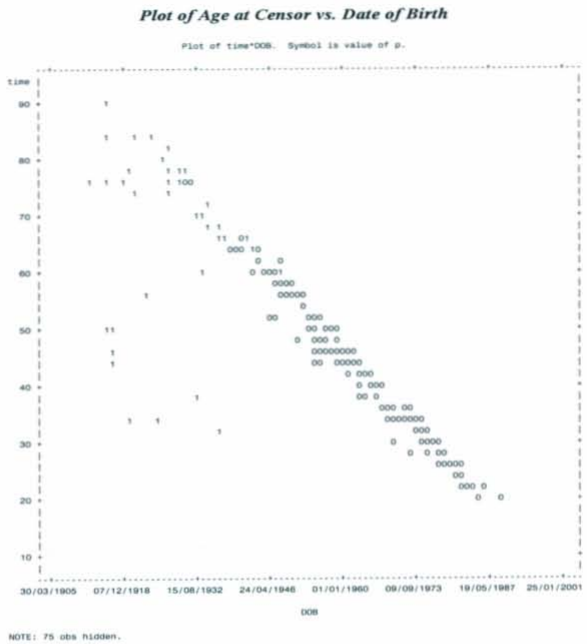


Figure 4.3: Plot of age at censoring vs. date of birth. "1" represents parent and "0" represents offspring

4.2.2 An analysis of the survival data

From Figure 4.3, age at censoring, *Age.Death.or.Last.Known.Age*, can be approximated by

$$\text{age} = 2009 - \text{birth year}$$

for the mutation carriers who were born after 1940. The age at censoring can be predicted by birth year. Actually, these ages at censoring are created by the closing date of the study and are artifacts. However, "A censored observation is one whose

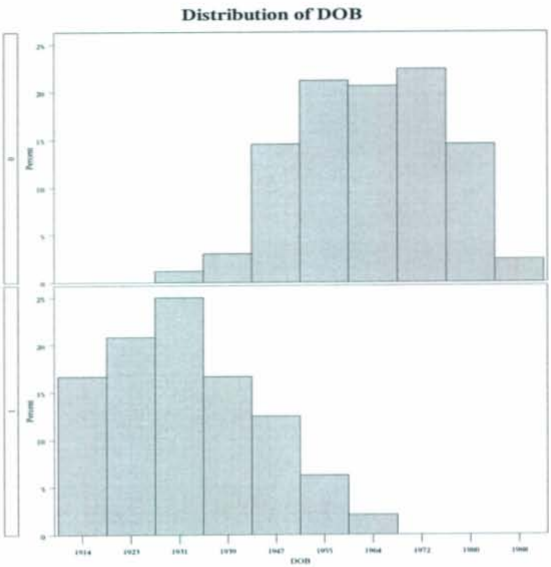


Figure 4.4: The distribution of date of birth: "censored"

value is incomplete due to factors that are random for each subject" [David W Hosmer et al (2007)]. It is obvious that this definition is not met here, especially for the second generation. If we use *Age.Death.or.Last.Known.Age* for those mutation carriers without disease as censoring time (time, event index=0), what will be the consequence?

As an illustrative example, let us consider the following two data sets. Both of them have the same distribution of time-to-event, but a different censoring schema.

Schema 1 and Schema 2 : time to event with event index = 1

mean=median = 35.5

11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35

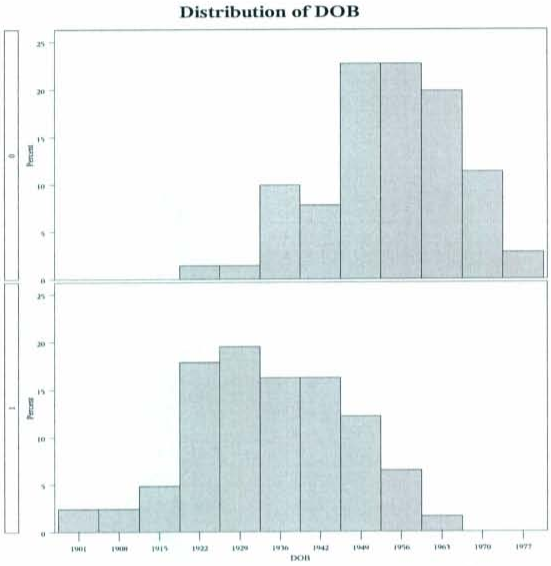


Figure 4.5: The distribution of date of birth: affected

36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60

Schema 1 (early censoring) : time to event with event index = 0

10.5 11.0 11.5 12.0 12.5 13.0 13.5 14.0 14.5 15.0 15.5 16.0 16.5 17.0 17.5
18.0 18.5 19.0 19.5 20.0 20.5 21.0 21.5 22.0 22.5 23.0 23.5 24.0 24.5 25.0

Schema 2 (late censoring) : time to event with event index = 0

44.0 45.5 46.0 46.5 47.0 47.5 48.0 48.5 49.0 49.5 50.0 50.5 51.0 51.5 52.0
52.5 53.0 53.5 54.0 54.5 55.0 55.5 56.0 56.5 57.0 57.5 58.0 58.5 59.0 59.5

We present here the results of the Kaplan-Meier estimators of two survival functions and the log-rank test. The log-rank test in univariate survival analysis is employed to

test the difference in time to event between two samples. Specifically, it is to compare the survival distributions of two samples. Though the two samples have identical distributions of the time to event with event $index = 1$, the two different censoring schemas make the two survival distributions significantly different. The outcome of the log-rank test is to rejects $\mathcal{H}_0$ and in favor of $\mathcal{H}_\alpha$ with p=0.000568. The outcome of the Kaplan-Meier estimators of the two survival functions gives median estimates of the two samples, early vs. late, being 37 vs. 50, Both are larger than 35.5 which is the mean and median of time to event with event $index = 1$.

Survival distributions and test for difference of the two distributions

```
survfit(Surv(time, event index) ~ sample index, data=x)
         records n.max n.start events median 0.95LCL 0.95UCL
sample 1      80    80      80     50     37      32      45
sample 2      80    80      80     50     50      43      57

survdiff(Surv(time, event index) ~ sample index, data=x)
           N Observed Expected (O-E)^2/E (O-E)^2/V
sample 1  80       50     34.2      7.27      11.9
sample 2  80       50     65.8      3.78      11.9
Chisq= 11.9  on 1 degrees of freedom, p= 0.000568
```

The above example demonstrates that a censoring scheme affects the survival function. Returning to the survival data set A (for the description of data set A, please see page 46), since the age at censoring is created by the closing date of the study, the survival estimates based on this data are also artifacts. If this group of individuals is followed up for a longer time, say up to the year 2020, one could expect that the age at censoring would be cut off by

$$\text{age} = 2020 - \text{birth year},$$

thus would have a later censoring scheme. This late censoring scheme alone would result in a survival function with a smaller risk. However, the truth should not be

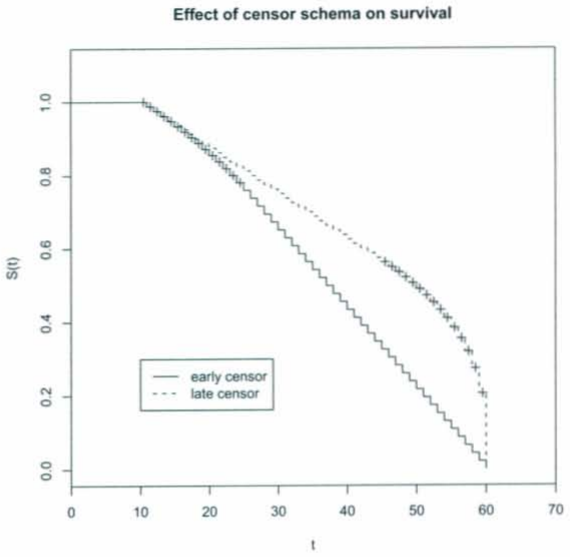


Figure 4.6: Effect of censoring schema on survival

affected by the way the data is collected. We can conclude that applying standard survival analysis directly to survival data set A (that is, MUT_0923_pair_merge_new) without accounting for the sampling bias will result in a biased result.

Standard survival methods are not appropriate for the data on hand

Standard survival analysis methods can give a more accurate result by including censoring time, which provides partial information about time to event into the analysis only if censoring time is random in nature and time to event is not subject to sampling bias. If time to event is subject to sampling bias, even though censoring time is random in nature, standard survival analysis without accounting for the sampling bias will result in a biased result. Censoring time only provides partial information

about an unobserved time to event, but can not correct the bias which exists in the observed time to event. If censoring time is not random in nature, even though time to event represents the population from the previous analysis, the result obtained by standard survival analysis is questionable. From Figures 4.2 and 4.3, we see that both age at diagnosis and age at censoring are subject to the sampling bias and the sampling bias has different effects on different generations. Therefore, it is not adequate to use standard survival analysis directly to analyze this data without accounting for the sampling bias. Furthermore, even if the data are free of sampling bias, to test a decrease in age at diagnosis over successive generations, it is better to consider parent-offspring pairs as bivariate survival data instead of parent generation vs. offspring generation. Tests based on parent-offspring pairs are more sensitive to the decrease in age at diagnosis over successive generations. As a comparison with the result obtained by the symmetry test method in section 4.1, we list results of a survival analysis with censoring time being *Age.Death.or.Last.Known.Age* as follows. Also, please compare the results obtained below with the corresponding ones listed in Table 3.1.

Analysis by the Kaplan-Meier estimators of two survival functions with age at censoring offspring vs. parent

The median age at diagnosis is **48** for offspring vs. the median age at diagnosis is **56** for parent; the decrease in the median age at diagnosis over successive generations amounts to **8** years

```
survfit(Surv(time, colorectal_CA) ~ p, data=MUT_0923_pair_merge_new)
    records n.max n.start events median 0.95LCL 0.95UCL
p=0     315   315     315    143     48      45      52
p=1     181   181     181    128     56      51      61
```

Analysis by the log-rank test with age at censoring offspring vs. parent

The two survival functions are significantly different with p= 0.00152

```
survdiff(Surv(time, colorectal_CA) ~ p, data=MUT_0923_pair_merge_new)
      N Observed Expected (O-E)^2/E (O-E)^2/V
```

```
p=0 315       143       119       4.92       10.1
p=1 181       128       152       3.84       10.1
Chisq= 10.1   on 1 degrees of freedom, p= 0.00152
```

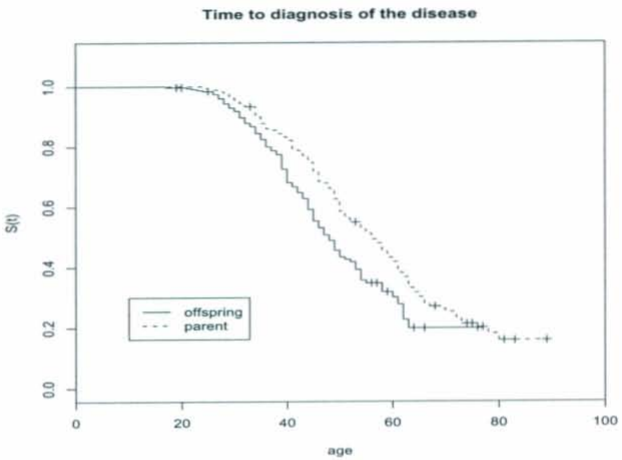


Figure 4.7: The Kaplan-Meier estimators of the survival function with age at censoring: parents vs. offspring

Analysis by the Kaplan-Meier estimators of two survival functions without age at censoring

offspring vs. parent

The median age at diagnosis is **39** for offspring vs. the median age at diagnosis is **49** for parent; the decrease in the median age at diagnosis over successive generations amounts to **10** years

```
survfit(Surv(time, colorectal_CA) ~ p, data=MUT_0923_pair_merge_new,
                                       subset=colorectal_CA==1)
    records n.max n.start events median 0.95LCL 0.95UCL
p=0     143   143     143    143     39      38      41
p=1     128   128     128    128     49      46      51
```

Analysis by the log-rank test without age at censoring

Offspring vs. parent

The two survival functions are significantly different with p=1.95e-13

```
survdiff(Surv(time, colorectal_CA) ~ p, data=MUT_0923_pair_merge_new,
                                        subset=colorectal_CA==1)

      N Observed Expected (O-E)^2/E (O-E)^2/V
p=0 143      143     90.6      30.4        54
p=1 128      128    180.4      15.2        54
Chisq= 54.1  on 1 degrees of freedom, p= 1.95e-13
```

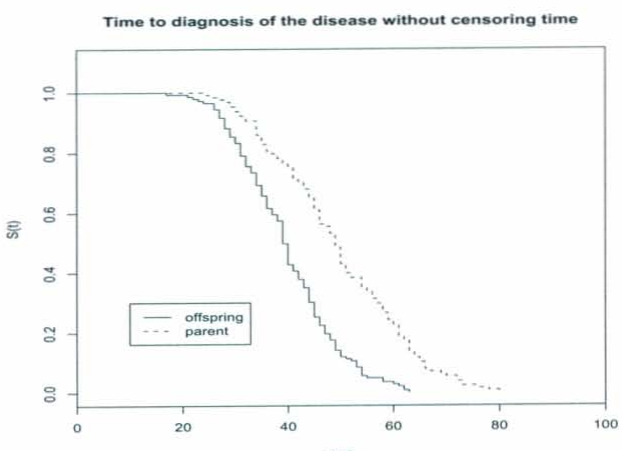


Figure 4.8: The Kaplan-Meier estimators of the survival function without age at censoring: parents vs. offspring

Analysis by the Kaplan-Meier estimators of the survival functions for parent without age at censoring vs. with age at censoring

The median age at diagnosis is **49** vs. the median age at diagnosis is **56**;

the median age at diagnosis increases **7** years by including age at censoring into the analysis

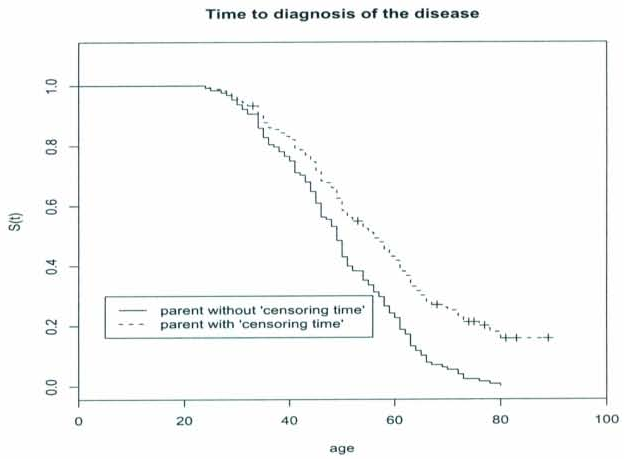


Figure 4.9: The Kaplan-Meier estimators of survival function for parent: without age at censoring vs. with age at censoring

Analysis by the Kaplan-Meier estimators of the survival functions for offspring without age at censoring vs. with age at censoring

The median age at diagnosis is **39** vs. the median age at diagnosis is **48**; the median age at diagnosis increases **9** years by including age at censoring into the analysis

From the above survival analysis (Kaplan-Meier estimate), we find that a survival function corresponding to a smaller risk (of the disease onset) is obtained by including age at censoring, and thus results in a larger estimate of the median is produced. For the parent generation, compared with the distribution of age at diagnosis, the censoring scheme corresponds to an early censoring scheme, while for the offspring generation, compared with the distribution of age at diagnosis, the censoring scheme corresponds to a late censoring scheme, and thus the increase in time to diagnosis of the disease by including age at censoring is larger for offspring than for parents

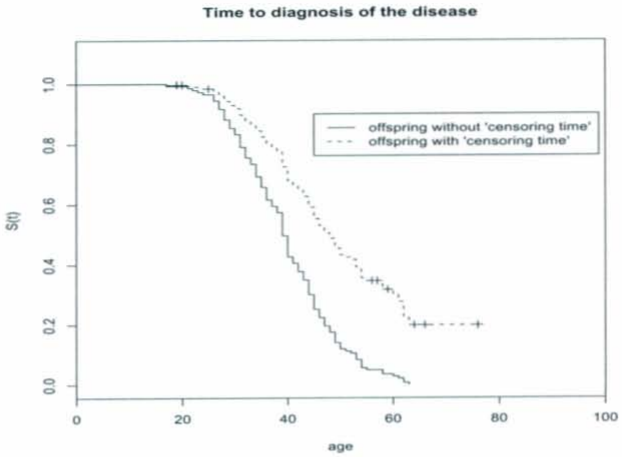


Figure 4.10: The Kaplan-Meier estimators of the survival function for offspring: without age at censoring vs. with age at censoring

(9 years vs. 7 years). If this group of individuals is followed up for a longer time, compared with the parent generation, the data set would have more late age at diagnosis observations and more late age at censoring observations for the offspring generation. As a result, we would expect that the difference between the two survival functions (offspring vs. parents) would decrease further, that is, the decrease of age at diagnosis over successive generations would become smaller.

Analysis by Cox-ph Model

The **risk**† **ratio** in each family **is** $\frac{R(p=0)}{R(p=1)} = \mathbf{1.503807}$ with p-value $= 0.0016$.

The median age at diagnosis is **49** for offspring vs. the median age at diagnosis is **56** for parents;

the decrease in median age at diagnosis over successive generations amounts to **7** years

Note: †Risk of the disease onset.

```
coxph(formula = Surv(time, colorectal_CA) ~ p, data = MUT_0923_pair_merge_new)
```

```
    coef   exp(coef)  se(coef)      z       p
p -0.408      0.665    0.130      -3.15  0.0016
Likelihood ratio test=10  on 1 df, p=0.00156  n= 496
```

```
fit <-coxph(Surv(time, colorectal_CA) ~ p, data = MUT_0923_pair_merge_new)
survfit(fit, newdata=data.frame(p=c(0,1)))
  records n.max n.start events median 0.95LCL 0.95UCL
1     496   496     496    271     49      46      51
2     496   496     496    271     56      51      61
```

The survival analysis used above implicitly assumes a homogenous population for each generation to be studied. This means that all individuals sampled from the same generation in the study are subject, in principle, to the same risk (risk of disease onset). However, the data set under study has a family-clustered structure as shown below

The distribution of family ID

```
table(MUT_0923_pair_merge_new$FamilyID)

    10987 110004550 110006763 110007107 110008508 110011009 110022394 110025272 110038202 121003039 125000010 125000029
        2         4         2         2         3         3         3         2         2         2         3         3
125000047 125000146 125000236 125000308 125000777 125001190 125001217 125001235 125001622 125001767 130151007 130151008
        4         2         5         2         4         2         3         4         5         3         7         3
130151010 130151011 130151013 130151018 130151020 130151021 130151023 130151026 130151031 130151035 130151036 130151039
        4         6         6         2         2         3         3         3         2         2         2         3
130151049 130151053 130151059 130151060 130151062 130151064 130151071 130151077 130151082 130151087 130151106 130252000
        2         4         5         2         2         4         4         2         3         2         2         5
130252001 130252007 130252024 130350018 130350029 130350033 130350050 130350052 130350059 130350061 130350067 130350071
        3         4         2         2        11         3         2         5         7         3         3         4
130350101 130350110 130350123 130350158 130350166 130350224 130350281 130350283 130350292 130350293 130350294 130350297
        3         3         3         3         5         4         2         3         3         2         3         2
130350299 130350313 130350353 130454002 130454003 130454004 130454005 130454008 130454009 130454012 130454015 130454016
        4         3         2         3         2         2         4         3         4         4         5         2
130454017 130454026 130454027 130454030 130454034 130454036 130454041 130454050 130454064 130454065 130454068 130454076
        2         2         2         4         4         2         3         3         2         2         4         5
130454077 130656001 130656002 130656004 130656006 130656007 130656009 130656010 130656029 130656039 130656042 130656050
        2         7         3         2         3         2         2         3         2         4         2         2
130656061 130656062 130656066 130656070 130656072 130656087 130656093 130757006 130757007 130757010 130757016 130757019
        3         2         3         3         2         4         2         2         2         2         2         4
130757020 130757023 130757044 130757046 130757080 130757081 130939003 139910178 139920485 139920628 139920762 139920768
        5         2         2         2         3         2         2         3         2         2         4         2
139920790 140000001 140000158 150550024 150550067 150550130 150550221 150550223 150550229 150550284 150550638 150550641
        3         2         3         2         3         2         2         4         5         2         3         5
150550642 150550643 150550644 150550653 150550744 150560152 150560310 150570132 150570273 150570386 150570476 150570504
        3         2         2         3         2         4         4         2         5         2         3         4
150570623 150570650 150570664 150651076 150651177 155947303 160001077 160001456
        2         2         3         2         2         3         4         2
```

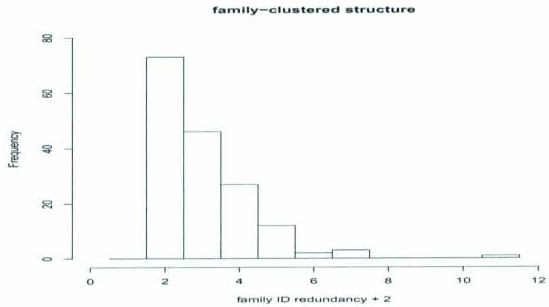


Figure 4.11: Family ID frequency distribution. Note: two individuals of a parent-offspring pair have same family ID. Each Family ID is counted for both parent and his/her offspring, so the count of each Family ID = family ID redundancy + 2

Due to shared genotype and/or environment, individuals from the same generation but from different families may be under different risk of disease onset, while individuals from the same generation as well as from the same family share some common frailty risk. The current studying population may not be assumed to be homogeneous but may more appropriately be considered as a heterogeneous sample for each generation, that is, a mixture of individuals with different hazards for each generation. A natural way to model the dependence of (family) clustered event times is through the introduction of a cluster-specific random effects, the frailty. These random effects explain the dependence in the sense that had we known the frailty, the events would be independent. In other words, the survival times are conditionally independent given the frailty. This approach can be used for survival times of related individuals like family members. The frailty approach is a statistical modeling concept which aims to account for heterogeneity caused by unmeasured covariates. In statistical terms, a frailty model is a random effects model for time to event data, where the random effects (the frailty) have a multiplicative effect on the baseline hazard function [see Andersen et al. (1999)].

To handle multiple affected parents and offsprings within the same family, and to adjust for their mutual dependence, we apply a frailty model, a random effects model. The result of the analysis is given below.

Analysis by Frailty Model

The **risk**† **ratio** in each family **is** $\frac{R(p=0)}{R(p=1)} = \mathbf{1.989732}$ with p-value $= 1.5e-06$.

Note: † Risk of the disease onset.

```
coxph(formula = Surv(time, colorectal_CA) ~ p + frailty(FamilyID,
                dist = "gamma"), data = MUT_0923_pair_merge_new)
                        coef   se(coef)   se2   Chisq    DF        p
p                     -0.688      0.143  0.138   23.1   1.0  1.5e-06
frailty(FamilyID, dist =                        115.2  64.5  1.1e-04
Variance of random effect= 0.402   I-likelihood = -1430.8
Degrees of freedom for terms=  0.9 64.5
Likelihood ratio test=178  on 65.4 df,  p=2.52e-12   n= 496
```

Here we do not intend to fit a model to the data. Finding the model which best fits a biased sample makes no sense. We just want to get a sense of what kind of results can be obtained by the survival methods mentioned above.

4.2.3 Discussion and conclusion

By the Cox proportional hazards model, the risk ratio for offspring to parent is $\frac{R(p=0)}{R(p=1)} = \frac{1}{exp(-0.408)} = 1.503807$ with p-value $= 0.0016$. By the shared frailty Cox proportional hazards model, the risk ratio for offspring to parent in each family is $\frac{R(p=0)}{R(p=1)} = \frac{1}{exp(-0.688)} = 1.989732$ with p-value $=$ 1.5e-06 (Risk of the disease onset will double over successive generations). The increased risk of the disease onset over successive generations implies a decreased age at diagnosis of the disease over successive generations.

To summarize, the previous various survival analyses of this biased sample indicates there exists a more than seven year decrease in median age at diagnosis of LS over successive generations and the decrease is significantly larger than zero. This decrease in median age at diagnosis over successive generations is much larger than the estimate obtained by the symmetry test. The symmetry test suggests a two year decrease in observed median age at diagnosis over successive generations and the decrease is larger than zero with much less significance. The results obtained by these survival analyses reflect the sampling bias more than anything else and they should be regarded as an example of how an inappropriate analysis that fails to properly account for the duration of observation in all individuals being studied will result in a biased result, which can mislead researchers to a wrong conclusion.

Chapter 5

Conclusions

This thesis assessed the decrease in age at diagnosis of LS over successive generations as well as parent-of-origin effects in age at diagnosis of LS based on the data provided by the Colon Cancer Family Registry. We examined the data from parent-child pairs who are known to carry a mutation in one of four causal genes *MSH2*, *MLH1*, *MSH6*, and *PMS2*.

Chapter 2 demonstrated that the variable age at diagnosis in the sample is right truncated by the closing date of the study. As a result, the variable age at diagnosis was subjected to sampling bias, more specifically, persons with early age at diagnosis are over-represented, especially for individuals born later in this sample.

Chapter 3 examined and improved the method used in Lindor et al. (2010) to test parent-of-origin effects. Our preliminary analysis does not support that anticipation for age at diagnosis is more pronounced when the disease allele was transmitted through the father than through the mother or when the disease allele was transmitted from the father to daughter than father to son after accounting for the sampling bias. In summary, we found no evidence for parent-of-origin effects on age at diagnosis of LS in this sample.

Chapter 4 assessed the decrease in age at diagnosis of LS over successive generations. To account for the bias caused by the right truncation of both the parent's and child's age at diagnosis, we employed the symmetry test proposed by Tsai et al. (2005) to detect the decrease in age at diagnosis of LS over successive generations. We found that the observed mean decrease of the paired age at diagnosis is 2.688 years and the observed median decrease of the paired age at diagnosis is 2 years for comparable parent-child pairs. For the null hypothesis that the true mean decrease of paired age at diagnosis is equal to 0 vs. the alternative hypothesis that the true mean decrease of paired age at diagnosis is larger than 0, the t-test and bootstrap test applied to 32 comparable parent-child pairs gave a p-values of 0.0379 and 0.03408, respectively. For the null hypothesis that the true median decrease of paired age at diagnosis is equal to 0 vs. the alternative hypothesis that the true median decrease of paired age at diagnosis is larger than 0, the Wilcoxon signed-rank test and bootstrap test applied to 32 comparable parent-child pairs gave a p-values of 0.03991 and 0.02311, respectively. In summary, the outcome of the symmetry test suggested that there exists evidence in this sample for age at diagnosis anticipation in LS case. This result should be valid and free of ascertainment bias caused by the right truncation if the underlying assumption of the symmetry test that the ordered pairs (T_p, T_c) and (C_p, C_c) are independent is satisfied by this sample.

We also examined standard survival methods for appropriateness to be used to analyze data subject to the sampling bias in Chapter 4. Our analysis demonstrated that the standard survival methods yield biased results in this case due to these methods not accounting for the different duration of observation in all persons being studied which was caused by the closing date of the study

The evidence for anticipation for age at diagnosis presented here, however, could (1) result from selection bias other than that caused by the right truncation, for example, under-representation of "younger parent-older child" pairs in which the parent

had died before producing a "complete" family, (2) reflect changes in environmental factors such as dietary and life style habits, and (3) be attributable, at least in part, to earlier and better diagnosis progressively over time and greater awareness in descendants. Therefore, whether and how much true genetic anticipation contributes to the decrease in age at diagnosis over successive generations observed in this disease still remains uncertain. Genetic anticipation in LS is an interesting issue for biologists.

Our results are preliminary and more work on further developing both the database and the statistical methods used are called for. As large data sets which are free of sampling bias become available in the future, more reliable results of assessment of both anticipation for age at diagnosis and parent-of-origin effects on age at diagnosis of LS can be obtained. With such data, we can also refine the analysis by incorporating other risk factors into the analysis so that the issue of genetic anticipation in LS can be clarified.

Appendix A

R code and Output

A.1 Match Parent-Offspring Pair

We use the following method to obtain the 315 parent-offspring pairs in Section 2.2.

```
dim( MUT_0923)
[1] 778  25

MUT_0923 <- subset(MUT_0923,MotherID !="NA")
 dim(MUT_0923)
[1] 767  25

attach(MUT_0923)
s <- 0
I <- c(0,0)
J <- c(0,0)
o_PersonID <- c(0,0)
p_PersonID <- c(0,0)
...
...
...
```

```
o_PROBAND_FLAG <- c(0,0)
p_PROBAND_FLAG <- c(0,0)

for (i in 1: dim(MUT_0923)[1])
 {for (j in 1 : dim(MUT_0923)[1])
    { if ( PersonID[j]==MotherID[i]|PersonID[j]==FatherID[i])
      { s <- s+1
      I[s] <- i
      J[s] <- j
      o_PersonID[s] = as.character(PersonID[i])
      p_PersonID[s] = as.character(PersonID[j])
      o_MotherID[s] = MotherID[i]
      o_FatherID[s] = FatherID[i]
      o_Sex[s] <- Sex[i]
      p_Sex[s] <- Sex[j]
      o_DOB[s] <- as.character(DOB[i])
      p_DOB[s] <- as.character(DOB[j])
      o_Age_Death.or.Last.Known.Age[s] <- Age_Death.or.Last.Known.Age[i]
      p_Age_Death.or.Last.Known.Age[s] <- Age_Death.or.Last.Known.Age[j]
      o_colorectal_CA[s] <- colorectal_CA [i]
      p_colorectal_CA[s] <-colorectal_CA [j]
      o_age_at_colorectalCA[s] <- age_at_colorectalCA[i]
      p_age_at_colorectalCA[s] <- age_at_colorectalCA[j]
      o_nonlynch[s] <- nonlynch[i]
      p_nonlynch[s] <- nonlynch[j]
      o_age_at_nonlynch[s] <- age_at_nonlynch[i]
      p_age_at_nonlynch[s] <- age_at_nonlynch[j]
      o_GENE[s] <- as.character(GENE[i])
      p_GENE[s] <- as.character(GENE[j])
      o_MUT_STATUS[s] <- as.character (MUT_STATUS[i])
      p_MUT_STATUS[s] <- as.character (MUT_STATUS[j])
```

```
      o_ObligateCarrier[s] <- as.character (ObligateCarrier[i])
      p_ObligateCarrier[s] <- as.character(ObligateCarrier[j])
      o_mut_descrip[s] <- as.character (mut_descrip[i])
      p_mut_descrip[s] <- as.character( mut_descrip[j])
      o_centre_no[s] <- centre_no[i]
      p_centre_no[s] <- centre_no[j]
      o_FamilyID[s] <- FamilyID[i]
      p_FamilyID[s] <- FamilyID[j]
      o_FSRC[s] <- FSRC[i]
      p_FSRC[s] <- FSRC[j]
      o_PROBAND_FLAG[s] <- PROBAND_FLAG[i]
      p_PROBAND_FLAG[s] <- PROBAND_FLAG[j]
      }
   }
 }

o_PersonID <- as.factor(o_PersonID)
p_PersonID <- as.factor(p_PersonID)
o_GENE <- as.factor(o_GENE)
p_GENE <- as.factor(p_GENE)
o_DOB <- as.character(o_DOB)
p_DOB <- as.character(p_DOB)
o_MUT_STATUS <- as.factor(o_MUT_STATUS)
p_MUT_STATUS <- as.factor(p_MUT_STATUS)
o_ObligateCarrier <- as.factor(o_ObligateCarrier)
p_ObligateCarrier <- as.factor(p_ObligateCarrier)
o_mut_descrip <- as.factor(o_mut_descrip)
p_mut_descrip <-as.factor(p_mut_descrip)

detach(MUT_0923)
```

```
MUT_0923_pair<- data.frame(o_PersonID, p_PersonID, o_MotherID, o_FatherID,
o_Sex, p_Sex, o_DOB, p_DOB, o_Age_Death.or.Last.Known.Age,
p_Age_Death.or.Last.Known.Age, o_colorectal_CA, p_colorectal_CA,
o_age_at_colorectalCA, p_age_at_colorectalCA, o_nonlynch, p_nonlynch,
o_age_at_nonlynch, p_age_at_nonlynch, o_GENE, p_GENE, o_MUT_STATUS,
p_MUT_STATUS, o_ObligateCarrier, p_ObligateCarrier, o_mut_descrip,
p_mut_descrip, o_centre_no, p_centre_no, o_FamilyID, p_FamilyID,
o_FSRC, p_FSRC,  o_PROBAND_FLAG,  p_PROBAND_FLAG)

dim(MUT_0923_pair)
[1] 315  34
```

A.2 Match Mother-Offspring Pair

We use the following method to get the 58 mother-offspring pairs in Section 3.2.

```
attach(colorectal_CA_0923)
s <- 0
I <-c(0,0)
J <-c(0,0)
o_PersonID <- c(0,0)
m_PersonID <- c(0,0)
...
...
...
o_PROBAND_FLAG <- c(0,0)
m_PROBAND_FLAG <- c(0,0)

for(i in 1:478)
  {for(j in 1 : 478)
       {if(PersonID[j] == MotherID[i])
          {s <- s+1
```

```
I[s] <- i
J[s] <- j
o_PersonID[s] = as.character(PersonID[i])
m_PersonID[s] = as.character(PersonID[j])
o_MotherID[s] = MotherID[i]
m_MotherID[s] = MotherID[j]
o_Sex[s] <- Sex[i]
m_Sex[s] <- Sex[j]
o_DOB[s] <- as.character(DOB[i])
m_DOB[s] <- as.character(DOB[j])
o_Age_Death.or.Last.Known.Age[s] <- Age_Death.or.Last.Known.Age[i]
m_Age_Death.or.Last.Known.Age[s] <- Age_Death.or.Last.Known.Age[j]
o_colorectal_CA[s] <- colorectal_CA[i]
m_colorectal_CA[s] <- colorectal_CA[j]
o_age_at_colorectalCA[s] <- age_at_colorectalCA[i]
m_age_at_colorectalCA[s] <- age_at_colorectalCA[j]
o_nonlynch[s] <- nonlynch[i]
m_nonlynch[s] <- nonlynch[j]
o_age_at_nonlynch[s] <- age_at_nonlynch[i]
m_age_at_nonlynch[s] <- age_at_nonlynch[j]
o_GENE[s] <- as.character(GENE[i])
m_GENE[s] <- as.character(GENE[j])
o_MUT_STATUS[s] <- as.character(MUT_STATUS[i])
m_MUT_STATUS[s] <- as.character(MUT_STATUS[j])
o_ObligateCarrier[s] <- as.character(ObligateCarrier[i])
m_ObligateCarrier[s] <- as.character(ObligateCarrier[j])
o_mut_descrip[s] <- as.character(mut_descrip[i])
m_mut_descrip[s] <- as.character(mut_descrip[j])
o_centre_no[s] <- centre_no[i]
m_centre_no[s] <- centre_no[j]
o_FamilyID[s] <- FamilyID[i]
```

```
            m_FamilyID[s] <- FamilyID[j]
            o_FSRC[s] <- FSRC[i]
            m_FSRC[s] <- FSRC[j]
            o_PROBAND_FLAG[s] <- PROBAND_FLAG[i]
            m_PROBAND_FLAG[s] <- PROBAND_FLAG[j]
           }
      }
   }

o_PersonID <- as.factor(o_PersonID)
m_PersonID <- as.factor(m_PersonID)
o_DOB <- as.character(o_DOB)
m_DOB <- as.character(m_DOB)
o_GENE <- as.factor(o_GENE)
m_GENE <- as.factor(m_GENE)
o_MUT_STATUS <- as.factor(o_MUT_STATUS)
m_MUT_STATUS <- as.factor(m_MUT_STATUS)
o_ObligateCarrier <- as.factor(o_ObligateCarrier)
m_ObligateCarrier <- as.factor(m_ObligateCarrier)
o_mut_descrip <- as.factor(o_mut_descrip)
m_mut_descrip <-as.factor(m_mut_descrip)

detach(colorectal_CA_0923)

col_CA_MO_pair<- data.frame(o_PersonID, m_PersonID, o_MotherID, m_MotherID,
o_Sex, m_Sex, o_DOB, m_DOB, o_Age_Death.or.Last.Known.Age,
m_Age_Death.or.Last.Known.Age, o_colorectal_CA, m_colorectal_CA,
o_age_at_colorectalCA, m_age_at_colorectalCA,o_nonlynch, m_nonlynch,
o_age_at_nonlynch, m_age_at_nonlynch, o_GENE, m_GENE, o_MUT_STATUS,
m_MUT_STATUS, o_ObligateCarrier, m_ObligateCarrier, o_mut_descrip,
m_mut_descrip, o_centre_no, m_centre_no, o_FamilyID, m_FamilyID,
```

```
o_FSRC, m_FSRC, o_PROBAND_FLAG, m_PROBAND_FLAG)

dim(col_CA_MO_pair)
[1] 58 34
```

A.3 Match Father-Offspring Pair

We use the following method to get the 40 father-offspring pairs in Section 3.2.

```
attach(colorectal_CA_0923)

s <- 0
II <-c(0,0)
JJ <-c(0,0)
of_PersonID <- c(0,0)
 f_PersonID <- c(0,0)
...
...
...
of_PROBAND_FLAG <- c(0,0)
 f_PROBAND_FLAG <- c(0,0)

for (i in 1:478)
  {for (j in 1 : 478)
    {if (PersonID[j] == FatherID[i])
       {s <- s+1
        II[s] <- i
        JJ[s] <- j
        of_PersonID[s] = as.character(PersonID[i])
         f_PersonID[s] = as.character(PersonID[j])
        of_FatherID[s] = FatherID[i]
         f_FatherID[s] = FatherID[j]
```

```
of_Sex[s] <- Sex[i]
 f_Sex[s] <- Sex[j]
of_DOB[s] <- as.character(DOB[i])
 f_DOB[s] <- as.character(DOB[j])
of_Age_Death.or.Last.Known.Age[s] <- Age_Death.or.Last.Known.Age[i]
 f_Age_Death.or.Last.Known.Age[s] <- Age_Death.or.Last.Known.Age[j]
of_colorectal_CA[s] <- colorectal_CA[i]
 f_colorectal_CA[s] <- colorectal_CA[j]
of_age_at_colorectalCA[s] <- age_at_colorectalCA[i]
 f_age_at_colorectalCA[s] <- age_at_colorectalCA[j]
of_nonlynch[s] <- nonlynch[i]
 f_nonlynch[s] <- nonlynch[j]
of_age_at_nonlynch[s] <- age_at_nonlynch[i]
 f_age_at_nonlynch[s] <- age_at_nonlynch[j]
of_GENE[s] <- as.character(GENE[i])
 f_GENE[s] <- as.character(GENE[j])
of_MUT_STATUS[s] <- as.character(MUT_STATUS[i])
 f_MUT_STATUS[s] <- as.character(MUT_STATUS[j])
of_ObligateCarrier[s] <- as.character(ObligateCarrier[i])
 f_ObligateCarrier[s] <- as.character(ObligateCarrier[j])
of_mut_descrip[s] <- as.character(mut_descrip[i])
 f_mut_descrip[s] <- as.character(mut_descrip[j])
of_centre_no[s] <- centre_no[i]
 f_centre_no[s] <- centre_no[j]
of_FamilyID[s] <- FamilyID[i]
 f_FamilyID[s] <- FamilyID[j]
of_FSRC[s] <- FSRC[i]
 f_FSRC[s] <- FSRC[j]
of_PROBAND_FLAG[s] <- PROBAND_FLAG[i]
 f_PROBAND_FLAG[s] <- PROBAND_FLAG[j]
}
```

```
    }
  }

of_PersonID <- as.factor(of_PersonID)
 f_PersonID <- as.factor(f_PersonID)
of_DOB <- as.character(of_DOB)
 f_DOB <- as.character(f_DOB)
of_GENE <- as.factor(of_GENE)
 f_GENE <- as.factor(f_GENE)
of_MUT_STATUS <- as.factor(of_MUT_STATUS)
 f_MUT_STATUS <- as.factor(f_MUT_STATUS)
of_ObligateCarrier <- as.factor(of_ObligateCarrier)
 f_ObligateCarrier <- as.factor(f_ObligateCarrier)
of_mut_descrip <- as.factor(of_mut_descrip)
 f_mut_descrip <- as.factor(f_mut_descrip)

detach(colorectal_CA_0923)

col_CA_FO_pair <- data.frame(of_PersonID, f_PersonID, of_FatherID, f_FatherID,
of_Sex, f_Sex, of_DOB, f_DOB, of_Age_Death.or.Last.Known.Age,
f_Age_Death.or.Last.Known.Age, of_colorectal_CA, f_colorectal_CA,
of_age_at_colorectalCA, f_age_at_colorectalCA, of_nonlynch, f_nonlynch,
of_age_at_nonlynch, f_age_at_nonlynch, of_GENE, f_GENE, of_MUT_STATUS,
f_MUT_STATUS, of_ObligateCarrier, f_ObligateCarrier, of_mut_descrip,
f_mut_descrip, of_centre_no, f_centre_no, of_FamilyID, f_FamilyID, of_FSRC,
f_FSRC, of_PROBAND_FLAG, f_PROBAND_FLAG)

dim(col_CA_FO_pair)
[1] 40 34
```

A.4 Bootstrap Test

The code for the bootstrap test in Table 4.6 and Table 4.7.

```
x<-amf $p_age_at_colorectalCA - amf$o_age_at_colorectalCA
nboot <- 1e+05

x
 4  6  2  6 12  2 -2  14 -6  3  20   2  2  2 10 -4
-5 -7  8  5  1  8  14 -16  -4 15 -16  12  2 -4  0  0

mean.star <- double(nboot)
n <- length(x)
for (i in 1:nboot) {
    k.star <- sample(n, replace = TRUE)
    mean.star[i] <- mean(x[k.star])
}

mean(mean.star)
[1] 2.689320
lower-tailed test of mean = 0
ltpv <- mean(mean.star <= 0)
 ltpv
[1] 0.03408
sort.mean.star<-sort(mean.star)
sort.mean.star[5000]
[1] 0.28125
sort.mean.star[3408]
[1] 0


median.star <- double(nboot)
```

```
for (i in 1:nboot) {
    k.star <- sample(n, replace = TRUE)
    median.star[i] <- median(x[k.star])
 }
mean(median.star)
[1] 2.376735
lower-tailed test of median = 0
mean(median.star<=0)
[1] 0.02311
sort.median.star<-sort(median.star)
sort.median.star[5000]
[1] 1
sort.median.star[2311]
[1] 0
```

A.5 Output of Model Fitting

Regressors: p_sex (1-father, 2-mother)

The GLM Procedure

Dependent Variable: dage

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	93.166964	93.166964	0.62	0.4367
Error	49	7422.989899	151.489590		
Corrected Total	50	7516.156863			

R-Square	Coeff Var	Root MSE	dage Mean
0.012396	79.96352	12.30811	15.39216

Source	DF	Type I SS	Mean Square	F Value	Pr > F
p_Sex	1	93.16696376	93.16696376	0.62	0.4367

Source	DF	Type III SS	Mean Square	F Value	Pr > F
p_Sex	1	93.16696376	93.16696376	0.62	0.4367

Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	14.39393939	B	2.14256707	6.72	<.0001
p_Sex 1	2.82828283	B	3.60647778	0.78	0.4367
p_Sex 2	0.00000000	B	.	.	.

Note: The X'X matrix has been found to be singular, and a generalized inverse was used to solve the normal equations. Terms whose estimates are followed by the letter 'B' are not uniquely estimable.

Figure A.1: Test 2: Effect of the parent on daughter's age at diagnosis depends on the gender of the parent: parameter estimates and significance

Regressors: p_sex (1-father, 2-mother)

The GLM Procedure

Dependent Variable: dage

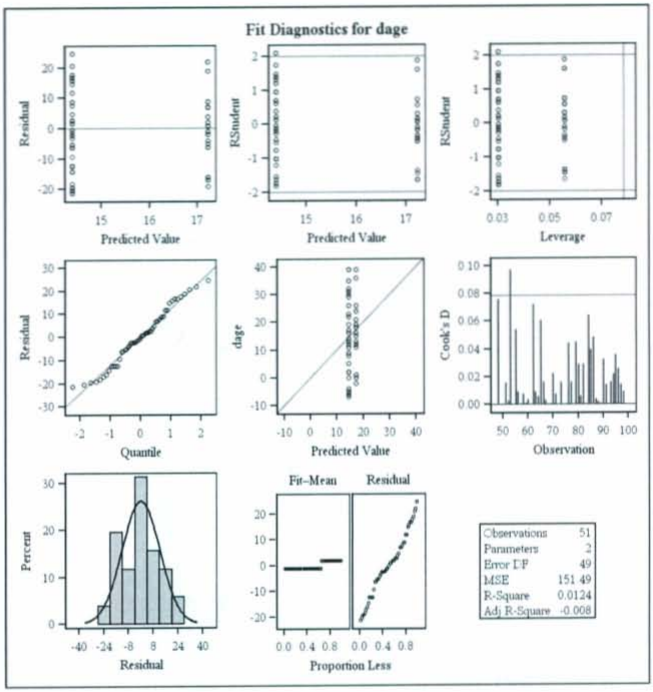


Figure A.2: Test 2: Effect of the parent on daughter's age at diagnosis depends on the gender of the parent: fit diagnostics

Regressors: p_sex (1-father, 2-mother)

The GLM Procedure

Dependent Variable: dage

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	379.47482	379.47482	1.73	0.1956
Error	45	9894.01455	219.86699		
Corrected Total	46	10273.48936			

R-Square	Coeff Var	Root MSE	dage Mean
0.036937	128.1088	14.82791	11.57447

Source	DF	Type I SS	Mean Square	F Value	Pr > F
p_Sex	1	379.4748162	379.4748162	1.73	0.1956

Source	DF	Type III SS	Mean Square	F Value	Pr > F
p_Sex	1	379.4748162	379.4748162	1.73	0.1956

Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	14.24000000	B	2.96558251	4.80	<.0001
p_Sex 1	-5.69454545	B	4.33458576	-1.31	0.1956
p_Sex 2	0.00000000	B	.	.	.

Note: The X'X matrix has been found to be singular, and a generalized inverse was used to solve the normal equations. Terms whose estimates are followed by the letter 'B' are not uniquely estimable.

Figure A.3: Test 3: Effect of the parent on son's age at diagnosis depends on the gender of the parent: parameter estimates and significance

Regressors: p_sex (1-father, 2-mother)

The GLM Procedure

Dependent Variable: dage

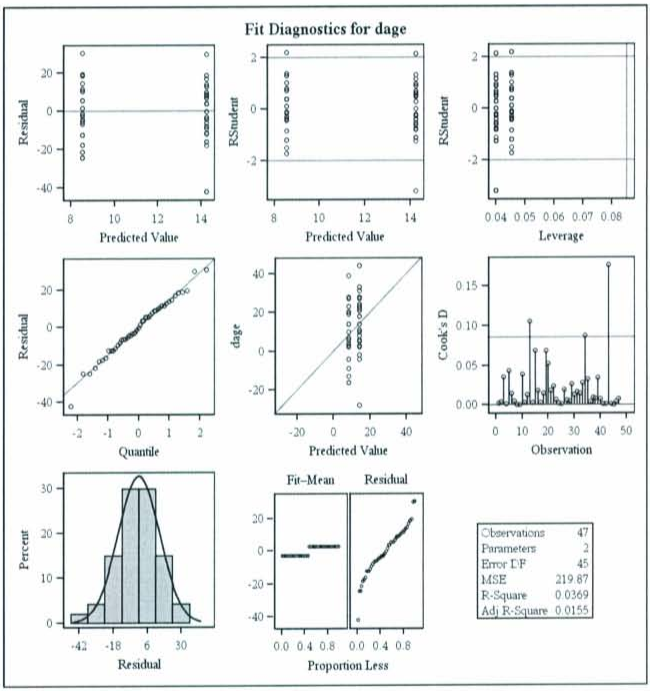


Figure A.4: Test 3: Effect of the parent on son's age at diagnosis depends on the gender of the parent: fit diagnostics

Regressor: o_sex (1-son, 2-daughter)

The GLM Procedure

Dependent Variable: dage

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	0.33707	0.33707	0.00	0.9664
Error	56	10566.43879	188.68641		
Corrected Total	57	10566.77586			

R-Square	Coeff Var	Root MSE	dage Mean
0.000032	95.87321	13.73632	14.32759

Source	DF	Type I SS	Mean Square	F Value	Pr > F
o_Sex	1	0.33707419	0.33707419	0.00	0.9664

Source	DF	Type III SS	Mean Square	F Value	Pr > F
o_Sex	1	0.33707419	0.33707419	0.00	0.9664

Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	14.39393939	B	2.39118588	6.02	<.0001
o_Sex 1	-0.15393939	B	3.64214582	-0.04	0.9664
o_Sex 2	0.00000000	B	.	.	.

Note: The X'X matrix has been found to be singular, and a generalized inverse was used to solve the normal equations. Terms whose estimates are followed by the letter 'B' are not uniquely estimable.

Figure A.5: Test 4 : Effect of the mother on offspring's age at diagnosis depends on the gender of the offspring: parameter estimates and significance

Regressor: o_sex (1-son, 2-daughter)

The GLM Procedure

Dependent Variable: dage

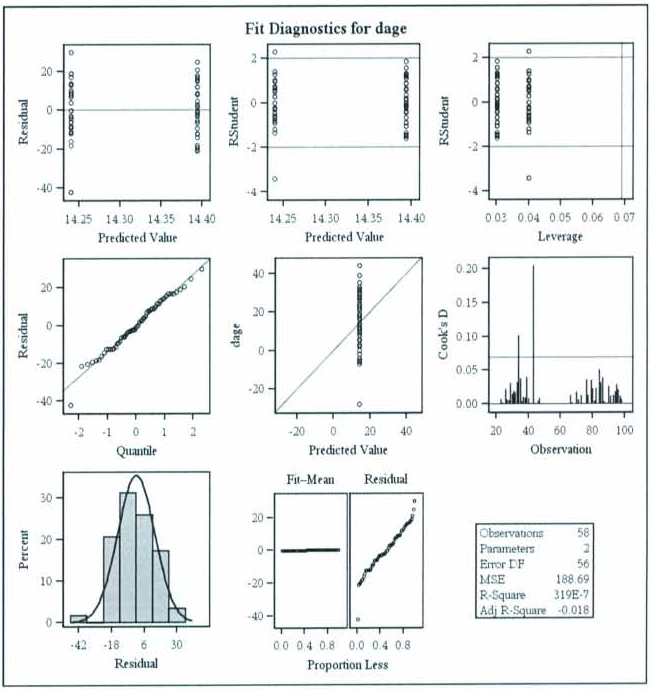


Figure A.6: Test 4 : Effect of the mother on offspring's age at diagnosis depends on the gender of the offspring: fit diagnostics

Regressors: o_sex (1-son, 2-daughter)

The GLM Procedure

Dependent Variable: dage

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	745.334343	745.334343	4.20	0.0475
Error	38	6750.565657	177.646465		
Corrected Total	39	7495.900000			

R-Square	Coeff Var	Root MSE	dage Mean
0.099432	107.0555	13.32841	12.45000

Source	DF	Type I SS	Mean Square	F Value	Pr > F
o_Sex	1	745.3343434	745.3343434	4.20	0.0475

Source	DF	Type III SS	Mean Square	F Value	Pr > F
o_Sex	1	745.3343434	745.3343434	4.20	0.0475

Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	17.22222222	B	3.14153594	5.48	<.0001
o_Sex 1	-8.67676768	B	4.23604619	-2.05	0.0475
o_Sex 2	0.00000000	B	.	.	.

Note: The X'X matrix has been found to be singular, and a generalized inverse was used to solve the normal equations. Terms whose estimates are followed by the letter 'B' are not uniquely estimable.

Figure A.7: Test 5: Effect of the father on offspring's age at diagnosis depends on the gender of the offspring: parameter estimates and significance

Regressors: o_sex (1-son, 2-daughter)

The GLM Procedure

Dependent Variable: dage

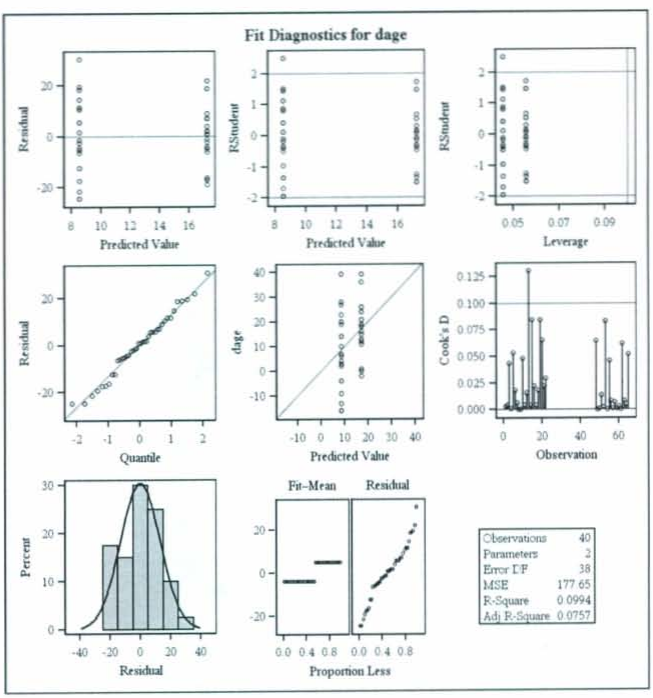


Figure A.8: Test 5: Effect of the father on offspring's age at diagnosis depends on the gender of the offspring: fit diagnostics

Regressors: odob pdob p_sex (1-father, 2-mother)

The GLM Procedure

Dependent Variable: dage

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	2292.918173	764.306058	7.25	0.0006
Error	39	4112.988804	105.461251		
Corrected Total	42	6405.906977			

R-Square	Coeff Var	Root MSE	dage Mean
0.357938	68.25125	10.26943	15.04651

Source	DF	Type I SS	Mean Square	F Value	Pr > F
odob	1	2160.940692	2160.940692	20.49	<.0001
pdob	1	131.683921	131.683921	1.25	0.2706
p_Sex	1	0.293560	0.293560	0.00	0.9582

Source	DF	Type III SS	Mean Square	F Value	Pr > F
odob	1	837.7225149	837.7225149	7.94	0.0075
pdob	1	129.4840080	129.4840080	1.23	0.2746
p_Sex	1	0.2935595	0.2935595	0.00	0.9582

Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	-1485.984875	B	359.3374489	-4.14	0.0002
odob	1.182974		0.4197315	2.82	0.0075
pdob	-0.421787		0.3806551	-1.11	0.2746
p_Sex 1	0.184266	B	3.4925590	0.05	0.9582
p_Sex 2	0.000000	B	.	.	.

Note: The X'X matrix has been found to be singular, and a generalized inverse was used to solve the normal equations. Terms whose estimates are followed by the letter 'B' are not uniquely estimable.

Figure A.9: Test 2: Effect of the parent on daughter's age at diagnosis depends on the gender of the parent: parameter estimates and significance

Regressors: odob pdob p_sex (1-father, 2-mother)

The GLM Procedure

Dependent Variable: dage

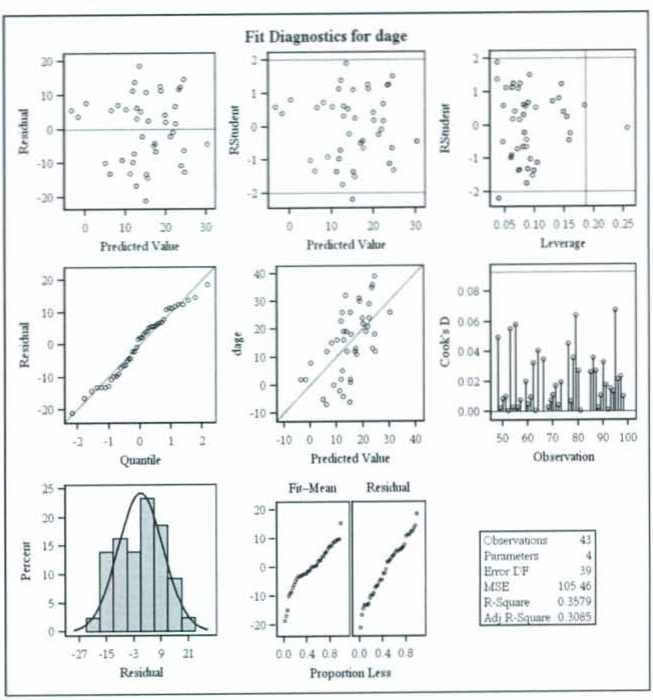


Figure A.10: Test 2: Effect of the parent on daughter's age at diagnosis depends on the gender of the parent: fit diagnostics

Regressors: odob pdob p_sex (1-father, 2-mother)

The GLM Procedure

Dependent Variable: dage

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	1340.386484	446.795495	4.38	0.0104
Error	34	3469.955621	102.057518		
Corrected Total	37	4810.342105			

R-Square	Coeff Var	Root MSE	dage Mean
0.278647	83.27318	10.10235	12.13158

Source	DF	Type I SS	Mean Square	F Value	Pr > F
odob	1	356.8870163	356.8870163	3.50	0.0701
pdob	1	628.4011195	628.4011195	6.16	0.0182
p_Sex	1	355.0983480	355.0983480	3.48	0.0708

Source	DF	Type III SS	Mean Square	F Value	Pr > F
odob	1	1129.342544	1129.342544	11.07	0.0021
pdob	1	822.694111	822.694111	8.06	0.0076
p_Sex	1	355.098348	355.098348	3.48	0.0708

Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	-753.4469448	B	373.6280126	-2.02	0.0517
odob	1.5141806		0.4551844	3.33	0.0021
pdob	-1.1368786		0.4004215	-2.84	0.0076
p_Sex 1	-6.3688524	B	3.4143604	-1.87	0.0708
p_Sex 2	0.0000000	B	.	.	.

Note: The X'X matrix has been found to be singular, and a generalized inverse was used to solve the normal equations. Terms whose estimates are followed by the letter 'B' are not uniquely estimable.

Figure A.11: Test 3: Effect of the parent on son's age at diagnosis depends on the gender of the parent: parameter estimates and significance

Regressors: odob pdob p_sex (1-father, 2-mother)

The GLM Procedure

Dependent Variable: dage

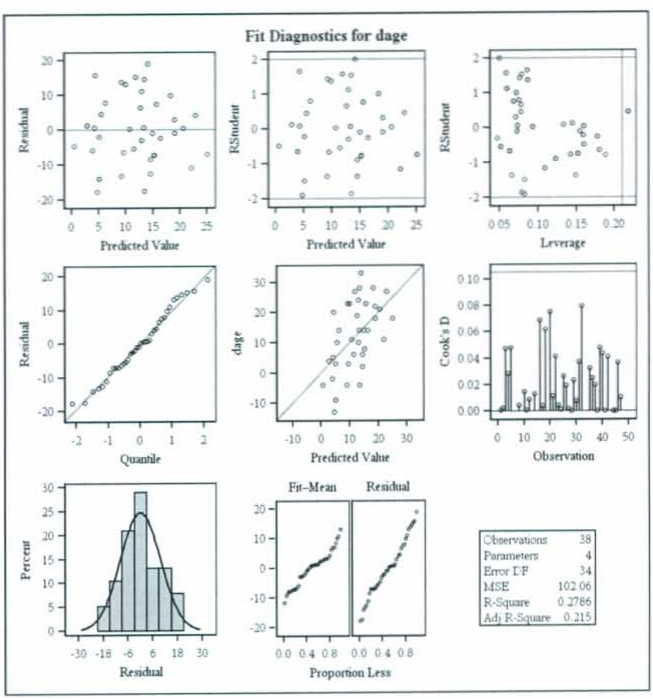


Figure A.12: Test 3: Effect of the parent on son's age at diagnosis depends on the gender of the parent: fit diagnostics

Regressors: odob pdob o_sex (1-son, 2-daughter)

The GLM Procedure

Dependent Variable: dage

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	2251.854419	750.618140	8.17	0.0002
Error	39	3582.424651	91.857042		
Corrected Total	42	5834.279070			

R-Square	Coeff Var	Root MSE	dage Mean
0.385970	66.57850	9.584208	14.39535

Source	DF	Type I SS	Mean Square	F Value	Pr > F
odob	1	1813.987384	1813.987384	19.75	<.0001
pdob	1	435.985186	435.985186	4.75	0.0355
o_Sex	1	1.881849	1.881849	0.02	0.8869

Source	DF	Type III SS	Mean Square	F Value	Pr > F
odob	1	1510.589998	1510.589998	16.45	0.0002
pdob	1	435.781301	435.781301	4.74	0.0355
o_Sex	1	1.881849	1.881849	0.02	0.8869

Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	-1531.478521	B	368.4015157	-4.16	0.0002
odob	1.573164		0.3879331	4.06	0.0002
pdob	-0.792553		0.3638732	-2.18	0.0355
o_Sex 1	-0.420464	B	2.9375996	-0.14	0.8869
o_Sex 2	0.000000	B	.	.	.

Note: The X'X matrix has been found to be singular, and a generalized inverse was used to solve the normal equations. Terms whose estimates are followed by the letter 'B' are not uniquely estimable.

Figure A.13: Test 4 : Effect of the mother on offspring's age at diagnosis depends on the gender of the offspring: parameter estimates and significance

Regressors: odob pdob o_sex (1-son, 2-daughter)

The GLM Procedure

Dependent Variable: dage

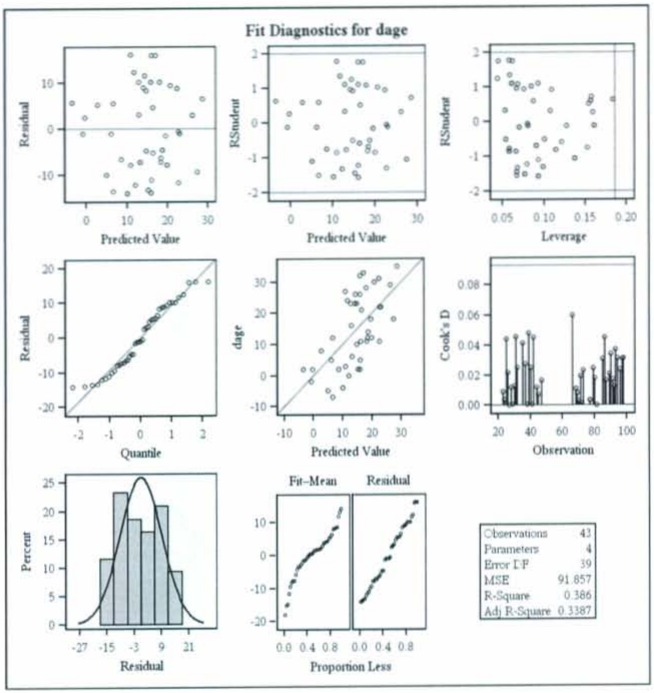


Figure A.14: Test 4 : Effect of the mother on offspring's age at diagnosis depends on the gender of the offspring: fit diagnostics

Regressors: odob pdob o_sex (1-son, 2-daughter)

The GLM Procedure

Dependent Variable: dage

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	1499.514154	499.838051	3.55	0.0261
Error	30	4226.015258	140.867175		
Corrected Total	33	5725.529412			

R-Square	Coeff Var	Root MSE	dage Mean
0.261900	99.88550	11.86875	11.88235

Source	DF	Type I SS	Mean Square	F Value	Pr > F
odob	1	1192.274467	1192.274467	8.46	0.0068
pdob	1	11.522617	11.522617	0.08	0.7768
o_Sex	1	295.717070	295.717070	2.10	0.1577

Source	DF	Type III SS	Mean Square	F Value	Pr > F
odob	1	442.8099376	442.8099376	3.14	0.0864
pdob	1	41.4670031	41.4670031	0.29	0.5914
o_Sex	1	295.7170698	295.7170698	2.10	0.1577

Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	-1126.272549	B	539.8809872	-2.09	0.0456
odob	0.773246		0.4361280	1.77	0.0864
pdob	-0.192777		0.3553113	-0.54	0.5914
o_Sex 1	-6.523543	B	4.5024640	-1.45	0.1577
o_Sex 2	0.000000	B	.	.	.

Note: The X'X matrix has been found to be singular, and a generalized inverse was used to solve the normal equations. Terms whose estimates are followed by the letter 'B' are not uniquely estimable.

Figure A.15: Test 5: Effect of the father on offspring's age at diagnosis depends on the gender of the offspring: parameter estimates and significance

Regressors: odob pdob o_sex (1-son, 2-daughter)

The GLM Procedure

Dependent Variable: dage

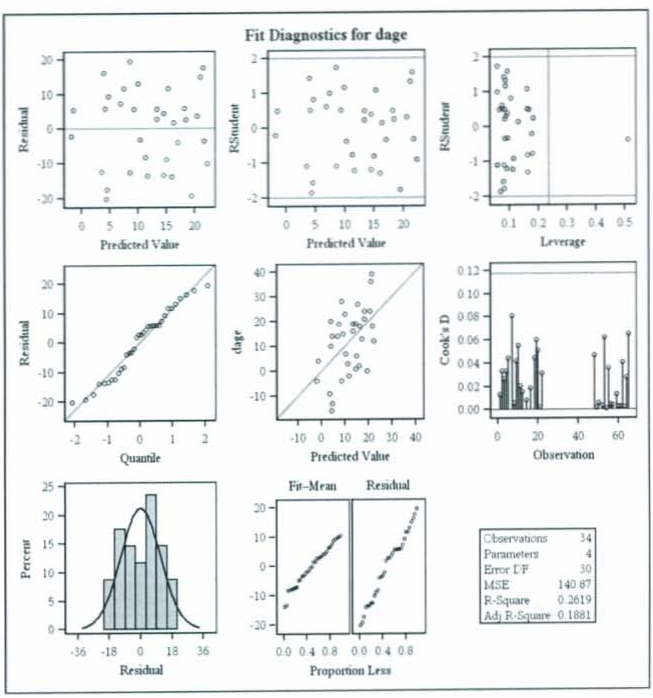


Figure A.16: Test 5: Effect of the father on offspring's age at diagnosis depends on the gender of the offspring: fit diagnostics

Bibliography

[1] Binder D.A. (1992) Fitting Cox's proportional hazards models from survey data. *Biometrika*, **79**, 139-147.

[2] Bowerman B.L. and O'Connell R.T. (1993) Forecasting and time series: an applied approach (third edition).

[3] Daugherty S.E., Pfeiffer R.M., Mellemkjaer L., Hemminki K., and Goldin1 L.R. (2005) No evidence for anticipation in lymphoproliferative tumors in population-based samples. *Cancer Epidemiology, Biomarkers & Prevention*, **14**, 1245-1250.

[4] Forthofer R.N., Lee E.S., and Hernandez M. (2007) Biostatistics: a Guide to design, analysis and discovery (second edition).

[5] Gruber S. and Mukherjee B. (2009) Anticipation in lynch syndrome: still waiting for the answer. *Journal of Clinical Oncology*, **27**, 326-327.

[6] Haynatzki G.R., Brand R.E., Haynatzka V.R., and Lynch H.T. (2007) A comparition of statistical approaches for genetic anticipation with application to pancreatic cancer. *Proceedings of the 40th Hawaii International Conference on System Sciences - 2007*, **IEEE 2007**.

[7] Hosmer D.W., Lemeshow S., and May S. (2007) Applied survival analysis : regression modeling of time-to-event data (second edition).

[8] Hsu L., Zhao L.P., Malone K.E., and Daling J.R. (2000) Assessing changes in ages at onset over successive generation: an application to breast cancer. *Genetic Epidemiology*, **18**, 17-32.

[9] Huang J. and Vieland V. (1997) A new ttatistical test for age-of-onset anticipation: application to bipolar disorder. *Genetic Epidemiology*, **14**, 1091-1096.

[10] Huang J., Vieland V., and Wang K. (2001) Nonparametric estimation of marginal distributions under bivariate truncation with application to testing for age-of-onset anticipation. *Statistica Sinica*, **11**, 1047-1068.

[11] Klein J.P. (1992) Semiparametric estimation of random effects using the Cox model based on the EM algorithm. *Biometrics*, **48**, 795-806.

[12] Lindor N.M., Rabe K.G., Petersen G.M., Chen H., Bapat B., Hopper J., Young J., Jenkins M., Potter J., Newcomb P., Templeton A., Lemarchand L., Grove J., Burgio M.R., Haile R., Green J., Woods M.O., Seminara D., Limburg P.J., and Thibodeau S.N. (2010) Parent of origin effects on age at colorectal cancer diagnosis. *International Journal of Cancer*, **127(2)**, 361-366.

[13] Lynch H.T. and de la Chapelle, A. (1999) Genetic susceptibility to non-polyposis colorectal cancer. *Journal of Medical Genetics*, **36**, 801-818.

[14] Menko F.H., te Meerman G.J., and Sampson J.R. (1993) Variable age of onset in hereditary nonpolyposis colorectal cancer: clinical implications. *Gastroenterology*, **104**, 946-947.

[15] Mérette C., Roy-Gagnon M.H., Ghazzali N., Savard F., Boutin P., Roy M.A., and Maziade M. (2000) Anticipation in schizophrenia and bipolar disorder controlling for an information bias. *American Journal of Medical Genetics*, **96**, 61-68.

[16] Myers R.H., Cupples L.A., Schoenfeld M., D'Agostino R.B., Terrin N.C., Goldmakher N., and Wolf P.A. (1985) Maternal factors in onset of Huntington disease. *American Journal of Human Genetics*, **37**, 511-523.

[17] Nilbert M., Timshel S., Bernstein I., and Larsen K. (2009) Role for genetic anticipation in lynch syndrome. *Journal of Clinical Oncology*, **27**, 360-364.

[18] Picco M.F., Goodman S., Reed J., and Bayless T.M. (2001) Methodologic pitfalls in the determination of genetic anticipation: the case of crohn disease. *Annals of Internal Medicine*, **134**, 1124-1129.

[19] Rabinowitz D. and Yang Q. (1999) Testing for age-at-onset anticipation with affected parent-child pairs. *Biometrics*, **55**, 834-838.

[20] Rodriguez-Bigas M.A., Lee P.H., Malley L., Weber T.K., Suh O., Anderson G.R., and Petrelli N.J. (1996) Establishment of a hereditary nonpolyposis colorectal cancer registry. *Diseases of the Colon & Rectum*, **39**, 649-653.

[21] Tsai W.Y., Heiman G.A., and Hodge S.E. (2005) New simple test for age-at-onset anticipation: application to panic disorder. *Genetic Epidemiology*, **28**, 256-262.

[22] Tsai Y.Y., Petersen G.M., Booker S.V., Bacon J.A., Hamilton S.R., and Giardiello F.M. (1997) Evidence against genetic anticipation in familial colorectal cancer. *Genet Epidemiol*, **14**, 435-446.

[23] Scherer S.J., Avdievich E., and Edelmann W. (2005) Functional consequences of DNA mismatch repair missense mutations in murine models and their impact on cancer predisposition. *Biochemical Society Transactions* **33**, 689-693.

[24] Strachan T. and Read A.P. (1999). Human molecular genetics. Wiley, New York.

[25] Vasen H.F., Taal B.G., Griffioen G., Nagengast F.M., Cats A., Menko F.H., Oskam W., Kleibeuker J.H., Offerhaus G.J., and Khan P.M. (1994) Clinical heterogeneity of familial colorectal cancer and its influence on screening protocols. *Gut*, **35**, 1262-1266.

[26] Westphalen A.A., Russell A.M., Buser M., Berthod C.R., Hutter P., Plasilova M., Mueller H., and Heinimann K. (2005) Evidence for genetic anticipation in hereditary non-polyposis colorectal cancer. *Human Genetics*, **116**, 461-465.