Probability of Carrying a Mutation of Colorectal Cancer Gene hMSH2/hMLH1 Based on Family History

M.Math essay

Qing Li

Supervisor: Professor R. Wayne Oldford

Department of Statistics and Actuarial Science

University of Waterloo,

Waterloo, Ontario, Canada

August 2004

Acknowledgement

I greatly acknowledge my supervisor – Professor R. Wayne Oldford for his guidance on every part of my Master's essay. I also thank Dr. Roger Green and Dr. Ban Younghusband for their valuable suggestions and supplying the data. Thank Professor Ming Li for being my second reader and giving me valuable suggestions on my essay.

Abstract

Background:

The presence of heritable mutations in hMSH2 or hMLH1 genes is highly predictive of the development of colorectal cancer although it is rare. One published estimate is 0.03% of the Scottish population for mutations in the hMSH2 and hMLH1 genes. (5, MG Dunlop et al., 2000, p.1643)

Purpose:

A mathematical model for calculating the probability that a person with a family history of colorectal cancer carries a mutation of hMSH2 or hMLH1 is developed and applied. This model is the analogy of the one developed in the BRACAPRO paper (2, D.A. Berry et al., 1997, pp.227-237).

Methods and Results:

Mendelian genetics and probability theory are used to model the information of the family history of colorectal cancer. The pedigree of the family members is of great importance. For example, a person with a long family history of colorectal cancer will have a greater probability of carrying the mutation than a person with no family history of colorectal cancer. And so, the relationships of all family members can have a substantial impact on the probability of carrying the mutation gene. Other important determinants which might affect this probability include the ages at diagnosis of cancer of affected family members, the number and relationships of members free of cancer, and their current ages or ages at death.

Estimates of hMSH2 and hMLH1 mutation frequencies in the general population are available as are age and sex specific incidence rates of colorectal cancer, for both carriers and non-carriers. These are used to estimate the probability that a particular member of the family carries a mutation.

The model is illustrated for a few simple cases. First the case of a single individuals(with and without cancer) having no family history available. Then two artificial and

one actual family histories are to be considered and the sensitivity of the calculations to various assumptions is addressed.

Conclusion:

This model gives informative and specific probabilities that a particular person carries mutation in hMSH2 or hMLH1 genes.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Colorectal cancer is the third most common cancer and the third most common cause of death from cancer for both men and women. In 2003, there were 18100 new cases, 9800 new cases and 4400 deaths among men, and 8300 new cases and 3800 deaths among women (3, Canadian Cancer Statistics 2003, Table 1, p.19). There exists some genetic mutation genes (e.g. hMSH2 and hMLH1) among a very small proportion of general population. Those with mutation genes have a high probability of developing colorectal cancer. The life time risk of colorectal cancer was 80% for hMSH2/hMLH1 mutation gene carriers (4, H.F.A Vasen et al., 1996, p.1020). Those who have mutation genes are well advised to undergo colonoscopies every six months. Because a colonoscopy is an invasive procedure, it is not likely to be entertained unless there is strong evidence that the person is in fact a carrier. Unfortunatedly it is not always possible to determine with certainty that a person is a carrier. And the potential risks such as psychological distress, restriction of life, and disability insurance and so on, which are associated with genetic testing, make a decision of whether to proceed with testing difficult. The objective of the model is to produce a probability of being a carrier given the available information on the people so that people are supplied with more information before they go for a genetic testing. Because the mutation is inheritable, and the high penetrance of colorectal cancer among people with hMSH2 or hMLH1 mutations means that family history of this disease

is a strong indicator of whether a mutation is present in the family.

The family structure is given by a pedigree as shown in Figure (1.1). Those square nodes represent male, circle nodes represent female. Color of the nodes represents cancer status of the person. A node in grey means this person had been diagnosed with colorectal cancer, while a node in white means this person is free of colorectal cancer. If there is a diagonal slash across the node, it means that individual is dead. Otherwise means they are alive. The number following a capital letter "C" is the age at diagnosis with colorectal cancer. The number following the word "age" is the current age if this person is alive or age at death if this person had died already. Straight line connecting a male and a female means that they are a couple. The vertical line leads to the nodes of their children. For example, in Figure (1.1), member 9 is male, who had been diagnosed with colorectal cancer at the age of 36, and had died at the age of 40. He married with member 10 and had a daughter who is member 6. Among the three children of member 6 and 7, two of them (member 3 and 4) had been diagnosed with colorectal cancer.

A model will be developed to produce the probability that a particular family member carries a mutation at hMSH2 or hMLH1 based on their faimly history. More specifically, the model will produce:

$$Pr(M_i = M|C_i, A_i, S_i, f_C(i), f_A(i), f_S(i), T_i)$$

where M_i is the mutation status of individual i, $M_i = M$ means that individual i is a mutation carrier; C_i, A_i, S_i, $f_C(i)$, $f_A(i)$, $f_S(i)$ and T_i are the information on individual i and the family, which will be introduced later in Chapter 4. Thus, an estimate of the probability of being a carrier can be produced for any individual given the available information on themselves and their family history. If this probability estimate is large, then some or all of the following might be considered: 1. colonoscopies might be recommended (depending on probability and expense); 2. the patient might be tested for the known mutations (depending on expense and specificity and sensitivity of the genetic test which are to be introduced in Chapter 6); 3. If the person tests negative on the

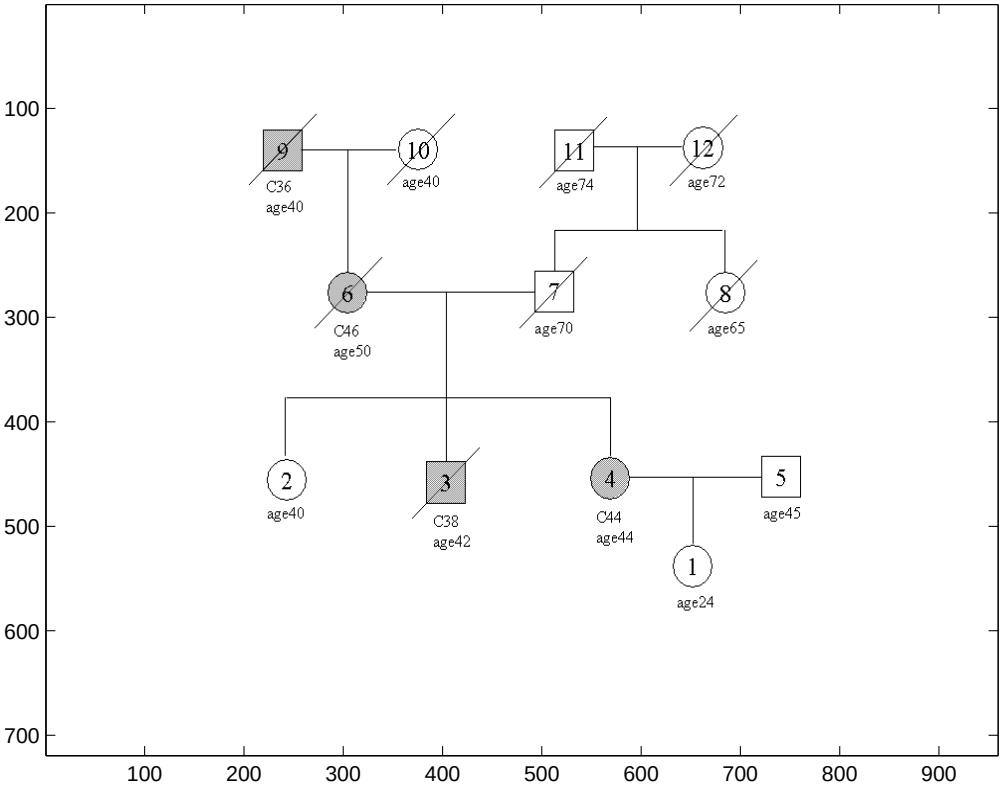

Figure 1.1: Artificial Family History 1 (H_1).

known mutations, the person (or other family members) might be considered for further (and possibly longer term) research to possibly discover a new mutation. Similarly if the probability is very small, no such precedure might be recommended.

Chapter 2

Modelling

The main purpose of this essay is to develop and apply a mathematical model for calculating the probability that a person with a family history of colorectal cancer is a mutation gene carrier. The model will first be described by considering single individuals: a person with unknown cancer status, a person free of colorectal cancer and also a person diagnosed with colorectal cancer. Then, two artificial and one actual family histories are to be considered in Chapters 4 and Chapter 5.

2.1 Three simple examples

2.1.1 Individual with unknown cancer status

A simple case would be a 40-year-old man whose cancer status is unknown. And we have no information about his other family members. The probability for him to be a hMSH2/hMLH1 mutation gene carrier with his cancer status unknown is same as the estimated population mutation gene carrier frequency $Pr(M_i = M) \approx 0.00032$ (5, MG Dunlop et al., 2000, p.1643) if the assumption has been made that the proportion of mutation carriers is same at all ages. This assumption may not be correct since those who are mutation carriers tend to die earlier compared with noncarriers. Unfortunately it is very hard to estimate the proportion of mutation carriers at all ages at this moment.

2.1.2 Individual free of Colorectal Cancer

Suppose we have a single individual who is free of colorectal cancer and beyond this all that is known about the person is their sex and their age. The probability that this individual is a mutation gene carrier given their current age or age of death will be calculated. This situation typically is of special interest since usually those who are cancer free tend to have more interest in knowing the probability for them to be mutation carriers.

Suppose the person is a 60-year-old woman and she is free of colorectal cancer. Suppose we have no information about the other family members. The probability for her to be free of colorectal cancer by the age of 60 given she is a hMSH2/hMLH1 mutation gene carrier is:

Pr(free of colorectal cancer by 60 | M,Female)=$Pr(H|M)$

Thus the probability for her to carry mutation genes given her gender and being free of colorectal cancer by the age of 60 could be calculated as follows:

$$Pr(M_i = M \mid H) = \frac{Pr(H|M_i = M) * Pr(M_i = M)}{Pr(H)} \tag{2.1}$$

where M_i is the mutation status of individual i, $M_i = M$ means this individual i is a mutation gene carrier, and H stands for "being free of colorectal cancer by the age of 60". Equation (2.1) could be rewritten as:

$$\begin{aligned} & Pr(M_i = M|H) \\ = & \frac{Pr(H|M_i = M) * Pr(M_i = M)}{Pr(H|M_i = M) * Pr(M_i = M) + Pr(H|M_i = \overline{M}) * Pr(M_i = \overline{M})} \qquad (2.2) \\ = & \frac{\frac{Pr(H|M_i=M)}{Pr(H|M_i=\overline{M})}}{\frac{Pr(H|M_i=M)}{Pr(H|M_i=\overline{M})} + \frac{1-Pr(M_i=M)}{Pr(M_i=M)}} \qquad (2.3) \end{aligned}$$

Dividing both numerator and denominator of equation (2.2) by $Pr(H|M_i = \overline{M}) * Pr(M_i = M)$ will get equation (2.3).

We need to determine three individual probabilities: $Pr(H|M_i = M)$, $Pr(H|M_i = \overline{M})$ and $Pr(M_i = M)$. Estimates of the cumulative proportion of being diagnosed with colorectal cancer by any given ages are available in the literature of Dunlop (6, MG Dunlop et al., 1997, p.107). He gave these for each sex and each mutation carrier status. Let $B_M(Age, Gender)$ be the cumulative proportion of being diagnosed with colorectal cancer by the given age for mutation carriers, and $B_{\overline{M}}(Age, Gender)$ be the cumulative proportion of being diagnosed with colorectal cancer by the given age for non-mutation carriers. These cumulative risk curves are shown in Figure (2.1) and (2.2) – A, C. The estimation of Dunlop shows that male carriers has higher risk of colorectal cancer than female carriers if they are at the same age.

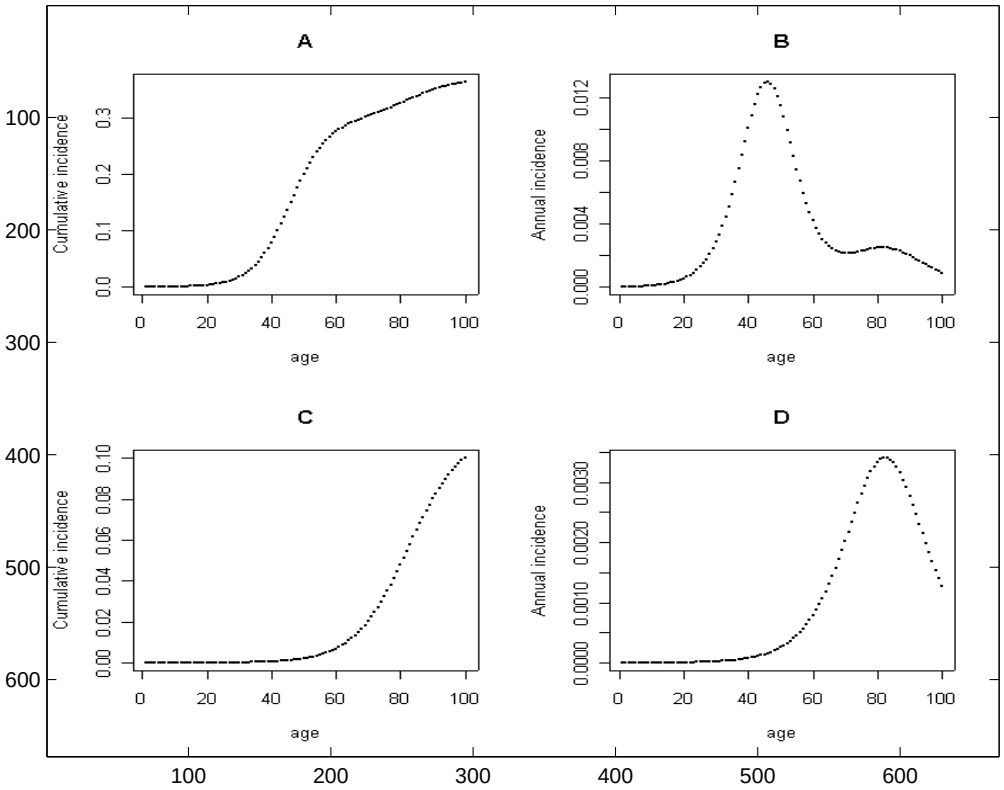

Figure 2.1: A) Cumulative risk of colorectal cancer for female mutation carriers. B) Age-specific incidence of colorectal cancer for female mutation carriers. C) Cumulative risk of colorectal cancer for female noncarriers. D) Age-specific incidence of colorectal cancer for female noncarriers.

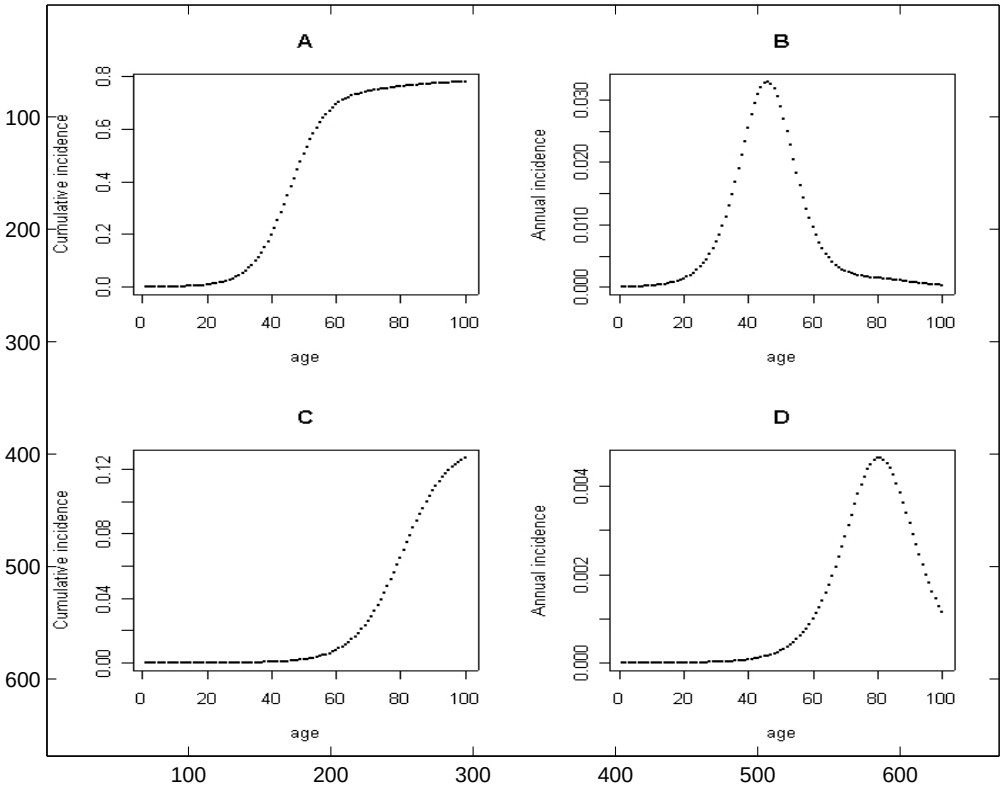

Figure 2.2: A) Cumulative risk of colorectal cancer for male mutation carriers. B) Age-specific incidence of colorectal cancer for male mutation carriers. C) Cumulative risk of colorectal cancer for male noncarriers. D) Age-specific incidence of colorectal cancer for male noncarriers.

Thus equation (2.3) can be rewritten as:

$$= \frac{\frac{1-B_M(60,Female)}{1-B_{\overline{M}}(60,Female)}}{\frac{1-B_M(60,Female)}{1-B_{\overline{M}}(60,Female)} + \frac{1-Pr(M_i=M)}{Pr(M_i=M)}}$$

Where $1 - B_M(60, Female)$ is the probability for a female to be free of colorectal cancer by the age of 60 given she is a hMSH2/hMLH1 mutation gene carrier, and $1 - B_{\overline{M}}(60, Female)$ is the same probability except that she is NOT a hMSH2/hMLH1 mutation gene carrier.

The hMSH2 or hMLH1 gene mutation frequency $Pr(M_i = M)$ in the general population is estimated to be $1/3139 \approx 0.00032$ with the 95% CI:$[1/7626, 1/1247] \approx [0.00013, 0.00080]$ (5, MG Dunlop et al., 2000, p.1643).

For illustration, if we take $B_M(60, female) \approx 0.2751$ and $B_{\overline{M}}(60, female) \approx 0.0069$ (from the cumulative risk curves estimated by Dunlop as shown in Figure (2.1) and (2.2)), we will get $Pr(M_i = M|H) \approx 0.0002$. Thus the evidence that a 60-year-old woman is free of cancer has decreased the probability of carrying mutations in hMSH2 or hMLH1 gene from $Pr(M_i = M) = 0.00032$ at birth to $Pr(M_i = M \mid H) \approx 0.0002$.

2.1.3 Individual with Colorectal Cancer

We then calculate the probability that an individual carries a mutation at hMSH2 or hMLH1 gene given he/she was diagnosed with colorectal cancer. Same as above, suppose we have no information about the other family members. The probability of carrying hMSH2/hMLH1 mutation genes for a man who is diagnosed with colorectal cancer at the age of 35 can be estimated by using formula (2.1), where H stands for "being diagnosed with colorectal cancer at the age of 35". Thus we need to calculate three things:

First, the probability for a male hMSH2/hMLH1 mutation carrier to be diagnosed with colorectal cancer at the age of 35 is:

$$Pr(H \mid M_i = M) = b_M(35, Male)$$

Where H stands for "being diagnosed with colorectal cancer at the age of 35", $b_M(Age, Gender)$ is the estimated age and sex specific incidence rates of colorectal cancer for mutation carriers, and $b_{\overline{M}}(Age, Gender)$ is the same incidence rates for non-mutation carriers. These curves are shown in Figure (2.1) and (2.2) – B, D.

It can be easily inferred that $b_M(Age, Gender)$ is the increase in the cumulative proportion of mutation carriers being diagnosed with cancer by the given age, namely, the derivative of $B_M(Age, Gender)$; And $b_{\overline{M}}(Age, Gender)$ is the increase in the cumulative proportion of non-mutation carriers being diagnosed with cancer by the given age, namely, the derivative of $B_{\overline{M}}(Age, Gender)$.

Second, the probability for a male non-mutation carrier to be diagnosed with colorectal cancer at the age of 35 is:

$$Pr(H \mid M_i = \overline{M}) = b_{\overline{M}}(35, Male)$$

Where $b_{\overline{M}}(35, Male)$ is the probability for a male non-mutation carrier to be diagnosed with colorectal cancer at the age of 35.

Third, by using formula (2.1) we will get:

$$\begin{aligned}
& Pr(M_i = M|H) \\
= & \frac{Pr(H|M_i = M) * Pr(M_i = M)}{Pr(H)} \\
= & \frac{Pr(H|M_i = M) * Pr(M_i = M)}{Pr(H|M_i = M) * Pr(M_i = M) + Pr(H|M_i = \overline{M}) * Pr(M_i = \overline{M})} \\
= & \frac{\frac{Pr(H|M_i=M)}{Pr(H|M_i=\overline{M})}}{\frac{Pr(H|M_i=M)}{Pr(H|M_i=\overline{M})} + \frac{1-Pr(M_i=M)}{Pr(M_i=M)}} \\
= & \frac{\frac{b_M(35,Male)}{b_{\overline{M}}(35,Male)}}{\frac{b_M(35,Male)}{b_{\overline{M}}(35,Male)} + \frac{1-Pr(M_i=M)}{Pr(M_i=M)}}
\end{aligned}$$

Putting it all together by using the cumulative risks and age and sex specific incidence rates estimated by Dunlop (as shown in Figure (2.1) and (2.2)) gives: $Pr(M_i = M|H) \approx 0.1151$. It shows the evidence that this individual had colorectal cancer at a young age

increases the probability that he carries a mutation from $Pr(M_i = M) = 0.00032$ by about 360 folds.

Chapter 3

Mendelian Genetics

Think of $M_i = M$ meaning that a dominant damaged gene (G) can appear on one or another chromosome and this damaged gene appears in the population with frequency f. The population mutation gene carrier frequency $Pr(M_i = M)$ has been estimated to be 0.00032(95%CI: [0.00013,0.0008]) (5, MG Dunlop et al., 2000, p.1643) in the general population, but it may vary, depending on racial or ethnic group. The corresponding estimate of allelic frequency f will be 0.00016 with a 95%CI: [0.000066,0.0004]. Table 3.1 shows the frequencies for the 4 possible gene pairs:

	G	g
G	f^2	$f(1-f)$
g	$f(1-f)$	$(1-f)^2$

Table 3.1: Gene Frequency

Thus:

$$\begin{aligned} Pr(M_i = M) &= Pr(G_i = \text{“}Gg\text{”}) + Pr(G_i = \text{“}GG\text{”}) \\ &= 2f(1-f) + f^2 \\ &= 2f - f^2 \end{aligned} \tag{3.1}$$

where G_i represents the gene type of member i.

$G_{pa(i)}$	G_i	$Pr(G_i \mid G_{pa(i)})$
"GG,GG"	"GG"	1
	"Gg"	0
	"gg"	0
"GG,Gg"	"GG"	0.5
	"Gg"	0.5
	"gg"	0
"GG,gg"	"GG"	0
	"Gg"	1
	"gg"	0
"Gg,Gg"	"GG"	0.25
	"Gg"	0.5
	"gg"	0.25
"Gg,gg"	"GG"	0
	"Gg"	0.5
	"gg"	0.5
"gg,gg"	"GG"	0
	"Gg"	0
	"gg"	1

Table 3.2: The probability for i to get his/her gene type given his/her parents' gene types

Let $pa(i)$ represent the parents of i; $G_{pa(i)}$ represent the gene types of parents of i. The probability for i to get their gene type given their parents' gene types are calculated in Table (3.2).

Let M_i represent the gene status of member i in the family ($M_i = M$ indicates that i is a mutation carrier, $M_i = \overline{M}$ indicates that i is a noncarrier), $Mo(i)$ be the indicator of the mother of member i, and $Fa(i)$ be the indicator of the father of member i. Repeatedly applying Bayes theorem and Mendelian principles, gives the following results:

2.1.1 $Pr(M_i = M | M_{Fa(i)} = M) = \frac{1+f-f^2}{2-f}$

2.1.2 $Pr(M_i = \overline{M} | M_{Fa(i)} = M) = \frac{1-2f+f^2}{2-f}$

2.1.3 $Pr(M_i = M | M_{Fa(i)} = \overline{M}) = f$

2.1.4 $Pr(M_i = \overline{M} | M_{Fa(i)} = \overline{M}) = 1 - f$

2.1.5 $Pr(M_{Fa(i)} = M | M_i = M) = \frac{1+f-f^2}{2-f}$

2.1.6 $Pr(M_{Fa(i)} = \overline{M} | M_i = M) = \frac{1-2f+f^2}{2-f}$

2.1.7 $Pr(M_{Fa(i)} = M | M_i = \overline{M}) = f$

2.1.8 $Pr(M_{Fa(i)} = \overline{M} | M_i = \overline{M}) = 1 - f$

2.1.9 $Pr(M_{Fa(i)} = M, M_{Mo(i)} = M | M_i = M) = \frac{3f-2f^2}{2-f}$

2.1.10 $Pr(M_{Fa(i)} = M, M_{Mo(i)} = \overline{M} | M_i = M) = \frac{(1-f)^2}{2-f}$

2.1.11 $Pr(M_{Fa(i)} = \overline{M}, M_{Mo(i)} = M | M_i = M) = \frac{(1-f)^2}{2-f}$

2.1.12 $Pr(M_{Fa(i)} = \overline{M}, M_{Mo(i)} = \overline{M} | M_i = M) = 0$

2.1.13 $Pr(M_{Fa(i)} = M, M_{Mo(i)} = M | M_i = \overline{M}) = f^2$

2.1.14 $Pr(M_{Fa(i)} = M, M_{Mo(i)} = \overline{M} | M_i = \overline{M}) = f(1 - f)$

2.1.15 $Pr(M_{Fa(i)} = \overline{M}, M_{Mo(i)} = M | M_i = \overline{M}) = f(1 - f)$

2.1.16 $Pr(M_{Fa(i)} = \overline{M}, M_{Mo(i)} = \overline{M} | M_i = \overline{M}) = (1 - f)^2$

2.1.17 $Pr(M_i = M | M_{Fa(i)} = M, M_{Mo(i)} = M) = \frac{3-2f}{(2-f)^2}$

2.1.18 $Pr(M_i = M | M_{Fa(i)} = M, M_{Mo(i)} = \overline{M}) = \frac{1}{2-f}$

2.1.19 $Pr(M_i = M | M_{Fa(i)} = \overline{M}, M_{Mo(i)} = M) = \frac{1}{2-f}$

2.1.20 $Pr(M_i = M | M_{Fa(i)} = \overline{M}, M_{Mo(i)} = \overline{M}) = 0$

For illustration, here is an example of the mathematical reasoning of formula 2.1.1:

$$
\begin{aligned}
& Pr(M_i = M | M_{Fa(i)} = M) \\
= \; & \frac{Pr(M_i = M, M_{Fa(i)} = M)}{Pr(M_{Fa(i)} = M)} \\
= \; & \frac{Pr(M_i = M, G_{Fa(i)} = \text{“}GG\text{”}) + Pr(M_i = M, G_{Fa(i)} = \text{“}Gg\text{”})}{Pr(G_{Fa(i)} = \text{“}GG\text{”}) + Pr(G_{Fa(i)} = \text{“}Gg\text{”})} \\
= \; & \frac{Pr(M_i = M | G_{Fa(i)} = \text{“}GG\text{”}) * Pr(G_{Fa(i)} = \text{“}GG\text{”})}{Pr(G_{Fa(i)} = \text{“}GG\text{”}) + Pr(G_{Fa(i)} = \text{“}Gg\text{”})} \\
& + \frac{Pr(M_i = M | G_{Fa(i)} = \text{“}Gg\text{”}) * Pr(G_{Fa(i)} = \text{“}Gg\text{”})}{Pr(G_{Fa(i)} = \text{“}GG\text{”}) + Pr(G_{Fa(i)} = \text{“}Gg\text{”})} \\
= \; & \frac{1 * f^2 + \frac{1}{2}(1+f) * 2f(1-f)}{f^2 + 2f(1-f)} \\
= \; & \frac{1 + f - f^2}{2 - f}
\end{aligned}
$$

where $Pr(M_i = M | G_{Fa(i)} = \text{“}GG\text{”}) = 1$, $Pr(G_{Fa(i)} = \text{“}GG\text{”}) = f^2$, $Pr(M_i = M | G_{Fa(i)} = \text{“}Gg\text{”}) = \frac{1}{2} + \frac{1}{2} * f$, and $Pr(G_{Fa(i)} = \text{“}Gg\text{”}) = 2f(1-f)$. Other formulas listed above are calculated in the similar way. These formulas are implemented in the following algorithms.

$Pr(M_j | M_i, T_i)$ represent the probability for individual j to have mutation status M_j given that the mutation status M_i of individual i is known. Thus $Pr(M_j | M_i, T_i)$ can be calculated by using Mendelian principles if we know the pedigree for that family. The algorithms for calculating $Pr(M_j | M_i, T_i)$ described in this essay are based on the assumption that any couples in the family do not have common ancestors.

Algorithm1 Prob1(j, M_j, i, M_i): Calculating the probability for member j to have mutation status M_j ($M_j = M$ or $M_j = \overline{M}$) given member i's mutation status M_i ($M_i = M$ if member i is a mutation carrier or $M_i = \overline{M}$ if member i is a non-mutation carrier) and family pedigree.

The relationship between i and j can be one of the following three situations:

1. *parental:*

   formula 2.1.1 to 2.1.8 can be used according to the mutation statuses of M_j and

M_i;

2. *one is the lineal ancestor of the other, but is not a parent:*

let the one with later generation be denoted by l;

let the one with earlier generation be denoted by e;

find the parent of l *who is also the lineal offspring of* e, *and denote this individual by* k. *Then:*

$$\begin{aligned} Prob1(j, M_j, i, M_i) &= Prob1(j, M_j, k, M) * Prob1(k, M, i, M_i) \\ &+ Prob1(j, M_j, k, \overline{M}) * Prob1(k, \overline{M}, i, M_i) \end{aligned}$$

where M *stands for the mutation status of being a carrier, and* $\overline{M}$ *stands for the mutation status of being a noncarrier.*

The mathematical reasoning is shown as follows:

$$\begin{aligned} & Pr(M_j|M_i) \\ = \; & Pr(M_j, M_k = M|M_i) + Pr(M_j, M_k = \overline{M}|M_i) \\ = \; & Pr(M_j|M_k = M, M_i) * Pr(M_k = M|M_i) + Pr(M_j|M_k = \overline{M}, M_i) * Pr(M_k = \overline{M}|M_i) \\ = \; & Pr(M_j|M_k = M) * Pr(M_k = M|M_i) + Pr(M_j|M_k = \overline{M}) * Pr(M_k = \overline{M}|M_i) \end{aligned}$$

since k *is the one inbetween,* M_j *is conditionally independent of* M_i *given* M_k, *that is:* $Pr(M_j|M_k, M_i) = Pr(M_j|M_k)$.

3. *relationship other than 1 and 2:*

find all the lineal ancestors of j *who are also the lineal ancestors of* i,*we call them common ancestors set of* i *and* j;

In common ancestors set find those who are most recent in time; there can be three situations:

(1) cannot find any common ancestor of i *and* j; *This means that the mutation status of* j *is independent of* i*'s mutation status.(e.g. in family history* H_1 *shown in*

Chapter 1, can not find any common ancestors of member 5 and 3.) Thus:

$$Prob1(j, M_j, i, M_i) = \begin{cases} 2 * f - f^2 & \text{if } M_j = M, \\ (1-f)^2 & \text{if } M_j = \overline{M}. \end{cases}$$

since M_j *is independent of* M_i*:* $Pr(M_j|M_i) = Pr(M_j)$.

(2) can only find one common ancestor of i *and* j*, be denoted by* k*; Then:*

$$\begin{aligned} Prob1(j, M_j, i, M_i) &= Prob1(j, M_j, k, M) * Prob1(k, M, i, M_i) \\ &+ Prob1(j, M_j, k, \overline{M}) * Prob1(k, \overline{M}, i, M_i) \end{aligned}$$

The family pedigree tree shown below gives a simple example: Member 2 married twice. Member 4 is the half-sister of member 5, thus they only have one most recent common ancestor – member 2.

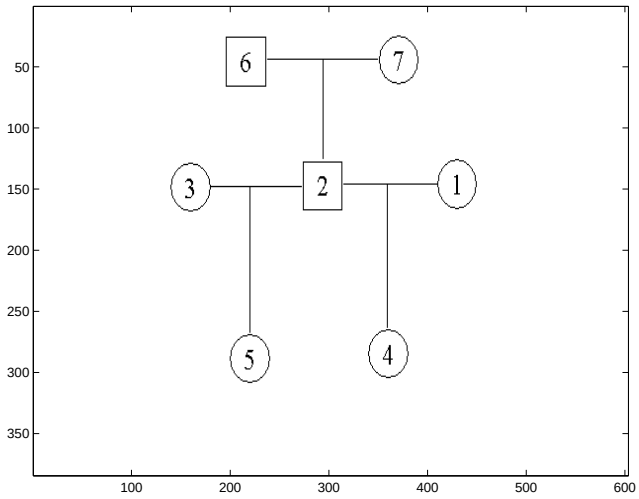

Figure 3.1: Example of one most recent common ancestor

The mathematical reasoning is shown as follows:

$$
\begin{aligned}
& Pr(M_j|M_i) \\
= \;& Pr(M_j, M_k = M|M_i) + Pr(M_j, M_k = \overline{M}|M_i) \\
= \;& Pr(M_j|M_k = M, M_i) * Pr(M_k = M|M_i) \\
& + \; Pr(M_j|M_k = \overline{M}, M_i) * Pr(M_k = \overline{M}|M_i) \\
= \;& Pr(M_j|M_k = M) * Pr(M_k = M|M_i) + Pr(M_j|M_k = \overline{M}) * Pr(M_k = \overline{M}|M_i)
\end{aligned}
$$

since k is the one inbetween, M_j is conditionally independent of M_i given M_k: $Pr(M_j|M_k, M_i) = Pr(M_j|M_k)$.

(3) can find one couple who are common ancestors of i and j, be denoted by k_1 and k_2; Then:

$$
\begin{aligned}
& Prob1(j, M_j, i, M_i) \\
= \;& Prob2(j, M_j, k_1, M, k_2, M) * Prob3(k_1, M, k_2, M, i, M_i) \\
& + \; Prob2(j, M_j, k_1, M, k_2, \overline{M}) * Prob3(k_1, M, k_2, \overline{M}, i, M_i) \\
& + \; Prob2(j, M_j, k_1, \overline{M}, k_2, M) * Prob3(k_1, \overline{M}, k_2, M, i, M_i) \\
& + \; Prob2(j, M_j, k_1, \overline{M}, k_2, \overline{M}) * Prob3(k_1, \overline{M}, k_2, \overline{M}, i, M_i)
\end{aligned}
$$

The mathematical reasoning is shown as follows:

$$
\begin{aligned}
& Pr(M_j|M_i) \\
= \;& Pr(M_j, M_{k_1}=M, M_{k_2}=M|M_i) \\
& + \; Pr(M_j, M_{k_1}=M, M_{k_2}=\overline{M}|M_i) \\
& + \; Pr(M_j, M_{k_1}=\overline{M}, M_{k_2}=M|M_i) \\
& + \; Pr(M_j, M_{k_1}=\overline{M}, M_{k_2}=\overline{M}|M_i) \\
= \;& Pr(M_j|M_{k_1}=M, M_{k_2}=M, M_i) * Pr(M_{k_1}=M, M_{k_2}=M|M_i) \\
& + \; Pr(M_j|M_{k_1}=M, M_{k_2}=\overline{M}, M_i) * Pr(M_{k_1}=M, M_{k_2}=\overline{M}|M_i) \\
& + \; Pr(M_j|M_{k_1}=\overline{M}, M_{k_2}=M, M_i) * Pr(M_{k_1}=\overline{M}, M_{k_2}=M|M_i) \\
& + \; Pr(M_j|M_{k_1}=\overline{M}, M_{k_2}=\overline{M}, M_i) * Pr(M_{k_1}=\overline{M}, M_{k_2}=\overline{M}|M_i) \\
= \;& Pr(M_j|M_{k_1}=M, M_{k_2}=M) * Pr(M_{k_1}=M, M_{k_2}=M|M_i) \\
& + \; Pr(M_j|M_{k_1}=M, M_{k_2}=\overline{M}) * Pr(M_{k_1}=M, M_{k_2}=\overline{M}|M_i) \\
& + \; Pr(M_j|M_{k_1}=\overline{M}, M_{k_2}=M) * Pr(M_{k_1}=\overline{M}, M_{k_2}=M|M_i) \\
& + \; Pr(M_j|M_{k_1}=\overline{M}, M_{k_2}=\overline{M}) * Pr(M_{k_1}=\overline{M}, M_{k_2}=\overline{M}|M_i)
\end{aligned}
$$

since $Pr(M_j|M_{k_1}, M_{k_2}, M_i) = Pr(M_j|M_{k_1}, M_{k_2})$.

Algorithm2 $Prob2(j, M_j, k_1, M_{k_1}, k_2, M_{k_2})$*: Calculating the probability for member* j *to have mutation status* M_j ($M_j = M$ *or* $M_j = \overline{M}$) *given* k_1 *and* k_2*'s mutation statuses as well as family pedigree.*

Since k_1 *and* k_2 *are couple and are the lineal ancestors of* j*, the relationship between* j *and* k_1,k_2 *may have the following two situations:*

1. *parental: (*k_1 *and* k_2 *are* j*'s parents)*

   formula 2.1.17 to 2.1.20 can be used according to the mutation statuses of M_{k_1},M_{k_2} *and* M_j*;*

2. k_1 *and* k_2 *are the lineal ancestors of* j*, but not parents:*

find the parent of j who is also the lineal offspring of k_1 and k_2, be denoted by m. Then:

$$\begin{aligned} & Prob2(j, M_j, k_1, M_{k_1}, k_2, M_{k_2}) \\ = \; & Prob1(j, M_j, m, M) * Prob2(m, M, k_1, M_{k_1}, k_2, M_{k_2}) \\ + \; & Prob1(j, M_j, m, \overline{M}) * Prob2(m, \overline{M}, k_1, M_{k_1}, k_2, M_{k_2}) \end{aligned}$$

The mathematical reasoning is shown as follows:

$$\begin{aligned} & Pr(M_j | M_{k_1}, M_{k_2}) \\ = \; & Pr(M_j, M_m = M | M_{k_1}, M_{k_2}) + Pr(M_j, M_m = \overline{M} | M_{k_1}, M_{k_2}) \\ = \; & Pr(M_j | M_m = M, M_{k_1}, M_{k_2}) * Pr(M_m = M | M_{k_1}, M_{k_2}) \\ + \; & Pr(M_j | M_m = \overline{M}, M_{k_1}, M_{k_2}) * Pr(M_m = \overline{M} | M_{k_1}, M_{k_2}) \\ = \; & Pr(M_j | M_m = M) * Pr(M_m = M | M_{k_1}, M_{k_2}) \\ + \; & Pr(M_j | M_m = \overline{M}) * Pr(M_m = \overline{M} | M_{k_1}, M_{k_2}) \end{aligned}$$

since M_j is conditionally independent of M_{k_1} and M_{k_2} given M_m: $Pr(M_j | M_m, M_{k_1}, M_{k_2}) = Pr(M_j | M_m)$.

Algorithm3 $Prob3(k_1, M_{k_1}, k_2, M_{k_2}, i, M_i)$: Calculating the probability for member k_1 to have mutation status M_{k_1}, member k_2 to have mutation status M_{k_2} given member i's mutation status and family pedigree.

Since k_1 and k_2 are couple and are the lineal ancestors of i, the relationship between i and k_1,k_2 may have the following two situations:

1. *parental: (k_1 and k_2 are i's parents)*

   formula 2.1.9 to 2.1.16 can be used according to the mutation statuses of M_{k_1},M_{k_2} and M_i;

2. *k_1 and k_2 are the lineal ancestors of i, but not parents:*

find the parent of i who is also the lineal offspring of k_1 and k_2, be denoted by m. Then:

$$\begin{aligned} & Prob3(k_1, M_{k_1}, k_2, M_{k_2}, i, M_i) \\ = \; & Prob3(k_1, M_{k_1}, k_2, M_{k_2}, m, M) * Prob1(m, M, i, M_i) \\ & + \; Prob3(k_1, M_{k_1}, k_2, M_{k_2}, m, \overline{M}) * Prob1(m, \overline{M}, i, M_i) \end{aligned}$$

The mathematical reasoning is shown as follows:

$$\begin{aligned} & Pr(M_{k_1}, M_{k_2} | M_i) \\ = \; & Pr(M_{k_1}, M_{k_2}, M_m = M | M_i) + Pr(M_{k_1}, M_{k_2}, M_m = \overline{M} | M_i) \\ = \; & Pr(M_{k_1}, M_{k_2} | M_m = M, M_i) * Pr(M_m = M | M_i) \\ & + \; Pr(M_{k_1}, M_{k_2} | M_m = \overline{M}, M_i) * Pr(M_m = \overline{M} | M_i) \\ = \; & Pr(M_{k_1}, M_{k_2} | M_m = M) * Pr(M_m = M | M_i) \\ & + \; Pr(M_{k_1}, M_{k_2} | M_m = \overline{M}) * Pr(M_m = \overline{M} | M_i) \end{aligned}$$

since M_{k_1} and M_{k_2} are conditionally independent of M_i given M_m: $Pr(M_{k_1}, M_{k_2} | M_m, M_i) = Pr(M_{k_1}, M_{k_2} | M_m)$.

Chapter 4

Application to Two Artificial Family Histories

4.1 Two Artificial Family Histories

Consider two family histories H_1 (as shown in Figure (1.1) in Chapter 1) and H_2 (as shown in Figure (4.1)). Calculation in this part will extend to family members other than the individual of interest. The conditional probability for member i ($i = 1, \ldots, n$. n is the family size) to be a mutation gene carrier given his/her cancer status, age, gender, other family member's cancer statuses, ages and genders, as well as the family relations is of interest:

$$Pr(M_i = M | C_i, A_i, S_i, f_C(i), f_A(i), f_S(i), T_i) \tag{4.1}$$

where $M_i, C_i, A_i, S_i, f_C(i), f_A(i), f_S(i), T_i$ are defined as the following.

For a particular family, let M_i represent the mutation status of individual i. There are two possible values for M_i:

$$M_i = \begin{cases} M & \text{if individual } i \text{ has mutation genes in hMSH2/hMLH1,} \\ \overline{M} & \text{if individual } i \text{ does not have mutation genes in hMSH2/hMLH1.} \end{cases}$$

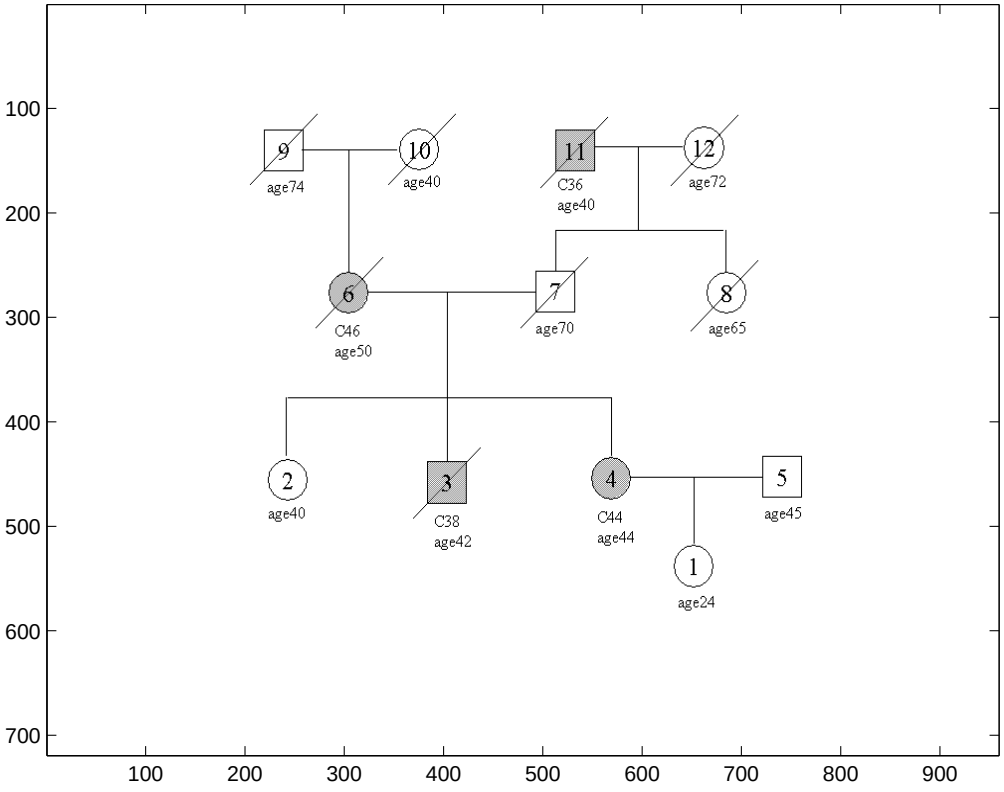

Figure 4.1: Artificial Family History 2 (H_2), almost same as H_1 but switches 9 and 11.

Let S_i represent the gender of individual i, thus:

$$S_i = \begin{cases} M & \text{if individual } \ i \ \text{ is male,} \\ F & \text{if individual } \ i \ \text{ is female.} \end{cases}$$

Let A_i represent age associated with individual i. A_i could be: (1) age of diagnosed with cancer if individual i has been diagnosed with colorectal cancer at a certain age, or (2) current age if individual i is alive and is free of colorectal cancer, or (3) age of death if individual i is dead and was free of colorectal cancer all their life.

Let T_i represent the pedigree tree for individual i. This is simply the structure of the family tree. For all the family members in a particular family, they all have the same pedigree tree for the whole family. That is for any two members i and j in the same family, $T_i = T_j$.

Let $f_A(i)$ denote the ages associated with each family member excluding member i.

Similarly, $f_S(i)$ denote the genders associated with each family member excluding member i.

Let C_i represent the target cancer status of individual i in the family; C_i can take values:

$$C_i = \begin{cases} C & \text{if individual } \ i \ \text{ has been diagnosed with a target cancer (e.g. colorectal cancer),} \\ \overline{C} & \text{if individual } \ i \ \text{ is cancer free by current age or by the age of death.} \end{cases}$$

Let $f_C(i)$ denote the known cancer statuses associated with each family member excluding member i.

For the moment, we will take C_i to indicate whether individual i has been diagnosed with colorectal cancer or not. In the data set which will be considered later in the essay, $C_i = C$ could be the combination of the following 4 situations: (1) diagnosed with colorectal cancer; (2) HNPCC Tumour has been found; (3) Other Tumour has been found; (4) Other cancer has been diagnosed.

Let O_i represent any other information of individual i, this might contains other can-

cer statuses (e.g. endometrial cancer, small bowel cancer, urinary tract cancer, stomach cancer and ovaries cancer), smoking history, alcohol drinking history, dietary intake, etc.

Let $f_O(i)$ represent any other information associated with each family member excluding member i.

Since the data set used later in this essay don't have the information as stated in O_i, O_i and $f_O(i)$ will not be considered in this model. As more information be abtained, an extended model could be considered by incorporating family history of other cancers which are highly predictive of the presence of hMSH2/hMLH1 mutation genes.

The probability for member i to be a hMSH2/hMLH1 mutation gene carrier given his/her age, gender, other family members' ages and genders, as well as the family relations is written as:

$$Pr(M_i = M \mid A_i, S_i, f_A(i), f_S(i), T_i) \tag{4.2}$$

If for any other individual j in the family ($j \neq i$ where i is the individual we are interested in), A_j, S_j, C_j any one of them is missing, although j will still appear in the family pedigree tree, their cancer statuses will not contribute in calculating the conditional probability for individual i to be a mutation carrier given family history.

We begin by applying Bayes' theorem to express equation (4.1) in terms of other probabilities for which we are able to model:

$$\begin{aligned}
& Pr(M_i = M \mid C_i, A_i, S_i, f_C(i), f_A(i), f_S(i), T_i) \\
= \; & \frac{Pr(M_i = M, C_i, f_C(i) \mid A_i, S_i, f_A(i), f_S(i), T_i)}{Pr(C_i, f_C(i) \mid A_i, S_i, f_A(i), f_S(i), T_i)} \\
= \; & \frac{Pr(C_i, f_C(i) \mid M_i = M, A_i, S_i, f_A(i), f_S(i), T_i) * Pr(M_i = M \mid A_i, S_i, f_A(i), f_S(i), T_i)}{Pr(C_i, f_C(i) \mid A_i, S_i, f_A(i), f_S(i), T_i)} \\
= \; & \frac{Pr(C_i, f_C(i) \mid M_i = M, A_i, S_i, f_A(i), f_S(i), T_i) * (4.2)}{Pr(C_i, f_C(i) \mid A_i, S_i, f_A(i), f_S(i), T_i)}
\end{aligned} \tag{4.3}$$

where the left part of the numerator of equation (4.3):

$$Pr(C_i, f_C(i) \mid M_i = M, A_i, S_i, f_A(i), f_S(i), T_i) \tag{4.4}$$

is the joint probability of cancer statuses of all family members given that member i is a mutation gene carrier, ages and genders of all the family members and the family relations. Similarly,

$$Pr(C_i, f_C(i) \,|\, M_i = \overline{M}, A_i, S_i, f_A(i), f_S(i), T_i) \tag{4.5}$$

is the joint probability of cancer statuses of all family members given that member i is NOT a mutation gene carrier, ages and genders of all the family members and the family relations. (4.2) in the mathematical reasoning denotes the expression in equation (4.2). Same rule applies in this essay.

The denominator of equation (4.3) is the joint probability of cancer statuses of all family members given member i's age, gender, other family members' ages and genders, as well as the family relations. It can be expressed in terms of equation (4.4),(4.5) and (4.2), namely:

$$\begin{aligned}
& Pr(C_i, f_C(i) \,|\, A_i, S_i, f_A(i), f_S(i), T_i) \\
= \;& Pr(C_i, f_C(i), M_i = M | A_i, S_i, f_A(i), f_S(i), T_i) \\
& + \; Pr(C_i, f_C(i), M_i = \overline{M} | A_i, S_i, f_A(i), f_S(i), T_i) \\
= \;& Pr(C_i, f_C(i) \,|\, M_i = M, A_i, S_i, f_A(i), f_S(i), T_i) \\
& \quad * Pr(M_i = M \,|\, A_i, S_i, f_A(i), f_S(i), T_i) \\
& + \; Pr(C_i, f_C(i) \,|\, M_i = \overline{M}, A_i, S_i, f_A(i), f_S(i), T_i) \\
& \quad * Pr(M_i = \overline{M} \,|\, A_i, S_i, f_A(i), f_S(i), T_i) \\
= \;& (4.4) * (4.2) + (4.5) * (1 - (4.2))
\end{aligned} \tag{4.6}$$

Combining (4.3), (4.4), (4.5) and (4.6) gives:

$$\begin{aligned}
& Pr(M_i = M \,|\, C_i, A_i, S_i, f_C(i), f_A(i), f_S(i), T_i) \\
= \;& \frac{(4.4) * (4.2)}{(4.4) * (4.2) + (4.5) * [1 - (4.2)]}
\end{aligned} \tag{4.7}$$

So the three probabilities we actually need are (4.4), (4.5) and (4.2). Since (4.4) and (4.5) are similar in how we approach them, let's look at (4.2) first.

In the model, the assumption has been made that other family members' relations, ages and genders provide no information on the probability that member i being a mutation carrier. Available estimates of hMSH2 and hMLH1 mutation frequency in the general population can be used to estimate the probability for an individual to be a mutation carrier when absent the gene status of the other family members. Thus (4.2) can be calculated by:

$$\begin{aligned} & Pr(M_i = M \mid A_i, S_i, f_A(i), f_S(i), T_i) \\ = \;& Pr(M_i = M) \\ = \;& 2f - f^2 \end{aligned}$$

And so, from (4.7) we know that (4.1) can be rewritten as:

$$\begin{aligned} & Pr(M_i = M | C_i, A_i, S_i, f_C(i), f_A(i), f_S(i), T_i) \\ = \;& \frac{(4.4) * (4.2)}{(4.4) * (4.2) + (4.5) * [1 - (4.2)]} \\ = \;& \frac{\frac{(4.4)}{(4.5)}}{\frac{(4.4)}{(4.5)} + \frac{1-(4.2)}{(4.2)}} \\ = \;& \frac{LR}{LR + \frac{1-(4.2)}{(4.2)}} \\ = \;& \frac{LR}{LR + \frac{Pr(M_i=\overline{M})}{Pr(M_i=M)}} \end{aligned}$$

Where the likelihood ratio, LR, is:

$$\begin{aligned} LR &= \frac{(4.4)}{(4.5)} \\ &= \frac{Pr(C_i, f_C(i) | M_i = M, A_i, S_i, f_A(i), f_S(i), T_i)}{Pr(C_i, f_C(i) | M_i = \overline{M}, A_i, S_i, f_A(i), f_S(i), T_i)} \end{aligned} \tag{4.8}$$

and $\frac{Pr(M_i=\overline{M})}{Pr(M_i=M)}$ is the unconditional odds against being a carrier. From equation (3.1) we know that $Pr(M_i = M) = 2f - f^2$, so $Pr(M_i = \overline{M}) = 1 - Pr(M_i = M) = (1 - f)^2$. By using the population mutation gene allelic frequency estimate of 0.00016 (5, MG Dunlop et al., 2000, p.1643), we can calculate the unconditional odds against being a carrier to be $\frac{Pr(M_i=\overline{M})}{Pr(M_i=M)} = (1 - 0.00032)/0.00032 \approx 3124$. The conditional probability of being a carrier is then:

$$Pr(M_i = M|H) \approx \frac{LR}{LR + 3124} \tag{4.9}$$

The information in a family history enters through the LR, which compares the probability of the individual i's actual family history assuming member i is a mutation carrier with the probability of i's family history assuming that member i is a non-mutation carrier. From (4.9) we can see that $Pr(M_i = M|H)$ is a monotonic function which increases as LR increases (LR takes the value from 0 to infinity). $LR = 1$ is the break-even point at which family history is noninformative as regards M. Or, in other words, the cancer status of entire family is not informed by mutation status of i. Thus, $Pr(M_i = M|H) = Pr(M_i = M) \approx 0.00032$ when $LR = 1$.

In order to calculate LR, we need to find out (4.4) and (4.5). The numerator of the likelihood ratio, $Pr(C_i, f_C(i)|M_i = M, A_i, S_i, f_A(i), f_S(i), T_i)$, is the probability for observing the current family cancer history given the rest of the family information and that member i is indeed a mutation carrier.

According to Professor Oldford's conversation with Dr. Roger Green and Dr. Ban Younghusband (1), the assumption that the occurrences of colorectal cancer in a family are conditionally independent given the family members' genetic statuses, is reasonable. So when we assume that certain family members carry mutations while others do not, we can multiply the corresponding probabilities of these observations. Without conditioning on the family members' genetic statuses, occurrences of colorectal cancer are not independent. Observing colorectal cancer in one family member makes it more likely to observe colorectal cancer in another member.

By the conditional independence assumption of the occurrences of colorectal cancer

in a family given the family members' genetic statuses, (4.4) can be rewritten as the following:

$$
\begin{aligned}
& Pr(C_i, f_C(i) | M_i = M, A_i, S_i, f_A(i), f_S(i), T_i) \\
= \; & Pr(C_i | M_i = M, A_i, S_i, f_A(i), f_S(i), T_i) \prod_{j=1, j \neq i}^{n} Pr(C_j | M_i = M, A_i, S_i, f_A(i), f_S(i), T_i) \\
= \; & Pr(C_i | M_i = M, A_i, S_i) \prod_{j=1, j \neq i}^{n} Q_{Mj}
\end{aligned}
$$

Where

$$
\begin{aligned}
& Q_{Mj} \\
= \; & Pr(C_j | M_i = M, A_i, S_i, f_A(i), f_S(i), T_i) \\
= \; & Pr(C_j | M_i = M, A_j, S_j, f_A(j), f_S(j), T_i) \\
= \; & Pr(C_j, M_j = M | M_i = M, A_j, S_j, f_A(j), f_S(j), T_i) \\
& + \; Pr(C_j, M_j = \overline{M} | M_i = M, A_j, S_j, f_A(j), f_S(j), T_i) \\
= \; & Pr(C_j | M_j = M, M_i = M, A_j, S_j, f_A(j), f_S(j), T_i) \\
& \quad * Pr(M_j = M | M_i = M, A_j, S_j, f_A(j), f_S(j), T_i) \\
& + \; Pr(C_j | M_j = \overline{M}, M_i = M, A_j, S_j, f_A(j), f_S(j), T_i) \\
& \quad * Pr(M_j = \overline{M} | M_i = M, A_j, S_j, f_A(j), f_S(j), T_i) \\
= \; & Pr(C_j | M_j = M, A_j, S_j) * Pr(M_j = M | M_i = M, T_i) \\
& + \; Pr(C_j | M_j = \overline{M}, A_j, S_j) * Pr(M_j = \overline{M} | M_i = M, T_i), \qquad for j = 1, 2, \ldots, n, j \neq i
\end{aligned}
$$

and

$$Pr(C_j|M_j = M, A_j, S_j) = \begin{cases} b_M(A_j, S_j) & \text{if } C_j\text{=diagnosed with cancer,} \\ 1 - B_M(A_j, S_j) & \text{if } C_j\text{=cancer free.} \end{cases}$$

$$Pr(C_j|M_j = \overline{M}, A_j, S_j) = \begin{cases} b_{\overline{M}}(A_j, S_j) & \text{if } C_j\text{=diagnosed with cancer,} \\ 1 - B_{\overline{M}}(A_j, S_j) & \text{if } C_j\text{=cancer free.} \end{cases}$$

$Pr(M_j|M_i, T_i)$ can be calculated by using Mendelian principles if we know the pedigree for that family. The algorithms for calculating $Pr(M_j|M_i, T_i)$ are shown in Chapter 3.

Similarly, the denominator of the likelihood ratio, $Pr(C_i, f_C(i)|M_i = \overline{M}, A_i, S_i, f_A(i), f_S(i), T_i)$ can also be rewritten as the following by the conditional independence assumption of the occurrences of colorectal cancer in a family given the family members' genetic statuses.

$$\begin{aligned} & Pr(C_i, f_C(i)|M_i = \overline{M}, A_i, S_i, f_A(i), f_S(i), T_i) \\ = \; & Pr(C_i|M_i = \overline{M}, A_i, S_i, f_A(i), f_S(i), T_i) \prod_{j=1, j\neq i}^{n} Pr(C_j|M_i = \overline{M}, A_i, S_i, f_A(i), f_S(i), T_i) \\ = \; & Pr(C_i|M_i = \overline{M}, A_i, S_i) \prod_{j=1, j\neq i}^{n} Q_{\overline{M}j} \end{aligned}$$

Where

$$
\begin{aligned}
& Q_{\overline{M}j} \\
= \; & Pr(C_j|M_i = \overline{M}, A_i, S_i, f_A(i), f_S(i), T_i) \\
= \; & Pr(C_j|M_i = \overline{M}, A_j, S_j, f_A(j), f_S(j), T_i) \\
= \; & Pr(C_j, M_j = M|M_i = \overline{M}, A_j, S_j, f_A(j), f_S(j), T_i) \\
& + \; Pr(C_j, M_j = \overline{M}|M_i = \overline{M}, A_j, S_j, f_A(j), f_S(j), T_i) \\
= \; & Pr(C_j|M_j = M, M_i = \overline{M}, A_j, S_j, f_A(j), f_S(j), T_i) \\
& \quad * Pr(M_j = M|M_i = \overline{M}, A_j, S_j, f_A(j), f_S(j), T_i) \\
& + \; Pr(C_j|M_j = \overline{M}, M_i = \overline{M}, A_j, S_j, f_A(j), f_S(j), T_i) \\
& \quad * Pr(M_j = \overline{M}|M_i = \overline{M}, A_j, S_j, f_A(j), f_S(j), T_i) \\
= \; & Pr(C_j|M_j = M, A_j, S_j) * Pr(M_j = M|M_i = \overline{M}, T_i) \\
& + \; Pr(C_j|M_j = \overline{M}, A_j, S_j) * Pr(M_j = \overline{M}|M_i = \overline{M}, T_i), \qquad for j = 1, 2, \ldots, n, j \neq i
\end{aligned}
$$

For each member in the family, whether he/she develops with cancer by a certain age is independent of other members cancer status given his/her mutation status. Thus, by breaking into little pieces, equation (4.4) and (4.5) can be easily calculated, and $Pr(M_i = M|H_1)$ could be obtained then.

The probability of member i to carry mutation genes given family history H_1 or H_2 can be calculated in the similar way. The results by using the cumulative risks and age and sex specific incidence rates estimated by Dunlop (as shown in Figure (2.1) and (2.2), Chapter 2) are shown in table 4.1.

The population mutation carrier proportion estimate $Pr(M_i = M) \approx 0.00032$ (5, MG Dunlop et al., 2000, p.1643) is used. The conditional probability that member 1 is a mutation carrier given all the information of her families members is of interest. Mendelian principles of inheritance have to be applied in this case. Both H_1 and H_2 provide convincing evidence that member 3 and 4 are carriers. In H_1, the cancers observed

PID	$Pr(M_i = M\|H_1)$	$Pr(M_i = M\|H_2)$
1	0.179188	0.189890
2	0.432829	0.470590
3	0.958853	0.964468
4	0.864203	0.881131
5	0.000206	0.000206
6	0.961001	0.652312
7	0.002532	0.045534
8	0.002118	0.038353
9	0.981401	0.095214
10	0.087311	0.087311
11	0.001262	0.663080
12	0.003461	0.003461

Table 4.1: Conditional Probabilities of having a mutation in hMSH2/hMLH1 gene for members of families 1 and 2, $f = 0.00016$, using the cumulative risks and age-specific incidence rates estimated by Dunlop

in member 3 and 4 strongly support the hypothesis that their mother (member 6) carries a mutation, and member 9 provides a likely source for the mutation of genes in his daughter (member 6). The information in H_2 supports the hypothesis that at least one of the couples (member 6 or 7) carries mutation in hMSH2/hMLH1 genes. The father (member 7) is a likely source of a mutation in H_2 but not in H_1. The second possible source in H_2 increases the probability that member 3 and 4 carry mutation genes. The conditional probabilities for member 9 or 11 to be a mutation carrier in family history 1 is quite different from those in family history 2 due to their cancer status and age being switched.

4.2 Uncertainty Concerning Prevalence of Mutations

The calculation of LR involves f, and the calculations in table (4.1) are based on estimate of population mutation frequency $f = 0.00016$. The corresponding estimate of prevalence is about 0.00032 which has the 95% CI: [0.00013,0.0008] (5, MG Dunlop et al., 2000, p.1643). So the corresponding 95% CI for the population mutation frequency f would be: [0.000066,0.0004]. The unconditional probability of carrying a mutation is

PID	$Pr(M_i = M\|H_1)$		$Pr(M_i = M\|H_2)$	
	$f = 0.000066$	$f = 0.0004$	$f = 0.000066$	$f = 0.0004$
1	0.083329	0.347628	0.088926	0.363931
2	0.240889	0.651485	0.269871	0.685278
3	0.906286	0.982868	0.918463	0.985257
4	0.725372	0.939988	0.754693	0.948038
5	0.000085	0.000516	0.000085	0.000516
6	0.910884	0.983803	0.437178	0.822905
7	0.001050	0.006247	0.019405	0.105219
8	0.000880	0.005208	0.016299	0.089192
9	0.956362	0.992345	0.041877	0.205433
10	0.038214	0.190306	0.038214	0.190306
11	0.000524	0.003107	0.449356	0.829228
12	0.001438	0.008495	0.001438	0.008495

Table 4.2: Conditional Probabilities of having a mutation in hMSH2/hMLH1 gene for members of families 1 and 2, $f = 0.000066$ or $f = 0.0004$, use the cumulative risks and age-specific incidence rates estimated by Dunlop

$Pr(M_i = M) = 2f - f^2$, it increases approximately proportionally to f when f is small. Suppose the conditional probabilty for member i to be a mutation carrier given his/her family history $Pr(M_i = M|H)$ is close to 0 for a particular f. Then an increase in f will result in an increase in $Pr(M_i = M|H)$ by approximately the same proportion. But if $Pr(M_i = M|H)$ is large, then changing f has little impact on the conditional probability. Consider values of f at the upper and lower 95% confidence interval values, the effect on the conditional probabilty $Pr(M_i = M|H)$ is shown in table 4.2 (use cumulative risks and age and sex specific incidence rates estimated by Dunlop as shown in Figure (2.1) and (2.2), Chapter 2).

Chapter 5

Application to One Actual Family History

This chapter gives the application on a real family history which comes from the data supplied by Dr. Roger Green and Dr. Ban Younghusband. The real family history H_3 is shown in Figure (5.1).

Data came with four different cancer categories: (1) diagnosed with colorectal cancer; (2) HNPCC Tumour has been found; (3) Other Tumour has been found; (4) Other cancer has been diagnosed. Strictly speaking, only situation (1) should be considered. But for most of the families in the dataset, only a few people belong to situation (1) – diagnosed with colorectal cancer. Thus the conditional probability calculated would be almost the same as the general population carrier estimate $Pr(M_i = M) \approx 0.00032$ for most of the people in the dataset. Here all four were grouped together so that occurrence of any one of the four for person i is treated as $C_i = C$.

As can be seen from the figure, there are nine members in the family who have $C_i = C$. Four of them lack the information of age at diagnosis as well as their current age, thus they do not contribute in calculating the conditional probability that a particular member carries mutation genes given their family history.

The results for H_3 by using the cumulative risks and age and sex specific incidence

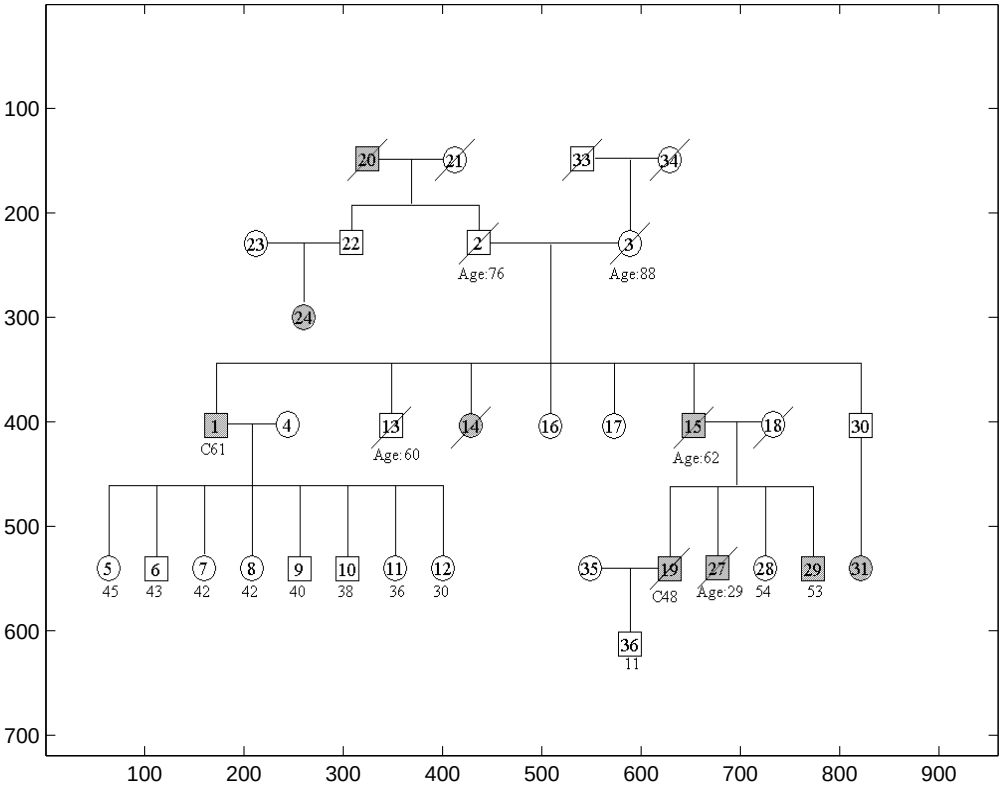

Figure 5.1: An Actual Family History 3 (H_3).

rates estimated by Dunlop (as shown in Figure (2.1) and (2.2), Chapter 2) are shown in Table (5.1).

As what appears in Table (5.1), member 19, 27, and 29 have larger probability of being a mutation carrier compared with the other members in the family since they were all diagnosed with colorectal cancer. Among them, member 29 has the largest probability of carrying mutation genes since he was diagnosed with colorectal cancer at a young age. Although member 1 was also diagnosed with colorectal cancer at almost the same age as member 29, the fact that all his children are free of cancer greatly decreases his probability of being a mutation carrier. Similarly, the fact that all the children of member 4 are free of cancer decreases the probability for member 4 to be a mutation carrier from 0.00032 at birth to 0.000119.

Member 36 is the only one in the family whose information will influence the conditional probability of member 35, and the fact that member 36 is free of colorectal by the age of 11 supplies very weak evidence that member 35 could be a mutation carrier. Thus the probability for member 35 to be a mutation carrier is approximately equal to the estimate of general population mutation frequency.

For member 28 and 36, the probability that they are mutation carriers respectively is of particular interest since the siblings of member 28 were all diagnosed as $C_i = C$, and there is strong evidence to support the hypothesis that member 36's father – member 19, is a mutation carrier. The probability calculated in the table is 0.1641 for member 28 and 0.0586 for member 36. Depending on available medical resources these probabilities may be large enough to warrant further investigation on them.

PID	$Pr(M_i = M \mid H_1)$	
	$f = 0.00016$	95%CI=[0.000066, 0.0004]
1	0.034446	[0.014571, 0.080977]
2	0.013794	[0.005765, 0.033383]
3	0.037099	[0.015719, 0.086867]
4	0.000119	[0.000049, 0.000298]
5	0.005119	[0.002127, 0.012548]
6	0.004119	[0.001771, 0.010111]
7	0.005088	[0.002115, 0.012473]
8	0.005088	[0.002115, 0.012473]
9	0.004411	[0.001833, 0.010824]
10	0.004554	[0.001893, 0.011174]
11	0.005018	[0.002086, 0.012303]
12	0.004950	[0.002058, 0.012140]
13	0.009679	[0.004036, 0.023559]
14	0.020293	[0.008514, 0.048645]
15	0.189726	[0.088365, 0.367133]
16	0.020293	[0.008514, 0.048645]
17	0.020293	[0.008514, 0.048645]
18	0.093338	[0.040866, 0.203307]
19	0.742106	[0.543319, 0.877332]
20	0.006770	[0.002817, 0.016555]
21	0.006770	[0.002817, 0.016555]
22	0.006769	[0.002817, 0.016552]
23	0.000320	[0.000132, 0.000800]
24	0.002338	[0.000969, 0.005768]
27	0.895336	[0.779634, 0.955045]
28	0.164111	[0.075182, 0.327108]
29	0.618593	[0.401541, 0.800992]
30	0.020293	[0.008514, 0.048645]
31	0.007157	[0.002979, 0.017490]
33	0.006746	[0.002807, 0.016499]
34	0.006746	[0.002807, 0.016499]
35	0.000319	[0.000132, 0.000799]
36	0.058609	[0.025145, 0.133391]

Table 5.1: Conditional Probabilities of having a mutation in hMSH2/hMLH1 gene for members of families 3, with the allelic frequency estimates to be $f = 0.00016$ and its 95% CI lower and upper bound respectively ($f = 0.000066$ or $f = 0.0004$). The cumulative risks and age-specific incidence rates estimated by Dunlop are used.

Chapter 6

Impact of Genetic Testing

Here we consider the impact of genetic testing following that of Berry (2, D.A. Berry et al., 1997, p.234). A person who has a family history of colorectal cancer may consider genetic testing. Those who have been diagnosed with colorectal cancer and who have a family history of colorectal cancer may be concerned about the possibility of developing a second or even more cancers. Many unaffected relatives of these people who are at increased risk of developing cancer will also worry about developing colorectal cancer. But the potential risks such as psychological distress, restriction of life, and disability insurance and so on, which are associated with testing, make a decision of whether to proceed with testing difficult.

The estimate of the conditional probability for a person to carry a mutation based on their family history will help them to make a more informed decision of whether or not to proceed with genetic testing. There are two possible results for the test: $+/-$. A positive test result predicts that the tested individual is a mutation carrier and a negative test result predicts that the tested individual is not a mutation carrier. But the prediction might not always be correct.

The false-positive rate is the probability that the tested individual gets a positive test result but he/she is not a mutation carrier. Then: false-positive rate=$Pr(+|M_i = \overline{M})$. The false-negative rate is the probability that the tested individual gets a negative test

result while he/she is actually a mutation carrier, or false-negative rate=$Pr(-|M_i = M)$.

An important consideration in deciding whether to be tested is the test's specificity (the probability that the tested individual gets a negative test result given he/she is not a mutation carrier) and sensitivity (the probability that the tested individual gets a positive test result given he/she is a mutation carrier). We know that:

specificity= $Pr(-|M_i = \overline{M}) = 1-$false-positive rate

sensitivity= $Pr(+|M_i = M) = 1-$false-negative rate

For simplicity, let a=specificity and b=sensitivity. $Pr(M_i = M|H, +)$ is the probability for individual i to be a mutation carrier given their family history and he gets a positive test result; $Pr(M_i = M|H, -)$ is the probability for individual i to be a mutation carrier given their family history and he gets a negative test result. $Pr(M_i = M|H)$ now plays as a prior probability and the test result $(+/-)$ is the evidence. By applying Bayes' theorem, the carrier probability of a positive test result would be:

$$\begin{aligned} & Pr(M_i = M|H, +) \\ = & \frac{Pr(+|M_i = M) * Pr(M_i = M|H)}{Pr(+|M_i = M) * Pr(M_i = M|H) + Pr(+|M_i = \overline{M}) * Pr(M_i = \overline{M}|H)} \\ = & \frac{b * Pr(M_i = M|H)}{b * Pr(M_i = M|H) + (1-a) * Pr(M_i = \overline{M}|H)} \end{aligned} \tag{6.1}$$

And the carrier probability of a negative test result would be:

$$\begin{aligned} & Pr(M_i = M|H, -) \\ = & \frac{Pr(-|M_i = M) * Pr(M_i = M|H)}{Pr(-|M_i = M) * Pr(M_i = M|H) + Pr(-|M_i = \overline{M}) * Pr(M_i = \overline{M}|H)} \\ = & \frac{(1-b) * Pr(M_i = M|H)}{(1-b) * Pr(M_i = M|H) + a * Pr(M_i = \overline{M}|H)} \end{aligned} \tag{6.2}$$

As an example, consider member 1 of family 1(as shown in Figure (1.1)) and suppose a=95% and b=85%. Her carrier probability based on her family history is: $Pr(M_1 = M|H_1) = 0.179188$ (assuming population mutation frequency to be 0.00016, as shown in

Table (4.1)). If she is tested, the probability of getting a positive test result is:

$$\begin{aligned}
& Pr(+|H_1) \\
= \; & Pr(+|M_1 = M) * Pr(M_1 = M|H_1) + Pr(+|M_1 = \overline{M}) * Pr(M_1 = \overline{M}|H_1) \\
= \; & b * Pr(M_1 = M|H_1) + (1-a) * Pr(M_1 = \overline{M}|H_1) \\
= \; & 0.85 * 0.179188 + 0.05 * (1 - 0.179188) \\
\approx \; & 0.161
\end{aligned}$$

And the probability of a negative test is: $Pr(-|H_1) = 1 - Pr(+|H_1) = 0.839$. If her test result was positive, the her carrier probability given her family history and test result would be:

$$\begin{aligned}
Pr(M_1 = M|H_1, +) &= \frac{Pr(+|M_1 = M) * Pr(M_i = M|H)}{Pr(+|H_1)} \\
&= \frac{b * Pr(M_i = M|H)}{Pr(+|H_1)} \\
&= \frac{0.85 * 0.179188}{0.161} \\
&\approx 0.946
\end{aligned}$$

If she were to test negative, then her carrier probability given her family history and test result would be:

$$\begin{aligned}
Pr(M_1 = M|H_1, -) &= \frac{Pr(-|M_1 = M) * Pr(M_i = M|H)}{Pr(-|H_1)} \\
&= \frac{(1-b) * Pr(M_i = M|H)}{Pr(-|H_1)} \\
&= \frac{0.15 * 0.179188}{0.839} \\
&\approx 0.032
\end{aligned}$$

For this woman, if she gets a positive test result, the probability for her to carry mutation genes is much larger than that given a negative test result. In this situation, the genetic

testing is worthwhile.

As another example, consider member 2 of family history 2 (as shown in Figure (4.1)) and still suppose a=95% and b=85%. The carrier probability of her is $Pr(M_2 = M|H_2) = 0.470590$ (assuming allelic mutation frequency $f = 0.00016$, as shown in Table (4.1)). So the probability of a positive test is:

$$Pr(+|H_2) = b * Pr(M_2 = M|H_2) + (1 - a) * Pr(M_2 = \overline{M}|H_2) = 0.426472$$

and the probability of a negative test is: $Pr(-|H_2) = 1 - Pr(+|H_2) = 0.573528$. If she get a positive test result:

$$Pr(M_2 = M|H_2, +) = \frac{b * Pr(M_2 = M|H_2)}{Pr(+|H_2)} \approx 0.938.$$

If she get a negative test result:

$$Pr(M_2 = M|H_2, -) = \frac{(1 - b) * Pr(M_2 = M|H_2)}{Pr(-|H_2)} \approx 0.123.$$

Should this 40-year-old, disease-free woman take a genetic testing? There are a variety of considerations. Any benefit of testing depends on the woman's probability of being a mutation carrier based on family history and on the sensitivity and specificity of the testing, as well as on the effectiveness of available prophylactic interventions. The woman might choose them if the available prophylactic interventions are deemed as highly effective even if her probability of being a mutation carrier is small. But if the test result of genetic testing would lead her to two different interventions, then the testing might has some value. By using this model, physicians and counselors may be able to help women determine what probability of identifying a mutation would lead them to testing.

Chapter 7

Limitation and Further Directions

7.1 Estimates of population carrier frequency of hMSH2 and hMLH1 mutations

Estimates of population carrier frequency of hMSH2 and hMLH1 mutations derived empirically from HNPCC families induce ascertainment bias. Such families are relatively small in number. Not all gene carriers have a family history of the disease, and families fulfilling the Amsterdam criteria are relatively uncommon (8, SM Farrington et al., 1998, p.749). Hence, population prevalence of mismatch repair gene mutations cannot be calculated from studies employing family history ascertainment (6, MG Dunlop et al., 1997, p.105). MG Dunlop had provided an estimate of carrier frequency of mutations in hMSH1 and hMLH2 based on systematically collected data that is not subject to bias due to family history ascertainment. Analysis are restricted to people aged 15-74 years, who are all surviving relatives of Scottish probands with early-onset colorectal cancer, and these probands with documented mutations were ascertained on the basis of being affected by colorectal cancer when aged less than 30 years at diagnosis, irrespective of family history. Thus ascertainment bias was minimised since these people are not from pre-selected HNPCC families.

7.2 Population based cancer risk estimation

Similarly, cancer risk estimates derived empirically from pre-selected known HNPCC kindreds inevitably induce ascertainment bias. Again MG Dunlop targeted genetic analysis to patients with both early-onset colorectal cancer and an RER tumour as a means of identifying kindreds with germline MMR gene mutations. The lifetime cancer risk assessed in the cohort of relatives of probands, in whom ascertainment bias has been minimised, provide a rational population-wide basis. Although the least biased of all possible samples would involve a population screening approach regardless of disease state, but at present this is not practical. It might not be rational to use the estimates obtained from Scottish population on the cases from Canada since mutation frequency may vary with ethnicity.

The probability that a person carries mutation genes at hMSH2/hMLH1 given their family history is calculated. This person may be affected or unaffected of colorectal cancer. This model can be used to calculate the cumulative probability of colorectal cancer before a given age by averaging the cumulative incidence probabilities for carriers and noncarriers. For example, member 1 of family history 1 is 24 years old. The probability that she carries a mutation at hMSH2/hMLH1 is 0.179188. The probability that she develops colorectal cancer by age 45 is: $B_M(45) - B_M(24) \approx 0.337$ if she is a mutation gene carrier, and it is: $B_{\overline{M}}(45) - B_{\overline{M}}(24) \approx 0.016$ if she is a noncarrier. Thus the unconditional probability that she will develop colorectal gene by age 45 is: $0.337 * 0.179188 + 0.016 * (1 - 0.179188) \approx 0.074$. In addition, this model considers the number, relationship, and ages of unaffected individuals. Having many unaffected family members can substantially lower a carrier probability.

7.3 Other mutation genes might be influential to HNPCC

This model focuses on mutations in hMSH2 or hMLH1 and assumes that all other colorectal cancer is sporadic. The estimate of carrier frequency used in this essay is based

on hMSH2 and hMLH1 mutations. Similarly the age and sex specific incidence rates by Dunlop (6, MG Dunlop et al., 1997, p.107) is based on these same two mutation genes. However, sometimes there are other families with a great deal of cancer history for which there is no detectable presence of the known mutations hMSH2/hMLH1. Some of these new influential genes have been identified. Recent data indicate that mutations of hMSH6 account for an appreciable proportion of HNPCC-like families (7, J Wijnen et al., 1999, p.143). Usually it is expensive to find new mutations. And it is hoped that the same probability model will work for mutations not yet determined (depending on the correct value of f and the applicability of the incidence curves).

7.4 Other cancer related to hMSH2/hMLH1 can be considered

The occurance of various other early-onset cancers are strong evidence of the presence of mutation in hMSH2 or hMLH1 genes. Studies indicate that carrying a mutation in hMSH2/hMLH1 genes will increase the risk of endometrial cancer, small bowel cancer, urinary tract cancer, stomach cancer and ovaries cancer (4, H.F.A Vasen et al., 1996, p.1020). To incorporate family history of these diseases might improve the estimate of a particular member in the family carrying a mutation gene in hMSH2/hMLH1.

Appendix

```
Private Function FindCommonAncestor(iPid1, iPid2) As CommonAncestors
    Dim cAncestor1 As Ancestors
    Dim cAncestor2 As Ancestors
    Dim inFlag, i, j, jindex As Integer
    With FindCommonAncestor
        .Pid1 = iPid1
        .Pid2 = iPid2
        .iNum = 0
        .PidList = Array()
        .GenerList1 = Array()
        .GenerList2 = Array()
        .minr1 = FamilySize
        .minr2 = FamilySize
        cAncestor1 = FindAncestor(iPid1, 0)
        cAncestor2 = FindAncestor(iPid2, 0)
        For i = LBound(cAncestor1.PidList) To UBound(cAncestor1.PidList)
            inFlag = 0
            For j = LBound(cAncestor2.PidList) To UBound(cAncestor2.PidList)
                If cAncestor1.PidList(i) = cAncestor2.PidList(j) Then
                    inFlag = 1
                    jindex = j
                End If
            Next j
            If inFlag = 1 Then
                .iNum = .iNum + 1
                .PidList = Concatenate(.PidList, Array(cAncestor1.PidList(i)))
                .GenerList1 = Concatenate(.GenerList1,
```

```
Array(cAncestor1.rGeneration(i)))
                .GenerList2 = Concatenate(.GenerList2,
Array(cAncestor2.rGeneration(jindex)))
                If cAncestor1.rGeneration(i) < .minr1 Then
                    .minr1 = cAncestor1.rGeneration(i)
                End If
                If cAncestor2.rGeneration(jindex) < .minr2 Then
                    .minr2 = cAncestor2.rGeneration(jindex)
                End If
            End If
        Next i
    End With
End Function

Private Function FindAncestor(iPid, Rg) As Ancestors
    Dim cAnces As Ancestors
    Dim fAnces As Ancestors
    Dim mAnces As Ancestors
    Dim i As Integer
    Dim Fid, Mid As String
    cAnces.iNum = 0
    cAnces.PidList = Array()
    cAnces.rGeneration = Array()
    FindAncestor = cAnces
    fAnces = cAnces
    mAnces = cAnces
    For i = 1 To FamilySize
        If ListView1.ListItems.Item(i).SubItems(1) = iPid Then
```

```
            Fid = ListView1.ListItems.Item(i).SubItems(2)
            If Fid = 0 Then
                fAnces.iNum = 0
            Else
                fAnces = FindAncestor(Fid, Rg + 1)
            End If
            Mid = ListView1.ListItems.Item(i).SubItems(3)
            If Mid = 0 Then
                mAnces.iNum = 0
            Else
                mAnces = FindAncestor(Mid, Rg + 1)
            End If
            FindAncestor = MergeAnces(fAnces, mAnces)
            If Fid <> 0 Then
                FindAncestor.iNum = FindAncestor.iNum + 1
                FindAncestor.PidList = Concatenate(Array(Fid),
FindAncestor.PidList)
                FindAncestor.rGeneration = Concatenate(Array(Rg + 1),
FindAncestor.rGeneration)
            End If
            If Mid <> 0 Then
                FindAncestor.iNum = FindAncestor.iNum + 1
                FindAncestor.PidList = Concatenate(Array(Mid),
FindAncestor.PidList)
                FindAncestor.rGeneration = Concatenate(Array(Rg + 1),
FindAncestor.rGeneration)
            End If
        End If
```

```
    Next i
End Function

Private Function IsFather(Opid, fpid) As Boolean
    Dim i As Integer
    Dim iPid, Fid As String
    IsFather = False
    For i = 1 To FamilySize
        iPid = ListView1.ListItems.Item(i).SubItems(1)
        Fid = ListView1.ListItems.Item(i).SubItems(2)
        If iPid = Opid Then
            If Fid = fpid Then
                IsFather = True
            End If
        End If
    Next i
End Function

Private Function IsMother(Opid, mpid) As Boolean
    Dim i As Integer
    Dim iPid, Mid As String
    IsMother = False
    For i = 1 To FamilySize
        iPid = ListView1.ListItems.Item(i).SubItems(1)
        Mid = ListView1.ListItems.Item(i).SubItems(3)
        If iPid = Opid Then
            If Mid = mpid Then
                IsMother = True
```

```
            End If
        End If
    Next i
End Function

Private Function IsKid(Opid, PPid) As Boolean
    Dim i As Integer
    Dim Fid, Mid, iPid As String
    IsKid = False
    For i = 1 To FamilySize
        iPid = ListView1.ListItems.Item(i).SubItems(1)
        Fid = ListView1.ListItems.Item(i).SubItems(2)
        Mid = ListView1.ListItems.Item(i).SubItems(3)
        If iPid = Opid Then
            If Fid = PPid Or Mid = PPid Then
                IsKid = True
            End If
        End If
    Next i
End Function

Private Function FindKids(PPid) As Ancestors
    Dim i As Integer
    Dim Fid, Mid, iPid As String
    With FindKids
        .iNum = 0
        .PidList = Array()
        .rGeneration = Array()
```

```
        For i = 1 To FamilySize
            iPid = ListView1.ListItems.Item(i).SubItems(1)
            Fid = ListView1.ListItems.Item(i).SubItems(2)
            Mid = ListView1.ListItems.Item(i).SubItems(3)
            If Fid = PPid Or Mid = PPid Then
                .iNum = .iNum + 1
                .PidList = Concatenate(.PidList, Array(iPid))
                .rGeneration = Concatenate(.rGeneration, Array(-1))
            End If
        Next i
    End With
End Function

Private Function FindParents(Opid) As Ancestors
    Dim i As Integer
    Dim iPid, Fid, Mid As String
    With FindParents
        .iNum = 2
        For i = 1 To FamilySize
            iPid = ListView1.ListItems.Item(i).SubItems(1)
            Fid = ListView1.ListItems.Item(i).SubItems(2)
            Mid = ListView1.ListItems.Item(i).SubItems(3)
            If iPid = Opid Then
                .PidList = Array(Fid, Mid)
                .rGeneration = Array(1, 1)
            End If
        Next i
    End With
```

```
End Function

Private Function IsAncestor(offspringPid, AncList As Variant) As Boolean
    Dim i, j As Integer
    Dim cAnces As Ancestors
    Dim inFlag As Boolean
    cAnces = FindAncestor(offspringPid, 0)
    IsAncestor = True
    For i = LBound(AncList) To UBound(AncList)
        inFlag = False
        For j = LBound(cAnces.PidList) To UBound(cAnces.PidList)
            If AncList(i) = cAnces.PidList(j) Then
                inFlag = True
            End If
        Next j
        If inFlag = False Then
            IsAncestor = False
        End If
    Next i
End Function

Private Function IsOffspring(AncesPid, Offsprings As Variant) As Boolean
    Dim i As Integer
    IsOffspring = True
    For i = LBound(Offsprings) To UBound(Offsprings)
        If IsAncestor(Offsprings(i), Array(AncesPid)) = False Then
            IsOffspring = False
        End If
```

```
    Next i
End Function

Private Function aGb(apid, aMt As String, bPid, bMt As String) As Double
    Dim f, prob, re1, re2 As Double
    Dim cAnces As Ancestors
    Dim kAnces As Ancestors
    Dim i, j, k, n As Integer
    Dim kpid As String
    Dim fAnces As CommonAncestors
    Dim lAnces As CommonAncestors
If apid <> bPid Then
    If IsKid(apid, bPid) Then       ' apid is the kid of bpid
        aGb = KidG1Pa(aMt, bMt)
    ElseIf IsKid(bPid, apid) Then   ' apid is one of the parents of bpid
        aGb = Pa1GKid(aMt, bMt)
    Else                            ' they are not filiation
        prob = 0
        kAnces.iNum = 0
        kAnces.PidList = Array()
        kAnces.rGeneration = Array()
        If IsAncestor(apid, Array(bPid)) Then   ' bpid is the ancestor of apid
            cAnces = FindParents(apid)
            For i = LBound(cAnces.PidList) To UBound(cAnces.PidList)
            ' find parents of apid who are offsprings of bpid
                kpid = cAnces.PidList(i)
                If IsOffspring(bPid, Array(kpid)) Then
                    kAnces.iNum = kAnces.iNum + 1
```

```
                    kAnces.PidList = Concatenate(kAnces.PidList, Array(kpid))
                    kAnces.rGeneration =
Concatenate(kAnces.rGeneration, Array(cAnces.rGeneration(i)))
                End If
            Next i
            n = kAnces.iNum
            Select Case n
            Case 1
                kpid = kAnces.PidList(0)
                prob = aGb(apid, aMt, kpid, "M") * aGb(kpid, "M", bPid, bMt)
                prob = prob + aGb(apid, aMt, kpid, "N")
* aGb(kpid, "N", bPid, bMt)
            Case 2
                For j = 0 To ((2 ^ n) - 1)
                    prob = prob +
P1G2(apid, aMt, kAnces.PidList, j)*P2G1(kAnces.PidList, j, bPid, bMt)
                Next j
            End Select
        ElseIf IsOffspring(apid, Array(bPid)) Then
' apid is the ancestor of bpid
            cAnces = FindParents(bPid)
            For i = LBound(cAnces.PidList) To UBound(cAnces.PidList)
                kpid = cAnces.PidList(i)
                If IsOffspring(apid, Array(kpid)) Then
                    kAnces.iNum = kAnces.iNum + 1
                    kAnces.PidList = Concatenate(kAnces.PidList, Array(kpid))
                    kAnces.rGeneration = Concatenate(kAnces.rGeneration,
Array(cAnces.rGeneration(i)))
```

```
                End If
            Next i
            n = kAnces.iNum
            Select Case n
            Case 1
                kpid = kAnces.PidList(0)
                prob = aGb(apid, aMt, kpid, "M") * aGb(kpid, "M", bPid, bMt)
                prob = prob + aGb(apid, aMt, kpid, "N")
* aGb(kpid, "N", bPid, bMt)
            Case 2
                For j = 0 To ((2 ^ n) - 1)
                    prob = prob +
P1G2(apid, aMt, kAnces.PidList, j) * P2G1(kAnces.PidList, j, bPid, bMt)
                Next j
            End Select
        Else
            fAnces = FindCommonAncestor(apid, bPid)
            lAnces = FindLatestAncestor(fAnces)
            cAnces = AncesTypeChange(lAnces)
            Select Case cAnces.iNum
            Case 0          'no relationship between apid & bpid
                prob = PrMN(aMt)
            Case 1
                kpid = cAnces.PidList(0)
                prob = aGb(apid, aMt, kpid, "M") * aGb(kpid, "M", bPid, bMt)
                prob = prob +
aGb(apid, aMt, kpid, "N") * aGb(kpid, "N", bPid, bMt)
            Case 2
```

```
                n = cAnces.iNum
                For j = 0 To ((2 ^ n) - 1)
                    re1 = P1G2(apid, aMt, cAnces.PidList, j)
                    re2 = P2G1(cAnces.PidList, j, bPid, bMt)
                    prob = prob + re1 * re2
                Next j
            Case Else
                MsgBox "Error in aGb!", vbExclamation, "Error"
            End Select
        End If
        aGb = prob
    End If
Else
    If aMt = bMt Then
        aGb = 1
    Else
        aGb = 0
    End If
End If
End Function

Private Function P1G2(ByVal apid, aMt As String, ByVal PidList As Variant,
ByVal idx As Integer) As Double
    Dim fpid, mpid, kpid0, kpid1 As String
    Dim bf, bm, ba0, ba1, bo0, bo1 As Boolean
    Dim cAnces As Ancestors
    Dim kAnces As Ancestors
    Dim re As Double
```

```
    fpid = PidList(0)
    mpid = PidList(1)
    bf = IsKid(apid, fpid)
    bm = IsKid(apid, mpid)
    Select Case Abs(CInt(bf + bm))
    Case 0
        If IsAncestor(apid, PidList) Then
            cAnces = FindParents(apid)
            kpid0 = cAnces.PidList(0)
            kpid1 = cAnces.PidList(1)
            bo0 = IsAncestor(kpid0, PidList)
            bo1 = IsAncestor(kpid1, PidList)
            Select Case Abs(CInt(bo0 + bo1))
            Case 0      'impossible
                MsgBox "Error in P1G2!", vbExclamation, "Error"
                P1G2 = 0
            Case 1
                If bo0 Then
                    re = aGb(apid, aMt, kpid0, "M")
* P1G2(kpid0, "M", PidList, idx)
                    re = re + aGb(apid, aMt, kpid0, "N")
* P1G2(kpid0, "N", PidList, idx)
                Else
                    re = aGb(apid, aMt, kpid1, "M")
* P1G2(kpid1, "M", PidList, idx)
                    re = re + aGb(apid, aMt, kpid1, "N")
* P1G2(kpid1, "N", PidList, idx)
                End If
```

```
                P1G2 = re
            Case 2
                re = P1G2(apid, aMt, cAnces.PidList, 0)
* P1G2(kpid0, "N", cAnces.PidList, idx) * P1G2(kpid1, "N", cAnces.PidList, idx)
                re = re + P1G2(apid, aMt, cAnces.PidList, 1)
* P1G2(kpid0, "N", cAnces.PidList, idx) * P1G2(kpid1, "M", cAnces.PidList, idx)
                re = re + P1G2(apid, aMt, cAnces.PidList, 2)
* P1G2(kpid0, "M", cAnces.PidList, idx) * P1G2(kpid1, "N", cAnces.PidList, idx)
                re = re + P1G2(apid, aMt, cAnces.PidList, 3)
* P1G2(kpid0, "M", cAnces.PidList, idx) * P1G2(kpid1, "M", cAnces.PidList, idx)
                P1G2 = re
            End Select
        Else
            MsgBox "Error in P1G2!", vbExclamation, "Error"
            P1G2 = 0
        End If
    Case 1
        P1G2 = 0
    Case 2
        P1G2 = KidG2Pa(aMt, idx)
    End Select
End Function

Private Function P2G1(ByVal PidList As Variant, ByVal idx As Integer,
ByVal bPid, bMt As String) As Double
    Select Case idx
    Case 0
        P2G1 = P1G2(bPid, bMt, PidList, idx)*PrMN("N")*PrMN("N")/PrMN(bMt)
```

```
    Case 1
        P2G1 = P1G2(bPid, bMt, PidList, idx)*PrMN("N")*PrMN("M")/PrMN(bMt)
    Case 2
        P2G1 = P1G2(bPid, bMt, PidList, idx)*PrMN("M")*PrMN("N")/PrMN(bMt)
    Case 3
        P2G1 = P1G2(bPid, bMt, PidList, idx)*PrMN("M")*PrMN("M")/PrMN(bMt)
    End Select
End Function

Private Sub FindDistinctFMember(iFam As String)
    Dim strSQL As String
    Dim i As Integer
    strSQL = "SELECT DISTINCT PID FROM cancer WHERE FAM=" & iFam
    With Adodc1
        .ConnectionString = strAccessConnect
        .RecordSource = strSQL
        .Refresh
        Set AdodcRs = .Recordset
    End With
    If AdodcRs.RecordCount > 0 Then
        DistinctFamSize = AdodcRs.RecordCount
        AdodcRs.MoveFirst
        For i = 0 To (AdodcRs.RecordCount - 1)
            FMembers(i) = AdodcRs("PID")
            AdodcRs.MoveNext
        Next i
'        MsgBox "There are " & DistinctFamSize & " members in the family."
    Else
```

```
        DistinctFamSize = 0
        MsgBox "Find record error, FAM=" & iFam
    End If
End Sub

Private Function ProbMgivenH(iPid As String) As Double
    Dim priorM, priorN, cancerM, cancerN As Double
    priorM = PrMN("M")
    priorN = PrMN("N")
    cancerM = FCancer(iPid, "M")
    cancerN = FCancer(iPid, "N")
    ProbMgivenH = cancerM * priorM / (cancerM * priorM + cancerN * priorN)
End Function

Private Function FCancer(iPid As String, iMutStatus As String) As Double
    Dim jPid As String
    Dim i As Integer
    FCancer = 1
    For i = 0 To (DistinctFamSize - 1)
        jPid = FMembers(i)
        FCancer = FCancer * JCancer(jPid, iPid, iMutStatus, 0)
    Next i
End Function

Private Function JCancer(jPid As String, iPid As String, iMutStatus As String,
msgFlag As Integer) As Double
    Dim result, prm, prn, pagbm, pagbn As Double
    Dim str As String
```

```
Call FindInfo(jPid)
If (PersonJ.age = 0) Then
    If PersonJ.DOB <> "" Then
        If PersonJ.DOD <> "" Then
            PersonJ.age = Left(PersonJ.DOD, 4) - Left(PersonJ.DOB, 4)
        Else
            PersonJ.age = Year(Date) - Left(PersonJ.DOB, 4)
        End If
    End If
End If
If jPid = iPid Then
    If iMutStatus = "M" Then
        result = Carrier(PersonJ.STAT, PersonJ.AgeDx, PersonJ.age,PersonJ.Sex)
    ElseIf iMutStatus = "N" Then
        result = NonCarrier(PersonJ.STAT, PersonJ.AgeDx, PersonJ.age, PersonJ.Sex)
    Else
        result = 0
    End If
Else
    prm = Carrier(PersonJ.STAT, PersonJ.AgeDx, PersonJ.age, PersonJ.Sex)
    prn = NonCarrier(PersonJ.STAT, PersonJ.AgeDx, PersonJ.age, PersonJ.Sex)
    pagbm = aGb(jPid, "M", iPid, iMutStatus)
    pagbn = aGb(jPid, "N", iPid, iMutStatus)
    result = prm * pagbm + prn * pagbn
End If
JCancer = result
If msgFlag <> 0 Then
```

```
        str = "Member " & jPid
        Select Case PersonJ.Sex
        Case 0
            str = str & ", whose gender is unknown,"
        Case 1
            str = str & ",female,"
        Case 2
            str = str & ",male,"
        End Select
        If PersonJ.STAT = 1 Then
            str = str & " diagnosed of cancer at the age of "
            If PersonJ.AgeDx > 0 Then
                str = str & PersonJ.AgeDx
            Else
                str = str & PersonJ.age
            End If
        Else
            str = str & " is cancer free by the age of " & PersonJ.age
        End If
        str = str & " given member " & iPid
        If iMutStatus = "M" Then
            str = str & " has mutation is: " & JCancer
        Else
            str = str & " does not have mutation is: " & JCancer
        End If
        MsgBox str
    End If
End Function
```

```
Private Function FX(ByVal b0 As Double, ByVal b1 As Double, ByVal b2 As Double, ByVal a0
    Dim tempa As Double

    tempa = Exp(-a1 * (age - a2))
    FX = 1 - (1 - GX(b0, b1, b2, age)) * (1 - a0 / (1 + tempa))
End Function

Private Function GX(ByVal b0 As Double, ByVal b1 As Double, ByVal b2 As Double, ByVal ag
    GX = b0 / (1 + Exp(-b1 * (age - b2)))
End Function

Private Function FIncidence(ByVal b0 As Double, ByVal b1 As Double, ByVal b2 As Double, I
    Dim tempa, tempb, parta, partb As Double

    tempb = Exp(-b1 * (age - b2))
    tempa = Exp(-a1 * (age - a2))
    parta = b0 * b1 * tempb / ((1 + tempb) ^ 2) * (1 - a0 / (1 + tempa))
    partb = (1 - b0 / (1 + tempb)) * a0 * a1 * tempa / ((1 + tempa) ^ 2)
    FIncidence = parta + partb
End Function

Private Function GIncidence(ByVal b0 As Double, ByVal b1 As Double, ByVal b2 As Double, I
    Dim temp As Double

    temp = Exp(-b1 * (AgeDx - b2))
    GIncidence = b0 * b1 * temp / ((1 + temp) ^ 2)
End Function
```

```
Private Function Carrier(CStat As Integer, AgeDx As Integer, age As Integer,
Sex As Integer) As Double
    Dim b0, b1, b2, a0, a1, a2 As Double
    Select Case Sex
    Case 1      'female
        b0 = b0Female
        b1 = b1Female
        b2 = b2Female
        a0 = a0Female
        a1 = a1Female
        a2 = a2Female
    Case 2      'male
        b0 = b0Male
        b1 = b1Male
        b2 = b2Male
        a0 = a0Male
        a1 = a1Male
        a2 = a2Male
    End Select
    If (age = 0 And AgeDx = 0) Or (Sex = 0) Then
        Carrier = 1
    Else
        Select Case CStat
        Case 0      ' free of cancer
            Carrier = 1 - FX(b0, b1, b2, a0, a1, a2, age)
        Case 1      'diagnosed of cancer
            If AgeDx = 0 Then
```

```
                AgeDx = age
            End If
            Carrier = FIncidence(b0, b1, b2, a0, a1, a2, AgeDx)
        End Select
    End If
End Function

Private Function NonCarrier(CStat As Integer, AgeDx As Integer,
age As Integer, Sex As Integer) As Double
    Dim temp As Double
    Dim b0, b1, b2, a0, a1, a2 As Double
    Select Case Sex
    Case 1      'female
        b0 = b0Female
        b1 = b1Female
        b2 = b2Female
        a0 = a0Female
        a1 = a1Female
        a2 = a2Female
    Case 2      'male
        b0 = b0Male
        b1 = b1Male
        b2 = b2Male
        a0 = a0Male
        a1 = a1Male
        a2 = a2Male
    End Select
    If (age = 0 And AgeDx = 0) Or (Sex = 0) Then
```

```
        NonCarrier = 1
    Else
        Select Case CStat
        Case 0      ' free of cancer
            NonCarrier = 1 - GX(b0, b1, b2, age)
        Case 1      ' diagnosed of cancer
            If AgeDx = 0 Then
                AgeDx = age
            End If
            NonCarrier = GIncidence(b0, b1, b2, AgeDx)
        End Select
    End If
End Function

Private Function FindInfo(jPid As String)
    Dim i As Integer
    With PersonJ
        .STAT = 0
        .AgeDx = 0
        .Fam = txtFAM.Text
        .Pid = jPid
        For i = 1 To FamilySize
            If ListView1.ListItems.Item(i).SubItems(1) = jPid Then
                .Fid = ListView1.ListItems.Item(i).SubItems(2)
                .Mid = ListView1.ListItems.Item(i).SubItems(3)
                .Sex = ListView1.ListItems.Item(i).SubItems(4)
                Select Case ListView1.ListItems.Item(i).SubItems(5)
                Case "1", "2", "3", "4"
```

```
                    .STAT = 1       'cancer diagnosed
                    If ListView1.ListItems.Item(i).SubItems(6) = "" Then
                        .AgeDx = 0
                    Else
                        .AgeDx = ListView1.ListItems.Item(i).SubItems(6)
                    End If
                Case Else           'blank
                End Select
                .DOB = ListView1.ListItems.Item(i).SubItems(7)
                .age = ListView1.ListItems.Item(i).SubItems(8)
                .DOD = ListView1.ListItems.Item(i).SubItems(9)
                .DEAD = ListView1.ListItems.Item(i).SubItems(10)
            End If
        Next i
    End With
End Function

Public Function PrMN(Mt) As Double
    Select Case Mt
    Case "M"
        PrMN = 2 * fMutation - fMutation * fMutation
    Case "N"
        PrMN = (1 - fMutation) ^ 2
    End Select
End Function

Public Function Pa1GKid(MtPa As String, MtKid As String) As Double
    Select Case MtPa
```

```
    Case "M"
        Select Case MtKid
        Case "M"
            Pa1GKid = (1+fMutation-fMutation*fMutation)/(2-fMutation)
        Case "N"
            Pa1GKid = fMutation
        End Select
    Case "N"
        Select Case MtKid
        Case "M"
            Pa1GKid = (1 - fMutation) ^ 2 / (2 - fMutation)
        Case "N"
            Pa1GKid = 1 - fMutation
        End Select
    End Select
End Function

Public Function Pa2Kid(idx As Integer, MtKid As String) As Double
    Select Case MtKid
    Case "M"
        Select Case idx
        Case 0
            Pa2Kid = 0
        Case 1
            Pa2Kid = (1 - fMutation) ^ 2 / (2 - fMutation)
        Case 2
            Pa2Kid = (1 - fMutation) ^ 2 / (2 - fMutation)
        Case 3
```

```
            Pa2Kid = (3 * fMutation - 2 * fMutation ^ 2) / (2 - fMutation)
        End Select
    Case "N"
        Select Case idx
        Case 0
            Pa2Kid = (1 - fMutation) ^ 2
        Case 1
            Pa2Kid = fMutation(1 - fMutation)
        Case 2
            Pa2Kid = fMutation(1 - fMutation)
        Case 3
            Pa2Kid = fMutation ^ 2
        End Select
    End Select
End Function

Public Function KidG1Pa(MtKid As String, MtPa As String) As Double
    Select Case MtKid
    Case "M"
        Select Case MtPa
        Case "M"
            KidG1Pa = (1 + fMutation - fMutation ^ 2) / (2 - fMutation)
        Case "N"
            KidG1Pa = fMutation
        End Select
    Case "N"
        Select Case MtPa
        Case "M"
```

```
            KidG1Pa = 1 - (1 + fMutation - fMutation ^ 2) / (2 - fMutation)
        Case "N"
            KidG1Pa = 1 - fMutation
        End Select
    End Select
End Function

Public Function KidG2Pa(MtKid As String, idx As Integer) As Double
    Select Case MtKid
    Case "M"
        Select Case idx
        Case 0
            KidG2Pa = 0
        Case 1
            KidG2Pa = 1 / (2 - fMutation)
        Case 2
            KidG2Pa = 1 / (2 - fMutation)
        Case 3
            KidG2Pa = (3 - 2 * fMutation) / ((2 - fMutation) ^ 2)
        End Select
    Case "N"
        Select Case idx
        Case 0
            KidG2Pa = 1
        Case 1
            KidG2Pa = 1 - 1 / (2 - fMutation)
        Case 2
            KidG2Pa = 1 - 1 / (2 - fMutation)
```

```
        Case 3
            KidG2Pa = 1 - (3 - 2 * fMutation) / ((2 - fMutation) ^ 2)
        End Select
    End Select
End Function
```

Bibliography

[1] R. Wayne Oldford, Personal communication from conversation with Dr. Roger Green and Dr. Ban Younghusband of the Faculty of Medicine, Memorial University of Newfoundland.

[2] Donald A. Berry, Giovanni Parmigiani, Juana Sanchez, Joellen Schildkraut, Eric Winer et al. (Feb 5, 1997) *Probability of Carrying a Mutation of Breast-Ovarian Cancer Gene BRCA1 Based on Family History.* Journal of the National Cancer Institute, **Vol. 89, No. 3**

[3] Canadian Cancer Statistics 2003

[4] H.F.A Vasen, Wijnen, J. Th., Menko, F.H. et al. (1996) *Cancer Risk in Families With Hereditary Nonpolyposis Colorectal Cancer Diagnosed by Mutation Analysis.* Gastroenterology, **110**, 1020-1027

[5] MG Dunlop, SM Farrington, I Nicholl, L Aaltonen, G Petersen et al. (2000) *Population Carrier Frequency of hMSH2 and hMLH1 Mutations.* British Journal of Cancer, **83**(23),1643-1645

[6] Malcolm G. Dunlop, Susan M. Farrington, Andrew D. Carothers, Andrew H. Wyllie, Linda Sharp et al. (1997) *Cancer Risk associated with Germline DNA Mismatch Repair Gene Mutations.* Human Molecular Genetics, **Vol.6, No. 1**, 105-110

[7] J Wijnen, W D Leeuw, H Vasen, H V D Klift, P Mxller et al. (1999) *Familial*

endometrial cancer in female carriers of MSH6 germline mutations. Nature Genetics, **23**, 142 - 144.

[8] Farrington SM, Lin-Goerke J, Ling J, Wang Y, Burczak JD, Robbins DJ, Dunlop MG et al. (1998) *Systematic analysis of hMSH2 and hMLH1 in young colon cancer patients and controls* American Journal of Human Genetics **63(3)**:749-59