Mathematical Analysis of Genetic Risk and Equilibrium in Clinical Genetics

Dávid Straka, Emma Erkočević, Per Schrijver, and Carmen Oliver

20 June 2020

1 Introduction

The importance of genes is becoming more and more evident. Since 1865, when Gregor Mendel wrote the paper "Experiments on Plant Hybridization" that laid the foundation for genetics, there have been many advances. For instance, it is now possible to study the structure of chromosomes in detail and even manipulate it, to identify criminals by samples of their DNA, or to determine how likely a woman is to suffer from breast cancer by analysing her parental clinical record.

The aim of this project is to mathematically study clinical genetics, a field of medicine that examines risks of hereditary diseases. The focus will be on constructing statistical models to determine the risk of hereditary forms of breast cancer. First, we will briefly explain the principles of genetics and introduce the mathematics associated with it. Next, we will move on to the subject of breast cancer and constitute two distinct statistical models.

2 The principles of genetics

The hereditary information of all organisms, such as humans, plants or fungi, is stored in a large number of genes, segments of DNA that determine an organism's traits. DNA, in turn, is one of the main components of chromosomes, which are located in the nuclei of cells. Furthermore, a gene functions as a code to produce proteins that bring about an organism's characteristics, e.g. the petal color of a flower or the susceptibility of a human to breast cancer. See Figure 1 for clarification.

In particular, chromosomes always go in pairs: one originates from the mother and the other from the father. As is shown in Figure 2, humans have 23 pairs of chromosomes for example. A consequence of chromosomes occurring in pairs is that each gene also exists in twofold. The two variants of a gene, called alleles, can differ from each other due to copying errors during cell division, called mutations. In this case we speak of mutated and normal alleles. When a mutation occurs in such large numbers that it is no longer possible to distinguish which allele is mutated and which is normal, both alleles, although different, are considered normal. Lastly, a mutation does not necessarily have to cause a problem: the phenotype (the outward appearance) of the organism may remain the same, even though the genotype (the genetic information) has changed. However, it is possible for offspring to be affected by the mutation even if the parent is not.

Additionally, reproductive cells (gametes) have a single set of chromosomes in their nuclei instead of pairs (23 instead of 46 chromosomes). That is why the union of two reproductive cells will give a normal cell. Moreover, for every gene it is random which allele of the parent will be passed to the offspring, which is defined as Mendel's first law.

Let us look at an example to make matters simpler. Consider the genotype of the petal color of a flower. A genotype is written as a pair of alleles (one from the father and the other from the mother). If the allele for "red petal color" is denoted by R and the allele for "white petal color" by W, then there are three different possible genotypes: RR, WW and RW, where the order of the alleles is irrelevant. Furthermore, a flower with RR or WW is called a homozygote, whereas a flower with RW is a heterozygote.

Following the notation for genotypes presented above, it is possible to determine the ratios of offspring genotypes given the genotypes of the parents. Consider, for example, the case where two heterozygous RW

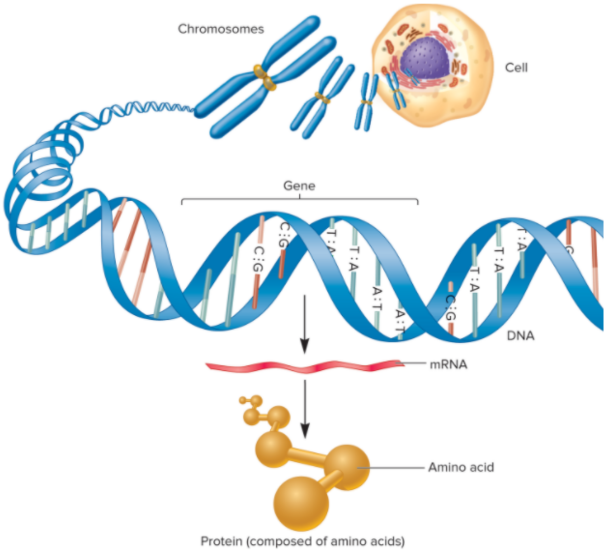


Figure 1: An overview of the structure of the hereditary material.

plants are bred. From the cross table below it follows that 50% of the offspring is heterozygous RW and 25% is homozygous WW.

	R	**W**
R	RR	RW
W	RW	WW

Now that the concept of genotype is clear, let us briefly touch upon the notion of phenotype, i.e. the outward appearance. If flowers with genotype RW are red, then red is said to be dominant over white and white is called recessive with respect to red, implying that the R allele is expressed and the W allele is not. In the case that RW plants produce pink petals, we speak of codominant alleles.

Having explained the fundamental principles concerning the hereditary material, it is now time to start with the corresponding mathematics. Suppose that there is a field of plants whose flowers are either red or white and that the allele for red petals is dominant over the allele for white petals. Furthermore, denote the probability that an arbitrary allele on the petal colour gene of a random plant on the field is R by p_R, that it is W by p_W and observe that $p_R + p_W = 1$. Similarly, write p_{RR}, p_{RW} or p_{WW} for the probability that an arbitrary plant has genotype RR, RW or WW respectively.

In general, a population is said to be in Hardy-Weinberg equilibrium for a particular gene if the two alleles on the gene of an arbitrary individual in the population are independently and identically distributed. In other words, a Hardy-Weinberg equilibrium implies that $p_{WW} = p_W^2$, $p_{RR} = p_R^2$ and $p_{RW} = 2p_R p_W$ in the case of the field with red and white flowers. It is claimed that once a population reaches the Hardy-Weinberg equilibrium, it remains in equilibrium, under the condition that random mating takes place. The remainder of this section is dedicated to proving the claim, i.e. showing that if one generation is in equilibrium, the offspring will be in equilibrium as well.

First, denote the probability that the genotype of the offspring is WW by q_{WW}. The table at the end of this section subsequently gives $q_{WW} = \frac{1}{4} * p_{RW}^2 + \frac{1}{2} * 2p_{RW}p_{WW} + p_{WW}^2$. Simplifying the right hand side, we obtain

$$q_{WW} = \left(\frac{1}{2}p_{RW} + p_{WW}\right)^2 .$$

Similarly, we have $q_{RR} = p_{RR}^2 + \frac{1}{2} * 2p_{RR}p_{RW} + \frac{1}{4} * p_{RW}^2$, which simplifies to

$$q_{RR} = \left(p_{RR} + \frac{1}{2}p_{RW}\right)^2 ,$$

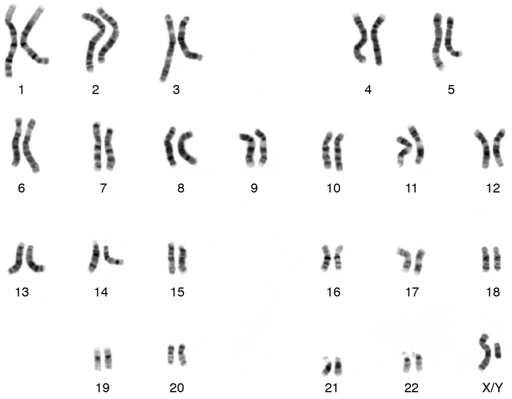


Figure 2: The 23 chromosome pairs of humans.

and $q_{RW} = \frac{1}{2} * 2p_{RR}p_{RW} + 2p_{RR}p_{WW} + \frac{1}{2} * p_{RW}^2 + \frac{1}{2} * 2p_{RW}p_{WW}$, which can be written as

$$q_{RW} = 2\left(p_{RR} + \frac{1}{2}p_{RW}\right)\left(p_{WW} + \frac{1}{2}p_{RW}\right).$$

Using that the parents are in Hardy-Weinberg equilibrium, i.e. $p_{WW} = p_W^2$, $p_{RR} = p_R^2$ and $p_{RW} = 2p_Rp_W$, we can further simplify the expressions of q_{WW}, q_{RR} and q_{RW}. For q_{WW}, we write $q_{WW} = \left(\frac{1}{2} * 2p_Rp_W + p_W^2\right)^2$. Taking out the factor p_W yields $q_{WW} = p_W^2\left(p_R + p_W\right)^2$. Recall that $p_R + p_W = 1$ to conclude that

$$q_{WW} = p_W^2. \tag{1}$$

Similarly, writing $q_{RR} = \left(p_R^2 + \frac{1}{2} * 2p_Rp_W\right)^2$, taking out the factor p_R to obtain $q_{RR} = p_R^2\left(p_R + p_W\right)^2$, and using that $p_R + p_W = 1$, results in

$$q_{RR} = p_R^2. \tag{2}$$

For q_{RW}, we have $q_{RW} = 2\left(p_R^2 + \frac{1}{2} * 2p_Rp_W\right)\left(p_W^2 + \frac{1}{2} * 2p_Rp_W\right)$, which can be simplified as $q_{RW} = 2p_Rp_W\left(p_R + p_W\right)^2$. As a result,

$$q_{RW} = 2p_Rp_W. \tag{3}$$

Finally, Equations (1), (2) and (3) show that once a population is in Hardy-Weinberg equilibrium, it remains in equilibrium, provided that there is random mating.

Genotypes parents	**Probability of breeding**	**Genotype offspring**		
		RR	**RW**	**WW**
RR x RR	p_{RR}^2	1	0	0
RR x RW	$2p_{RR}p_{RW}$	1/2	1/2	0
RR x WW	$2p_{RR}p_{WW}$	0	1	0
RW x RW	p_{RW}^2	1/4	1/2	1/4
RW x WW	$2p_{RW}p_{WW}$	0	1/2	1/2
WW x WW	p_{WW}^2	0	0	1

3 Breast Cancer

In modern society, breast cancer is occurring more and more frequently: see Figure 3 for the breast cancer incidence in the Netherlands over the years. More specifically, 14862 breast cancer cases among Dutch women have been recorded in 2019, of which 10508 concerns women of age 45 to 74 [4]. Due to the large number of cases among older women, in the Netherlands all women aged 50 to 70 are invited annually to undergo breast cancer screening. As a result, tumours can be detected and treated earlier, which increases survival rates.

Approximately 5 to 10% of breast cancer cases have a hereditary cause. In 1994 and 1995, two genes on which a mutation implies a high risk of breast cancer were discovered: BRCA1 and BRCA2. However, there are examples of families with relatively many breast cancer cases, but without mutations in BRCA1 or BRCA2. Therefore, more genes are suspected to be involved.

Before a woman is allowed to get checked for breast cancer in a clinical genetics centre, her general practitioner must first determine the breast cancer risk on the basis of the medical history of her family. There are some problems regarding this risk assessment that need to be addressed. First, having breast cancer in the family does not necessarily imply the presence of a mutation in the family. Secondly, due to the general trend of families becoming smaller, mutations express themselves less frequently, making it harder to trace them. This is especially the case in families with many men. In the third place, more genes than BRCA1 and BRCA2 may take part in hereditary breast cancer, as was already mentioned before. Lastly, environmental factors such as the age at which a woman has her first menstruation and the age at which her first child is born also play a role.

The goal in the following sections is to set up mathematical models to estimate breast cancer risks. We will start with the so-called Claus model, which was invented before the discovery of the BRCA1 and BRCA2 genes, but is nevertheless of great use. Thereafter, we will proceed by constructing a newer model that does take into account BRCA1 and BRCA2. Finally, we will include ovarian cancer related to hereditary breast cancer in the newer model and show some practical applications.

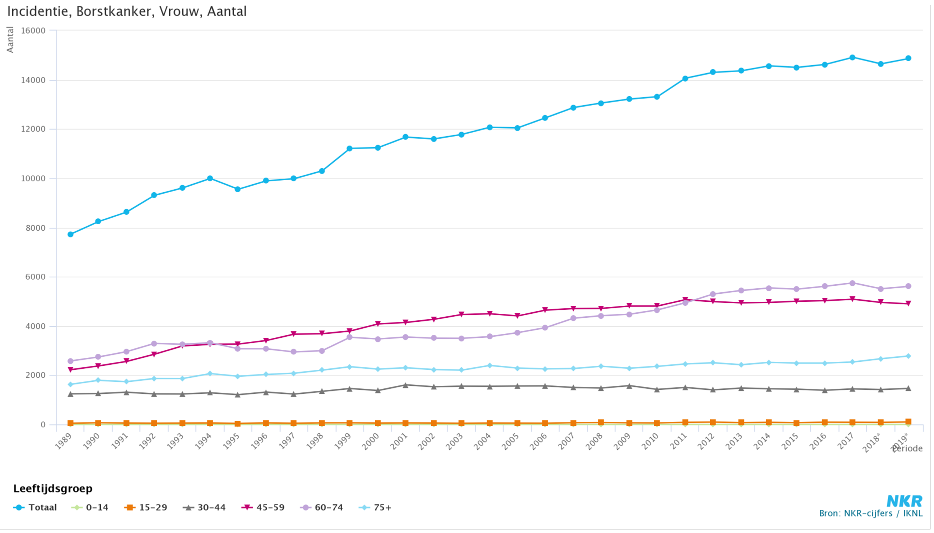


Figure 3: The number of breast cancer cases among Dutch women over the years, divided over different age categories.

4 The Claus model

In this section we construct a mathematical model called the Claus model. Its aim is to asses the risk of breast cancer in a specific patient. The model dates from 1991 and is named after its inventor Elisabeth Claus.

First, let us introduce the basic concepts of the Claus model. The model presumes that one fictitious gene relates to hereditary breast cancer (and thus leaves out the discovery of BRCA genes) and that the mutated allele of this gene increases breast cancer risks. The mutated allele is denoted by A and the normal allele by a, where A is dominant over a. Additionally, it is assumed that only women can get breast cancer and environmental factors are disregarded. Moreover, the probability that a woman gets breast cancer before a certain age is given by

$$F_A(t) = P(T \leq t \mid \text{genotype } Aa \text{ or } AA)$$

if she possesses the mutated allele, and by

$$F_a(t) = P(T \leq t \mid \text{genotype } aa)$$

if she has two normal alleles. Here T stands for the age at which a woman gets breast cancer. Finally, we suppose that the population is in Hardy-Weinberg equilibrium, i.e. $p_{AA} = p_A^2$, $p_{aa} = p_a^2$ and $p_{Aa} = 2p_A p_a$ with i.i.d. alleles.

Using the notation above, it is now possible to elegantly express various probabilities. For example, let us express the probability that an arbitrary woman from the population gets breast cancer before the age of 70, i.e. $P(T \leq 70)$. By probability theory,

$$\begin{aligned} P(T \leq 70) = P(T \leq 70 \mid \text{genotype } aa)&P(\text{genotype } aa) \\ &+ P(T \leq 70 \mid \text{genotype } Aa \text{ or } AA)P(\text{genotype } Aa \text{ or } AA). \end{aligned}$$

This expression can now be written as $P(T \leq 70) = p_{aa}F_a(70) + (p_{Aa} + p_{AA})F_A(70)$, and subsequently as

$$P(T \leq 70) = p_a^2 F_a(70) + (2p_A p_a + p_A^2)F_A(70).$$

In the following example, we find an expression for the probability that a daughter, whose father and mother have genotypes aa and Aa respectively, has not had breast cancer before the age of 50, i.e. $1 - P(T \leq 50)$. Again, from probability theory it follows that

$$\begin{aligned} 1 - P(T \leq 50) = 1 - P(T \leq 50 \mid \text{genotype } aa)&P(\text{genotype } aa) \\ &- P(T \leq 50 \mid \text{genotype } Aa \text{ or } AA)P(\text{genotype } Aa \text{ or } AA). \end{aligned}$$

Since $P(\text{genotype } aa) = \frac{1}{2} = P(\text{genotype } Aa \text{ or } AA)$, the expression above simplifies to

$$1 - P(T \leq 50) = 1 - \frac{1}{2}F_a(50) - \frac{1}{2}F_A(50).$$

To limit the length of calculations, let us work out the model by using first-degree family phenotype information only. More specifically, the focus will be on the woman for whom we want to determine the probability of breast cancer, called the proband, and her mother. Denote the ages at which the proband and her mother develop breast cancer by T_p and T_m, and the events that they are carriers of the mutation by M_p and M_m respectively. As a result,

$$P(T_p \leq t \mid M_p) = F_A(t) \text{ and } P(T_p \leq t \mid M_p^c) = F_a(t)$$

for the proband, and

$$P(T_m \leq t \mid M_m) = F_A(t) \text{ and } P(T_m \leq t \mid M_m^c) = F_a(t)$$

for the mother. In the following two paragraphs, the cumulative distribution functions $F_{T_p}(t) = F_{T_m}(t) = P(T_m \leq t)$ and $F_{T_p,T_m}(t,s) = P(T_p \leq t, T_m \leq s)$ are determined.

To find an expression for $F_{T_p}(t) = F_{T_m}(t) = P(T_m \leq t)$, first write $F_{T_p}(t) = F_{T_m}(t) = P(T_m \leq t) = P(T_m \leq t \mid M_m)P(M_m) + P(T_m \leq t \mid M_m^c)P(M_m^c)$. Next, observe that $P(M_m) = P(M_p) = p_{AA} + p_{Aa}$, which can be written as

$$P(M_m) = P(M_p) = -p_A^2 + 2p_A p_a$$

by using that $p_{AA} = p_A^2$, $p_{Aa} = 2p_A p_a$ and $p_A + p_a = 1$. For $P(M_m^c) = P(M_p^c)$, notice that $P(M_m^c) = P(M_p^c) = 1 - P(M_p)$ to conclude that

$$P(M_m^c) = P(M_p^c) = 1 + p_A^2 - 2p_A p_a.$$

Genotypes	**Prob. of breeding**	**Approximate prob.**	$\mathbf{M_p}$	$\mathbf{M_p^c}$
AA x AA	p_{AA}^2	0	-	-
AA x Aa	$p_{AA}p_{Aa}$	0	-	-
AA x aa	$p_{AA}p_{aa}$	0	-	-
Aa x AA	$p_{Aa}p_{AA}$	0	-	-
Aa x Aa	p_{Aa}^2	0	-	-
Aa x aa	$p_{Aa}p_{aa}$	$2p_A$	1/2	1/2
aa x AA	$p_{aa}p_{AA}$	0	-	-
aa x Aa	$p_{aa}p_{Aa}$	$2p_A$	1/2	1/2
aa x aa	p_{aa}^2	$1-4p_A$	0	1

Table 1: A help to compute $P(M_m \cap M_p)$, $P(M_m \cap M_p^c)$, $P(M_m^c \cap M_p)$ and $P(M_m^c \cap M_p^c)$.

Now it follows that

$$F_{T_p}(t) = F_{T_m}(t) = P(T_m \leq t) = (-p_A^2 + 2p_A)F_A(t) + (1 + p_A^2 - 2p_A)F_a(t).$$

Finally, for computational purposes, let us leave out all terms p_A^n with $n \geq 2$ in the expression of $F_{T_p}(t) = F_{T_m}(t) = P(T_m \leq t)$ above to obtain

$$F_{T_p}(t) = F_{T_m}(t) = P(T_m \leq t) \approx 2p_A F_A(t) + (1 - 2p_A)F_a(t). \tag{4}$$

This is reasonable for sufficiently small values of p_A, because p_A^n with $n \geq 2$ then becomes arbitrarily small.

To determine the joint distribution function $F_{T_p,T_m}(t,s) = P(T_p \leq t, T_m \leq s)$, let us first write

$$\begin{aligned} F_{T_p,T_m}(t,s) = P(T_p \leq t, T_m \leq s) &= P(T_p \leq t, T_m \leq s \mid M_m \cap M_p)P(M_m \cap M_p) \\ &+ P(T_p \leq t, T_m \leq s \mid M_m \cap M_p^c)P(M_m \cap M_p^c) \\ &+ P(T_p \leq t, T_m \leq s \mid M_m^c \cap M_p)P(M_m^c \cap M_p) \\ &+ P(T_p \leq t, T_m \leq s \mid M_m^c \cap M_p^c)P(M_m^c \cap M_p^c). \end{aligned} \tag{5}$$

Thereafter, we use Table 1 at the top of this page to find that

$$P(M_m \cap M_p) \approx \frac{1}{2} * 2p_A = p_A, \;\; P(M_m \cap M_p^c) \approx \frac{1}{2} * 2p_A = p_A,$$

$$P(M_m^c \cap M_p) \approx \frac{1}{2} * 2p_A = p_A \;\text{ and }\; P(M_m^c \cap M_p^c) \approx 1 - 4p_A + \frac{1}{2} * 2p_A = 1 - 3p_A.$$

Note that we again left out all terms p_A^n with $n \geq 2$ in the third column of the table and that this results in excluding the genotypes AA x AA, AA x Aa, AA x aa, Aa x AA, Aa x Aa and aa x AA. The final step is to determine the conditional probabilities in Equation (5). Since it is given whether or not the proband and mother are carriers, the random variables T_p and T_m in these conditional probabilities are independent. Therefore, we conclude that

$$\begin{aligned} F_{T_p,T_m}(t,s) &= P(T_p \leq t, T_m \leq s) \\ &\approx p_A F_A(t)F_A(s) + p_A F_a(t)F_A(s) + p_A F_A(t)F_a(s) + (1 - 3p_A)F_a(t)F_a(s). \end{aligned} \tag{6}$$

Having determined the cumulative distribution functions F_{T_m} and F_{T_p,T_m}, we deduce the probability density f_{T_m} of T_m and the joint probability density f_{T_m,T_p} of T_p and T_m by differentiating Equations (4) and (6). We denote the derivatives of F_a and F_A by f_a and f_A. It follows that

$$f_{T_m}(s) = \frac{d}{dt}F_{T_m}(t) \approx 2p_A f_A(t) + (1 - 2p_A)f_a(t)$$

and

$$f_{T_p,T_m}(t,s) = \frac{\partial^2}{\partial t \partial s}F_{T_p,T_m}(t,s) \approx p_A f_A(t)f_A(s) + p_A f_a(t)f_A(s) + p_A f_A(t)f_a(s) + (1 - 3p_A)f_a(t)f_a(s).$$

Now we can look at the probability of the proband having breast cancer if the mother has breast cancer. In general, the conditional probability of T_p given T_m is written as

$$f_{T_p|T_m}(t|s) = \frac{f_{T_p,T_m}(t,s)}{f_{T_m}(s)}.$$

Substituting the expressions for f_{T_m} and f_{T_m,T_p} we just obtained then yields

$$f_{T_p|T_m}(t|s) \approx \frac{p_A f_A(t) f_A(s) + p_A f_a(t) f_A(s) + p_A f_A(t) f_a(s) + (1-3p_A) f_a(t) f_a(s)}{2p_A f_A(s) + (1-2p_A) f_a(s)}.$$

In addition, let us find an expression for $P(T_p \le x \mid T_m = s)$. From probability theory it follows that $P(T_p \le x \mid T_m = s) = \int_{-\infty}^{x} f_{T_p|T_m}(t|s)dt$. Using the expression of $f_{T_p|T_m}(t|s)$ above and carrying out the integration then results in

$$\begin{aligned} P(T_p \le x \mid T_m = s) \approx{}& \frac{p_A F_A(x) f_A(s) + p_A F_a(x) f_A(s) + p_A F_A(x) f_a(s) + (1-3p_A) F_a(x) f_a(s)}{2p_A f_A(s) + (1-2p_A) f_a(s)} \\ &- \frac{p_A F_A(-\infty) f_A(s) + p_A F_a(-\infty) f_A(s) + p_A F_A(-\infty) f_a(s) + (1-3p_A) F_a(-\infty) f_a(s)}{2p_A f_A(s) + (1-2p_A) f_a(s)}. \end{aligned}$$

However, since $F_A(-\infty) = 0 = F_a(-\infty)$, the second term vanishes. Therefore, we conclude that

$$P(T_p \le x \mid T_m = s) \approx \frac{p_A F_A(x) f_A(s) + p_A F_a(x) f_A(s) + p_A F_A(x) f_a(s) + (1-3p_A) F_a(x) f_a(s)}{2p_A f_A(s) + (1-2p_A) f_a(s)}. \tag{7}$$

Claus applied her model with the fictitious gene to reality in order to acquire estimates for the model parameters. Let us now introduce these estimates. First of all, $\hat{p}_A = 0.0033$. Moreover, $F_A(t)$ and $F_a(t)$ are defined as normal distributions: $F_A(t)$ has expectation 55.432, standard deviation 15.387 and a correction of 0.928 as a mutation does not imply breast cancer, and $F_a(t)$ has expectation 68.900, standard deviation 15.387 and a correction of 0.100. Hence the probability that a woman gets breast cancer before age t (in years) is given by

$$F_A(t) = 0.928\Phi\left(\frac{t - 55.432}{15.387}\right) \tag{8}$$

if she has at least one mutation and by

$$F_a(t) = 0.100\Phi\left(\frac{t - 68.900}{15.387}\right). \tag{9}$$

if she has no mutations at all.

The probability of cancer is greater if there is a mutation in the Claus gene. Substitution of $t = 50$ into Equations (8) and (9) gives the probability of cancer before turning 50 based on the presence of the mutated allele. If the mutated allele is present, then the probability of cancer is given by

$$F_A(50) \approx 0.3370. \tag{10}$$

If there is no mutation, the probability of cancer is estimated by

$$F_a(50) \approx 0.0109. \tag{11}$$

Hence, as $F_A(50) > F_a(50)$, a mutation in the Claus gene indeed increases the probability of getting cancer.

Low frequency p_A of the mutated allele implies that an arbitrary proband is almost as likely to get cancer as a proband with no mutation. Assuming $\hat{p}_A = 0.0033$ and substituting Results (10) and (11) into Equation (4) yield the probability of cancer

$$F_{T_p}(50) \approx 0.0131$$

for an arbitrary proband younger than 50. Notice that the probability of cancer for an arbitrary proband $F_{T_p}(50)$ is almost equal to the probability of cancer for a proband with no mutation $F_a(50)$. In particular, the fact that no mutation is present in the gene lowers the probability by approximately 0.022 compared to not knowing if a mutation is present or not.

Recall that Dutch women are eligible for breast cancer screening from the age of 50. However, as was mentioned before, there is a higher risk of breast cancer if a mutation has occurred on the Claus gene. That is why a woman with breast cancer in her family would benefit from undertaking a screening even if she is younger than 50. Therefore, women younger than 50 with breast cancer in their families are also allowed to get screened, under the criterion that their risk of developing breast cancer in the next 5 years (computed with the Claus model) is at least as high as the risk that an arbitrary woman who has not had breast cancer before her 50th birthday gets it in the 5 years following it.

In light of the criterion, let us calculate the probability that a 50-year old woman who has not had breast cancer gets it within the next 5 years, i.e. $P(50 < T_p \leq 55 \mid T_p > 50)$. To begin, write

$$P(50 < T_p \leq 55 \mid T_p > 50) = \frac{P(50 < T_p \leq 55)}{P(T_p > 50)} = \frac{P(T_p \leq 55) - P(T_p \leq 50)}{1 - P(T_p \leq 50)}.$$

Using Equation (4) to express $P(T_p \leq 55)$ and $P(T_p \leq 50)$ subsequently yields

$$P(50 < T_p \leq 55 \mid T_p > 50) \approx \frac{2p_A F_A(55) + (1-2p_A)F_a(55) - 2p_A F_A(50) - (1-2p_A)F_a(50)}{1 - 2p_A F_A(50) - (1-2p_A)F_a(50)}.$$

Now compute the values of $F_A(55)$ and $F_a(55)$ using Equations (8) and (9) to finally substitute them, $F_A(50) \approx 0.3370$, $F_a(50) \approx 0.0109$ and $\hat{p}_A = 0.0033$ in the expression above. The final result then is

$$P(50 < T_p \leq 55 \mid T_p > 50) \approx 0.0083.$$

Now consider a 30-year old proband whose mother got breast cancer at age 35. To determine if the proband is eligible for breast cancer screening, it is necessary to calculate the probability that the proband will get breast cancer within the next 5 years, i.e. the probability $P(T_p \leq 35 \mid T_p > 30, T_m = 35)$. In the first place, notice that

$$\begin{aligned} P(T_p \leq 35 \mid T_p > 30, T_m = 35) &= \frac{P(30 < T_p \leq 35 \mid T_m = 35)}{P(T_p > 30 \mid T_m = 35)} \\ &= \frac{P(T_p \leq 35 \mid T_m = 35) - P(T_p \leq 30 \mid T_m = 35)}{1 - P(T_p \leq 30 \mid T_m = 35)}. \end{aligned} \tag{12}$$

Here the first equality holds because in general $P(A|(B \cap C)) = \frac{P((A \cap B)|C)}{P(B|C)}$. Secondly, using Equation (7) yields

$$P(T_p \leq 35 \mid T_m = 35) \approx \frac{p_A F_A(35) f_A(35) + p_A F_a(35) f_A(35) + p_A F_A(35) f_a(35) + (1-3p_A) F_a(35) f_a(35)}{2p_A f_A(35) + (1-2p_A) f_a(35)}$$

and a similar expression for $P(T_p \leq 30 \mid T_m = 35)$. The values of $F_A(35)$, $F_a(35)$, $F_A(30)$ and $F_a(30)$ are computed by applying Equations (8) and (9). To calculate $f_A(35)$ and $f_a(35)$, observe that

$$f_A(t) = \frac{0.928}{\sqrt{2\pi * 15.387^2}} e^{-\frac{(t-55.435)^2}{2*15.387^2}}$$

and

$$f_a(t) = \frac{0.100}{\sqrt{2\pi * 15.387^2}} e^{-\frac{(t-68.900)^2}{2*15.387^2}}.$$

Lastly, use $\hat{p}_A = 0.0033$ and the values of $F_A(35)$, $F_a(35)$, $F_A(30)$, $F_a(30)$, $f_A(35)$ and $f_a(35)$ to obtain

$$P(T_p \leq 35 \mid T_p > 30, T_m = 35) \approx 0.0053.$$

Since $0.0053 < 0.0083$, it follows that the proband is not eligible for breast cancer screening.

To determine from what age the proband in question becomes eligible for breast cancer screening according to the set criterion, the equation

$$P(T_p \leq x+5 \mid T_p > x, T_m = 35) = 0.0083 \tag{13}$$

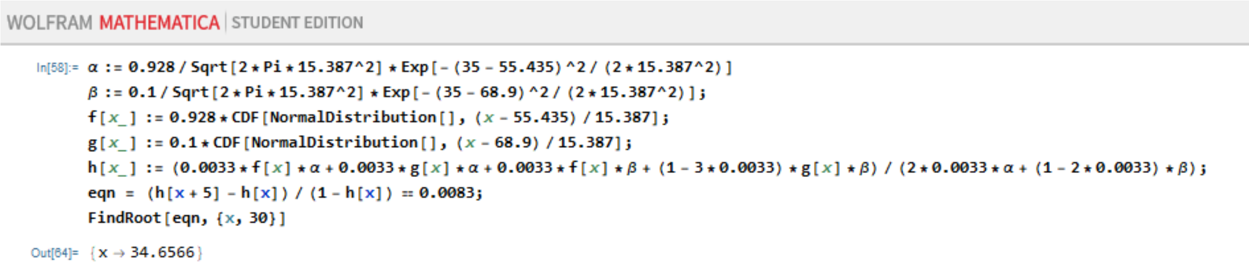
must be solved for x. To do so, first rewrite the left hand side in a similar manner as was done in Equation (12) to obtain

$$\frac{P(T_p \leq x+5 \mid T_m = 35) - P(T_p \leq x \mid T_m = 35)}{1 - P(T_p \leq x \mid T_m = 35)} = 0.0083.$$

From Equation (7) it then follows that

$$P(T_p \leq x \mid T_m = 35) \approx \frac{p_A F_A(x) f_A(35) + p_A F_a(x) f_A(35) + p_A F_A(x) f_a(35) + (1-3p_A) F_a(x) f_a(35)}{2p_A f_A(35) + (1-2p_A) f_a(35)}.$$

Writing the expression of $P(T_p \leq x \mid T_m = 35)$ as a function of x in `Mathematica`, toghether with the expressions of $F_A(x)$, $F_a(x)$, $f_A(35)$ and $f_a(35)$, and using the `FindRoot` function subsequently gives $x \approx 34.6566$. See Figure 4 for the code and Figure 5 for a plot of $P(T_p \leq x+5 \mid T_p > x, T_m = 35)$ as a function of x and the boundary condition. In conclusion, the age from which the proband is eligible for screening is 35.



```
In[58]:= α := 0.928 / Sqrt[2 * Pi * 15.387^2] * Exp[- (35 - 55.435) ^2 / (2 * 15.387^2)]
         β := 0.1 / Sqrt[2 * Pi * 15.387^2] * Exp[- (35 - 68.9) ^2 / (2 * 15.387^2)];
         f[x_] := 0.928 * CDF[NormalDistribution[], (x - 55.435) / 15.387];
         g[x_] := 0.1 * CDF[NormalDistribution[], (x - 68.9) / 15.387];
         h[x_] := (0.0033 * f[x] * α + 0.0033 * g[x] * α + 0.0033 * f[x] * β + (1 - 3 * 0.0033) * g[x] * β) / (2 * 0.0033 * α + (1 - 2 * 0.0033) * β);
         eqn = (h[x + 5] - h[x]) / (1 - h[x]) == 0.0083;
         FindRoot[eqn, {x, 30}]

Out[64]= {x → 34.6566}
```

Figure 4: The numerical solution of Equation (13) for x. Here $f_A(35)$ is denoted by α, $f_a(35)$ by β, $F_A(x)$ by `f(x)`, $F_a(x)$ by `g(x)` and $P(T_p \leq x \mid T_m = 35)$ by `h(x)`.

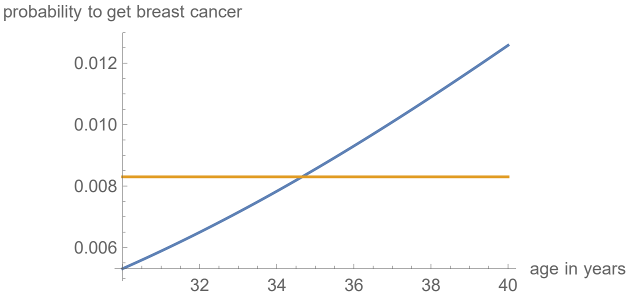


Figure 5: The left hand side of Equation (13) in blue plotted against its right hand side in orange.

5 A more elaborate model

The Claus model is interesting to start to get in touch with the subject, but there also exist more accurate models. Therefore, this section is devoted to creating a more refined model that takes into account more deterministic factors, such as the presence of ovarian cancer that is related to breast cancer and the existence of BRCA1 and BRCA2 genes. This model will eventually enable us to make a more precise estimate on the risk of breast and ovarian cancer, based on the breast and ovarian cancer history of the proband's family.

To begin, let us declare the fundamental assumptions. The model includes the fact that dominant mutations in the BRCA1 and BRCA2 genes increase risks of breast cancers and related ovarian cancers. However, as mentioned previously, more unknown genes are involved in hereditary breast cancer. Thus, we define a fictitious gene, the BRCA3 gene, that acts as an explanation of what the other two genes cannot describe. A mutation in this gene is assumed to be dominant and, contrary to the other two genes, has nothing to do with ovarian cancer. Furthermore, the three genes are located on different chromosomes, resulting in independent inheritance of their alleles.

Now it is time to consider probabilities of mutations. It is helpful to denote the mutation frequencies for BRCA1, BRCA2 and BRCA3 by p_1, p_2 and p_3 respectively. Experts have arrived at the estimates $\hat{p}_1 = 0.0006$ and $\hat{p}_2 = 0.0006$. In addition, a Hardy-Weinberg equilibrium is assumed. It follows that the probability of no mutations in the BRCA1 gene is given by

$$(1-\hat{p}_1)^2 = 0.99880036,$$

the probability of one mutation by

$$2\hat{p}_1(1-\hat{p}_1) = 0.0011928$$

and of two by

$$\hat{p}_1^2 = 3.6 * 10^{-7}.$$

Likewise, the probability that a person has at least one mutation in the BRCA1 gene and at least one mutation in the BRCA2 gene is written as

$$\left(2\hat{p}_1(1-\hat{p}_1)+\hat{p}_1^2\right)\left(2\hat{p}_2(1-\hat{p}_2)+\hat{p}_2^2\right) = 1.43913613 * 10^{-6}.$$

Looking at the probability that there is more than one mutation, we realise that it is very small with respect to the probability of just one mutation. Therefore, moving forward, we choose to ignore the probability of more than one mutation in the same gene or in different genes.

Since this model considers multiple genes and cancers, there are many distribution functions to be defined. In general, the distribution function for the age at which an arbitrary woman gets breast cancer is given by F. We denote the distribution functions for breast cancer given a mutation on BRCA1, BRCA2 or BRCA3 by F_1, F_2 and F_3 respectively and the distribution function for breast cancer with no mutations by F_0. Similarly, we write the distribution functions for ovarian cancer given a mutation in BRCA1 or BRCA2 as H_1 and H_2 respectively and the distribution function for ovarian cancer with no mutations as H_0. Moreover, we presume that given the genotype on BRCA1, BRCA2 and BRCA3, the ages at which breast and ovarian cancer arise are independent.

The risk of breast cancer can be expressed by the so-called relative risk, which is the probability of breast cancer given family history information divided by this probability without such information. Using only information on the proband's mother, the relative risk is defined as

$$\lambda(t|s) := \frac{P(T_p \leq t | T_m \leq s)}{P(T_p \leq t)} \tag{14}$$

and can be explained as the probability that the proband gets breast cancer before age t if her mother had breast cancer before age s, with respect to the probability that the proband gets breast cancer before age t in case of no information on her mother. For example, $\lambda(50|30) = 2$ means that the proband is two times more likely to get breast cancer before the age of 50 given that her mother developed breast cancer before the age of 30, than to get breast cancer before the age of 50 without any information on her mother. On the other hand, $\lambda(t|s) = 1$ if there were no hereditary forms of breast cancer. This is because in the absence of hereditary breast cancers, T_p and T_m are independent, implying that $P(T_p \leq t \mid T_m \leq s) = \frac{P(T_p \leq t)P(T_m \leq s)}{P(T_m \leq s)}$. Substituting this result in Definition (14), the numerator and denominator cancel each other out, yielding $\lambda(t|s) = 1$.

Let us now find expressions for $F(t)$ and $F(s)F(t)$. For $F(t)$, we first write

$$\begin{aligned} F(t) &= P(T \leq t) \\ &\approx P(T \leq t \mid \text{no mut.'s})P(\text{no mut.'s}) + \sum_{i=1,2,3} P(T \leq t \mid \text{mut. in BRCA}i)P(\text{mut. in BRCA}i). \end{aligned}$$

As stated earlier, probabilities of more than one mutation are left out here. Next, we use that

- $P(\text{no mut.'s}) \approx 1 - P(\text{mut. in BRCA1}) - P(\text{mut. in BRCA2}) - P(\text{mut. in BRCA3})$,
- $P(T \leq t \mid \text{no mut.'s}) = F_0(t)$,
- $P(\text{mut. in BRCA}i) \approx 2p_i$ for $i = 1, 2, 3$ and
- $P(T \leq t \mid \text{mut. in BRCA}i) = F_i(t)$ for $i = 1, 2, 3$

to conclude that

$$F(t) \approx (1 - 2p_1 - 2p_2 - 2p_3)F_0(t) + \sum_{i=1,2,3} 2p_i F_i(t). \tag{15}$$

For $F(s)F(t)$, we start by writing

$$\begin{aligned} F(s)F(t) \approx {} & [(1 - 2p_1 - 2p_2 - 2p_3)F_0(s) + 2p_1F_1(s) + 2p_2F_2(s) + 2p_3F_3(s)] \\ & * [(1 - 2p_1 - 2p_2 - 2p_3)F_0(t) + 2p_1F_1(t) + 2p_2F_2(t) + 2p_3F_3(t)]. \end{aligned}$$

Multiplying out the pairs of brackets and eliminating all terms with more than one factor p_i for $i = 1, 2, 3$ results in

$$\begin{aligned} F(s)F(t) \approx {} & (1 - 4p_1 - 4p_2 - 4p_3)F_0(s)F_0(t) + 2p_1F_0(s)F_1(t) + 2p_2F_0(s)F_2(t) + 2p_3F_0(s)F_3(t) \\ & + 2p_1F_1(s)F_0(t) + 2p_2F_2(s)F_0(t) + 2p_3F_3(s)F_0(t). \end{aligned}$$

Lastly, this can be written more elegantly as

$$F(s)F(t) \approx (1 - 4p_1 - 4p_2 - 4p_3)F_0(s)F_0(t) + \sum_{i=1,2,3} 2p_i \left[F_i(s)F_0(t) + F_0(s)F_i(t)\right].$$

Having determined an expression for F, we proceed by finding a formula for the joint distribution $F_{T_p,T_m}(t, s)$. First of all, because probabilities of more than one mutation are discounted, the proband and her mother can only each have a mutation in the same gene and not in two different genes. Therefore,

$$\begin{aligned} F_{T_p,T_m}(t, s) = P(T_p \leq t, T_m \leq s) \approx {} & P(T_p \leq t, T_m \leq s \mid \text{no mut.'s})P(\text{no mut.'s}) \\ & + \sum_{i=1,2,3} \big[P(T_p \leq t, T_m \leq s \mid \text{both mut. in BRCA}i)P(\text{both mut. in BRCA}i) \\ & + P(T_p \leq t, T_m \leq s \mid \text{only } m \text{ mut. in BRCA}i)P(\text{only } m \text{ mut. in BRCA}i) \\ & + P(T_p \leq t, T_m \leq s \mid \text{only } p \text{ mut. in BRCA}i)P(\text{only } p \text{ mut. in BRCA}i)\big]. \end{aligned}$$

From

- $P(\text{both mut. in BRCA}i) \approx p_i$, $P(\text{only } m \text{ mut. in BRCA}i) \approx p_i$, $P(\text{only } p \text{ mut. in BRCA}i) \approx p_i$,
- $P(\text{no mut.'s})$
  $\approx 1 - \sum_{i=1,2,3} \left[P(\text{both mut. in BRCA}i) + P(\text{only } m \text{ mut. in BRCA}i) + P(\text{only } p \text{ mut. in BRCA}i)\right]$
  $\approx 1 - 3\sum_{i=1,2,3} p_i$,
- $P(T_p \leq t, T_m \leq s \mid \text{no mut.'s}) = F_0(t)F_0(s)$,
  $P(T_p \leq t, T_m \leq s \mid \text{both mut. in BRCA}i) = F_i(t)F_i(s)$,
  $P(T_p \leq t, T_m \leq s \mid \text{only } m \text{ mut. in BRCA}i) = F_0(t)F_i(s)$ and
  $P(T_p \leq t, T_m \leq s \mid \text{only } p \text{ mut. in BRCA}i) = F_i(t)F_0(s)$

it then follows that

$$\begin{aligned} & F_{T_p,T_m}(t, s) = P(T_p \leq t, T_m \leq s) \\ & \approx \left(1 - 3\sum_{i=1,2,3} p_i\right) F_0(t)F_0(s) + \sum_{i=1,2,3} p_iF_i(t)F_i(s) + \sum_{i=1,2,3} p_iF_0(t)F_i(s) + \sum_{i=1,2,3} p_iF_i(t)F_0(s). \end{aligned}$$

Here the approximations in the first dot are obtained by using Table 1 and writing p_i instead of p_A.

To conclude this section, let us express $\lambda(t|s) - 1$ in terms of the mutation frequencies and distribution functions. The expression will be needed in the following section. By Definition (14),

$$\lambda(t|s) - 1 = \frac{P(T_p \le t \mid T_m \le s)}{P(T_p \le t)} - 1 = \frac{P(T_p \le t, T_m \le s)}{P(T_p \le t)P(T_m \le s)} - 1.$$

Writing this in terms of the distribution functions, it follows that

$$\lambda(t|s) - 1 = \frac{F_{T_p,T_m}(t,s)}{F(t)F(s)} - 1 = \frac{1}{F(t)F(s)}\Big(F_{T_p,T_m}(t,s) - F(t)F(s)\Big).$$

Using the just obtained approximations of $F_{T_p,T_m}(t,s)$ and $F(t)F(s)$ subsequently yields

$$\begin{aligned}\lambda(t|s) - 1 \approx{}& \frac{1}{F(t)F(s)}\Bigg(\sum_{i=1,2,3} p_i\Big(F_i(t)F_i(s) + F_0(t)F_i(s) + F_i(t)F_0(s)\Big) + (1 - 3p_1 - 3p_2 - 3p_3)F_0(t)F_0(s)\\ &- (1 - 4p_1 - 4p_2 - 4p_3)F_0(t)F_0(s) - \sum_{i=1,2,3} 2p_i\Big(F_i(s)F_0(t) + F_0(s)F_i(t)\Big)\Bigg)\\ \approx{}& \frac{1}{F(t)F(s)}\left(\sum_{i=1,2,3} p_i\Big(F_i(t) - F_0(t)\Big)\Big(F_i(s) - F_0(s)\Big)\right) \approx \sum_{i=1,2,3} p_i\left(\frac{F_i(t)}{F(t)} - \frac{F_0(t)}{F(t)}\right)\left(\frac{F_i(s)}{F(s)} - \frac{F_0(s)}{F(s)}\right).\end{aligned}$$

In short, the final result is given by

$$\lambda(t|s) - 1 \approx \sum_{i=1,2,3} p_i\left(\frac{F_i(t)}{F(t)} - \frac{F_0(t)}{F(t)}\right)\left(\frac{F_i(s)}{F(s)} - \frac{F_0(s)}{F(s)}\right). \tag{16}$$

6 Estimating the parameters

Estimating the model's parameters is key to proceed with the investigation of breast cancer risk. In this section, we will first determine point estimates of the functions $F_0(t)$, $F_1(t)$, $F_2(t)$, $F_3(t)$, $F(t)$ and $\lambda(t|t)$ at times $t = 40, 50, 60, 70$ using breast cancer data, to then find formulas for F_0, F_1, F_2 and F_3 by fitting normal distributions to the point estimates.

Let us first introduce the estimates of $F_1(t)$, $F_2(t)$, $F(t)$ and $\lambda(t|t)$ at $t = 40, 50, 60, 70$. They are made with data gathered from various countries and are collected in the following table.

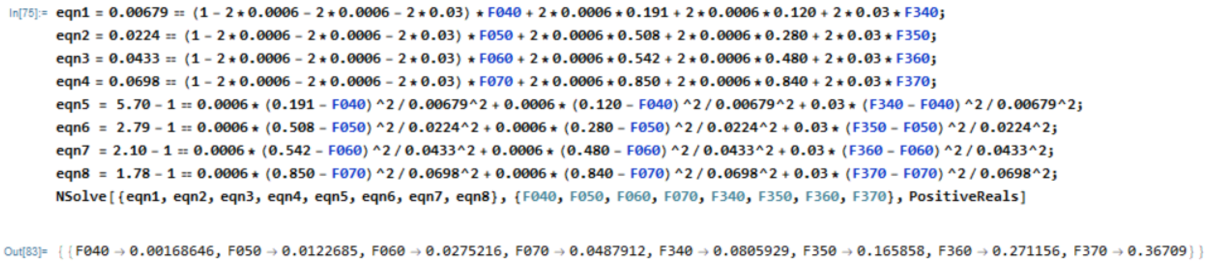
Age t	$\hat{\mathbf{F}}_\mathbf{1}(\mathbf{t})$	$\hat{\mathbf{F}}_\mathbf{2}(\mathbf{t})$	$\hat{\mathbf{F}}(\mathbf{t})$	$\hat{\lambda}(\mathbf{t}\vert\mathbf{t})$
40	0.191	0.120	0.00679	5.70
50	0.508	0.280	0.0224	2.79
60	0.542	0.480	0.0433	2.10
70	0.850	0.840	0.0698	1.78

More specifically, the estimations of F_1, F_2 and F follow from Dutch breast cancer data and those of λ from international data. To obtain the data, experts make observations and studies on many women. Nevertheless, the mutations on the BRCA1 and BRCA2 genes are very rare and hence careful estimates on p_1 and p_2 have to be made. In particular, $\hat{p}_1 = 0.0006 = \hat{p}_2$, as stated in the previous section.

Substituting the values given in the table and $\hat{p}_1 = 0.0006 = \hat{p}_2$ into Equations (15) and (16) results in a system of eight equations with nine unknowns. Because there are more unknowns than equations this system cannot be solved. To eliminate one unknown, we use that in the literature $\hat{p}_3 = 0.03$ is given as an estimate of the parameter p_3. As shown in Figure 6, the `NSolve` function in `Mathematica` then solves the system for each unknown. In the below the estimates of the distribution functions $F_0(t), F_1(t), F_2(t)$ and $F_3(t)$ for the ages $t = 40, 50, 60, 70$ are given.

$$\hat{F}_0(40) = 0.00169 \qquad \hat{F}_0(50) = 0.0123 \qquad \hat{F}_0(60) = 0.0275 \qquad \hat{F}_0(70) = 0.0488$$

$$\hat{F}_3(40) = 0.0806 \qquad \hat{F}_3(50) = 0.166 \qquad \hat{F}_3(60) = 0.271 \qquad \hat{F}_3(70) = 0.367$$



```
In[75]:= eqn1 = 0.00679 == (1 - 2 * 0.0006 - 2 * 0.0006 - 2 * 0.03) * F040 + 2 * 0.0006 * 0.191 + 2 * 0.0006 * 0.120 + 2 * 0.03 * F340;
         eqn2 = 0.0224 == (1 - 2 * 0.0006 - 2 * 0.0006 - 2 * 0.03) * F050 + 2 * 0.0006 * 0.508 + 2 * 0.0006 * 0.280 + 2 * 0.03 * F350;
         eqn3 = 0.0433 == (1 - 2 * 0.0006 - 2 * 0.0006 - 2 * 0.03) * F060 + 2 * 0.0006 * 0.542 + 2 * 0.0006 * 0.480 + 2 * 0.03 * F360;
         eqn4 = 0.0698 == (1 - 2 * 0.0006 - 2 * 0.0006 - 2 * 0.03) * F070 + 2 * 0.0006 * 0.850 + 2 * 0.0006 * 0.840 + 2 * 0.03 * F370;
         eqn5  =  5.70 - 1 == 0.0006 * (0.191 - F040) ^2 / 0.00679^2 + 0.0006 * (0.120 - F040) ^2 / 0.00679^2 + 0.03 * (F340 - F040) ^2 / 0.00679^2;
         eqn6  =  2.79 - 1 == 0.0006 * (0.508 - F050) ^2 / 0.0224^2 + 0.0006 * (0.280 - F050) ^2 / 0.0224^2 + 0.03 * (F350 - F050) ^2 / 0.0224^2;
         eqn7  = 2.10 - 1 == 0.0006 * (0.542 - F060) ^2 / 0.0433^2 + 0.0006 * (0.480 - F060) ^2 / 0.0433^2 + 0.03 * (F360 - F060) ^2 / 0.0433^2;
         eqn8  =  1.78 - 1 == 0.0006 * (0.850 - F070) ^2 / 0.0698^2 + 0.0006 * (0.840 - F070) ^2 / 0.0698^2 + 0.03 * (F370 - F070) ^2 / 0.0698^2;
         NSolve[{eqn1, eqn2, eqn3, eqn4, eqn5, eqn6, eqn7, eqn8}, {F040, F050, F060, F070, F340, F350, F360, F370}, PositiveReals]

Out[83]= {{F040 → 0.00168646, F050 → 0.0122685, F060 → 0.0275216, F070 → 0.0487912, F340 → 0.0805929, F350 → 0.165858, F360 → 0.271156, F370 → 0.36709}}
```

Figure 6: The system of equations that follows from substituting the data from this section's table into Equations (15) and (16). Here `F040`, `F050`, ..., `F370` represent $\hat{F}_0(40)$, $\hat{F}_0(50)$, ..., $\hat{F}_3(70)$.

Having determined the point estimates of F_0, F_1, F_2 and F_3, let us fit functions of the form $\alpha\Phi\left(\frac{t-\mu}{\sigma}\right)$ to the estimations. In other words, for each F_i with $i = 0, 1, 2, 3$ we aim to find values of α_i, μ_i and σ_i such that the function $F_i(t) = \alpha_i\Phi\left(\frac{t-\mu_i}{\sigma_i}\right)$ with $i = 0, 1, 2, 3$ best fits the point estimates. See Figure 7 for the corresponding `Mathematica` code. In the code the function `FindFit` is applied, which uses the least squares method by default. As a result,

- $F_0(t) = 0.181\Phi\left(\frac{t-83.214}{21.874}\right)$,
- $F_1(t) = 1.000\Phi\left(\frac{t-53.751}{18.567}\right)$,
- $F_2(t) = 1.000\Phi\left(\frac{t-58.674}{13.952}\right)$ and
- $F_3(t) = 0.476\Phi\left(\frac{t-56.917}{17.692}\right)$.

Lastly, Figures 8-11 display the point estimates plotted together with the distribution functions just defined.

This model is accurate enough to be used to estimate breast cancer risks. Moreover, it will have a greater accuracy than the previous model, the Claus Model. Nevertheless, we are still missing the information on ovarian cancer in the proband's mother.

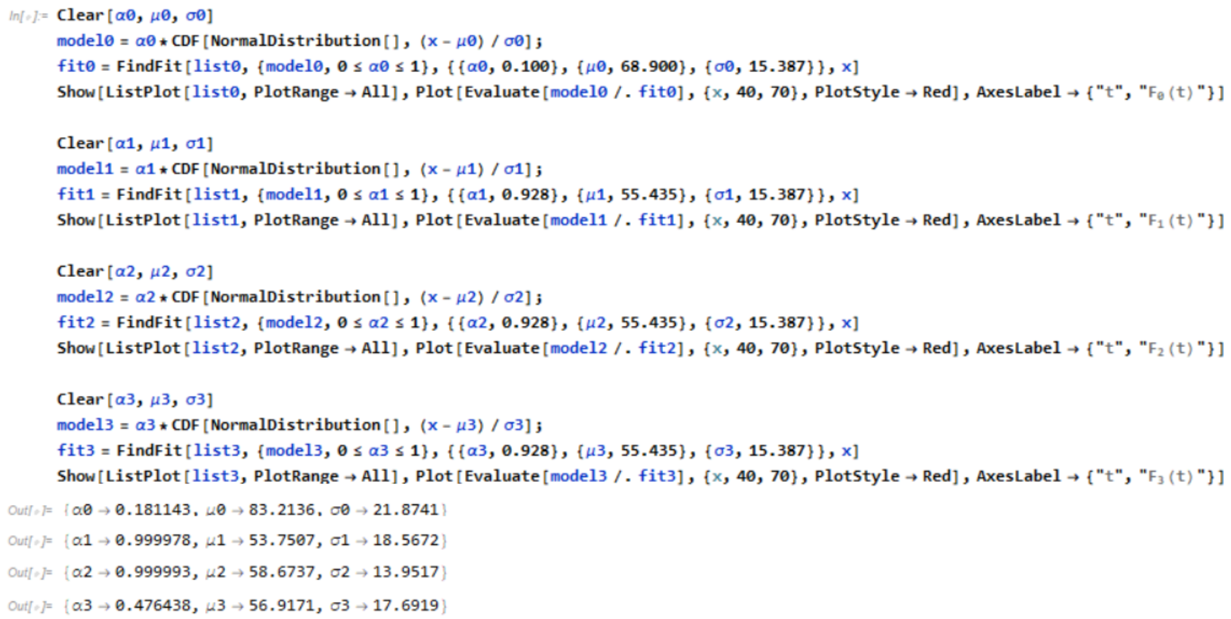
7 Extension with ovarian cancer

In this section, we will study an expanded model that takes into account what was previously missing: ovarian cancer. First of all, the estimates for the distribution functions H_0, H_1 and H_2 for ovarian cancer are described in the literature as

$$\begin{aligned} H_0(t) &= 0.01\Phi\left(\frac{t-70}{15.387}\right), \\ H_1(t) &= 0.45\Phi\left(\frac{t-45}{15.387}\right), \\ H_2(t) &= 0.125\Phi\left(\frac{t-50}{15.387}\right). \end{aligned}$$

Furthermore, we indicate the ages at which the mother and the proband get ovarian cancer by S_m and S_p respectively. For notational purposes, the random variables T_m, T_p, S_m and S_p can attain any value in the interval $(-\infty, \infty]$, where $\{S_m = \infty\}$ means the mother does not get ovarian cancer. Thus, each distribution function at infinity equals one: $H_0(\infty) = H_1(\infty) = H_2(\infty) = 1$ and similarly for F_0, F_1, F_2 and F_3.

Let us first establish formulas for the joint distribution functions F_{T_p,S_m,T_m} and F_{S_m,T_m}. Previous assumptions about possible genotypes hold, i.e. a person cannot have two mutations simultaneously and a



```
In[∘]:= Clear[α0, μ0, σ0]
        model0 = α0 * CDF[NormalDistribution[], (x - μ0) / σ0];
        fit0 = FindFit[list0, {model0, 0 ≤ α0 ≤ 1}, {{α0, 0.100}, {μ0, 68.900}, {σ0, 15.387}}, x]
        Show[ListPlot[list0, PlotRange → All], Plot[Evaluate[model0 /. fit0], {x, 40, 70}, PlotStyle → Red], AxesLabel → {"t", "F0(t)"}]

        Clear[α1, μ1, σ1]
        model1 = α1 * CDF[NormalDistribution[], (x - μ1) / σ1];
        fit1 = FindFit[list1, {model1, 0 ≤ α1 ≤ 1}, {{α1, 0.928}, {μ1, 55.435}, {σ1, 15.387}}, x]
        Show[ListPlot[list1, PlotRange → All], Plot[Evaluate[model1 /. fit1], {x, 40, 70}, PlotStyle → Red], AxesLabel → {"t", "F1(t)"}]

        Clear[α2, μ2, σ2]
        model2 = α2 * CDF[NormalDistribution[], (x - μ2) / σ2];
        fit2 = FindFit[list2, {model2, 0 ≤ α2 ≤ 1}, {{α2, 0.928}, {μ2, 55.435}, {σ2, 15.387}}, x]
        Show[ListPlot[list2, PlotRange → All], Plot[Evaluate[model2 /. fit2], {x, 40, 70}, PlotStyle → Red], AxesLabel → {"t", "F2(t)"}]

        Clear[α3, μ3, σ3]
        model3 = α3 * CDF[NormalDistribution[], (x - μ3) / σ3];
        fit3 = FindFit[list3, {model3, 0 ≤ α3 ≤ 1}, {{α3, 0.928}, {μ3, 55.435}, {σ3, 15.387}}, x]
        Show[ListPlot[list3, PlotRange → All], Plot[Evaluate[model3 /. fit3], {x, 40, 70}, PlotStyle → Red], AxesLabel → {"t", "F3(t)"}]
Out[∘]= {α0 → 0.181143, μ0 → 83.2136, σ0 → 21.8741}
Out[∘]= {α1 → 0.999978, μ1 → 53.7507, σ1 → 18.5672}
Out[∘]= {α2 → 0.999993, μ2 → 58.6737, σ2 → 13.9517}
Out[∘]= {α3 → 0.476438, μ3 → 56.9171, σ3 → 17.6919}
```

Figure 7: The parameter values that give the best fit for the point estimates.

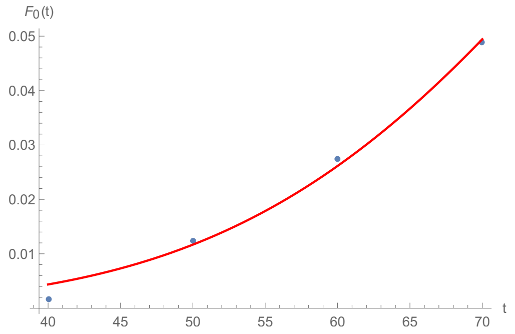


Figure 8: The point estimates of F_0 and $F_0(t) = 0.181\Phi\left(\frac{t-83.214}{21.874}\right)$.

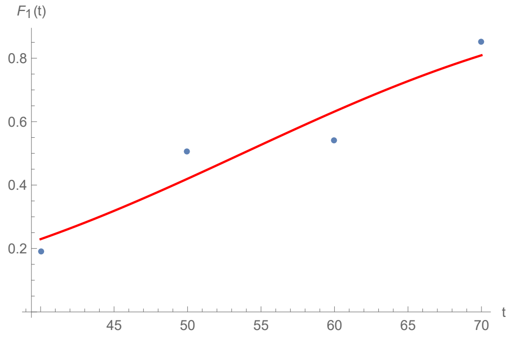


Figure 9: The point estimates of F_1 and $F_1(t) = 1.000\Phi\left(\frac{t-53.751}{18.567}\right)$.

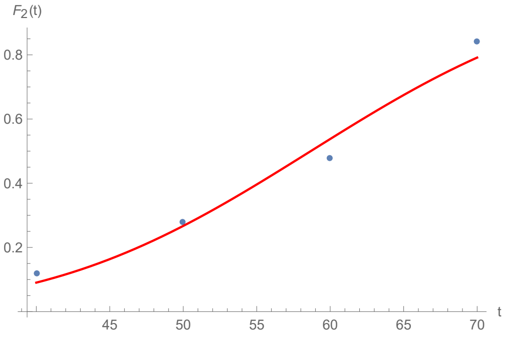


Figure 10: The point estimates of F_2 and $F_2(t) = 1.000\Phi\left(\frac{t-58.674}{13.952}\right)$.

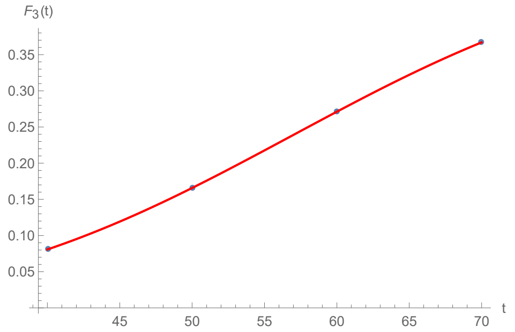


Figure 11: The point estimates of F_3 and $F_3(t) = 0.476\Phi\left(\frac{t-56.917}{17.692}\right)$.

proband cannot possess a different mutation than her mother. Conditioning on the considered genotypes then gives

$$\begin{aligned}F_{T_p,S_m,T_m}(t,s,r) &= P(T_p \le t, S_m \le s, T_m \le r)\\ &= P(\text{no mut.'s})P(T_p \le t, S_m \le s, T_m \le r \mid \text{no mut.'s})\\ &+ \sum_{i=1,2,3}\Big[P(\text{only } p \text{ mut. in BRCA}i)P(T_p \le t, S_m \le s, T_m \le r \mid \text{only } p \text{ mut. in BRCA}i)\\ &+ P(\text{only } m \text{ mut. in BRCA}i)P(T_p \le t, S_m \le s, T_m \le r \mid \text{only } m \text{ mut. in BRCA}i)\\ &+ P(\text{both mut. in BRCA}i)P(T_p \le t, S_m \le s, T_m \le r \mid \text{both mut. in BRCA}i)\Big].\end{aligned}$$

Using that T_p, S_m and T_m are independent when the genotypes are given and that a mutation in the BRCA3 gene does not increase the risk of ovarian cancer, we obtain

$$\begin{aligned}F_{T_p,S_m,T_m}(t,s,r) &= P(T_p \le t, S_m \le s, T_m \le r)\\ &= (1-3p_1-3p_2-3p_3)F_0(t)H_0(s)F_0(r)\\ &+ \left[\sum_{i=1,2} p_i\big(F_i(t)H_0(s)F_0(r) + F_0(t)H_i(s)F_i(r) + F_i(t)H_i(s)F_i(r)\big)\right]\\ &+ p_3\big(F_3(t)H_0(s)F_0(r) + F_0(t)H_0(s)F_3(r) + F_3(t)H_0(s)F_3(r)\big).\end{aligned}$$

An expression for $F_{S_m,T_m}(s,r) = P(S_m \le s, T_m \le r)$ can be found analogously: we first write

$$\begin{aligned}F_{S_m,T_m}(s,r) = P(S_m \le s, T_m \le r) &= P(\text{no mut.'s})P(S_m \le s, T_m \le r|\text{no mut.'s})\\ &+ \sum_{i=1,2,3} P(\text{mut. BRCA}i)P(S_m \le s, T_m \le r|\text{mut. BRCA}i)\end{aligned}$$

to conclude that

$$\begin{aligned}F_{S_m,T_m}(s,r) = P(S_m \le s, T_m \le r) &= (1-2p_1-2p_2-2p_3)H_0(s)F_0(r)\\ &+ \left[\sum_{i=1,2} 2p_iH_i(s)F_i(r)\right] + 2p_3H_0(s)F_3(r).\end{aligned}$$

Now that all the necessary distribution functions are determined, it is time to calculate some probabilities. For example, consider a 44-year old proband who has not yet had breast or ovarian cancer, while it is known that her mother has had ovarian cancer before the age of 43 and has never had breast cancer. Let us compute the conditional probability that the proband will get breast cancer before the age of 49, given all the information on the mother. By probability theory,

$$\begin{aligned}&P(44 \le T_p \le 49 \mid T_p \ge 44, S_m \le 43, T_m = \infty)\\ &= \frac{P(44 \le T_p \le 49, S_m \le 43, T_m = \infty)}{P(T_p \ge 44, S_m \le 43, T_m = \infty)}\\ &= \frac{P(T_p \le 49, S_m \le 43, T_m = \infty) - P(T_p \le 43, S_m \le 43, T_m = \infty)}{P(S_m \le 43, T_m = \infty) - P(T_p \le 43, S_m \le 43, T_m = \infty)}.\end{aligned}$$

Writing this in terms of the newly obtained distribution functions F_{T_p,S_m,T_m} and F_{S_m,T_m} yields

$$P(44 \le T_p \le 49 \mid T_p \ge 44, S_m \le 43, T_m = \infty) = \frac{F_{T_p,S_m,T_m}(49,43,\infty) - F_{T_p,S_m,T_m}(43,43,\infty)}{F_{S_m,T_m}(43,\infty) - F_{T_p,S_m,T_m}(43,43,\infty)}.$$

As shown in Figure 12 , `Mathematica` then gives the result

$$P(44 \le T_p \le 49 \mid T_p \ge 44, S_m \le 43, T_m = \infty) \approx 0.0322. \tag{17}$$

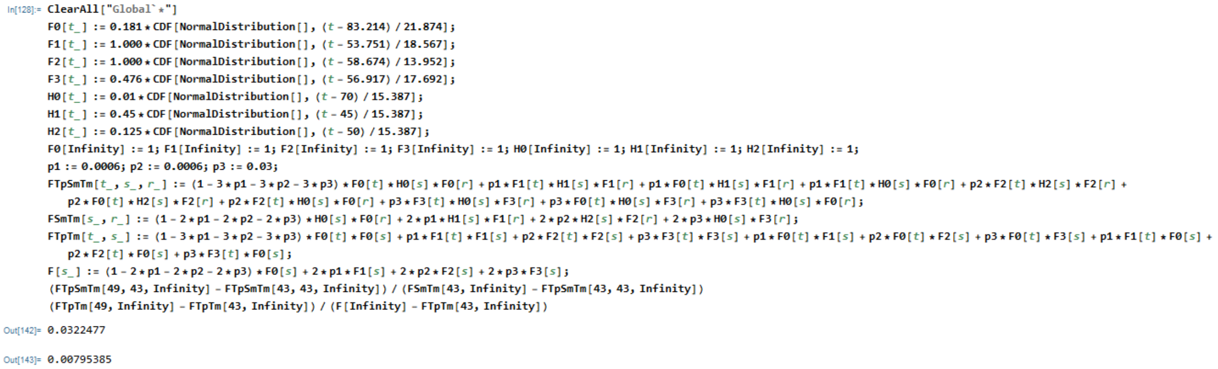
Let us compare this result to the probability that the proband gets breast cancer before the age of 49, without taking into account her mother's ovarian cancer. Defining the probability as

$$\begin{aligned} &P(44 \leq T_p \leq 49 \mid T_p \geq 44, T_m = \infty) \\ &= \frac{P(44 \leq T_p \leq 49, T_m = \infty)}{P(T_p \geq 44, T_m = \infty)} \\ &= \frac{P(T_p \leq 49, T_m = \infty) - P(T_p \leq 43, T_m = \infty)}{P(T_m = \infty) - P(T_p \leq 43, T_m = \infty)} \\ &= \frac{F_{T_p,T_m}(49, \infty) - F_{T_p,T_m}(43, \infty)}{F(\infty) - F_{T_p,T_m}(43, \infty)}, \end{aligned}$$

`Mathematica` returns

$$P(44 \leq T_p \leq 49 \mid T_p \geq 44, T_m = \infty) \approx 0.0080. \tag{18}$$

See Figure 12 for the corresponding code. Logically, it is correct that taking the mother's ovarian cancer into account increases the probability of the proband developing breast cancer.



```
In[128]:= ClearAll["Global`*"]
          F0[t_] := 0.181 * CDF[NormalDistribution[], (t - 83.214) / 21.874];
          F1[t_] := 1.000 * CDF[NormalDistribution[], (t - 53.751) / 18.567];
          F2[t_] := 1.000 * CDF[NormalDistribution[], (t - 58.674) / 13.952];
          F3[t_] := 0.476 * CDF[NormalDistribution[], (t - 56.917) / 17.692];
          H0[t_] := 0.01 * CDF[NormalDistribution[], (t - 70) / 15.387];
          H1[t_] := 0.45 * CDF[NormalDistribution[], (t - 45) / 15.387];
          H2[t_] := 0.125 * CDF[NormalDistribution[], (t - 50) / 15.387];
          F0[Infinity] := 1; F1[Infinity] := 1; F2[Infinity] := 1; F3[Infinity] := 1; H0[Infinity] := 1; H1[Infinity] := 1; H2[Infinity] := 1;
          p1 := 0.0006; p2 := 0.0006; p3 := 0.03;
          FTpSmTm[t_, s_, r_] := (1 - 3 * p1 - 3 * p2 - 3 * p3) * F0[t] * H0[s] * F0[r] + p1 * F1[t] * H1[s] * F1[r] + p1 * F0[t] * H1[s] * F1[r] + p1 * F1[t] * H0[s] * F0[r] + p2 * F2[t] * H2[s] * F2[r] +
             p2 * F0[t] * H2[s] * F2[r] + p2 * F2[t] * H0[s] * F0[r] + p3 * F3[t] * H0[s] * F3[r] + p3 * F0[t] * H0[s] * F3[r] + p3 * F3[t] * H0[s] * F0[r];
          FSmTm[s_, r_] := (1 - 2 * p1 - 2 * p2 - 2 * p3) * H0[s] * F0[r] + 2 * p1 * H1[s] * F1[r] + 2 * p2 * H2[s] * F2[r] + 2 * p3 * H0[s] * F3[r];
          FTpTm[t_, s_] := (1 - 3 * p1 - 3 * p2 - 3 * p3) * F0[t] * F0[s] + p1 * F1[t] * F1[s] + p2 * F2[t] * F2[s] + p3 * F3[t] * F3[s] + p1 * F0[t] * F1[s] + p2 * F0[t] * F2[s] + p3 * F0[t] * F3[s] + p1 * F1[t] * F0[s] +
             p2 * F2[t] * F0[s] + p3 * F3[t] * F0[s];
          F[s_] := (1 - 2 * p1 - 2 * p2 - 2 * p3) * F0[s] + 2 * p1 * F1[s] + 2 * p2 * F2[s] + 2 * p3 * F3[s];
          (FTpSmTm[49, 43, Infinity] - FTpSmTm[43, 43, Infinity]) / (FSmTm[43, Infinity] - FTpSmTm[43, 43, Infinity])
          (FTpTm[49, Infinity] - FTpTm[43, Infinity]) / (F[Infinity] - FTpTm[43, Infinity])

Out[142]= 0.0322477

Out[143]= 0.00795385
```

Figure 12: The code yielding Results (17) and (18).

8 Extension with practical applications

The model from Sections 5-7 can be extended even further: the amount of input can be enlarged by taking into account the information on ovarian and breast cancer of more family members. We will do so by considering two distinct family trees. The use of more information quickly gives complexity to the model, requiring the use of a computer program to perform the calculations.

Family trees 1 and 2 are displayed in Figures 13 and 14. In these images, squares represent men and circles stand for women. A circle whose left half is black indicates a woman who has (had) breast cancer at the age depicted above the black half. Analogously, a circle with a black right half implies ovarian cancer. The arrow in the trees marks the proband. The following table shows the probability that the proband gets breast cancer before the age of 80, computed for each family tree by using all the different methods studied. These probabilities were calculated by a computer program.

	Claus model	BRCA123 model without ov. cancer	BRCA123 model with ov. cancer
Family tree 1	0.2699	0.2434	0.4767
Family tree 2	0.2774	0.2374	0.2217

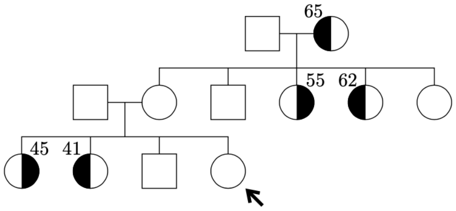


Figure 13: Family tree 1.

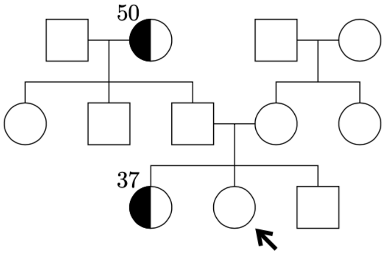


Figure 14: Family tree 2.

First, let us look at family number 1. A notable difference between the estimates is that in the BRCA123 model without ovarian cancer the estimate is significantly higher than in the BRCA123 model with ovarian cancer. Observing the family tree offers an explanation: ovarian cancer has occurred in the family (two times) and including it in the model therefore increases the risk of breast cancer.

In family 1, the proband's aunt who had breast cancer at age 62 is a carrier of the BRCA1 mutation. Because it is assumed that an individual can at most have one mutation and that mothers cannot have different mutations than their daughters, this means that the women in the family tree either have a mutation in the BRCA1 gene or no mutations. In addition, a woman is more probable to develop breast cancer if she possesses a mutation than when she has no mutations. It follows that the women in the family tree that had breast or ovarian cancer probably also have the mutation in the BRCA1 gene.

Now consider family number 2. The table shows that the estimated probability of breast cancer according to the BRCA123 model with ovarian cancer is lower than the same probability estimated using the BRCA123 model without ovarian cancer. Again, the family tree offers an explanation: no one in the proband's family has had ovarian cancer, so including ovarian cancer in the model decreases the probability of the proband getting breast cancer. As a matter of fact, the reasoning is similar to the reasoning when we compared Results (17) and (18), where including the information that the mother had had ovarian cancer increased the probability of developing breast cancer for the proband. Now it is the other way around: including the information that no one in the family has had ovarian cancer decreases the probability of getting breast cancer for the proband.

9 Conclusion

To conclude this project, we briefly summarise its main findings. In Section 2, we proved that a population in Hardy-Weinberg equilibrium remains in equilibrium, a statement used thoroughly throughout the rest of the project. At the end of Section 4, the first numerical results were obtained: according to the Claus model, a 50-year old woman without breast cancer has a 0.83% chance of getting it in the next 5 years and a woman whose mother got breast cancer at age 35 can be screened from the age of 35. Section 6 was dedicated to fitting functions of the form $\alpha\Phi\left(\frac{t-\mu}{\sigma}\right)$ to data, resulting in formulas for the distribution functions of the newer model presented in Section 5. In Section 7, we calculated that for a 44-year old woman without breast or ovarian cancer, but with a mother who has had ovarian cancer before the age of 43 and has never had breast cancer, including the mother's ovarian cancer yields a higher risk of breast cancer.

References

[1] KlinischeGeneticaEN.pdf

[2] *Source Figure 1:* Genetics: Analysis and Principles 7$^{\text{th}}$ Edition By Robert Brooker

[3] *Source Figure 2:* National Human Genome Research Institute

[4] *Source Figure 3:* Integraal Kankercentrum Nederland