Modeling in Medical Decision Making: A Bayesian Approach

First Edition

G. PARMIGIANI
Johns Hopkins University, Baltimore, MD

JOHN WILEY & SONS
Chichester · New York · Brisbane · Toronto · Singapore

This displays the posterior probabilities π^+ or π^- expressions (1.1) and (1.2) as a function of the prior probability, for a fixed Bayes factor. A Bayes factor of 1, which corresponds to an uninformative test, will leave the prior unchanged. This case corresponds to the diagonal line in Figure 2. If a test is informative, it will be characterized by two lines, one above the diagonal, corresponding to a positive result, and one below the diagonal, corresponding to a negative result.

For example, in Figure 2, the dashed lines correspond to the liver scan application. The dashed line above the diagonal reflects evidence in favor of disease (positive test) and leads to post-test probabilities that are higher than the pre-test. The dashed line below the diagonal reflects evidence against disease (negative test) and leads to post-test probabilities that are lower than the pre-test. Using this graph we can derive the clinical significance of positive and negative liver scans in other populations. For example, Drum and Christacapoulos, 1972 indicate that about 63% of the patient referred to a liver scan showed evidence of disease by either clinical or pathological findings. By setting the prior on the horizontal scale to .63 we can adjust the positive predictive value accordingly. Also, if we know the specific cause that lead to the indication of a liver scan (search for liver metastases vs search for parenchimal liver disease etc), we may be able to use prior probabilities that are specific to such subgroups.

Figure 2 also illustrates the effect of test accuracy on predictive values. The better the test, the further away the predictive value lines will be from the diagonal. For example, increasing the Bayes factors to 99 (dotted line) and 1/99 will push the graphs further away from the diagonal. Consider the case of a prior probability of .75, as in the liver scan application. All possible posterior probabilities are represented by the vertical line in Figure 2. The liver scan results will revise the prior to either $\pi^+ = .878$ or $\pi^- = .372$, while the hypothetical test represented by the dotted line would revise the prior more pronouncedly.

1.3 Genetic Counseling

Our next example of probabilistic inference arises in counseling individuals who are concerned about the possibility of being genetically predisposed to inheritable diseases, such as breast and colon cancer, or Alzheimer's disease. After some discussion of medical background, we examine in detail a model for genetic counseling of women with a family history of breast cancer, and use it to illustrate how one can extend the simple diagnostic paradigm described in the previous section to situations in which we have several unknowns and several related empirical observations that provide evidence about the diagnosis to be made. Our discussion is based on work by Berry, myself and others on the BRCAPRO model (Berry et al., 1997; Berry and Parmigiani, 1997; Parmigiani et al., 1998).

1.3.1 Genetic Susceptibility to Breast and Ovarian Cancer

Both breast and ovarian cancer cluster in families. In the 1980's researchers identified the hereditary nature of some breast and ovarian cancer clusters (Lynch et al., 1984), and suggested the presence of autosomal dominant susceptibility genes (Claus et al., 1990). In the 1990's our understanding of inherited susceptibility to breast cancer made further progress, with the identifications of two of these genes: BRCA1 and BRCA2. Approximately 2% of breast cancers and 10% of ovarian cancers are believed to occur in women who carry an inherited deleterious mutation of the BRCA1 gene, on chromosome 17q (Miki et al., 1994; Futreal et al., 1994). A smaller fraction of breast cancer is attributable to inherited deleterious mutations of BRCA2, on chromosome 13q (Wooster et al., 1995). Mutations are rare, occurring in less than .2% of women (Ford and Easton, 1995). But accruing evidence is confirming that women with mutations are very likely to develop one or both of breast and ovarian cancers, and to develop them at relatively young ages (Easton et al., 1995; Narod et al., 1995; Struewing et al., 1997).

The isolation of the BRCA1 and BRCA2 genes allows direct testing of individuals for the presence of a mutation (Weber, 1996). For women who have a family history of cancer but are not affected themselves, genetic testing may provide important information about cancer risk, and the risk of transmitting a high susceptibility to their offspring (Warmuth et al., 1997). For women with cancer, testing may also have implications for their future health. For example, women with breast cancer may be concerned about the risk of contralateral breast cancer and/or ovarian cancer.

Whether or not to be tested for BRCA1 and BRCA2 mutations is a complex decision. It depends on the woman's chance of carrying a mutation, as well as the effectiveness and cost of the testing procedure, the available prophylactic interventions, the effectiveness and negative effects of these interventions, the impact of testing on other family members, the impact on the women's ability to obtain insurance coverage (Meissen et al., 1991; Schwartz et al., 1995) and employment (Billings et al., 1992). A positive test or simply the perception of a high risk can lead to aggressive management, ranging from more frequent mammographies to bilateral mastectomy, again with substantive consequences on a woman's life. In this scenario, there is great interest in providing useful information that can help women decide if they indeed want to undergo testing.

Recognizing these problems, the National Action Plan on Breast Cancer Working Group on Hereditary Susceptibility and the American Society of Clinical Oncology have developed informative material for the education of physicians on issues of genetic testing. They stress that

> the decision whether to undergo predisposition genetic testing hinges on the adequacy of the information provided to the individual in regards to the risks and benefits of testing, possible testing outcomes,

> sensitivity and specificity of the test being performed, cancer risk management strategies, and the right of choice. Individuals should be provided with the necessary information to make an informed decision in a manner that they can understand (National Action Plan on Breast Cancer and American Society of Clinical Oncology, 1997).

A critical step in counseling a women facing these decisions is an accurate evaluation of the probability that she carries a mutation.

1.3.2 Individualized Probabilities

While there is evidence that other inheritable factors may affect breast cancer, BRCA1 and BRCA2 are associated with a large fraction of the cases that are attributable to genetic susceptibility (Newman et al., 1997; Weber, 1998). This means that despite the mutations' rarity, they are likely to be present in families that have multiple occurrences of breast or ovarian cancer. Family history of these diseases is a strong indicator of whether a mutation is present in the family, and in a particular family member.

In fact, the chance of carrying a genetic mutation varies markedly from woman to woman, depending on family history of breast and related cancers. Many aspects of a woman's family history are important in determining her chance of being a carrier, such as the exact relationships of all family members, including both affected and unaffected members, the ages at diagnosis of the affected members, and the current ages of the unaffected members. Particular relationships of family members with cancer and also without cancer can have a substantial impact on the probability of carrying a susceptibility gene. Ages at diagnosis of affected family members and their types of cancer are also important. An affected woman with several cancers in her family can have a probability of carrying a mutation that ranges from less than 5% to close to 100% (Berry et al., 1997). A woman with two primary cancers can have a probability of carrying a mutation in excess of 80%, even with no other information about family history.

This complexity makes it challenging to convey accurate carrier probability information. In response to this challenge, several quantitative models for determining carrier probabilities have been proposed. Model-based prediction is emerging as an efficient, potentially effective process that can enable the delivery of accurate individualized information about genetic testing to a larger audience. Quantitative models are currently being used in a range of counseling and clinical activities. Materials distributed to women that are considering genetic testing often include model-based predictions (Myriad Genetics, Inc., 1996; Bluman et al., 1999).

In addition to clinical and counseling applications, model-based approaches are being used in a variety of scientific investigations (Blackwood and Weber,

1998). Their roles include: to help in the study of characteristics of BRCA1 and BRCA2 carriers in population in which genetic testing has not been carried out, or has only partially been carried out (Schildkraut et al., 1997; Iversen et al., 1997; Smith et al., 1996; Hartmann et al., 1999); to guide the selection of high risk families for the study of risk, and for the assessment of prevention strategies; and the study of the quality of care and counseling of high risk women (Iglehart et al., 1998).

Broadly speaking, two general modeling approaches have been used so far; here we will label them "empirical" and "Mendelian". Empirical approaches build models using statistical modeling techniques directly on pedigree data for tested individuals: the test results constitute the response variable(s); pedigree features provide the predictor variables. Candidate features are often extracted from pedigrees based on clinical and epidemiological expertise. Highly predictive features are then identified using statistical variable selection techniques. Examples include Shattuck-Eidens et al., 1997; Couch et al., 1997; Frank et al., 1998 and Hartge et al., 1999.

By contrast, Mendelian models use background knowledge about the autosomal dominant inheritance of the genes to formulate explicit genetic models for predicting carrier status. This is done in two steps: estimation of "genetic parameters" comprising the mutations' population prevalences and penetrance functions; and application of Bayesian prediction to transform the genetic parameters into carrier probabilities (Murphy and Mutalik, 1969; Elston and J, 1971; Szolovits and Pauker, 1992; Offit and Brown, 1994). Penetrances and prevalences may be estimated directly or abstracted from published work.

Here we discuss in detail how to build a basic Mendelian model considering just a single gene, to ease the task of illustrating the fundamental concepts. Our discussion is based on a simplified version of BRCAPRO (Berry et al., 1997; Parmigiani et al., 1998), which is a Mendelian model and software for finding the probability that an individual carries a germline mutation at BRCA1 or BRCA2, based on his or her family's history of breast and ovarian cancer. In the current implementation of the model, the family history includes the counseland and her first- and second-degree relatives. For each member—including the counseland—the model processes information about whether the member has been diagnosed with breast cancer and, if so, age at diagnosis or, if cancer free, current age or age at death. Similar data is processed for ovarian cancer if the family member is female. Here, we focus on simple family structures.

Our discussion comprises estimates of genetic parameters, discussed in Section 1.3.3; structural assumptions, such as inheritance mechanism and conditional independence, presented in Section 1.3.4; computing principles and algorithms, also presented in Section 1.3.4; and sequential use of Bayes' rule to incorporate both family history and the results of genetic tests in counseling, discussed in Section 1.3.7.

1.3.3 Genetic Parameters

For simplicity, we consider a single hypothetical gene, which we call BRCA. We assume that there are only two types (or alleles) of BRCA genes: the normal type and the type conferring a genetic susceptibility, also called deleterious. As is the case with the real breast cancer susceptibility genes, we assume that BRCA is an autosomal gene. This means that each individual carries two copies of the BRCA gene, each of which is inherited from one from one of the parents, by drawing at random from each parent's two copies. So each individual could carry 0, 1, or 2 deleterious alleles. Here we assume that carrying 1 or 2 deleterious alleles confers the same risk. This assumption is not yet validated by empirical evidence because the BRCA genes are too rare to extensively study individuals with 2 deleterious copies. In view of this rarity this assumption is also not likely to influence the results.

We will use the notation g to indicate the BRCA genotype, and g_i to indicate the BRCA genotype of individual i. We will restrict attention to two possibilities: $g_i = 1$ if individual i has 1 or 2 deleterious alleles (carrier), and $g_i = 0$ if individual i has 2 normal copies (noncarrier).

The fraction of women carriers of BRCA genetic susceptibility mutations who develop breast and ovarian cancer by age is often called penetrance. We use the letter x for age of onset of breast cancer and y for the age of onset of ovarian cancer. So the penetrance curves are defined as

$$\begin{aligned} B_0(x) &= \Pr\{\text{Breast cancer by age } x \mid g = 0\} \\ B_1(x) &= \Pr\{\text{Breast cancer by age } x \mid g = 1\} \\ C_0(y) &= \Pr\{\text{Ovarian cancer by age } y \mid g = 0\} \\ C_1(y) &= \Pr\{\text{Ovarian cancer by age } y \mid g = 1\} \end{aligned}$$

Similarly, considering yearly increments of the curves above:

$$\begin{aligned} b_0(x) &= \Pr\{\text{Breast cancer at age } x \mid g = 0\} \\ b_1(x) &= \Pr\{\text{Breast cancer at age } x \mid g = 1\} \\ c_0(y) &= \Pr\{\text{Ovarian cancer at age } y \mid g = 0\} \\ c_1(y) &= \Pr\{\text{Ovarian cancer at age } y \mid g = 1\} \end{aligned}$$

Here, we take the penetrance functions of the hypothetical BRCA gene to be a weighted average of the penetrance functions for BRCA1 and BRCA2 in the current implementation of BRCAPRO. Those are derived by combining the estimates reported by Ford et al., 1998 and Struewing et al., 1997. Details are discussed by Iversen et al., 2000 The resulting curves are shown in Figure 3. For example, B_1 is the dashed curve in the top left panel, while b_1 is the dashed curve in the bottom left panel. These functions will help us constructing a likelihood function (that is the probability of the empirical evidence given the unknown genotype) for making inference on the genotype.

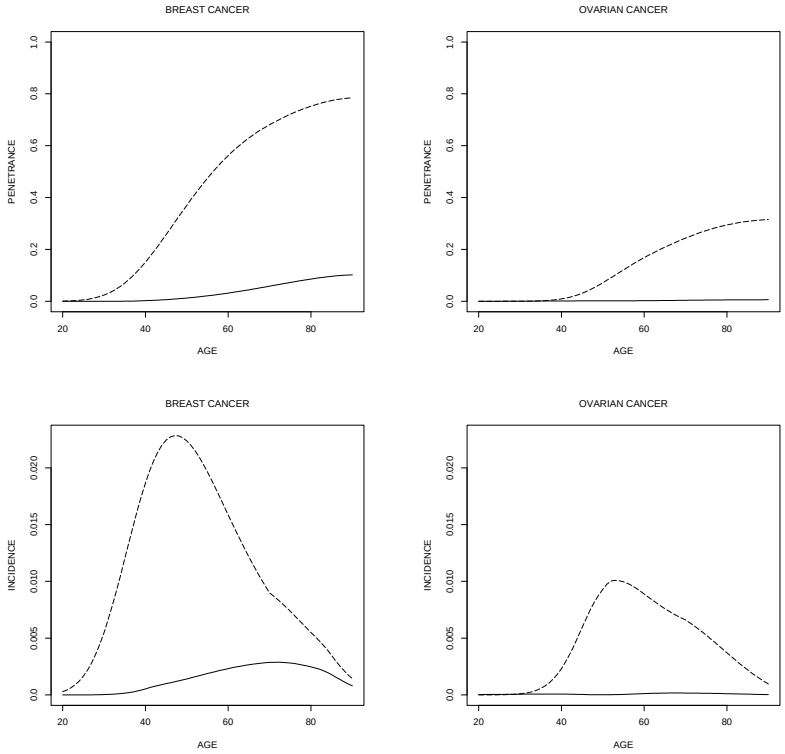

Figure 3 Hypothetical penetrance curves for the BRCA gene. The top row portrays the fraction of women developing breast and ovarian cancer by age, for carriers of deleterious BRCA mutations (dashed) and noncarriers (solid). The bottom rows expresses the same information in terms of the number of cases per year of age. The curves at the bottom are obtained by taking yearly increments of the curves at the top.

The other component that is needed for Bayes' rule is the prevalence of carriers of deleterious alleles in the population, $\pi = \Pr\{g = 1\}$. In our example we hypothesize that the women being counseled are Ashkenazi Jews, an ethnic group characterized by an especially high prevalence of deleterious alleles. We derive estimates of mutation prevalence among the Askenazim from Oddoux et al., 1996 and Roa et al., 1996, both of which are population–based studies. The combined allele frequency for deleterious mutations of BRCA1 and BRCA2 is $\phi = .01289$. The probability of carrying at least one copy of a deleterious allele is then $\pi = \phi^2 + 2\phi(1-\phi) = .02562$. The resulting prior odds of being a carrier are $\pi/(1-\pi) = .0263$.

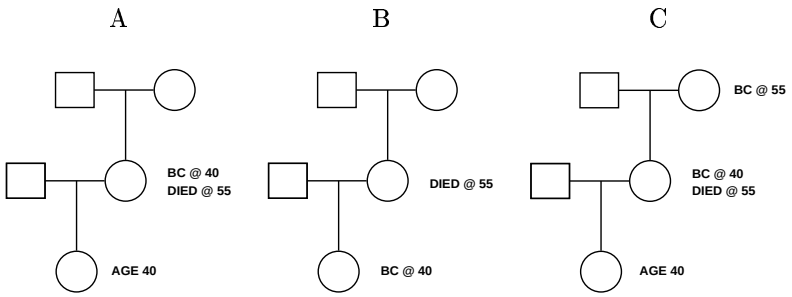

Figure 4 As customary in pedigree graphs, circles indicate females and squares indicate males. Offspring are below their ancestors. Horizontal links connect mates and vertical links represent parent-child relations. The graph also shows the complete cancer history and age information of female members of the family. Breast cancer cases are indicated by BC and ovarian cancer cases by OC.

1.3.4 Bayes' rule in genetic counseling

Our first example is that of a woman, say Anne, seeking counseling because of her mother's history of breast cancer. The family tree is Pedigree A of Figure 4. Anne is the individual at the bottom of the pedigree. We will use $i = 1$ for Anne and $i = 2$ for her mother. Our goal is to compute the probability that Anne carries a deleterious BRCA allele. We will build our carrier probability calculation from simple to more complex. Initially we ignore ovarian cancer, and suppose that deleterious BRCA alleles only affect breast cancer risk.

To set the stage, we begin by deriving the probability that Anne is a BRCA carrier based on her own history only. We have two facts to consider: 1) Anne is Ashkenazi, and we know the prevalence of deleterious BRCA alleles among Ashkenazim; 2) Anne lived to be 40 without a diagnosis of breast cancer —in our notation $x_1 > 40$. The quantity of interest is

$$\Pr\{ \text{ Anne has deleterious BRCA allele } | \text{ Anne has no breast cancer at 40 } \}$$

or, in the more formal notation,

$$\pi(g_1 = 1 \mid x_1 > 40).$$

To compute this probability we can use the information about ethnicity to specify a prior probability $\pi(g_1)$, and the information about family history to construct a Bayes factor. This situation mirrors that of Section 1.2: the presence of a deleterious allele is playing the same role as the presence of liver disease, and the personal history is playing the same role as the outcome of the liver scan. So, by analogy with expression (1.4), the posterior odds of carrying a deleterious allele are

$$\frac{\pi(g_1 = 1 \mid x_1 > 40)}{\pi(g_1 = 0 \mid x_1 > 40)} = \frac{\pi}{1-\pi} \; \frac{p(x_1 > 40 \mid g_1 = 1)}{p(x_1 > 40 \mid g_1 = 0)}, \tag{1.7}$$

where the right hand side of the equation is the product of the prior odds and the Bayes factor. Both the numerator and denominator of the Bayes factor can be determined directly from the information in Figure 3. Specifically, the top left panel provides the probabilities $B_1(40)$ and $B_0(40)$ of developing breast cancer by age 40 for carriers and noncarriers respectively. So the Bayes factor is

$$\frac{p(x_1 > 40 \mid g_1 = 1)}{p(x_1 > 40 \mid g_1 = 0)} = \frac{1 - B_1(40)}{1 - B_0(40)} = \frac{1 - .1509}{1 - .002318} \approx .85.$$

As expected, this provides weak, but not completely negligible, evidence against the hypothesis that the woman is a carrier. The prior odds decrease from .026 to .022 as a result of knowing that Anne was not diagnosed before age 40.

If instead Anne had been diagnosed at age 40 with breast cancer, we would need to consider the probabilities of diagnosis $b_1(40)$ and $b_0(40)$ for carriers and noncarriers, displayed in the bottom left panel of Figure 3. The Bayes factor would be

$$\frac{p(x_1 = 40 \mid g_1 = 1)}{p(x_1 = 40 \mid g_1 = 0)} = \frac{b_1(40)}{b_0(40)} \approx 34.96,$$

leading to close to even odds that the woman is a carrier.

We now consider how to incorporate the information that Anne's mother was diagnosed with breast cancer at age 40, that is $x_2 = 40$. The quantity of interest is

$$\Pr\{ \text{ woman has deleterious BRCA allele} \mid \text{family history } \}$$

or, in the more formal notation,

$$\pi(g_1 = 1 \mid x_1 > 40, x_2 = 40).$$

Similarly to (1.7), the posterior odds of carrying a deleterious allele are

$$\frac{\pi(g_1 = 1 \mid x_1 > 40, x_2 = 40)}{\pi(g_1 = 0 \mid x_1 > 40, x_2 = 40)} = \frac{\pi}{1 - \pi} \; \frac{p(x_1 > 40, x_2 = 40 \mid g_1 = 1)}{p(x_1 > 40, x_2 = 40 \mid g_1 = 0)}. \quad (1.8)$$

Here the Bayes factor incorporates evidence from the whole family history, and is no longer so simple to evaluate: it involves two events rather than one, and the events are likely to be connected. Before we can make progress we need to make important additional modeling assumption. This in turn requires a digression about the statistical concepts of independence and conditional independence.

1.3.5 Conditional Independence

Two events are called statistically independent if knowledge of one does not change the probability of the other. For example, in the situation of

Section 1.2, the outcome of a liver test may not depend on the presence of, say, hypertension. If the probability of a positive liver test in the overall population is the same as the probability of a positive liver test in the subpopulation of patients with hypertension, then knowing about a patient's hypertension does not affect the the probability of a positive liver test, and the two events are said to be independent.

If we use H for hypertension, and D for liver disease, independence translates into

$$P(D|H) = P(D).$$

Only apparently this is an asymmetric definition. Consider $P(H|D)$. By substituting $P(D, H)/P(H)$ for $P(D|H)$, we get that independence can also be rewritten as

$$P(D, H) = P(D)P(H) \tag{1.9}$$

or

$$P(H|D) = P(H),$$

illustrating that independence is symmetric.

It is not uncommon for events to be independent in subgroups of a population, but not in the population at large. For a simple fictional example, consider a population made of two ethnic groups. Imagine that within each of the groups hypertension and hair loss are independent. That means that within an ethnic group, the fraction of hypertensive patients is the same among patients with and without hair loss. This illustrates one of the most important concepts in statistical modeling: that of conditional independence. Hypertension and hair loss are independent when we condition on ethnicity. If L is hair loss, and E ethnicity conditional independence is formally expressed as

$$P(H, L|E) = P(H|E)P(L|E)$$

which is simply expression (1.9) applied to a specific subgroup.

Now imagine that one of the ethnic groups has a much higher prevalence of both hypertension and hair loss. Then, in the population at large, the fraction of hypertensive patients is larger among patients with hair loss. The different ethnicities in the population and the different prevalences determine a dependence between hypertension and hair loss in the population at large, or more technically in the marginal distribution of the two features. This happens because hair loss teaches us about ethnicity and thus, indirectly, about hypertension.

1.3.6 Bayes factors for family history

Returning to the genetic counseling example, we are going to use a specific conditional independence assumption to simplify the calculation of the Bayes factor of family history in expression (1.8). Cancer histories of mothers and

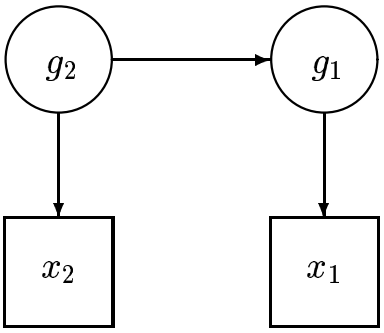

Figure 5 Graph of the conditional independence relationship among random variables in the genetic counseling problem, in pedigrees A and B of Figure 4. Quantities in rectangles are observed, while quantities in circles are unobserved.

daughters are clearly not independent in the population of all Ashkenazi families: we know that breast cancer tends to cluster in families. Is this clustering entirely due to inheritable susceptibility via the BRCA gene or are there other factors that determine clustering? One way of thinking about this question in terms of conditional independence is to look at the subpopulation of women with known deleterious alleles, and ask: is there a relationship between cancer in those women and cancer in their mothers? In what follows we assume that the answer is no. Formally, we assume that the cancer histories of family members are conditionally independent given the genotype. Anne's mother's history of cancer informs us about Anne's risk only by informing us about her own genotype, which in turn affects Anne's genotype, and thus Anne's cancer risk.

This assumption of conditional independence of phenotypes (cancer histories) given genotype at a single locus is unlikely to be literally true. There may be other genes, or shared environmental factors, such as exposure of an entire family to radiation or carcinogenic pesticides, that determine familial clustering of cancer cases. However, if the BRCA gene is responsible for the majority of such clustering, the assumption of conditional independence may be reasonable. As in all modeling exercises, no model is literally representing reality: some simplification is necessary to make progress, and some is useful in revealing important structure. How much simplification is too much is to be judged based on the accuracy of the model's results, and the costs of making the model more complex.

Our conditional independence relationships are illustrated in Figure 5. Circles represent the random variables in question. Arrows represent dependences among the variables. Arrows' directions represent natural ways

for specifying conditional probabilities. For example it is simple to specify conditional probabilities of the genotype of the daughter given the mother's, and to specify probability of cancer-related events given the genotype. Arrows' directions could also be attributed a causal connotation (which would be plausible in this specific example), but that does not have to be the case. Whichever the direction, arcs can be thought of as conduits of information. Our conditional independence assumption is reflected in the lack of a direct link between x_1 and x_2.

We are now ready to formalize the conditional independence conditions as:

$$p(x_1, x_2|g_1, g_2) = p(x_1|g_1, g_2)p(x_2|g_1, g_2) = p(x_1|g_1)p(x_2|g_2). \tag{1.10}$$

The first equality expresses the conditional independence of the individuals' phenotypes given the whole set of genotypes in the pedigree. It is analogous to the hypertension / hair loss example, with the two cancer histories replacing hypertension and hair loss, and the pair of genotypes replacing the ethnic group. The second equality makes reflect the additional assumption that an individual's phenotype depends on the relative's genotypes only via his or her own, as illustrated in Figure 5.

Using this factorization we can make progress towards evaluating the Bayes factor for family history. The two terms of the right hand side of (1.10) can be evaluated directly by using the information of Figure 3, as we did in the simpler case of Section 1.3.4. For example, for pedigree A, we have

$$\begin{aligned} p(x_1 > 40, x_2 = 40 \mid g_1 = 1, g_2 = 1) = \\ p(x_1 > 40 \mid g_1 = 1)p(x_2 = 40 \mid g_2 = 1) = \\ [1 - B_1(40)]b_1(40) = .849 \times .0187 = .0159 \end{aligned} \tag{1.11}$$

and

$$\begin{aligned} p(x_1 > 40, x_2 = 40 \mid g_1 = 1, g_2 = 0) = \\ p(x_1 > 40 \mid g_1 = 1)p(x_2 = 40 \mid g_2 = 0) = \\ [1 - B_1(40)]b_0(40) = .849 \times .000534 = .000453, \end{aligned} \tag{1.12}$$

and so forth.

Our last obstacle is that we cannot work with $p(x_1, x_2|g_1, g_2)$ directly because in computing the numerator and denominator of the Bayes factor we cannot condition on the mother's genotype. We need instead $p(x_1, x_2|g_1)$. In other words we have here two unknowns, but only one is the focus of investigation. What comes handy here is an argument similar to the one we have used to derive the denominator of Bayes rule in Section 1.2. The idea is to derive $p(x_1, x_2|g_1)$ by a weighted average of $p(x_1, x_2|g_1, g_2)$'s. Consider the

numerator of the Bayes factor first. Formally

$$p(x_1, x_2|g_1) = \\ p(x_1, x_2, g_2 = 0|g_1) + p(x_1, x_2, g_2 = 1|g_1) = \\ p(x_1, x_2|g_2 = 0, g_1)p(g_2 = 0|g_1) + p(x_1, x_2|g_2 = 1, g_1)p(g_2 = 1|g_1) \quad (1.13)$$

where the first equality is a marginalization over the g_2 dimension, and the second is an application of the definition of conditional probability.

This strategy is effective so long as we can determine the probabilities of the genotype of Anne's mother given Anne's own. Assuming Mendelian inheritance, we can determine these exactly. As we are studying the numerator of the Bayes factor, Let's assume that Anne is a carrier. What is the probability that her mother is as well? If we ignore the possibility that individuals in the pedigree carry two deleterious BRCA mutations, Anne can only have one deleterious allele, and this has the same probability of coming from her father or from her mother. So the probability that Anne's mother is a carrier is approximately one half. On the other hand, if Anne is not a carrier (as we assume when evaluating the denominator of the Bayes factor), then her mother is very unlikely to be, but could be and have passed a normal copy of the gene. The probability of that happening is the probability of her "other" allele being deleterious, that is the allele frequency ϕ of deleterious mutations.

In reality the possibility of carrying more than one deleterious allele is not so remote among Ashkenazi families, especially when there are cancer cases on both sides of the family, but the bookkeeping of the Mendelian probabilities would lead us too far astray. BRCAPRO incorporates the correct probabilities, and you can see Parmigiani et al., 1998 or Weiss, 1993 to see how that works.

In any event, using (1.13) and expressions (1.11) and (1.12) we get that the numerator of the Bayes factor is approximately

$$p(x_1 > 40, x_2 = 40 \mid g_1 = 1) \quad = \quad [1 - B_1(40)] \left\{ \frac{1}{2} b_1(40) + \frac{1}{2} b_0(40) \right\}$$

while the denominator is approximately

$$p(x_1 > 40, x_2 = 40 \mid g_1 = 0) \quad = \quad [1 - B_0(40)]\{\phi b_1(40) + (1 - \phi) b_0(40)\}.$$

The differences between the numerator and the denominator are in the conditioning population for the daughter (as before) and in the weights used to mix the two possible conditioning populations for the mother.

This exemplifies a very general feature of probabilistic inference, that is the possibility of specifying models in terms of as many unknowns as it is natural from a medical or scientific standpoint, and then getting rid of the unknowns that are not the focus of investigation by marginalization. The same approach could have been used to make inferences about the genotype of Anne's mother.

	Breast Cancer Only			Breast and Ovarian Cancer	
	A	B	C	A	B
Posterior Probability	0.219	0.460	0.372	0.203	0.415
Posterior Odds	0.280	0.852	0.591	0.255	0.710
Bayes Factor	10.64	32.414	22.477	9.678	27.014

Table 4 Posterior probabilities, posterior odds and Bayes factors for the pedigrees of Figure 4.

Evaluating the expressions above leads to the results of the first column of Table 4. The Bayes factor is about 10–fold. Compare this to the value of about 35–fold that we had obtained in Section 1.3.4 assuming that Anne herself had been diagnosed at 40. Here the impact of Anne's mother's diagnosis on her own genotype is the same as that of Anne's diagnosis on her own genotype in Section 1.3.4. But the uncertainty about whether Anne's mother has passed the gene to Anne, and Anne's health, all contribute to moderating the ultimate effect of this on the carrier probability.

Table 4 also considers pedigree A under a more elaborate model that incorporates ovarian cancer. Because Anne has not been diagnosed with ovarian cancer, and her mother lived to be 55 without a diagnosis of ovarian cancer, the relevant Bayes factor is

$$\frac{p(x_1 > 40, y_1 > 40, x_2 = 40, y_2 > 55 \mid g_1 = 1)}{p(x_1 > 40, y_1 > 40, x_2 = 40, y_2 > 55 \mid g_1 = 0)} = \frac{[1 - B_1(40)][1 - C_1(40)]\left\{\frac{1}{2}b_1(40)[1 - C_1(55)] + \frac{1}{2}b_0(40)[1 - C_0(55)]\right\}}{[1 - B_0(40)][1 - C_0(40)]\left\{\phi b_1(40)[1 - C_1(55)] + (1 - \phi)b_0(40)[1 - C_0(55)]\right\}}.$$

From Table 4, we see that the Bayes factor is slightly reduced as a result of the lack of ovarian cancer diagnoses. The expression above relies on one additional conditional independence assumption, and that is that conditional on genotype, the time to diagnosis of breast and ovarian cancer are independent. This assumption is supported in the literature (Ford et al., 1998), although further investigation of the function of the BRCA genes may reveal violations.

Table 4 lets us contrast pedigree A with pedigree B, in which the diagnoses are reversed. As expected, the impact on Anne's carrier probability of a diagnosis at age 40 is much greater (the Bayes factor is over 32–fold) than the impact of the same diagnosis in her mother. However, the value of 32–fold is less than the value of about 35–fold that we had obtained in Section 1.3.4 without incorporating information about the mother, who is healthy until age 55 in pedigree B. Again incorporating the information that neither of the two is diagnosed with ovarian cancer reduces the Bayes factor.

Our last calculation is for pedigree C, which is the same as pedigree A with the addition of a breast cancer diagnosis at 55 in Anne's maternal

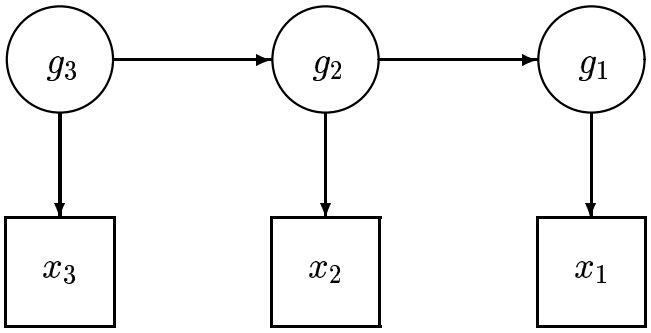

Figure 6 Graph of the conditional independence relationship among random variables in the genetic counseling problem, in pedigree C of Figure 4.

grandmother. Anne's grandmother will be individual 3. The conditional independence assumptions are now those of Figure 6. From the figure it is clear that the model has a recursive structure reflecting inheritance. So the same arguments that were used to relate Anne's mother's phenotype to Anne's genotype could be rephrased to relate Anne's grandmother's phenotype to Anne's mother's genotype. Mathematically, we need to recursively expand equation (1.13), by incorporating g_3 in the $p(x_1, x_2, x_3|g_2, g_1)$ terms in the same way as we had incorporated g_2 in the $p(x_1, x_2|g_1)$ term. For example, we can write

$$p(x_1, x_2, x_3|g_2 = 1, g_1) = p(x_1, x_2, x_3|g_3 = 0, g_2 = 1, g_1)p(g_3 = 0|g_2 = 0) + p(x_1, x_2, x_3|g_3 = 1, g_2 = 1, g_1)p(g_3 = 1|g_2 = 0).$$

Without going through further details, we directly give the expressions for the numerator and denominator of the Bayes factor, which are respectively:

$$p(x_1 > 40, x_2 = 40, x_3 = 55 \mid g_1 = 1) = [1 - B_1(40)] \left\{ \frac{1}{2} b_1(40) \left\{ \frac{1}{2} b_1(55) + \frac{1}{2} b_0(55) \right\} + \frac{1}{2} b_0(40) \{\phi b_1(55) + (1 - \phi) b_0(55)\} \right\}$$

and

$$p(x_1 > 40, x_2 = 40, x_3 = 55 \mid g_1 = 0) =$$
$$[1 - B_0(40)] \left\{ \phi b_1(40) \left\{ \frac{1}{2} b_1(55) + \frac{1}{2} b_0(55) \right\} + = \right.$$
$$\left. (1 - \phi) b_0(40) \left\{ \phi b_1(55) + (1 - \phi) b_0(55) \right\} \right\}$$

Again, results are in Table 4. The grandmother's diagnosis lends substantial support to the presence of a genetic susceptibility in the family, and results in a Bayes factor which is over 22–fold, more than doubling that of pedigree A. Note however that the addition of the mother's diagnosis had multiplied the original Bayes factor (.85) by over 10. This difference is the result of both the closer proximity of the mother and the earlier age at diagnosis.

In summary, using Bayes rule, clinicians and counselors can integrate and interpret pedigrees, historical data, physical findings and laboratory data, providing individualized probabilities of various outcomes (Pauker and Pauker, 1987). The interpretation of these individualized probability is that they are approximately the fraction of Ashkenazi women with the specified family history who are positive. These probabilities need to be communicated as such, for example using phrases like "among women with a family history comparable to yours in all important details, $\pi \times 100\%$ of women carry a deleterious allele."

In current genetic counseling practice, a single risk estimate is often quoted to a family rather than a range of risks. Such point estimates are predicated on knowing basic parameters like prevalences and penetrances, when there may be considerable uncertainty about them (Lange, 1986; Leal and Ott, 1995; Struewing et al., 1997). This uncertainty is incorporated in the model via probabilistic sensitivity analysis, which will be discussed in Chapter 3.

1.3.7 Sequential use of Bayes' rule

One of the reasons why it is considered important to ascertain family history in genetic counseling is to assist in interpreting the results of genetic testing for individuals who elect to be tested. Genetic tests are usually accurate, but not perfect. For example, in testing for the BRCA genes there are several technologies available, most of which have a specificity near one, but can miss some deleterious mutations. Depending on the technology, sensitivity ranges from about .8 to about .98. In this context a negative result provides a Bayes factor of $(1 - \beta)/\alpha$ that ranges from 1/5–fold ($\beta = .8$) to 1/50–fold ($\beta = .98$). These orders of magnitude are similar to the reciprocals of those of the Bayes factors for family histories in Table 4, emphasizing the importance of family history information.

To fix ideas, consider a test with specificity $\alpha = 1$ and sensitivity $\beta = .9$. Suppose that Anne, the woman of pedigree A, elects to get tested and tests

negative. Let's denote this outcome by T–, as in the liver scan example. What is the probability that Ann is a carrier? Formally, we need to compute

$$\pi(g_1 = 1 \mid x_1 > 40, x_2 = 40, \text{T}-).$$

To proceed we assume that the test outcome is conditionally independent of the family history given the genotype. This requires that the true positive rate is the same no matter what the family history is, or in other words, that the reasons why a genetic test may miss an existing deleterious mutation are unrelated to the presence of cancer in the family —an assumption which seems plausible. Then the posterior odds of being a carrier can be factored as:

$$\frac{\pi(g_1 = 1 \mid x_1 > 40, x_2 = 40, \text{T}-)}{\pi(g_1 = 0 \mid x_1 > 40, x_2 = 40, \text{T}-)} = \frac{\pi(g_1 = 1)}{\pi(g_1 = 0)} \frac{p(x_1 > 40, x_2 = 40 \mid g_1 = 1)}{p(x_1 > 40, x_2 = 40 \mid g_1 = 0)} \frac{1-\beta}{\alpha} = .0263 \times 10.64 \times \frac{1}{10} = 0.028. \quad (1.14)$$

The first two terms in the product are the same as those of our analysis of pedigree A in Section 1.3.6. The family history of pedigree A provides evidence in favor of the mutation which is of the about same weight as the evidence against a mutation provided by a negative test.

Expression (1.14) illustrates that, formally, the product .0263 10.64 can be though of in two ways: as the posterior odds after we acquired the information about family history; and as the prior odds for incorporating the information about the negative test. The posterior probability $\pi(g_1 = 1 \mid x_1 > 40, x_2 = 40, \text{T}-)$ can therefore be obtained directly by (1.14) or by sequentially applying Bayes' rule, first computing the posterior probability based on family history, or $\pi(g_1 = 1 \mid x_1 > 40, x_2 = 40)$, and than using the result as the prior in expression (1.3). Expression (1.14) also illustrates that the order in which we acquire the evidence from the test and the family history is not influencing the final result, so we could reverse the order of the application of Bayes' rule.

This illustrates a general property of Bayes' rule. Suppose hypothesis H is to be revised in the light of new evidence E consisting of two empirical observations E_1 and E_2, say the family history and the genetic test of the previous example. Suppose also that the likelihood of the evidence E factors as

$$P(\text{E}_1, \text{E}_2|\text{H}) = P(\text{E}_1|\text{H})P(\text{E}_2|\text{H}).$$

this corresponds to the conditional independence of family history and the genetic test discussed above. In this context we can use Bayes' rule to

iteratively updating information by

$$P(\mathrm{H}|\mathrm{E}_1) = \frac{P(\mathrm{H})P(\mathrm{E}_1|\mathrm{H})}{P(\mathrm{H}{=}\mathrm{true})P(\mathrm{E}_1|\mathrm{H}{=}\mathrm{true}) + P(\mathrm{H}{=}\mathrm{false})P(\mathrm{E}_1|\mathrm{H}{=}\mathrm{false})}$$

$$P(\mathrm{H}|\mathrm{E}_1,\mathrm{E}_2) = \frac{P(\mathrm{H}|\mathrm{E}_1)P(\mathrm{E}_2|\mathrm{H})}{P(\mathrm{H}{=}\mathrm{true}|\mathrm{E}_1)P(\mathrm{E}_2|\mathrm{H}{=}\mathrm{true}) + P(\mathrm{H}{=}\mathrm{false}|\mathrm{E}_1)P(\mathrm{E}_2|\mathrm{H}{=}\mathrm{false})}.$$

here the output of the first application of Bayes rule is the input in the second application. This property is sometimes described by saying that "today's posterior is tomorrow's prior".

Conditional independence of empirical findings given the hypothesis of interest is not always met in medical applications. For example Wiener et al., 1979 show that the true positive rate of exercise electrocardiogram (finding E_2) in diagnosing coronary artery disease (H) varied across subgroups of patients classified according the type of chest pain (E_1). Subgroups at high risk because of the type of chest pain also had a higher true positive rate. Examples of the sequential uses of Bayes' rule to interpret the results of several clinical findings is also discussed by Gorry and Barnett, 1968 and Sox et al., 1988.

The discussion of pedigree A in Section 1.3.6 can also be viewed as an examples of multiple findings by thinking about $x_1 > 40$ as E_1 and $x_2 = 40$ as E_2. however, these could not be considered independent conditionally on the hypothesis $g_1 = 1$ only. In fact we argued that we needed to condition on all genotypes on the pedigree before independence could be more realistically assumed. In Section 1.3.6 we also saw that conditional independence of empirical findings is not necessary for the application of Bayes' rule: it is only necessary for sequential updating.

1.4 Estimating Sensitivity and Specificity

In this section we reconsider the liver scan example of Section 1.2. We drop the assumption that the sensitivity and specificity of the scan are known with certainty, and turn to the problem of estimating them based on data such as those of Table 1. The values of α and β will now be treated as unknown. So far we used Bayes' rule to revise probabilities of unknowns about patients in the light of new empirical observations, either about test results or family history. In this Section we will extend this approach to unknowns that refer to populations of patients, such as are α and β. The sample of Table 1 is to these unknowns what a family history is to the patient's genotype.

1.4.1 Population and Statistical Model

To keep matters simple, we will start by focusing on inference about sensitivity only. Defining the unknown in this case requires thinking about the population of all patients that are eligible for a liver scan and have liver disease, and asking the question: what fraction of these patients would have a positive scan. Such