8-26-2004

BayesMendel: An R Environment for Mendelian Risk Prediction

Sining Chen
The Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University

Wenyi Wang
Johns Hopkins Bloomberg School of Public Health, Department of Biostatistics

Karl Broman
Johns Hopkins Bloomberg School of Public Health, Department of Biostatistics

Hormuzd A. Katki
Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health & Division of Cancer Epidemiology and Genetics National Cancer Institute, NIH, DHHS

Giovanni Parmigiani
The Sydney Kimmel Comprehensive Cancer Center, Johns Hopkins University & Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, gp@jimmy.harvard.edu

BayesMendel: an R environment for Mendelian Risk Prediction

Sining Chen[1],
Wenyi Wang[2],
Karl W. Broman[2],
Hormuzd A. Katki[2,3],
Giovanni Parmigiani[1,2]

Abstract

Several important syndromes are caused by deleterious germline mutations of individual genes. In both clinical and research applications it is useful to evaluate the probability that an individual carries an inherited genetic variant of these genes, and to predict the risk of disease for that individual, using information on his/her family history. Mendelian risk prediction models accomplish these goals by integrating Mendelian principles and state-of-the art statistical models to describe phenotype/genotype relationships. Here we introduce an R library called BayesMendel that allows implementation of Mendelian models in research and counseling settings. BayesMendel is implemented in an object–oriented structure in the language R and distributed freely as an open source library. In its first release, it includes two major cancer syndromes: the breast-ovarian cancer syndrome and the hereditary non-polyposis colorectal cancer syndrome, along with up-to-date estimates of penetrance and prevalence for the corresponding genes. Input genetic parameters can be easily modified by users. BayesMendel can also serve as a generic tool for genetic epidemiologists to flexibly implement their own Mendelian models for novel syndromes and local subpopulations, without reprogramming complex statistical analyses and prediction tools.

[1]The Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University
[2]Department of Biostatistics, Johns Hopkins University
[3]Division of Cancer Epidemiology and Genetics National Cancer Institute, NIH, DHHS

1 Introduction

Genetic research has identified a number of genes for which inherited mutations confer a significantly increased risk of disease (Weiss 1993, Jimenez-Sanchez et al. 2001). These mutations give rise to specific syndromes and to clustering of disease phenotypes within families. Examples of major public health interest occur in cancer (Foulkes and Hodgson 1998). The Breast-Ovarian Cancer Syndrome, which accounts for 5% to 10% of breast cancer in the US (Weber 1998), is caused by germline mutations of the BRCA1 or BRCA2 genes. Hereditary Nonpoliposis Colorectal Cancer (HNPCC) (Vogelstein and Kinzler 1998, Lynch and de la Chapelle 1999), which accounts for up to 5% of all diagnoses of colorectal cancer in the US, can be caused by a germline mutation of any one of a set of known DNA mismatch repair (MMR) genes including MSH2, MLH1 and others.

In relation to these syndromes, a common question concerns the probability that an individual carries a deleterious germline mutation of a disease gene, given a certain pattern of disease diagnoses in the individual's family history. This calculation is referred to here as carrier status prediction, or risk prediction. Its applications are in two areas. In clinical counseling of concerned individuals, risk prediction provides important support to decision making about genetic testing, disease prophylaxis, family planning and other issues. In research it provides a flexible approach to modeling and analyzing family data in situations in which testing is impractical but extensive family history is available.

Carrier status prediction in genetic counseling concerns inference on the genotype of an individual (the counselee) conditional on information about his/her disease history and his/her relatives' disease and genotype history (a pedigree). Two broad classes of modeling approaches have been used so far: the *Empirical* approaches model the conditional distribution of genotype given phenotype directly, by applying statistical or artificial intelligence techniques to pedigree data for tested individuals; in contrast, *Mendelian* models are built upon the conditional distributions of phenotypes given genotype (penetrance), and the marginal distributions of genotypes (prevalence). The probabilities required for counseling are then derived from these using Bayes' rule and Mendel's laws (Murphy and Mutalik 1969, Elston and Stewart 1971, Szolovits and Pauker 1992, Offit and Brown 1994, Parmigiani, Berry and Aguilar 1998, Antoniou et al. 2000).

Mendelian risk prediction models exploit domain knowledge of Mendelian inheritance and other biological characteristics of susceptibility genes and thus can incorporate pedigree features at higher resolution, provide intuitive parameterization in terms of penetrance and prevalence, and can be extended easily to arbitrary pedigrees. Validation studies in cancer models indicate that Mendelian models provide a well founded approach to genetic

counseling, and improved predictive performance compared to empirical approaches (Berry et al. 2002, Marroni et al. 2004).

In cancer a widely used Mendelian model is BRCAPRO, which assesses the probability that an individual carries a germline deleterious mutation of the BRCA1 and BRCA2 genes, based on his or her family's history of breast and ovarian cancer (Berry et al. 1997, Parmigiani, Berry, Iversen, Müller, Schildkraut and Winer 1998, Parmigiani, Berry and Aguilar 1998, Iversen et al. 2000). BRCAPRO assumes autosomal dominant inheritance, which is supported extensively by previous analyses (Newman et al. 1988). Following the template of BRCAPRO, CRCAPRO was later developed for the genes MSH2 and MLH1, involved in the HNPCC syndrome. CRCAPRO has a similar structure to BRCAPRO and uses information about colorectal and endometrial cancer, as well as microsatellite instability.

Here, expanding on the principles of BRCAPRO and CRCAPRO, we introduce BayesMendel, a generic tool for building Mendelian risk prediction models for autosomal dominant genes. Currently, the development of models merging Mendelian principles and state-of-the art statistical techniques requires substantial statistical and computational expertise. Our tool is designed to enable genetic epidemiologists to flexibly implement their own Mendelian models for novel syndromes and local subpopulations, without reprogramming complex statistical analyses and prediction tools. It will also allow other groups to contribute to our models by developing variants and input data for specific applications, countries, et cetera. We expect BayesMendel to increase the impact and usefulness of Mendelian models in cancer prevention. Applications of this tool will extend to inherited familial syndromes beyond cancer.

BayesMendel is distributed as a library under the open source environment R (Ihaka and Gentleman 1996), an intuitive, highly functional and extensible programming language. R provides users with a comprehensive, state-of-the-art statistical analysis toolbox based entirely on free and open source code.

In the remainder of this article, Section 2 reviews the theoretical basis of the Mendelian risk prediction approach. Section 3 presents the functionality and object–oriented structure of the library. Finally Section 4 discusses current limitations and possible future extensions.

2 Methods

2.1 Theory

Risk prediction calculations performed by the BayesMendel library are based on a general approach, which can be described as follows. Let γ_0 be the vector of genotypes of the counselee at each of the genes considered by the model. Each coordinate represents a different

locus. In the current formulation, all deleterious variants are assumed to have the same phenotypic implications. For each coordinate, γ_0 is either 0, 1, or 2 depending on whether the individual is non–carrier, heterozygous carrier, or homozygous carrier of a deleterious mutation at the locus. Let R be the number of relatives of the counselee, let r be a relative in the family, and let γ_r, for $r = 1, \ldots, R$ be the corresponding genotype vectors. Similarly, we denote by $h_0, h_1, \ldots, h_R$ the relevant phenotypes and ages of onset of the counselee and relatives. For example in BRCAPRO, for each relative r, the vector h_r includes information on affected status for all relevant cancer sites, with age of onset if affected, or current age or age at death if unaffected. In addition to this information, other individual specific covariates such as being of Ashkenazi Jewish origin are provided by $X_r, r = 0, 1, \ldots, R$, and the exact relationship of each relative to the counselee is known,

Carrier Probability Calculation

Our goal is to obtain the probability distribution of the counselee genotype given the family history, covariates and pedigree structure, that is

$$p(\gamma_0 | h_0, h_1, \ldots, h_R; X_0, X_1, \ldots, X_R) \tag{1}$$

We suppress the notation of $X_0, X_1, \ldots, X_R$ later on to keep the mathematical expressions concise. However conditioning on the covariates is implied in all calculations.

The genotype distribution of Expression (1) can be obtained using a two–step process: an updating step and an integration step via the law of total probability: the updating step is based on the mathematical identity

$$p(\gamma_0 | h_0, \ldots, h_R) = \frac{p(\gamma_0) p(h_0, h_1, \ldots, h_R | \gamma_0)}{\sum_\gamma p(\gamma) p(h_0, \ldots, h_R | \gamma)}, \tag{2}$$

which is an instance of Bayes' rule. The unconditional carrier probability $p(\gamma_0)$ (or prevalence) is updated to incorporate information from the pedigree. The term $p(h_0, h_1, \ldots, h_R | \gamma_0)$ is the probability of the phenotypes for the whole pedigree given the genotype of the counselee. This is complex to evaluate directly, but it can be simplified using the law of total probability:

$$p(h_0, h_1, \ldots, h_R | \gamma_0) = \sum_{\gamma_1 \cdots \gamma_R} p(h_0, \ldots, h_R | \gamma_1, \ldots \gamma_R) p(\gamma_1, \ldots, \gamma_R | \gamma_0), \tag{3}$$

which considers explicitly the unobserved genotypes of the relatives. In the current approach we make the additional assumption that individual histories are conditionally independent given the genotypes, and obtain:

$$p(h_0, \ldots, h_R | \gamma_0) = \sum_{\gamma_1 \cdots \gamma_R} \left[\prod_{r=0}^{R} p(h_r | \gamma_r) \right] p(\gamma_1, \ldots, \gamma_R | \gamma_0). \tag{4}$$

The term $p(\gamma_1, \ldots, \gamma_R|\gamma_0)$ is known for all genotype configurations from Mendel's laws, as long as the mode of inheritance is known. This set of relationships connects the carrier probability with penetrance and prevalence information that can be abstracted from the literature or estimated from cohort data, or both. This approach applies to arbitrary pedigree sizes.

BayesMendel is designed specifically to address the situation in which penetrance is incomplete and age at onset varies across individuals in both wild-type and mutants. Imagine we could observe the time to the development of a certain phenotype if there were no death or censoring. We call this the latent time to phenotype and we call its probability distribution net penetrance . One minus the Kaplan-Meier estimator is a consistent estimator of the net penetrance when the development of the phenotype is independent of other competing risks and censoring (Tsiatis 1998).

In the calculation of $p(h_r|\gamma_r)$, we assume that both the censoring process and deaths of causes unrelated to the syndrome is independent of the latent time to the phenotype. We also assume that censoring and deaths of other causes are non-informative, that is, the distribution of the latent time to deaths of other causes or time of censoring is the same for both wildtypes and mutation carriers. If the mutation affects deaths by other causes, then these assumptions do not hold. For details on the impact of censoring and competing risks on Mendelian models, see Katki et al. (2004).

Then $p(h_r|\gamma_r)$ can be written as a product of two terms: one that depends on the phenotype-specific net penetrance and another that depends on the latent time to death. Because the latter is the same for all genotypes, it cancels out in the evaluation of Expression (2) thus does not need to be considered. We code age in discrete one-year intervals. If we write the net penetrance of genotype γ by age t as $F(t;\gamma)$, then the likelihood contribution of a case individual diagnosed in the age interval $[t, t+1)$ is proportional to $f(t;\gamma) = F(t+1;\gamma) - F(t;\gamma)$, while the likelihood contribution of an asymptomatic individual of age t is proportional to $1 - F(t;\gamma)$.

Cancer Risk Prediction

Once the genotype distribution has been calculated for an asymptomatic counselee, we can predict the future risk that the counselee develops the phenotype by age t. There are two quantities that an asymptomatic counselee may be interested in. The first quantity of interest is the "net" probability that he/she develops a particular phenotype by a future age t, that is, the probability that one will develop the phenotype if death and other phenotypes are removed. The counselee may be interested in this quantity if he/she is solely concerned with reducing the net risk without considering other competing risk factors. This net probability

is a weighted average of the genotype-specific net penetrances with the weights being the genotype probabilities, that is

$$F(t|h_0,\ldots,h_R) = \sum_{\gamma_0} F(t|\gamma_0)\cdot p(\gamma_0|h_0,\ldots,h_R) \tag{5}$$

For more realistic purposes, we also provide a second quantity, that is, the probability that the counselee will develop the specific phenotype first at age t, surviving other causes and death. We call this the "crude" probability. We consider a case where the syndrome contains two phenotypes S_1 and S_2, and denote the latent time to diseases by T_{S_1} and T_{S_2} and latent time to death of other causes T_D. The calculation can be extended to syndromes containing more than two diseases. The net probability distributions of T_{S_1}, T_{S_2} and T_D are denoted by $F_1(t;\gamma_0)$, $F_2(t;\gamma_0)$ and $F_D(t;\gamma_0)$. Then the "crude" genotype-specific probability of developing disease S_1 at age t is

$$\begin{aligned} & P(t \le T_{S_1} < t+1, T_{S_2} \ge T_{S_1}, T_D \ge T_{S_1}|\gamma_0) && (6)\\ \approx\ & P(t \le T_{S_1} < t+1, T_{S_2} \ge t, T_D \ge t|\gamma_0) \\ =\ & (F_1(t+1|\gamma_0) - F_1(t|\gamma_0))\cdot(1-F_2(t|\gamma_0))\cdot(1-F_D(t|\gamma_0)) && (7)\end{aligned}$$

The $F_D(t|\gamma_0)$ in Expression (7) is the probability distribution of the latent time to dying of causes unrelated to the syndrome. The hazard corresponding to this net distribution can be derived from public domain data by the following procedure:

Let $\overline{F}_{all}(t)$ denote the cumulative mortality incidence rate (deaths of all causes) by age t in the population, and $\overline{F}_i(t)$ the cumulative incidence rate of death due to disease i. Then the cumulative incidence of death due to unrelated causes is $\overline{F}_D = \overline{F}_{all} - \overline{F}_1 - \overline{F}_2$. Then by assuming independence between the two competing risks, the net hazard of death from unrelated causes is equal to the cause-specific hazard $\lambda_D = \frac{d\overline{F}_D}{dt}/(1-\overline{F}_{all})$.

We convert the net hazard to $\overline{F}_D$, then use Equation (7) to obtain the crude probability of developing disease S_1.

2.2 A Simple Illustration

To illustrate the above calculations with a concrete example, we consider the case of a woman seeking counseling because of her mother's breast cancer history. Let us assume that the woman being counseled is of Ashkenazi ethnic origin, 40 years old and cancer free. Her mother was diagnosed with breast cancer at age 40, and subsequently died at 55 without additional cancer diagnoses. To simplify the exposition, we consider a single hypothetical gene called BRCA. For this gene, we will denote the woman's genotype with the variable

γ_0, where $\gamma_0 = 1$ when the woman is a heterozygous carrier of any deleterious BRCA allele, $\gamma_0 = 2$ when she is a homozygous carrier, and $\gamma_0 = 0$ otherwise. We denote the woman's breast cancer history with h_0, and her mother's with h_1. In our example, $h_0 =$ {breast and ovarian cancer free at 40}, $h_1 =$ {breast cancer diagnosed at 40, ovarian cancer free at 55}.

Our quantity of interest is the *a posteriori* probability that the counselee carries at least one deleterious BRCA mutation given her family history, that is

$$p(\gamma_0 = 1 \text{ or } \gamma_0 = 2 | h_0, h_1) = 1 - p(\gamma_0 = 0 | h_0, h_1)$$

By Equation (2),

$$\begin{aligned} & p(\gamma_0 = 0 | h_0, h_1) \\ = & \frac{p(\gamma_0 = 0) p(h_0, h_1 | \gamma_0 = 0)}{p(\gamma_0 = 2) p(h_0, h_1 | \gamma_0 = 2) + p(\gamma_0 = 1) p(h_0, h_1 | \gamma_0 = 1) + p(\gamma_0 = 0) p(h_0, h_1 | \gamma_0 = 0)} \\ = & \frac{p(\gamma_0 = 0) p(h_0, h_1 | \gamma_0 = 1)}{p(\gamma_0 \neq 0) p(h_0, h_1 | \gamma_0 \neq 0) + p(\gamma_0 = 0) p(h_0, h_1 | \gamma_0 = 0)} \end{aligned} \tag{8}$$

The bottom equation assumes that homozygous and heterozygous carriers have the same penetrance, an assumption that is currently made by BayesMendel. In the numerator of Equation (8), $p(\gamma_0 = 0)$ is the *a priori* probability that the counselee is a wild-type. In this illustration, we assume the allele frequency in the Ashkenazi Jewish population to be $f = 0.013$. Then $p(\gamma_0 = 0) = (1 - f)^2 = 0.974$. To calculate the probability of the observed family history conditional on the woman being a noncarrier, that is $p(h_0, h_1 | \gamma_0 = 0)$, we need to integrate out the genotype of the mother, denoted by γ_1, as is done in Expression (3) and (4). That is,

$$p(h_0, h_1 | \gamma_0 = 0) = \sum_{\gamma_1 = 0,1,2} \left[p(h_0 | \gamma_0 = 0) p(h_1 | \gamma_1) \right] p(\gamma_1 | \gamma_0 = 0) \tag{9}$$

$$= \left[p(h_0 | \gamma_0 = 0) p(h_1 | \gamma_1 = 0) \right] p(\gamma_1 = 0 | \gamma_0 = 0) + \tag{10}$$

$$\left[p(h_0 | \gamma_0 = 0) p(h_1 | \gamma_1 \neq 0) \right] p(\gamma_1 \neq 0 | \gamma_0 = 0) \tag{11}$$

The phenotype probabilities in Expression (9) can be taken directly from the known penetrance, as follows. We denote the net probability of getting breast cancer within the age interval $[t, t+1)$ for a mutation carrier by $f_b(t; \gamma = 1)$, and for a noncarrier, or wild-type, by $f_b(t; \gamma = 0)$. The net cumulative probabilities are denoted by $F_b(t; \gamma = 1)$ and $F_b(t; \gamma = 0)$. The corresponding probabilities for ovarian cancer are denoted by $f_o(t; \gamma = 1)$, $f_o(t; \gamma = 0)$, $F_o(t; \gamma = 1)$ and $F_o(t; \gamma = 0)$. In this illustration, we use the default penetrance for BRCA1 in BayesMendel. Here we make two further assumptions: conditional independence between the cancer sites given genotype, and independence of prognosis and genotype, once a given

cancer type is diagnosed. Using these assumptions,

$$\begin{aligned} p(h_1|\gamma_1) &= p(\text{breast cancer at 40, ovarian cancer free at 55}|\gamma_1) \\ &\propto f_b(40;\gamma_1)\cdot[1-F_o(55;\gamma_1)] = \begin{cases} 0.00054\cdot 0.998 = 0.00053, \text{ when } \gamma_1 = 0 \\ 0.024\cdot 0.817 = 0.020, \text{ when } \gamma_1 \neq 0 \end{cases} \end{aligned}$$

$$\begin{aligned} p(h_0|\gamma_0 = 0) &= p(\text{breast and ovarian cancer free at 40}|\gamma_0 = 0) \\ &\propto [1-F_b(40;\gamma_0=0)]\cdot[1-F_o(40;\gamma_0=0)] = 0.998\cdot 0.996 \approx 0.996 \end{aligned}$$

The mother's genotype given the daughter's genotype can be derived by using Mendel's Law and then applying Bayes' Theorem. The calculations involve integrating out the genotype of the counselee's father and leads to

$$\begin{aligned} p(\gamma_1 = 0|\gamma_0 = 0) &= \frac{p(\gamma_0=0|\gamma_1=0)p(\gamma_1=0)}{p(\gamma_0=0|\gamma_1=0)p(\gamma_1=0)+p(\gamma_0=0|\gamma_1\neq 0)p(\gamma_1\neq 0)} \\ &= 1-f = 0.987 \\ p(\gamma_1 \neq 0|\gamma_0 = 0) &= 1-p(\gamma_1=0|\gamma_0=0) = f = 0.013 \end{aligned}$$

Inserting the results into Expression (9) we get

$$p(h_0,h_1|\gamma_0=0) \propto [0.996\cdot 0.00053]0.987 + [0.996\cdot 0.020]0.013 = 0.00078$$

Following a similar procedure, we can get the family history likelihood when the counselee is a noncarrier, that is,

$$p(h_0,h_1|\gamma_0\neq 0) \propto 0.0082$$

Now we have all the pieces for Expression (8) and we get the *a posteriori* probability that the counselee is a noncarrier as

$$\frac{0.974\cdot 0.00078}{(1-0.974)\cdot 0.0082 + 0.974\cdot 0.00078} = 0.78$$

The probability that the counselee carries any deleterious BRCA mutation is then $1-0.78 = 0.22$.

3 Software

The BayesMendel R–library defines an object–oriented environment for Mendelian risk prediction. It provides functionality to **a)** evaluate Expression (4) for arbitrary syndromes, **b)** evaluate carrier probabilities for the breast-ovarian and HNPCC syndromes according to the BRCAPRO and CRCAPRO models, and **c)** process and check pedigree data.

The most challenging computational aspect of the approach of Section 2 is the evaluation of expression (4). The number of terms in the summation is 3^{GR}, where G is the number of genes and R is the number of untested relatives of the counselee plus the number of negative relatives if the genetic test has sensitivity less than one. In BayesMendel, this calculation is performed in C by a subroutine called MARGENE, which serves as the computational engine for all risk prediction models, including BRCAPRO and CRCAPRO. Details of the algorithm are given in Parmigiani, Berry and Aguilar (1998). In BayesMendel, the R function `aveG` serves as an interface to MARGENE. It takes as input the terms $p(h_r|\gamma_r)$ and $p(\gamma_0)$ in expression (4). The current implementation of expression (4) in the software considers two diseases and two genes, referred to as Disease1, Disease2 and Gene1, Gene2 in the text that follows. The two genes are assumed to be in Hardy-Weinberg and linkage equilibrium in the population. The pedigree may extend to first- and second-degree relatives of the counselee.

There are three major object classes in BayesMendel: pedigree objects, penetrance objects, and prediction objects.

Pedigree Objects

A pedigree object includes the pedigree structure and phenotype information for the counselee's family in matrix form. The definition of the variables is the same as in the original BRCAPRO and in CancerGene (Euhus 2001). The object is a matrix with one row for each family member and 12 or more columns with information of that member as shown in Table 1.

We explain how to prepare the pedigree object using the example family shown in Figure 1. Alternatively, the CancerGene package offers a user-friendly interface for forming these input files. The family is suspected to have the HNPCC syndrome.

Computing carrier probabilities for a different counselee within the same family requires creating a different input file. The input file corresponding to the family of Figure 1 is shown in Figure 2.

Let's consider now family member 1 in detail. Family member 1 is the counselee. She was diagnosed with colorectal cancer at age 47. She is alive and 57 years old. She has not undergone genetic testing. Her tumor was tested for microsatellite stability and the result was microsatellite instable. We enter 1 in the member identifier column; we enter 1 in the relation column, using Table 2; We enter 0 in the sex column; We enter 3 in the father's identifier number column — this will constrain us to input the father's information in the third row; We enter 2 in the mother's identifier number column — this will constrain us to input the mother's information in the second row; We enter 1 in the colorectal cancer status column; We enter 0 in the endometrial cancer status column; We enter 47 in the age column

Column	Content
1	Member identifier
2	Relation to the counselee
3	Sex (0=female, 1=male)
4	Father's identifier number
5	Mother's identifier number
6	Disease1 phenotype (0=unaffected, 1=affected, one breast only in breast-ovarian cancer syndrome; 2=bilateral breast cancer in breast-ovarian cancer syndrome)
7	Disease2 phenotype (0=unaffected, 1=affected)
8	Age of onset of Disease1 phenotype if affected. Current age or age of death if unaffected. 1 if unaffected and there is no age information.
9	Age of onset of Disease2 phenotype if affected. Current age or age of death if unaffected. 1 if unaffected and there is no age information.
10	Age at onset of breast cancer, second breast. Only useful in breast-ovarian cancer syndrome. For the rest enter a 0.
11	Gene1 testing result. (0=no test, 1=positive test, 2=negative test)
12	Gene2 testing result. (0=no test, 1=positive test, 2=negative test)
13 and up	Other model-specific risk factors. For example, as is implemented in the function `crcapro()` for suspected HNPCC families,the 13-th column is microsatellite status on each family member (0=no information, 1=microsatellite instable or "MSI", 2=microsatellite stable or "MSS").

Table 1: Column Codes for Pedigree Objects

for the colorectal cancer; We enter 57 — the current age — in the age column for endometrial cancer; We enter 0 in the age column which is not applicable to HNPCC syndrome; We enter 0 in the MLH1 test result column; We enter 0 in the MSH2 test result column; In the last column we enter 1 for MSI. Note that all of the counselee's affected first-degree relatives had their tumor samples tested for microsatellite instability. The father's tumor was tested microsatellite stable while those of the sister and mother were tested microsatellite instable. Thus we enter a 2 in the last column of the 3rd row and two 1s in the last column of the 2nd and 7th row.

There are some additional rules and restrictions that one should be aware of when preparing a pedigree information matrix:

Order. The only restrictions in the order of the family members is that the counselee's husband, if applicable, must be entered immediately after the counselee. The same applies for the brothers' and the sisters' husbands. If you are entering pedigrees by hand, we suggest that you begin by creating the first three columns for all the individuals, and then create the father's and mother's identifiers columns.

Missing Information. In general, if information about a family member, other than the counselee, is missing entirely, the member can be omitted without affecting the calculations.

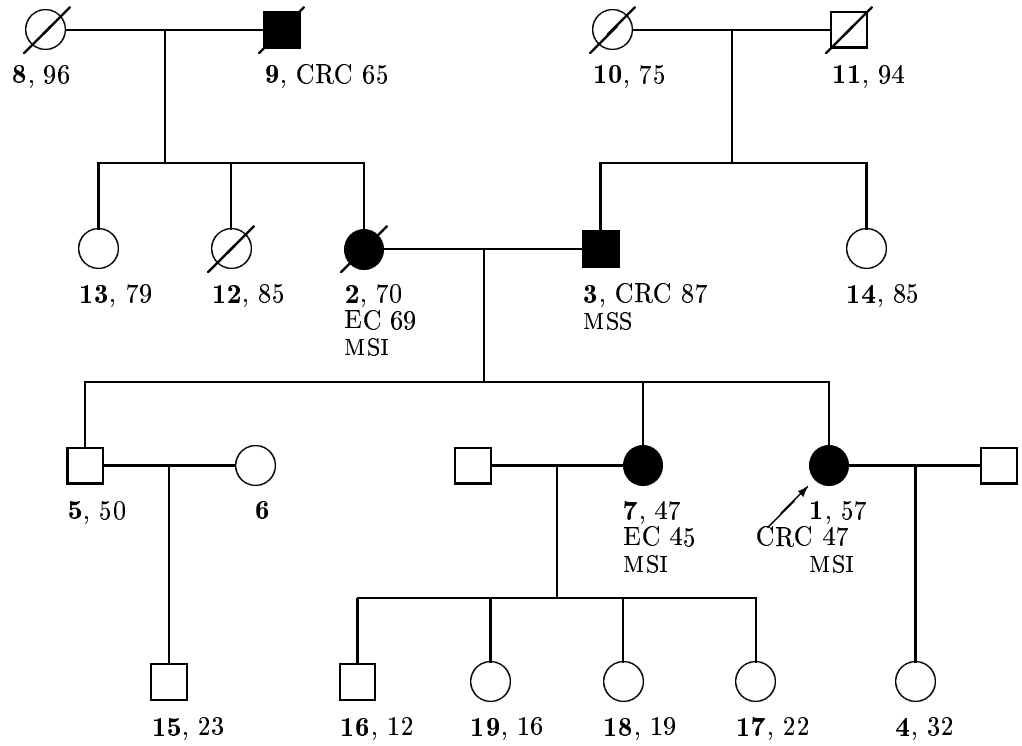

Figure 1: An example suspected HNPCC syndrome pedigree. The arrow indicates the individual to be counseled (counselee). The counselee can be either a man or a woman. A man is indicated by a square and a woman by a circle. For each relative, cancer of the colorectum (CRC) or endometrium (EC) are indicated by a solid square or circle with the type of cancer, age at onset and microsatellite status (MSI for microsatellite instability, MSS for microsatellite stability) below it. Boldface numbers on the left correspond to the relative ID's in Figure 2. Other numbers are ages.

In our example, we have no information about family member 6, the counselee's brother's wife, there are two ways to present this piece of missing information: first, we could omit the row in the family history matrix that corresponds to member 6; alternatively, we could treat that member's breast cancer status as being left censored at age 1, thus enter a 0 in the cancer status column and enter a 1 in the age at onset column. Both presentations will result in a likelihood contribution $p(h_r|\gamma_r)$ of 1 for all possible genotypes γ_r by that member in the full likelihood as shown in Expression (4). When the breast cancer column is not 0, the age at onset of breast cancer column must be specified. The same applies to ovarian cancer.

Unaffected relatives are very important in the calculation, they provide information as long as their current age or age at death or last contact is known. Effort should be made to incorporate such information. For example, if an aunt is known to be breast cancer free until age 40, the time when she was last in touch, but nothing is known about her ovarian cancer status, then a 0 should be entered in her breast cancer status column and a 40 in her

1	1	0	3	2	1	0	47	57	0	0	0	1
2	4	0	9	8	0	1	70	69	0	0	0	1
3	4	1	11	10	1	0	87	87	0	0	0	2
4	3	0	0	1	0	0	32	32	0	0	0	0
5	2	1	3	2	0	0	50	50	0	0	0	0
6	15	0	0	0	0	0	1	1	0	0	0	0
7	2	0	3	2	0	1	45	47	0	0	0	1
8	7	0	0	0	0	0	96	96	0	0	0	0
9	7	1	0	0	1	0	65	65	0	0	0	0
10	5	0	0	0	0	0	75	75	0	0	0	0
11	5	1	0	0	0	0	94	94	0	0	0	0
12	8	0	9	8	0	0	85	85	0	0	0	0
13	8	0	9	8	0	0	79	79	0	0	0	0
14	6	0	11	10	0	0	85	85	0	0	0	0
15	13	1	5	6	0	0	23	23	0	0	0	0
16	13	1	0	7	0	0	12	12	0	0	0	0
17	13	0	0	7	0	0	22	22	0	0	0	0
18	13	0	0	7	0	0	19	19	0	0	0	0
19	13	0	0	7	0	0	16	16	0	0	0	0

Figure 2: Pedigree information matrix corresponding to the family of Figure 1.

Number	Relation to the counselee
1	Counselee.
2	Brother or sister.
3	Son or daughter.
4	Parent.
5	Paternal grandparent.
6	Paternal aunt or uncle.
7	Maternal grandparent.
8	Maternal aunt or uncle.
13	Nephew or niece.
14	Husband.
15	Brother or sister in law.

Table 2: Relation codes

breast cancer age column, while a 0 should be entered in her ovarian cancer status column and a 1 in her ovarian cancer age column.

Nieces. There is one exception to the rule above. If there is information about a niece of the counselee, it is necessary to include a record (that is a row in the matrix) for the counselee's sibling that is a parent of the niece in question. This will come natural in most cases, but it must be done even in the case where there is no information about that sibling of the counselee, otherwise the family structure may not be unique.

Penetrance Objects

Penetrance objects include the net penetrance by age, gender, phenotype, and mutation status (wildtype at both loci, mutation on Gene1, mutation on Gene2, mutation on both).

The current default penetrance objects used by `brcapro` are `BRCApenet.nonAJ.2004` and `BRCApenet.AJ.2004`, containing the most up-to-date penetrance estimates based on Struewing et al. (1997), Antoniou et al. (2003) and King et al. (2003) and updated with the family data from the Cancer Genetics Network validation study (Parmigiani et al. 2004). The penetrance for the Ashkenazi Jewish(AJ) population and non-AJ population are estimated separately. Details of the penetrance estimation are discussed in Chen, Iversen Jr., Friebel, Finkelstein, Weber, Eisen and et al. (2004).

The user can choose to use an older version of the penetrance objects: `BRCApenet.nonAJ.2001` and `BRCApenet.AJ.2001`, which combined estimates from Ford et al. (1998) and Struewing et al. (1997) as described in Iversen et al. (2000).

In both versions of penetrances, the incidence rates for wild-types are derived by subtracting the carrier incidence times the population carrier prevalence from the (National Cancer Institute: Surveillance, Epidemiology, and End Results (SEER) Program 1997) population incidence. In the older version, the Ashkenazi Jewish population has the same carrier penetrance but a different phenocopy rate due to the significantly higher carrier prevalence compared to the non-AJ population.

The default penetrance object used by `crcapro` is called `HNPCCpenet.2004`, for the risk for colorectal and endometrial cancer of MLH1 and MSH2 by age and gender. The penetrance is derived from the literature (Lin et al. 1998, Vasen et al. 1996, Vasen et al. 2001). For noncarriers, incidence and mortality rates of colorectal and endometrial cancer by age groups are from the 1973–1995 (National Cancer Institute: Surveillance, Epidemiology, and End Results (SEER) Program 1997) Cancer Statistics Review and overall mortality rates from the `survexp.us` function in the `survival` package in R (Thernau 1996).

Prediction Objects

Prediction objects include the joint probability that the counselee carries an inherited deleterious mutation on the two genes in a 3 by 3 matrix. The three rows/columns signify the three genotypes at Gene1/Gene2: homozygous carrier, heterozygous carrier and wild-type. The prediction objects also include net and crude cumulative risk of developing disease in the future if the counselee is unaffected.

Functions

List of functions

- `ReadCaGeneFam`: reads an external pedigree file in CancerGene format and converts it into a BayesMendel pedigree object;
- `CheckFamStructure`: checks the pedigree object for errors and corrects them when possible; prints warnings and error messages if appropriate;
- `aveG`: is the R interface to the C routine MARGENE for evaluating the joint posterior probability of an individual (the counselee) carrying a phenotype-altering variant (mutation) of two autosomal dominant genes. Input are the conditional probability of phenotype(s) for each of the individual's first and second degree relatives, the variants' allelic frequencies, and the pedigree structure;
- `brcapro`: calculates the joint probability that an individual (the counselee) carries an inherited deleterious mutation of the BRCA1 and BRCA2 breast cancer susceptibility genes. Inputs are the corresponding penetrance object, pedigree object of the family history of breast and ovarian cancers and genetic testing results on BRCA1 and/or BRCA2. It calls `aveG` for its core computation.
- `crcapro`: similar to `brcapro` except the genes involved are MSH2 and MLH1, and the diseases involved are cancers at the colorectum and endometrium. Additionally, the probabilities are modified by MSI test results if provided, through individual contributions to the likelihood. Adjustments are made using literature based sensitivity and specificity (Chen, Watson and Parmigiani 2004);
- `disease.risk`: takes a pedigree object, uses one of the carrier probability models (*e.g.*,`brcapro`) to determine the joint carrier probabilities and the cumulative risks of developing phenotype for the counselee (the prediction is only meaningful if the counselee is asymptomatic), and stores the results in a prediction object. The cumulative risks of developing phenotype is calculated by taking the average of genotype specific

penetrances weighted by the marginal probability of each genotype as shown in Expression (5).

- `summary.disease.risk`: takes a prediction object and prints to the R terminal or an external file the summary of the probability of carrying mutations in susceptibility genes and the probability of developing the corresponding phenotypes.

Software is distributed from the website `http://astor.som.jhmi.edu/BayesMendel`.

4 Discussion

BayesMendel is designed to provide the genetic counseling and genetic epidemiology communities with a flexible tool for genetic susceptibility prediction of hereditary syndromes. It currently includes the BRCAPRO model for the breast-ovarian cancer syndrome and the CRCAPRO model for the hereditary non-polyposis colorectal cancer syndrome, and allows for easy modification of input genetic parameters of those models. BayesMendel can also serve as a generic tool for genetic epidemiologists to flexibly implement their own Mendelian models for novel syndromes, without reprogramming complex statistical analyses and prediction tools.

BayesMendel provides the backbone of the upcoming new release of the genetic counseling package CancerGene (Euhus 2001), a user-friendly environment for cancer risk prediction in breast and colorectal cancer that is currently licensed, free of charge, to over one thousand users. In addition to Mendelian model functionality, it includes a wide range of algorithms for risk prediction, and graphical interfaces for pedigree entry and updating.

Current limitations of the BayesMendel package reside primarily in the structure of the core computing engine MARGENE, which is at the moment confined to two unlinked autosomal dominant genes, to families without half-siblings or loops, and to pedigrees including first and second degree relatives. This is adequate in a wide spectrum of genetic counseling situations, in which the pedigree information is ascertained from a single counselee and information on third degree relative is not generally highly reliable. However, in controlled research studies that collect large pedigrees, important information can be lost. The theory behind the MARGENE calculation is conceptually extendible to arbitrary pedigrees.

Several studies have evaluated the errors associated with BayesMendel (Gilpin et al. 2000, Iversen et al. 1998, Berry et al. 2002). For example, Berry et al. (2002) compared the genetic test results for deleterious mutations of BRCA1 and BRCA2 to BayesMendel predictions for 301 individuals recruited from several breast cancer clinics and by self-referral. In this group, BRCAPRO predicted an average probability of 0.29 for the 150 probands with the

smallest predicted probabilities, and 0.952 for the 151 with the largest probabilities. The actual proportion of test positives in the two groups are 0.327 and 0.788. The authors have found BRCAPRO to be an adequately calibrated and to have better discrimination than its empirical counterparts. However, its ability of discriminating between a BRCA1 mutation and a BRCA2 mutation remains limited outside families with male breast cancer.

The release of BayesMendel described in this article is 1.2-1. The BayesMendel laboratory at Johns Hopkins plans a series of upgrades over the next several years. These will be made available at the lab's website (http://astor.som.jhmi.edu/BayesMendel). Penetrance and prevalence parameters for the major syndromes covered will be periodically updated to reflect major new publications. The functionality of the software will be expanded to generate prediction intervals based on a probabilistic sensitivity analysis approach, and to provide exceedance probabilities, that is the probability that a counselee's chance of carrying a gene exceeds a given threshold. Future version of the software will also migrate towards a more object oriented structure by incorporating, in the pedigree object, information about type of syndromes, the loci and cancers potentially involved, and the corresponding penetrance, thus free-ing the users from specifying the penetrance object to use and the prediction model to call.

Acknowledgment

The work of Sining Chen, Wenyi Wang and Giovanni Parmigiani was supported by the NCI under grants P30CA06973 (Hopkins Regional Oncology Research Center) and P50CA62924 (Hopkins GI SPORE). Authors thank David Euhus of the UTSW and Fabio Marroni of the University of Pisa for useful feedback.

References

Antoniou, A. C., Gayther, S. A., Stratton, J. F., Ponder, B. A. and Easton, D. F. (2000). Risk models for familial ovarian and breast cancer, *Genet. Epidemiol.* **18(2)**: 173–190.

Antoniou, A., Pharoah, P. D. P., Narod, S., Risch, H. A., Eyfjord, J. E., Hopper, J. L., Loman, N., Olsson, H., Johannsson, O., Borg, Å., Pasini, B., Radice, P., Manoukian, S., Eccles, D. M., Tang, N., Olah, E., Anton-Culver, H., Warner, E., Lubinski, J., Gronwald, J., Gorski, B., Tulinius, H., Thorlacius, S., Eerola, H., Nevanlinna, H., Syrjäkoski, K., Kallioniemi, O.-P., Thompson, D., Evans, C., Peto, J., Lalloo, F., Evans, D. G., and Easton1, D. F. (2003). Average risks of breast and ovarian cancer associated with BRCA1 or BRCA2 mutations detected in case series unselected for family history: A combined analysis of 22 studies, *Am. J. Hum. Genet* **72**: 1117–1130.

Berry, D. A., Iversen, E. S. J., Gudbjartsson, D. F., Hiller, E., Garber, J., Peshkin, B., Lerman, C., Watson, P., Lynch, H., Hilsenbeck, S., R., S., Hughes, K. and Parmigiani, G. (2002). Validation of BRCAPRO, sensitivity of genetic testing of BRCA1 and BRCA2, and implications for the existence of other breast cancer susceptibility genes, *J. Clin. Oncol.* **20**: 2701–2712.

Berry, D. A., Parmigiani, G., Sanchez, J., Schildkraut, J. and Winer, E. (1997). Probability of carrying a mutation of breast-ovarian cancer gene BRCA1 based on family history, *J Natl Cancer Inst* **89**: 227–238.

Chen, S., Iversen Jr., E. S., Friebel, T., Finkelstein, D., Weber, B., Eisen, A. and et al., L. E. P. (2004). Comprehensive evaluation of breast and ovarian cancer risks associated with BRCA1 and BRCA2 mutations. in progress.

Chen, S., Watson, P. and Parmigiani, G. (2004). Accuracy of msi testing in predicting germline mutations of MSH2 and MLH1: a case study in bayesian meta-analysis of diagnostic tests without a gold standard, *Technical report*, Johns Hopkins University, Department of Biostatistics.

Elston, R. C. and Stewart, J. (1971). A general model for the genetic analysis of pedigree data, *Hum. Hered.* **21**: 523–542.

Euhus, D. M. (2001). Understanding mathematical models for breast cancer risk assessment and counseling, *Breast J* **7(4)**: 224–232.

Ford, D., Easton, D. F., Stratton, M., Narod, S., Goldgar, D., Devilee, P., Bishop, D. T. et al. (1998). Genetic heterogeneity and penetrance analysis of the BRCA1 and BRCA2 genes in breast cancer families, *Am. J. Hum. Genet* **62**: 676–689.

Foulkes, W. D. and Hodgson, S. V. (eds) (1998). *Inherited Susceptibility to Cancer: Clinical, Predictive and Ethical Perspectives*, Cambridge University Press, Cambridge, UK.

Gilpin, C. A., Carson, N. and Hunter, A. G. (2000). A preliminary validation of a family history assessment form to select women at risk for breast or ovarian cancer for referral to a genetics center, *Clin Genet* **58(4)**: 299–308.

Ihaka, R. and Gentleman, R. (1996). R: A language for data analysis and graphics, *Journal of Computational and Graphical Statistics* **5**: 299–314.

Iversen, Jr, E. S., Parmigiani, G. and Berry, D. (1998). Validating Bayesian prediction models: a case study in genetic susceptibility to breast cancer, *Case Studies In Bayesian Statistics*, Vol. IV, pp. 321–338.

Iversen, Jr, E. S., Parmigiani, G., Berry, D. A. and Schildkraut, J. (2000). Genetic susceptibility and survival: Application to breast cancer, *Journal of the American Statistical Association* **95**: 28–42.

Jimenez-Sanchez, G., Childs, B. and Valle, D. (2001). Human disease genes, *Nature* **409**: 853–855.

Katki, H., Chen, S. and Parmigiani, G. (2004). Censoring and competing risks in Mendelian mutation prediction models, *manuscript* .

King, M. C., Marks, J. H., Mandell, J. B. and New York Breast Cancer Study Group (2003). Breast and ovarian cancer risks due to inherited mutations in BRCA1 and BRCA2, *Science* **302(5645)**: 643–6.

Lin, K. M., Shashidharan, M., Thorson, A G Ternent, C. A., Blatchford, G. F., Christensen, M. A. et al. (1998). Cumulative incidence of colorectal and extracolonic cancers in MLH1 and MSH2 mutation carriers of hereditary colorectal cancer, *J Gastrintest Surg* **2**: 67–71.

Lynch, H. T. and de la Chapelle, A. (1999). Genetic susceptibility to non-polyposis colorectal cancer, *J Med Genet* **36(11)**: 801–818.

Marroni, F., Aretini, P., D'Andrea, E., Caligo, M. A., Cortesi, L., Viel, A., Ricevuto, E., Montagna, M., Cipollini, G., Ferrari, S., Santarosa, M., Bisegna, R., Bailey-Wilson,

J. E., Bevilacqua, G., Parmigiani, G. and Presciuttini, S. (2004). Evaluation of widely used BRCA1/2-mutation-predicting models, *J Med Genet* **41(4)**: 278–285.

Murphy, E. A. and Mutalik, G. S. (1969). The application of Bayesian methods in genetic counseling, *Hum. Hered.* **19**: 126–151.

National Cancer Institute: Surveillance, Epidemiology, and End Results (SEER) Program (1997). SEER homepage, `http://www-seer.ims.nci.nih.gov`.

Newman, B., Austin, M. A., Lee, M. et al. (1988). Inheritance of human breast cancer: evidence for autosomal dominant transmission of high-risk families, *Proc Natl Acad Sci USA* **85**: 3044–3048.

Offit, K. and Brown, K. (1994). Quantitating familial cancer risk: a resource for clinical oncologists, *J Clin Oncol* **12**: 1724–1736.

Parmigiani, G., Berry, D. A. and Aguilar, O. (1998). Determining carrier probabilities for breast cancer susceptibility genes BRCA1 and BRCA2, *American Journal of Human Genetics* **62**: 145–158.

Parmigiani, G., Berry, D., Iversen, Jr, E. S., Müller, P., Schildkraut, J. and Winer, E. (1998). Modeling risk of breast cancer and decisions about genetic testing, *in* C. Gatsonis et al. (eds), *Case Studies In Bayesian Statistics*, Vol. IV, Springer, pp. 173–268.

Parmigiani, G., Friebel, T., Iversen, E. S., Chen, S., Finkelstein, D., Anton-Culver, H., Ziogas, A. and et al. (2004). Validity of models for prediction of BRCA1 and BRCA2 mutations: the cancer genetics network experience. manuscript.

Struewing, J. P., Hartge, P., Wacholder, S., Baker, S. M., Berlin, M., McAdams, M., Timmerman, M. M., Brody, L. C. and Tucker, M. A. (1997). The risk of cancer associated with specific mutations of BRCA1 and BRCA2 among Ashkenazi jews, *New England Journal of Medicine* **336**: 1401.

Szolovits, P. and Pauker, S. (1992). Pedigree analysis for genetic counseling, *in* K. C. Lun, P. Degoulet, T. E. Piemme and O. Rienhoff (eds), *MEDINFO-92 Proceedings of the Seventh Conference on Medical Informatics*, Elsevier, New York, pp. 679–683.

Thernau, T. M. (1996). *A package for survival analysis in S*, Mayo Foundation, Rochester, MN.

Tsiatis, A. (1998). *Encyclopedia of Biostatistics*, John Wiley and Sons, New York, pp. 824–834.

Vasen, H. F. A., Wijnen, J. T., Menko, F. H., Kleibeuker, J., Taal, B., Griffioen, G., Nagengast, F., Meijers-Heijboer, E., Bertario, L., Varesco, L., Bisgaard, M.-L., Mohr, J., Fodde, R. and Khan, P. (1996). Cancer risk in families with hereditary nonpolyposis colorectal cancer diagnosed by mutation analysis., *Gastroenterology* **110**: 1020–1027.

Vasen, H. F., Stormorken, A., Menko, F. H., Nagengast, F., Kleibeuker, J. H., Griffioen, G., Taal, B. G., Moller, P. and Wijnen, J. T. (2001). MSH2 mutation carriers are at higher risk of cancer than MLH1 mutation carriers: a study of hereditary nonpolyposis colorectal cancer families, *J Clin Oncol* **19(20)**: 4074–4080.

Vogelstein, B. and Kinzler, K. (1998). *The genetic basis of human cancer*, McGraw-Hill, New York.

Weber, B. L. (1998). Update on breast cancer susceptibility genes, *ASCO Educational Book* .

Weiss, K. M. (1993). *Genetic Variation and Human Disease*, Cambridge University Press, Cambridge.