

# PEDIG - A computer program for calculation of genotype probabilities using phenotype information

IVAR HEUCH<sup>1,2</sup> AND FRANCIS H. F. LI<sup>1</sup>

<sup>1</sup> Department of Human Genetics, University of Michigan Medical School, Ann Arbor, Michigan

<sup>2</sup> Institute of General Genetics, University of Oslo, Oslo, Norway

A program has been constructed for computer calculation of the probabilities of certain genotypes for a given propositus, conditional on the phenotype information in a pedigree without loops. The program can be applied directly for one autosomal or sex-linked locus, but is in fact more general and can easily be adapted to handle any situation with a completely known mode of inheritance.

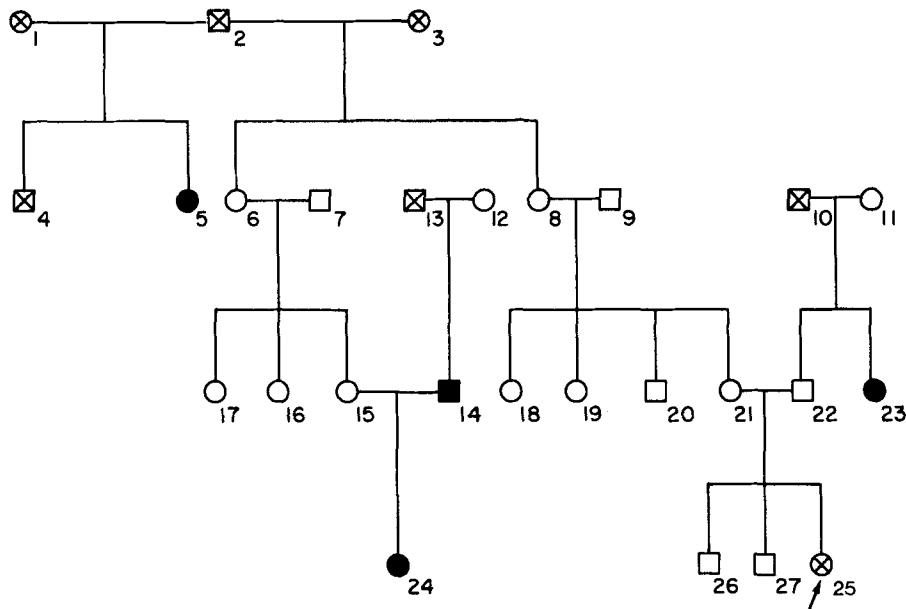
*Accepted for publication 21 August 1972*

In genetic counselling the following problem is encountered frequently: Given a pedigree with known phenotypes for certain relatives of a specified propositus, what is the probability distribution for the genotype of this propositus? Applying Bayesian methods it is, in principle, easy to solve many problems of this kind for given modes of inheritance, following, for instance, the approach suggested by Murphy & Mutalik (1969). However, even with a slightly complicated pedigree the actual calculations tend to be lengthy, and mistakes are made easily. Because of this, information from distant relatives often is ignored. These practical problems can be avoided by the use of computer programs (Chase et al. 1971, p. 143; Hilden 1970, p. 320). We have constructed such a FORTRAN program PEDIG, which to our knowledge is more general than any existing program of this kind. Starting with specified prior probability distributions for individuals included in the pedigree without ancestors, it derives the

probabilities for all possible genotypes of the propositus, conditional on the known phenotypes for the relatives.

## An Example

To illustrate the scope of the program, a pedigree with a certain amount of phenotype information is given in Fig. 1. The couple 21/22 with two normal sons 26 and 27 might want to know about the possibility for a future female child 25 to be affected by a hypothetical hereditary disease. PEDIG was run with this pedigree for three different modes of inheritance. 1) The affected state is due to homozygosity of a single autosomal allele *a* (with a corresponding dominant allele *A*), 2) It is caused by either of the genotypes *aa* or *Aa*. In the latter case, however, the chance of manifestation is only 0.5 (50% penetrance), 3) The disorder is caused by a recessive X-linked allele *a*. In all three cases it was assumed that the frequency of the allele *a* in the population



- ⊗ ⊗ NO INFORMATION ABOUT PHENOTYPE
- □ NORMAL
- ■ AFFECTED

Fig. 1. Pedigree used in the example. Note that phenotype information is known from three generations only.

was 5%. This is used by the program for constructing the prior distribution for the genotypes of the individuals 1, 2, 3, 7, 9, 10, 11, 12 and 13, whose parents are not included in the considerations. The output from PEDIG is given in Table 1. Except perhaps for case 3, the amount of work needed for performing the necessary calculations without a computer would be prohibitive.

Table 1

Probability distribution for the genotype of individual 25

	Case 1	Case 2	Case 3
aa	0.031	0.001	0.000
Aa	0.375	0.123	0.100
AA	0.594	0.876	0.900

Description of the Program

PEDIG is general in the sense that it accepts any probability distribution for the genotype of a child in given genotype combinations for the parents. The present version of the program includes the possibilities of one autosomal locus and one sex-linked locus (in both cases with any number of alleles), but the appropriate genotype distribution for the offspring is furnished by a FORTRAN FUNCTION, and the user may easily substitute his own version of this FUNCTION for any desired distribution (other interesting situations might include linked loci or mutations). Any relationship between genotypes and a finite number of phenotypes may be specified, also allowing probability distributions for the phenotype

given a certain genotype (as for  $Aa$  in case 2 in the example). These distributions may vary from one individual to another. This is useful when the penetrance coefficient for a trait depends on the age of the individuals involved.

The actual pedigree in each case is specified by giving the parents for each individual (if they are included at all). Certain restrictions apply if a pedigree is to be accepted by the program. It may not contain any loops, that is, no individuals may be produced by matings between related individuals included in the pedigree. An error message is printed out if this rule is violated. Half-sibs are taken into account, but half-sibs of half-sibs are ignored. Apart from this the pedigree may have branches everywhere, directed upwards leading to ancestors, or directed downwards, leading to offspring.

The program has been tested on an IBM 1130 computer (model 2C, 16K 16-bit words), but it can easily be adapted to any other system with a FORTRAN compiler. It does not make use of disks or tapes during execution. PEDIG itself, with sub-programs, requires 7214 core locations in our IBM version. The amount of core storage needed for data will vary with the tasks assigned to it. The total execution time on IBM 1130 was for each case in the example approximately 45 seconds, but most of this time was used for writing out the input information, and only 10 seconds were necessary for the actual calculations. Printing of all prior and conditional probabilities during calculations is optional. If inconsistencies are found in the input data, an error message is produced.

#### Method

The method applied for going through the pedigree and utilizing all phenotype information is similar to the one described in Murphy & Mutalik (1969), and also in

Murphy (1970) and Chase et al. (1971), although PEDIG does not transform the probabilities to other units. The calculations are performed recursively in a manner equivalent to using two procedures UP and DOWN, each of which is able to call itself and also the other one. The procedure UP will, for a specified individual, give the prior genotype distribution, using anterior information from the ancestors. DOWN will, for a specified parent combination, produce conditional probabilities for posterior phenotype information in the offspring (and, in some cases, in the parents themselves), given the possible genotypes in the parents. In fact, UP will use the appropriate input prior distribution if the specified individual has no parents. Otherwise it will call UP for both parents, then DOWN for these parents and their other children, convert the prior probabilities from UP to posterior probabilities by means of Bayes' theorem using the conditional probabilities from DOWN, and finally derive the prior distribution for the individual originally specified in the call to UP. If any one of the parents has children with other individuals, information from that part of the pedigree also is included in the final prior distribution. DOWN will go through the children in the given parent combination. For each such child it finds all combinations where the child itself is a parent, calls DOWN in every combination of this kind, calls UP for the spouse, and then constructs the conditional probabilities given the possible genotypes of the original child in question. These conditional probabilities for all children are combined in the output from the call to DOWN.

The execution starts with a call to UP for the propositus. Then calls are made to DOWN in each combination, if any, where the propositus is a parent. UP is called for each of the spouses of the propositus, and taking the marginals for the propositus'

genotypes, the program converts the prior distribution for the propositus into a posterior one, which gives the final result. For using the method outlined, certain basic assumptions are necessary, for instance that genotypes of sibs are stochastically independent, given the genotypes of their parents.

#### Applicability

The possible fields of application for the program are somewhat restricted by the fact that the mode of inheritance must be completely specified numerically. This excludes some practical cases in genetic counselling, where complete knowledge of the disease is lacking. It is also difficult to treat situations where theoretical considerations might reveal connections between the distributions that must be specified beforehand as, for instance, when individuals without ancestors can be considered as samples from a population where an equilibrium state exists between mutation pressure and selection (this is, for example, the basis for the treatment in Binet et al. (1958)). But in a number of situations, PEDIG should make it considerably easier to handle information from large pedigrees. It might also be of use outside genetic counselling, as in animal breeding, where it is often necessary to estimate the genotype of a given animal

from the phenotypes among its offspring. However, the requirement of no matings between related individuals will reduce the usefulness in such areas too.

The program listing and instructions for preparation of data cards are available from the authors on request. The construction of a program which is able to handle pedigrees with loops is now in progress.

#### References

- Binet, F. E., R. J. Sawyer & G. S. Watson (1958). Heredity counselling for sex-linked recessive deficiency diseases. *Ann. hum. Genet.* **22**, 144-152.
- Chase, G. A., E. A. Murphy & D. R. Bolling (1971). The ENCU scoring system. A strategy for solving a class of single-locus genetic counseling problems. *Clin. Genet.* **2**, 141-148.
- Hilden, J. (1970). GENEX- An algebraic approach to pedigree probability calculus. *Clin. Genet.* **1**, 319-348.
- Murphy, E. A. (1970). The Ensu scoring system in genetic counselling. *Ann. hum. Genet.* **34**, 73-78.
- Murphy, E. A. & G. S. Mutalik (1969). The application of Bayesian methods in genetic counselling. *Hum. Hered.* **19**, 126-151.

#### Address:

Ivar Heuch  
 Institute of General Genetics  
 University of Oslo  
 Box 1031, Blindern  
 Oslo 3, Norway