Addressing reproductive risk in consanguineous couples

Teeuw, M.E.

2015

document version
Publisher's PDF, also known as Version of record

CHAPTER 6

An improved method for estimating total pathogenic allele frequency of autosomal recessive disorders when parents are consanguineous

Jonker MA, Teeuw ME, Ten Kate LP.
Submitted

ABSTRACT

In this chapter we describe an improved method for estimating the total pathogenic allele frequency of an autosomal recessive disease based on mutational data and inbreeding coefficients of affected individuals. Furthermore, we explain how to construct confidence intervals for this frequency and describe the results of several simulation studies for investigating the unbiasedness and accuracy of the estimation method. Additionally, we consider the situation in which not all mutations for the disorder can be detected by the laboratory. An application of the method using real data is presented in Chapter 7.

INTRODUCTION

Suppose we want to estimate the total frequency of all pathogenic alleles of an autosomal recessive disease X. This can be done by, for instance, screening a sample of the population, but this is time-consuming. Since many laboratories already have mutational records of affected individuals, it would be convenient to use these data for estimating the total allele frequency in the population.

Gialluisi et al. (2012; 2013) describe an estimation method in accordance with a relationship that was formally derived in Ten Kate et al. (2010). They use data of mutational records and inbreeding coefficients of a sample of affected individuals. In this chapter we propose to use a maximum likelihood estimator that is based on the same type of data. We present details of this estimation method, and its performance (unbiasedness and accuracy) is investigated by means of simulation studies. The construction of a confidence interval for the total allele frequency is also described. In practice, it may happen that some individuals in the data set do have the disease and therefore must carry two disease alleles, but none or only one mutation has been observed. This may happen if for this disorder in the laboratory not all mutations are identified or routinely tested. We also consider this problem in this chapter.

This chapter does not contain an application of the method using real data. For an application we refer to Teeuw et al. where the proposed method is applied to data of recessive *MEFV* mutations in Tunisian and Moroccan Familial Mediterranean Fever (FMF) patients (Chapter 7).

METHODS

Let's suppose we want to estimate the total pathogenic allele frequency q of a recessive autosomal disease X in a particular population. For an individual with an inbreeding coefficient equal to a certain value F, the probability of him or her having disease X is given by

$$P\langle X \,|\, F\rangle = Fq + (1-F)q^2.$$

Suppose there are K different disease alleles with relative allele frequencies $\alpha_1,\alpha_2\ldots,\alpha_K$ in the population of interest
(so, $\sum_{j=1}^{K}\alpha_j = 1$).

Let variable Z indicate which pair of alleles is observed in a patient; $Z=(j,k)$ means that the alleles j and k are observed. The conditional probability or likelihood that an affected individual with inbreeding coefficient F has genotype $Z=(j,j)$, respectively, $Z=(j,k)$ with $j\neq k$, are equal to (1)

$$P\langle Z=(j,j)\,|\,X,F\rangle=\frac{P\langle Z=(j,j),X\,|\,F\rangle}{P\langle X\,|\,F\rangle}=\frac{Fqa_j+(1-F)q^2a_j^2}{Fq+(1-F)q^2},$$

$$P\langle Z=(j,k)\,|\,X,F\rangle=\frac{P\langle Z=(j,k),X\,|\,F\rangle}{P\langle X\,|\,F\rangle}=\frac{(1-F)q^2 2a_j a_k}{Fq+(1-F)q^2}$$

From the observed genotype, it can be determined whether an individual is homozygous or compound heterozygous; this is indicated by Δ:$\Delta=1$ if he is homozygous and $\Delta=0$ if he is compound heterozygous.

Now suppose that we have a random sample of n affected individuals from the population of interest for estimating q. For every individual, we observe the genotype at the disease gene and the inbreeding coefficient. The total conditional likelihood for all individuals in the data set is the product of the individual conditional likelihoods (here we assume that the individuals in the data set are unrelated), and is given by

$$\prod_{i=1}^{n} P\langle Z_i=(z_{i,1},z_{i,2})\,|\,X,F_i\rangle,$$

where the i in the formula refers to individual i, so for instance, F_i and Δ_i equal the inbreeding coefficient and Δ-value of individual i. Furthermore, $z_{i,1}$ and $z_{i,2}$ are the allele numbers observed for individual i, so $z_{i,1}$ and $z_{i,2}$ equal a number from the set $\{1,\ldots,K\}$. The maximum likelihood estimates for q and the relative allele frequencies $\alpha_1,\ldots,\alpha_k$ are those values for q and $\alpha_1,\ldots,\alpha_k$ where the total conditional (log) likelihood attains its maximum. Since no formula can be found for the maximum likelihood estimator of q, maximization has to be performed numerically. An attractive alternative is to estimate the relative allele frequencies beforehand, to insert these estimates into the likelihood and then to maximize the likelihood with respect to q in order to find an estimate of q. In the remainder of this chapter we estimate the relative allele frequencies by the sample frequencies among the affected individuals in our sample and impute these estimates into the likelihood before maximizing the likelihood with respect to q. Maximization of the (log) likelihood can be performed with a quasi-Newton algorithm or a grid search in, for instance, the statistical package R (http://www.r-project.org/).

Construction of confidence interval

Since no formula for the maximum likelihood estimator of q was found, the construction of a confidence interval has to be done implicitly. A confidence interval for q can be constructed with the help of the likelihood ratio test: the confidence interval consists of all values q_0 for which the null hypothesis H_0:$q=q_0$ is not rejected for a fixed chosen level α; i.e. for which the value of the likelihood ratio test statistic is smaller than the (1- α) quantile of the (asymptotic) distribution of the likelihood ratio statistic. If the relative allele frequencies are known (and not estimated), this distribution equals the chi-squared distribution with one degree of freedom. However, they are not known, but

estimated and imputed in the likelihood function. In that case, a small mistake is made when using the *(1-α)* quantile of the chi-squared distribution with one degree of freedom. However, simulation studies showed that this error is very small and can be ignored when constructing a confidence interval in our situation (Jonker and van der Vaart, 2014).

Existing estimation method

Gialluisi et al. describe their estimation method of *q* very concisely (2012; 2013). Since the value that should be taken for *F* in their estimator for *q* was unclear, we describe the estimation method again, focusing on the choice of the value of *F*. We will use the notation that was introduced in this chapter. As mentioned before, the probability that an individual with inbreeding coefficient *F* has the disease *X* is given by

$$P\langle X \mid F\rangle = Fq + (1-F)q^2.$$

The probabilities that an individual with inbreeding coefficient *F* is affected and compound heterozygous *(Δ=0)*, respectively, homozygous *(Δ=1)* at the disease gene equal

$$P\langle \Delta = 0, X \mid F\rangle = (1-F)q^2\left(1-\sum_{j=1}^{K} a_j^2\right),$$

$$P\langle \Delta = 1, X \mid F\rangle = F_q + (1-F)q^2\sum_{j=1}^{K} a_j^2$$

The probabilities $P(\Delta=1 \mid X,F)$ and $P(\Delta=0 \mid X,F)$ equal the fractions of homozygous and compound heterozygous individuals respectively, in the population of affected individuals with inbreeding coefficient equal to *F*. These probabilities equal

$$P\langle \Delta = 0 \mid X, F\rangle = \frac{P\langle \Delta = 0, X \mid F\rangle}{P\langle X \mid F\rangle} = \frac{(1-F)q^2\left(1-\sum_{j=1}^{K} a_j^2\right)}{Fq + (1-F)q^2}$$

$$P\langle \Delta = 1 \mid X, F\rangle = \frac{P\langle \Delta = 1, X \mid F\rangle}{P\langle X \mid F\rangle} = \frac{Fq + (1-F)q^2\sum_{j=1}^{K} a_j^2}{Fq + (1-F)q^2}$$

Solving *q* from one of the two equations yields (2)

$$q = \frac{P\langle \Delta = 0 \mid X, F\rangle F}{(1-F)\left(P\langle \Delta = 1 \mid X, F\rangle - \sum_{j=1}^{K} a_j^2\right)}$$

This was also deducted in Ten Kate et al., although with different notation (2010). In (2), *F* is the inbreeding coefficient for an affected child with consanguineous parents. The relationship (2) holds for any value of *F* with $0<F<1$. A special case is $F=F_{pop}$, where F_{pop} is the population inbreeding coefficient in the total population of interest i.e. the population of origin of the affected individuals. Computations in the Appendix show

that $P(\Delta=0|X,F_{pop})$=P($\Delta=0|X$) and $P(\Delta=1|X,F)$= P($\Delta=1|X$). Then, for $F=F_{pop}$, the relation in (2) equals (3):

$$q = \frac{P\langle \Delta = 0 | X \rangle F_{pop}}{\left(1 - F_{pop}\right)\left(P\langle \Delta = 1 | X \rangle - \sum_{j=1}^{K} a_j^2\right)}$$

The probabilities P($\Delta=0|X$) and P($\Delta=1|X$) equal the fractions of compound heterozygous and homozygous individuals respectively, in the subpopulation of affected individuals (so not necessarily with a specific inbreeding coefficient). These probabilities can, therefore, be estimated by the fractions of individuals with $\Delta=0$ (compound heterozygous) and $\Delta=0$ (homozygous) observed in the sample of affected individuals. We denote these sample frequencies as $\overline{\Delta}$ and $1-\overline{\Delta}$ respectively. Furthermore, the sum $\sum_{j=1}^{K} a_j^2$ is estimated as $\sum_{j=1}^{K} \hat{a}_j^2$ with $\hat{a}_1,...,\hat{a}_K$ the relative allele frequencies in the sample as described before. Suppose we have an accurate estimate $\hat{F}_{pop}$ for the population inbreeding coefficient F_{pop}. Then, an estimate for q can be found by inserting all the estimates (4):

$$\hat{q} = \frac{\left(1 - \hat{\Delta}\right)\hat{F}_{pop}}{\left(1 - \hat{F}_{pop}\right)\left(\hat{\Delta} - \sum_{j=1}^{K} \hat{a}_j^2\right)}$$

Estimating the population inbreeding coefficient is, in general, very complicated due to a lack of appropriate data (see for instance (Gialluisi et al., 2013). An alternative estimator for F_{pop} is the average inbreeding coefficient in the sample, $\hat{F}$. This yields the estimator (5):

$$\tilde{q} = \frac{\left(1 - \overline{\Delta}\right)\overline{F}}{\left(1 - \overline{F}\right)\left(\overline{\Delta} - \sum_{j=1}^{K} \hat{a}_j^2\right)}$$

Note that $\overline{F}$ will be close to the average of the inbreeding coefficient in the population of affected individuals, whereas F_{pop} equals the average of the inbreeding coefficient in the total population. These values (the average inbreeding coefficient in the total population and in the subpopulation of affected individuals) are, in general, unequal especially if q is small, which is usually the case in practice. If q is small, individuals with high inbreeding coefficients will be overrepresented in the population of affected individuals compared to the total population, and the estimator in (5) will be biased for estimating q. In the next section we perform a simulation study to consider the bias of this estimator.

Performance of estimators

In order to study the performance of the maximum likelihood estimator and the bias of the estimator in (5), we performed several simulation studies. We simulated data for n

individuals in a chosen setting. So, n individuals are chosen randomly from a population with affected children, with a chosen pattern of inbreeding coefficients, value for q, number of disease alleles, and relative allele frequencies. Based on the simulated data, we estimate the value of q with the maximum likelihood estimator and the estimator in (5). For the maximum likelihood estimator, we first estimate the relative allele frequencies, insert them into the likelihood, and then maximize with respect to q. The whole procedure of simulating data and estimating q is repeated 1000 times. So for every estimating method,1000 estimates are found.

We considered different settings and sample sizes. Large sample sizes were chosen to investigate the systematic bias of the estimation methods and the small sample sizes were chosen to see how the methods perform in practice where the sample sizes are typically small. We considered the following, arbitrary, setting. In all cases the population with affected children consists of respectively 5%, 10%, 20%, 15%, 10% and 35% of individuals with inbreeding coefficients equal to 1/8, 1/16, 1/32, 1/64, 1/128 and 0. In total there are four disease alleles with relative allele frequencies equal to 0.1, 0.2, 0.3, and 0.4. We considered sample sizes of 50, 100 and 1000 and the values of q are taken equal to 0.01 and 0.005.

Undiscovered alleles

As long as for every individual in the sample two mutations have been identified, the maximum likelihood method as described in this chapter can be applied. However, if an individual in the sample has all the disease symptoms, but the diagnostic test revealed only one or no mutations, the individual either does not have the disease (only the symptoms) or for this disorder not all disease alleles have been detected (see for instance Chapter 7). In the following we assume that everyone with the symptoms indeed has the disease, but not all alleles have been detected yet. If in an individual no mutation has been identified, it is unknown whether the individual is homozygous or compound heterozygous.

This makes the estimation of q more complex; the maximum likelihood method as described in this chapter cannot be applied directly. The simplest solution for this problem is to assume that there is just one unknown mutation, as yet not detected. Then, all unobserved disease alleles are assigned to this allele and the parameters can be estimated as before. If there is indeed one undetected allele, the assumption of one missing mutation is correct and the maximum likelihood estimates for q are (asymptotically) unbiased. If two or more mutations are undetected, the model with one additional mutation is incorrect and may affect the (asymptotically) unbiasedness of the estimator of q. In order to investigate this bias of the maximum likelihood estimator of q when multiple mutations are unknown, but are modelled as only one missing mutation, several simulation studies are performed. We assume there are four known mutations with relative allele frequencies equal to 0.25, 0.15, 0.10, 0.05 and four

undetected mutations with relative allele frequencies equal to 0.2, 0.1, 0.1, 0.05. Furthermore, we took q equal to 0.01 and 0.005 as before. The distribution of inbreeding coefficients among the affected individuals was chosen as in the simulation study for checking the performance of the estimator. In a first study we simulate 10,000 individuals to investigate the amount of bias by wrongly assuming one unknown mutation; for large n, the variation in the estimates is small, making a possible bias more visible. Next we took n =100 to see the bias in practice. Even if bias exists, but the bias is small compared to the variance of the estimator, it cannot be determined whether the maximum likelihood is an over- or underestimate of q. After simulating the data, the alleles that belong to one of the four undetected alleles are indicated as being missing. Next the total pathogenic allele frequency is estimated by the maximum likelihood estimator under the assumption that just one mutation has not been detected (whereas in the simulation model four alleles were unknown). Thus, all missing alleles are assigned to this fictive allele when estimating q. We repeated the simulation and estimation 1000 times.

RESULTS

Performance of estimator

The histograms of the 50, 100 and 1000 estimates as in (5) and the maximum likelihood estimates are shown in figures 1 and 2. For q= 0.005 we plotted only the maximum likelihood estimates. From the histograms it can be seen that the estimator given in (5) is biased upwards and is therefore not suitable for estimating q. The 1000 maximum likelihood estimates are distributed symmetrically around the true value of q (the black dot in the histograms). For increasing sample size, the variation in the estimates decreases and no systematic error is seen; the histogram is still symmetrical around the true value of q. Under some assumptions this can also be proved based on statistical theory for large sample sizes, see for example Van der Vaart (1998).

Multiple simulation studies have been performed for different settings; i.e. different values of q, distributions of the inbreeding coefficient, the number of alleles and relative allele frequencies and different sample sizes. In all cases, the maximum likelihood estimation method performed well.

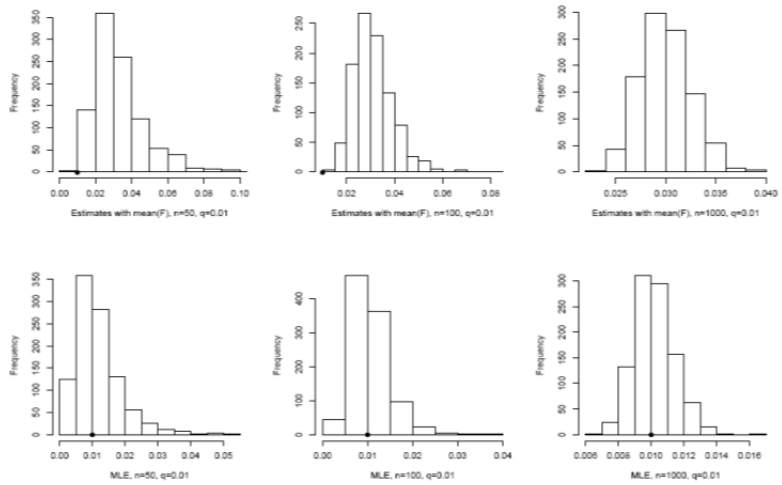

Figure 1. *Upper row: histograms of the estimates in (5), for sample sizes 50, 100, and 1000, from left to right Lower row: histograms of the maximum likelihood estimates, again for sample sizes 50, 100, 1000. In all cases q=0.01, represented as a black dot in the histograms. If no dot is visible, the dot is outside the plot.*

Undiscovered alleles

For q= 0.005 and q= 0.01, we computed the maximum likelihood estimators for q in the model with four known and four unknown mutations. In the latter model the four unknown mutations were modelled as one unknown mutation. There seems to be a slight overestimation of q in the four settings. We tried different settings, but in all cases the bias was small. We therefore conclude that the estimation method is quite robust against misspecification of the number of mutations.

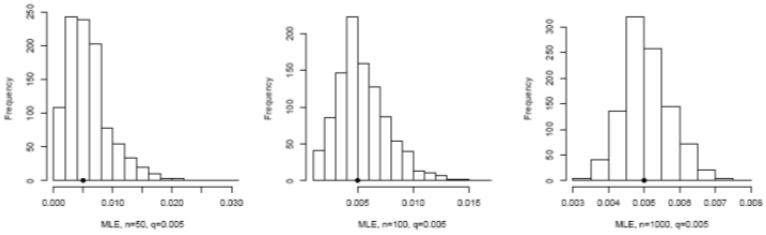

Figure 2. *Histograms of the maximum likelihood estimates for sample sizes 50, 100, 1000. In all cases q=0.05, represented as a black dot in the histograms.*

DISCUSSION

Estimation of the total pathogenic allele frequency based on mutational records of affected individuals can be done with the estimator in (4), but an estimate of the population inbreeding coefficient is needed. Finding an accurate estimate of this inbreeding coefficient is often very complicated in practice and may result in biased estimates of the allele frequency if no appropriate data are available. If for all affected individuals in the data set the inbreeding coefficient is available, the conditional maximum likelihood method (as described in this chapter) can be used for estimating the frequency of interest. This method does not require the population inbreeding coefficient to be estimated. Simulation studies showed that the maximum likelihood estimator is unbiased and accurate. It can also theoretically be proved that, in general, maximum likelihood estimators have desirable properties like asymptotic unbiasedness (i.e. for large data sets the estimated value is close to the true value) and optimality in the sense of small variance (see for instance van der Vaart, 1998, section 5.5;6 210). The confidence interval can be constructed with the help of the likelihood ratio test-statistic which has approximately a chi-squared distribution with one degree of freedom. In practice, even the seemingly unrelated parents of affected individuals can be actually related, but their inbreeding coefficient is small. Since in these cases their exact inbreeding coefficients are often unknown, it is convenient to take the inbreeding coefficient as equal to zero when computing the maximum likelihood estimates of q. Note that this procedure may affect the unbiasedness of the estimator. We performed a simulation study to consider the robustness of the method for setting small inbreeding coefficients as equal to zero. In this simulation study we assumed that all individuals (25% of the subset) with an inbreeding coefficient strictly smaller than 1/64 was falsely taken equal to zero. Although the estimator slightly overestimates the true value of q, in our simulation studies the amount of overestimation was always below two percent. It also may happen that closely related individuals do not know they are related or do not want to admit they are related (for whatever reason). In these cases the inbreeding coefficient is falsely taken as equal to zero. As long as this happens completely at random (i.e. it happens among all individuals regardless of the value of the inbreeding coefficient), there will be no bias in estimating q, since the individuals with $F=0$ do not affect the estimation of q. The data from these individuals are only used for estimating the relative allele frequencies (and the inbreeding coefficient is not used for this).

In the chapter we assume that all affected individuals are unrelated. By making this assumption, the likelihood for the total sample is equal to the product of the individual likelihoods. If there are multiple individuals from the same family, the assumption is violated and the likelihood will get a different form that includes the relationship between these affected individuals. However, ignoring the family bands and assuming

independence does not affect the (asymptotic) unbiasedness, but may affect the probability coverage of the confidence interval slightly.

The maximum likelihood estimator was constructed for autosomal recessive diseases; an individual needs to have two mutations to be affected. If the disease is dominant, the maximum likelihood can of course also be applied with a different likelihood. However, sample sizes will be very small in practice, because heterozygosity for autosomal dominant disorders is in general much rarer than it is for autosomal recessive disorders. Once the maximum likelihood estimate of q has been determined, an estimate of the population inbreeding coefficient can be found via the equation in (4) by filling in the estimate of q and solving F_{pop} from the equation. Although this was not the aim of our study, and there are probably better and more accurate methods that can be applied to estimate the population inbreeding coefficient, it is a pleasing result. We describe the estimation of the total pathogenic allele frequency in the case where the laboratory cannot detect all disease alleles, because not all mutations have been identified or routinely tested in the laboratory. However, it may also happen that individuals have the disease symptoms, but do not have the disease. In that case, no disease alleles are found and the individual can be discarded from the data. In the case where both situations happen simultaneously, i.e. symptoms but no disease, and undetected alleles, it is not possible to distinguish the underlying reason for not finding disease alleles. This problem is described in detail in Chapter 7.

In this chapter we described the methodological aspects of estimating the total pathogenic allele frequency for a recessive disorder, but we did not apply the method to real data. For an application of the method we refer to Chapter 7, where two data sets are described and the pathogenic allele frequency was estimated by the maximum likelihood estimate. Moreover, confidence intervals of the frequency are given.

APPENDIX (PROOF OF $P(\Delta=1|X,F)= P(\Delta=1|X)$)

In this appendix we will prove that $P(\Delta=1|X,F_{pop})$ is equal to $P(\Delta=1|X)$ by showing that $P(X|F_{pop})=P(X)$ and $P(\Delta=1,X|F_{pop})= P(\Delta=1,X)$.
Let F_{pop} be the population inbreeding coefficient in the population of interest. The probability

$$P\langle \Delta = 1 | X, F_{pop} \rangle = \frac{P\langle \Delta = 1, X | F_{pop} \rangle}{P\langle X | F_{pop} \rangle}$$

.

The denominator equals $P(X|F_{pop})= F_{pop}q+(1-F_{pop})q^2$. If F_{pop} is the population inbreeding coefficient in this population, the probabilities that an arbitrary individual from this population has two pathogenic alleles identical-by-descent (IBD) and non-IBD respectively, are F_{pop} and 1- F_{pop}. So the probability that this arbitrary individual has two disease alleles IBD and non-IBD respectively, are $F_{pop}q$ and $(1- F_{pop})q^2$. Combining these two yields the probability that an arbitrary individual has two disease alleles, or has disease X, is given by $P(X)= F_{pop}q+(1- F_{pop})q^2$: Conclude that $P(X|F_{pop})=P(X)$. This can also be seen as follows. Let G be the distribution function of the inbreeding coefficient of a random individual in the total population, then

$$P(X) = \int_0^1 P\langle X | F = f \rangle dG(f) = \int_0^1 \left(fq + (1-f)q^2 \right) dG(f)$$
$$= q\int_0^1 f dG(f) + q^2 \int_0^1 (1-f) dG(f) = F_{pop} q + \left(1 - F_{pop}\right) q^2$$
$$= P\langle X | F_{pop} \rangle,$$

since $\int_0^1 f dG(f)$ is the average inbreeding coefficient in the population and thus equal to F_{pop}. Similarly, it can be shown that $P(\Delta=1,X|F_{pop})= P(\Delta=1,X)$. This results in

$$P\langle \Delta = 1 | X, F_{pop} \rangle = \frac{P\langle \Delta = 1, X | F_{pop} \rangle}{P\langle X | F_{pop} \rangle} = \frac{P(\Delta = 1, X)}{P(X)} = P\langle \Delta = 1 | X \rangle$$

Similar computations hold for $P(\Delta=0|X,F_{pop})$.