A Note on Ascertainment Probability in the Allen/Hrubec Twin Model

Niels V. Holm

University Institute of Medical Genetics and the Danish Twin Register, Odense, Denmark

In their general model of twin concordance Allen and Hrubec recently introduced the concept of secondary ascertainment rate to consider partial ascertainment of secondary cases. This concept is found to be of limited interest because only pairs where valid information can be obtained from both twins should be included in genetic analyses. It is furthermore shown that the validity of the proband method and the Allen/Hrubec model depends on the equivalence of ascertainment in twins from concordant and discordant pairs. However, if this condition is fulfilled, it has been shown how the ascertainment probability in cotwins of probands is related to the overall ascertainment probability in the total twin population.

Key words: Twin study methodology, Ascertainment, Proband concordance rate

INTRODUCTION

The purpose of many twin studies is to evaluate the relative importance of genetic factors for the development of a given trait or disease. This is usually accomplished by the classical twin method, where the concordance rates in monozygotic (MZ) and dizygotic (DZ) twin pairs are compared. The most meaningful parameter here is the proband concordance rate as it gives the probability of the cotwin being affected [8]. Therefore, the proband concordance rate provides the best estimate of the risk of developing the disease in question in individuals who are partners of an affected twin [2]. Furthermore, the proband concordance rate may be interpreted as an estimate of the correlation coefficient in liability between relatives if the trait is not inherited in a simple mendelian manner [7]. Though the proband concordance rate in theory is simple to calculate, many complicating factors may arise in practice as discussed by Smith [7], eg, sex difference in occurrence of the trait/disease, varying age at onset, variation as to severity, and difficulties of diagnosis in addition to the problems related to the ascertainment process.

Among these factors, ascertainment has first and foremost been dealt with in many papers, of which the most recent was published by Allen and Hrubec [3]. These authors discuss previous contributions to the issue and then present what they call a more general

model of twin concordance. They introduce in their paper concepts of primary and secondary ascertainment rate.

The purpose of this note is in the first place to examine the consequences of Allen and Hrubec's "secondary ascertainment" and to show its limitations. Next, the conditions that must be fulfilled for the proband method as well as the Allen/Hrubec model to be valid will be examined and illustrated by a simple example.

Some additional points of major importance in the use of twin data in etiological research, such as the complications caused by variations with respect to age of onset, and in study of disease rate between groups differently exposed to exogenous noxious influences were disregarded by Allen and Hrubec; therefore, no discussion of these aspects are included here.

DEFINITION OF CONCEPTS

Studies of twins deal with pairs of individuals, where each pair is either MZ or DZ, but as demonstrated by Allen [1] it is usually advantageous to express frequencies in a given twin sample in terms of twin individuals. This approach will be followed throughout the present paper unless otherwise specified. As far as possible the same definitions and notation as those used by Allen and Hrubec will be adopted, but to prevent misunder-standings the definition of some basic concepts will be repeated here.

Applying Morton's definition of a proband [6] in twin studies, a proband is a twin with a given trait or disease who is ascertained independently of his/her cotwin. A secondary case is an affected twin individual whose detection is dependent on his/her cotwin being a proband; ie, a secondary case becomes known to the investigator only because his/her cotwin is already included in the study as a proband; thus a secondary case can never be a proband.

When the probands are independently ascertained, the ascertainment probability, or conditional probability that an individual of a population will be detected, given affection, is defined [4,6] as $\pi = P$ (proband/affection).

π is a parameter with an unknown value with the limits of zero and unity, but its value is estimated in a given twin population by

$$\hat{\pi} = \frac{\text{No. of probands}}{\text{No. of affected twin individuals in the target population}} \tag{1}$$

$\hat{\pi}$ thus equals m as used by Allen and Hrubec, and m will be used for estimates of the ascertainment probability in the present paper. As seen above, m is defined as a dimensionless proportion.

If m, as assumed by Allen and Hrubec, varies over different categories of twins (such as twins from discordant and concordant pairs) m may be thought of as a weighted average of category-specific ascertainment probabilities, m_i, with weights proportional to the number of affected twins in each category. With k categories this gives

$$m = \frac{\sum_{i=1}^{k} w_i m_i}{\sum_{i=1}^{k} w_i} \tag{2}$$

where w_i is the number of affected individuals in the i-th category.

In epidemiology the (point) prevalence of a disease is defined as the proportion of a specified population that exhibits the disease at a specified point of time [5]. Thus the prevalence, p, in a twin population is given by

$$p = \frac{\text{No. of affected twin individuals}}{\text{Total No. of twin individuals}} \tag{3}$$

The prevalence of the disease in cotwins of affected twins is denoted p_r and is given by

$$p_r = \frac{\text{No. of affected cotwins of affected twins}}{\text{Total No. of cotwins of affected twins}} \tag{4}$$

The proband concordance rate in a twin sample is denoted by c_{pb} and is given by

$$c_{pb} = \frac{\text{No. of affected cotwins of probands}}{\text{Total No. of cotwins of probands}} \tag{5}$$

If all affected twins in the target population are probands, the numerator and denominator of Equations 4 and 5 will be identical and so $p_r = c_{pb}$. If not all affected twins are probands, the relation between c_{pb} and p_r will depend on the ascertainment probability, as will be discussed later.

ALLEN/HRUBEC'S RATE OF SECONDARY ASCERTAINMENT

When a given study is started, the first stage is to find the affected individuals in the twin population, ie, the probands; only cases detected at this moment are by definition probands. Their proportion of the total number of affected twins in the twin population

TABLE 1. *Key to Abbreviations Used in Text*

p	Disease or trait prevalence among twin individuals
p_r	Disease prevalence among cotwins of affected twins
m	Ascertainment probability in the total twin population
m_c	Ascertainment probability among twins from concordant pairs
m_d	Ascertainment probability among twins from discordant pairs
m_r	Ascertainment probability among cotwins of probands from concordant pairs
m_0	Ascertainment probability among cotwins of nonproband twins from concordant pairs
m'	Allen/Hrubec's rate of secondary ascertainment among cotwins of primarily ascertained twins
c_{pb}	Proband concordant rate
N	Total number of twin pairs in the population including those with and those without the disease or trait
C	Total number of concordant pairs
C_0	Number of concordant pairs with zero probands
C_1	Number of concordant pairs with only one proband. The number of probands from these pairs is denoted $C_{1(P)}$, which is equal to the number of secondary cases $C_{1(S)}$ from these pairs
C_2	Number of concordant pairs with two probands
D	Total number of discordant pairs
D_0, D_1	Number of discordant pairs with zero and one proband, respectively
U	Total number of pairs where both twins are unaffected

is, according to Equation 1, equal to m. In the second stage of a twin study, information concerning presence/absence of the disease or trait in question must be obtained from all cotwins of probands. Cotwins who fulfil the diagnostic proband criteria are included in the study sample as affected, but no affected cotwin detected at this stage can ever become a proband.

When the twin investigator collects information on cotwins that are the individuals forming the basis of all genetic analyses, the following situations might occur:

A. The cotwin is a proband himself.
B. The cotwin is not a proband and
    (1a). has died before the manifestation period of the trait or disease;
    (1b). has emigrated or is for other reasons inaccessible;
    (2). has died before the time of the implementation of the twin study after having survived a certain amount of the manifestation period covered by the study;
    (3). is still alive and accessible to examination of the twin investigator.

Cotwins in group A need no further examination as they are already included in the study sample. Cotwins in group B1 cannot give any information concerning the risk in cotwins and therefore such pairs are excluded from genetic analyses. For cotwins in group B2 the twin investigator has to try to obtain valid information concerning presence/absence of the trait or disease. If valid information is obtained for either absence or presence of the trait or disease, then such pairs can be included in the genetic analyses. In cotwins where no valid information can be obtained, the pairs have to be excluded from genetic analyses as they give no useful information concerning the risk in cotwins and because no statistical procedure exists that can correct invalid data; so they are treated like group B1. The optimal situation exists with cotwins still alive and accessible to examination (group B3), as this most often will secure a correct classification into affected/unaffected. However, even here the twin investigator may feel that the information obtained from the cotwin cannot be considered valid, eg, the cotwin is mentally defective and no other sources of information are available; then such pairs have to be excluded from genetic analyses and be treated like group B1, as they give no useful information concerning the risk in cotwins.

Allen and Hrubec [3] introduce the measure m′, which they call secondary ascertainment rate, denoting the proportion of affected cotwins identified by the twin investigator among the total number of affected cotwins who are not probands; that is,

$$m' = \frac{\text{No. of identified affected cotwins who are not probands}}{\text{Total No. of affected cotwins who are not probands}}$$

As described by Allen and Hrubec the measure m′ is introduced to consider partial ascertainment of secondary cases, but it is not clear under which circumstances the secondary ascertainment is of importance and how to get valid estimates of concordance rates if secondary ascertainment must be considered incomplete.

Now, ascertainment in genetic studies traditionally relates to probands [4,6] as shown in Equation 1. The secondary rate, m′, does not have this property because m′ by definition is related to secondary cases. Furthermore, if a valid estimate of the importance of genetic factors should be obtained from twin studies, it is evident that only pairs in which information from the partner has been obtained with regard to the trait or disease in

question could be included in genetic analyses as discussed above. This appears also clearly from the fact that the classical twin method is based on the proband concordance rate, and as shown in Equation 5 this rate depends solely on the proportion of affected cotwins. Thus, if valid information concerning presence/absence of the trait or disease cannot be obtained from a cotwin whether he is alive or not, this pair must be excluded from genetic analyses as it gives no useful information. This is completely analogous to a family study where it is necessary to leave out probands when no information can be provided about the relatives included in the study. When this general principle is applied to twin studies, it becomes totally irrelevant to consider partial ascertainment of secondary cases. In fact, if no information is available on cotwins, the study is not a twin study but instead a study of trait or disease in individuals characterized by being born as twins.

As a consequence, the use of the concept of secondary ascertainment should be discontinued.

THE ESTIMATION PROCEDURES

Even if only pairs where cotwins' information is available are included in genetic analyses, there still remains an additional problem with the ascertainment probability for the proband method and the Allen/Hrubec twin model to be valid, namely, that the ascertainment probabilities (the Allen/Hrubec's primary ascertainment rates) are identical in twins from concordant and discordant pairs. This might be clarified by a simple example as follows.

For simplicity consider a congenital defect easily diagnosed at birth, eg, cleft lip with or without cleft palate [CL(P)], where the exact number of twin pairs, N, in the population is known through a complete twin register. A register of birth defects is in operation in this population, but for one reason or another the register is not complete. Now, an investigator wants to study CL(P) in twins. The affected twins are ascertained through a record linkage of the twin register and the register of birth defects. All affected twins ascertained through the record linkage are classified as probands. The investigator then examines all cotwins of the probands, and as CL(P) is easily diagnosed, all cotwins are correctly classified as affected or unaffected.

In Table 2 the total number of twin pairs in the population has been shown with regard to disease status and inclusion in the study. In C_2 of the C concordant pairs both twins are probands; in C_1 pairs only one twin is a proband, whereas the other twin in the pair is included as a secondary case. In C_0 pairs neither twin is a proband and therefore neither is included in the study material. Of the D discordant pairs only D_1 pairs are included.

TABLE 2. *Number of All Twin Pairs in the Population by Disease and Proband Status of Each Pair*

Number of probands in each pair	Number of twin pairs with at least one affected twin individual		Number of twin pairs with both twins unaffected	Total
	Concordant	Discordant		
2	C_2	. . .	. . .	C_2
1	C_1	D_1	. . .	$C_1 + D_1$
0	C_0	D_0	U	$C_0 + D_0 + U$
Total	C	D	U	N

TABLE 3. *Number of All Twin Individuals in the Population by Disease and Proband Status of Each Twin Individual in Relation to Type of Pairs*

Type of pairs	Individuals				Total number of twin individuals
	Number of affected individuals being				
	Probands	Secondary cases	Not ascertained	Unaffected	
Affected					
Concordant	$2C_2 + C_{1(P)}$	$C_{1(S)}$	$2C_0$	. . .	$2C$
Discordant	D_1	. . .	D_0	$D_1 + D_0$	$2D$
Unaffected concordant	. . .	. . .	. . .	$2U$	$2U$
Total	$2C_2 + C_{1(P)} + D_1$	$C_{1(S)}$	$2C_0 + D_0$	$D_1 + D_0 + 2U$	$2N$

Note that C, D, and U in this context refer to the true but unknown number in the total twin population of concordant, discordant, and unaffected pairs, respectively. This notation is different from the Allen/Hrubec notation which, however, can be calculated from Table 2. If the subscript A-H refers to Allen and Hrubec's definition, we have: $C_{A-H} = C_2 + C_1$, $D_{A-H} = D_1$, and $U_{A-H} = U + C_0 + D_0$.

In Table 3 the number of twin individuals in the population is given by disease and proband status of each twin individual in relation to the number of affected individuals in the twin pairs. Note that the number of probands from concordant pairs with only one proband is denoted $C_{1(P)}$, which equals the number of secondary cases $C_{1(S)}$ from these pairs. In all $(2C_2 + C_{1(P)} + C_{1(S)} + 2C_0) + (D_1 + D_0)$ twins are affected; of these, $2C_2 + C_{1(P)} + D_1$ are probands and $C_{1(S)}$ individuals are secondary cases, whereas $2C_0 + D_0$ are not ascertained. The number of unaffected twins equals $D_1 + D_0 + 2U$.

If it is assumed that we know all numbers in the different categories in Tables 2 and 3, it is possible to estimate all parameters of interest.

From Equation 3, the prevalence in the twin population, p, is

$$p = \frac{2C + D}{2(C + D + U)} = \frac{2C + D}{2N} \tag{6}$$

and, from Equation 4, the prevalence among cotwins of affected twins, p_r, is

$$p_r = \frac{2C}{2C + D} \tag{7}$$

The ascertainment probability can be estimated for different categories of twins. If m_c and m_d refer to the ascertainment probabilities in twins from concordant and discordant pairs, respectively, substituting from Equation 1 gives

$$m_c = \frac{2C_2 + C_{1(P)}}{2C_2 + C_{1(P)} + C_{1(S)} + 2C_0} = \frac{2C_2 + C_{1(P)}}{2C} \tag{8}$$

and

$$m_d = \frac{D_1}{D_1 + D_0} = \frac{D_1}{D} \tag{9}$$

The estimate of the overall ascertainment probability, m, in the twin population is, according to Equation 2,

$$m = \frac{m_c 2C + m_d D}{2C + D}$$

From Equation 5, the proband concordance rate, c_{pb}, can be estimated and, substituting from Equations 8 and 9,

$$c_{pb} = \frac{2C_2 + C_{1(S)}}{2C_2 + C_{1(S)} + D_1} \tag{10}$$

$$= \frac{m_c 2C}{m_c 2C + m_d D} = \frac{2C}{2C + \dfrac{m_d}{m_c} D}$$

because $C_{1(S)} = C_{1(P)}$.

Comparison of Equations 7 and 10 then shows that the proband concordance rate gives a valid estimate of the prevalence in cotwins of affected twins only if the ascertainment probability is identical in twins from concordant and discordant pairs. When $m_c > m_d$ then $c_{pb} > p_r$ and when $m_c < m_d$ then $c_{pb} < p_r$. As D in most twin studies is higher than 2C, a difference between m_c and m_d may be assumed to influence the estimate of p_r through c_{pb}.

In the Allen/Hrubec twin model no distinction was made between ascertainment probability in twins from concordant and discordant pairs as shown here. Concerning the ascertainment process they assumed that "in the first stage of ascertainment a certain proportion, m, of affected twins in the population are detected. However, ascertainment is correlated within pairs, so that the probability, m_r, of detecting the second partner of a concordant pair in the primary process is greater than m. The overall probability for twins remains m because under our assumption, excess ascertainment in some pairs is compensated by deficient ascertainment in others" [3]. The last sentence implies that the ascertainment probability in twins from concordant and discordant pairs is assumed to be the same.

When the overall ascertainment probability in the total twin population is m, if a subgroup of the twins have a higher ascertainment probability, m_r, then another subgroup of the twins must have a lower ascertainment probability than m. This lower probability will be called m_0 and so $m_r > m > m_0$. From Allen/Hrubec's Figure 2 [3] all true discordant pairs are assumed to have an estimated ascertainment probability equal to m. Therefore, m_r and m_0 only apply to twins from concordant pairs; the groups of twins from concordant pairs having the ascertainment probabilities m_r and m_0 will be specified below. However, if the ascertainment probability in twins from discordant pairs, m_d, is assumed to be identical to the overall ascertainment probability in twins, m, then, ac-

cording to Equation 2, the overall ascertainment probability in twins from concordant pairs, m_c, also equals m and then $m_c = m_d$

As the investigator only knows the exact number of twin pairs born, N, and the pairs with at least one proband, (ie, C_2, C_1, and D_1), he would never know if the ascertainment probability is identical in twins from concordant and discordant pairs. However, it is possible to test whether $m_c = m_d$ if the following conditions are fulfilled: The disease prevalence is assumed to be identical in twins and singletons and this prevalence is known. Furthermore, the ascertainment probability in singletons is known, and this probability can be assumed to be identical with the ascertainment probability in twins from discordant pairs. It is thus assumed that p and m_d are known. Solving for D in Equation 9,

$$D = \frac{D_1}{m_d} \tag{11}$$

Substituting from Equation 11 and solving for C in Equation 6,

$$C = pN - \frac{D_1}{2m_d} \tag{12}$$

Finally, from Equations 8 and 12,

$$m_c = \frac{2C_2 + C_1}{2pN - \dfrac{D_1}{m_d}}$$

which can then be compared with m_d.

If the ascertainment probability is the same in twins from concordant and discordant pairs, ie, $m_c = m_d = m$, the correlation between the ascertainment probability among cotwins of probands, m_r, and m is highly dependent on m_0, the ascertainment probability in cotwins of nonproband twins from concordant pairs.

Following Allen and Hrubec, all affected twins from concordant pairs can be divided into two groups according to whether their cotwin is a proband or not. In all, $2C_2 + C_{1(S)}$ affected twins have a proband as cotwin but only $2C_2$ of these are themselves probands. The ascertainment probability in cotwins of probands from concordant pairs, m_r, is then given by

$$m_r = \frac{2C_2}{2C_2 + C_{1(S)}} \tag{13}$$

The remaining number of affected twins from concordant pairs is $C_{1(P)} + 2C_0$, and as seen from Table 3, their cotwins are not probands. Of $C_{1(P)} + 2C_0$ affected twins only $C_{1(P)}$ are probands and therefore the ascertainment probability, m_0, among cotwins of nonproband twins from concordant pairs is

$$m_0 = \frac{C_{1(P)}}{C_{1(P)} + 2C_0} \tag{14}$$

The overall ascertainment probability, m_c, in twins from concordant pairs given in Equation 8 may therefore be thought of as a weighted average of m_r and m_0, where the weights are the number of twins having these probabilities, ie:

$$m_c = \frac{m_r(2C_2 + C_{1(S)}) + m_0(C_{1(P)} + 2C_0)}{(2C_2 + C_{1(S)}) + (C_{1(P)} + 2C_0)}$$

Substituting from Equation 8,

$$m_c = m_r m_c + m_0(1 - m_c)$$

Finally, solving for m_c,

$$m_c = \frac{m_0}{1 - m_r + m_0}$$

Therefore, when $m_c = m$,

$$m = \frac{m_0}{1 - m_r + m_0} \qquad (15)$$

In the Figure, the correlation between m_r and m has been shown for different values of m_0. These results were derived from Equation 15. There is a positive correlation between m_r and m but the magnitude of this correlation is highly dependent upon the value of m_0, which can be illustrated by Allen and Hrubec's example of schizophrenia in the Norwegian twin population [3]. From this study, an overall ascertainment probability, m, of 0.80 was estimated, whereas the ascertainment probability in cotwins of probands, m_r, was estimated as 0.903 in MZ cotwins and 0.857 in DZ cotwins. Substituting these values in Equation 15 and solving for m_0 gives $m_{0(MZ)} = 0.388$ and $m_{0(DZ)} = 0.572$. Allen and Hrubec assumed that m_r was higher than m, but neither in Equation 15 nor its derivation is this constraint necessary. The equation is valid in all situations, ie, $m_r > m$, $m_r = m$, and $m_r < m$.

CONCLUSIONS

A twin study implies that information concerning presence/absence of the trait or disease in question must be obtained about all cotwins of probands whether they are alive or not in order for them to be included in genetic analyses. If no information can be obtained about a cotwin, the pair must be excluded from genetic analyses, as it gives no useful information, just as in a family study probands must be left out when no information can be provided about relatives. Thus, neither the probands nor the cotwins have to be alive at the time of the implementation of the twin study (this of course might result in a biased twin sample in studies of diseases with any appreciable mortality), but the twin investigator must collect valid information about all twins if they are to be included in the analyses. Only this procedure will give valid estimates of parameters of interest. Following this general rule in genetics, the concept of incomplete secondary ascertainment among cotwins of probands introduced by Allen and Hrubec is irrelevant. Furthermore, ascertainment
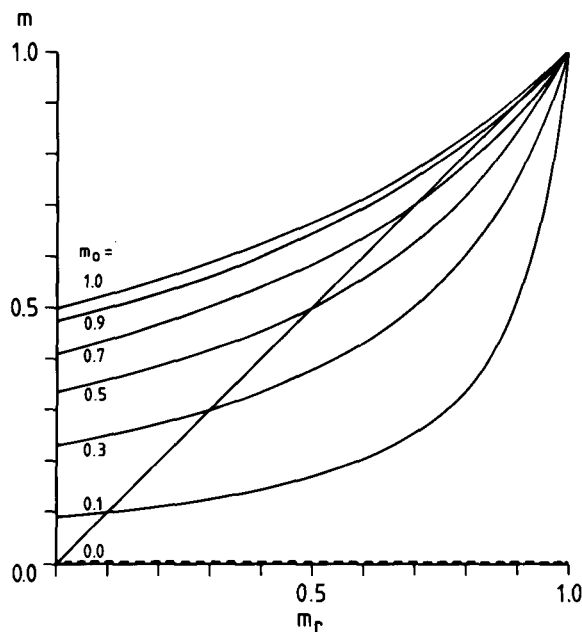
Figure.   Graph showing relation between the ascertainment probability in all twins combined, m, and cotwins of probands from concordant pairs, m_r, for different values of the ascertainment probability in cotwins of nonproband twins from concordant pairs, m_0. The diagonal represents the situation where $m = m_r = m_0$.

in genetic studies traditionally relates to probands, but secondary ascertainment does not have this property as it only relates to secondary cases. As a consequence, the use of the concept of secondary ascertainment should be discontinued.

An unrelated but important problem not discussed in detail by Allen and Hrubec is that the ascertainment probabilities must be identical in twins from concordant and discordant pairs if the proband concordance rate is to give an unbiased estimate of the disease prevalence in cotwins of affected twins, as demonstrated in Equation 10. This condition must also be fulfilled for the Allen/Hrubec twin model to be valid. However, if this condition is fulfilled it has been demonstrated how the ascertainment probability in cotwins of probands is related to the overall ascertainment probability in the total twin population.

Some further issues of major importance in twin studies, such as the complications caused by variation with respect to age of onset, and variation in disease rate between groups differently exposed to exogenous noxious influences, were not considered in detail by Allen and Hrubec, but these will be analyzed in a subsequent paper.

ACKNOWLEDGMENTS

REFERENCES

1. Allen G (1955): Comments on the analysis of twin samples. Acta Genet Med Gemellol 4:143–160.
2. Allen G, Harvald B, Shields J (1967): Measures of twin concordance. Acta Genet 17:475–481.
3. Allen G, Hrubec Z (1979): Twin concordance. A more general model. Acta Genet Med Gemellol 28:3–13.
4. Lalouel JM, Morton NE (1981): Complex segregation analysis with pointers. Hum Hered 31:312–321.
5. MacMahon B, Pugh TF (1970): "Epidemiology. Principles and Methods." Boston: Little, Brown, pp 61–66.
6. Morton NE (1959): Genetic tests under incomplete ascertainment. Am J Hum Genet 11:1–16.
7. Smith C (1974): Concordance in twins: Methods and interpretation. Am J Hum Genet 26:454–466.
8. WHO (1966): The use of twins in epidemiological studies. Report of the WHO meeting of investigators on methodology of twin studies. Acta Genet Med Gemellol 15:109–128.

Correspondence: Dr. Niels Holm, University Institute of Medical Genetics, Sygehuset, DK-5000 Odense, Denmark.