



Received 3 October 1978

Final 11 November 1978

Twin Concordance A More General Model

G. Allen¹, Z. Hrubec²

¹*Division of Biometry and Epidemiology, National Institute of Mental Health, US Department of Health, Education, and Welfare, Rockville, Maryland;* ²*Medical Follow-up Agency, National Academy of Sciences, Washington, DC, USA*

Estimation of the twin concordance rate for a disease often requires two stages. First, the disease is ascertained in a population or in a population sample, and such twins as are found with the disease become probands. Second, twin pairs with only one proband are further investigated and additional concordant pairs are thus discovered. A mathematical model is presented that allows for continuous variation in completeness of ascertainment in both stages, for correlation within pairs in the primary ascertainment process, and for correlation within pairs in occurrence of the disease. The concordance rate can be estimated by the proband method if secondary ascertainment is complete; other measures of concordance are accurate only if primary ascertainment is complete. A parameter analogous to the concordance rate but related to correlation in primary ascertainment can be estimated from the same data.

Key words: Ascertainment, Concordance rate, Estimation, Proband method, Schizophrenia, Twin model

Twins are often used to study rare dichotomous traits, especially diseases, in terms of presence or absence of the trait and in terms of concordance rates in monozygotic (MZ) and dizygotic (DZ) pairs. The objective is to assess the frequency with which similarity of genotype, as in MZ twins, determines similarity of expression of the trait or disease.

This kind of analysis contrasts with the study of variance and correlation that is convenient for quantitative traits. Quantitative methods can be applied to qualitative or threshold traits by postulating an underlying continuous variable, liability [2, 3], but even this method, when applied to twins, relies upon estimation of concordance rates [13].

It has long been recognized that the proportion of concordant twin pairs, known as the pairwise concordance rate, will vary for a given disease according to the completeness with which cases are ascertained in the sample [1, 15]. A fractional sample could result either from partial ascertainment throughout the population or from high ascertainment

in a geographically restricted segment of the population. Adult twins are likely to live in different cities or to attend different medical facilities, and in this situation concordant pairs are almost twice as likely as discordant pairs to be represented in a small fractional sample.

The problem is similar to one encountered in estimating genetic ratios from a sample of affected sibships, for which an early and still acceptable solution is to count a family once for each member who is independently ascertained [17]. Following Morton [11] we will refer to all independently ascertained members as probands, and reserve the term, index case, for the first proband in each sibship or twin pair. When correctly applied to twins, Weinberg's "proband method" should yield the same relatively high concordance rate under complete ascertainment as under minimal ascertainment, and this rate is the appropriate one for comparison with true prevalence in the general population. The problems of defining and estimating the true prevalence will be discussed below.

The literature on estimation of twin concordance is voluminous and continues to grow. To mention only some of the recent papers: Gedda and Brenci [5] and Selvin [12] proposed similar models of concordance that did not consider ascertainment problems. Vollmer [16] treated concordance concepts using set and probability theory. His parameters for completeness of ascertainment were expressed in terms of the ratio of concordant to discordant pairs, which does not lend itself to easy interpretation. Two papers that relate most directly to the formulations given here are those of Hrubec [7] and Smith [14].

Smith formulated a general model that took account of disease prevalence or trait frequency, rates of diagnosis and ascertainment, and the possibility that these parameters are elevated for the second member in twin pairs. He referred to secondary investigation of cotwins of probands and acknowledged that "... then the diagnosis rate may be higher than in the population ...", but his mathematical model failed to distinguish between correlated primary ascertainment of concordant pairs and secondary ascertainment of additional affected cotwins.

Hrubec studied the relative frequency of observed concordant pairs as a function of completeness of population screening, and compared these results with those obtained when such primary ascertainment is followed by "conditional" ascertainment of affected secondary cases in the same pairs. He showed that if secondary ascertainment is complete and the concordance rate high, the primary ascertainment rate has little influence on the estimate of pairwise concordance. His analysis did not consider partial ascertainment of secondary cases, and his formulas did not express in simple form the error or discrepancy between estimated and absolute concordance rates. The treatment proposed here addresses both of these topics.

THE MODEL

Again following Morton's usage [11] we shall differentiate probands, who are "independently ascertained", from secondary cases, whose ascertainment depends upon prior ascertainment of a relative. This terminology implies a stochastic, sequential process of ascertainment with two stages, the second conditional upon the first. Twin data may thus describe not only the distribution of disease within pairs, but the distribution of primary and secondary ascertainments. A detailed model of twin concordance should represent these two stages and should accommodate continuous variation of the parameters.

It must also allow for correlation between partners in likelihood of primary ascertainment, because in any twin study, even using the most complete twin register, geographic and social heterogeneity will influence ascertainment of twin partners in the same way. Such correlation does not necessarily invalidate the findings, as our model demonstrates.

In the model that we develop below, ascertainment is assumed to occur through a twin registry or a population survey and, in the primary stage, to detect cases by screening individuals without recourse to information about their cotwins. We have not differentiated ascertainment from diagnosis since they are mathematically similar [14]. The first-stage ascertainment may provide an estimate of population prevalence, and in that case the twin probands may be viewed as part of the disease experience in the sampled population.

The model can also be applied to other situations when certain assumptions are valid. For example, if twin probands are ascertained in an institution for which the source population is not defined, assumptions must be made about the equivalence of the disease as found in the institution and as ascertained for the estimate of prevalence in the population.

The disease or trait is assumed to be fully expressed at all ages, or at least in that age range counted as “the population”. When an age-dependent condition is studied, corrections are needed to convert prevalence to lifetime incidence, not only to facilitate genetic interpretation, but to provide valid concordance estimates. Even if singletons, MZ twins and DZ twins all have a similar age distribution, age dependence may differentially affect estimates for MZ and DZ twins. We have not tried to incorporate age corrections into our model.

We assume that for the disease in question there exists a “true” prevalence rate, p , in the population and that the same rate holds among twins and singletons. We also assume a positive correlation in twins with respect to the trait, ie, concordance in twins exceeds chance, so that prevalence in the cotwins of affected members, p_r , is greater than p .

These two variables alone, p and p_r , describe the distribution of disease among twins in a population, as shown in Figure 1. The central square area, divided in quarters, represents all twin pairs in the analysis (for example, all MZ pairs) cross-classified according to the status of the two partners, whether with “Disease” or “No Disease” as labelled at the top of the columns, Twin “A”, and at the left side of the rows, Twin “B”. Twins are assumed to be designated A and B randomly with respect to traits of interest, and the two aggregates are therefore entirely alike. Since p and p_r are proportions, the four squares add to unity as indicated at the lower right. As p is the true disease prevalence or trait frequency, so p_r is the true concordance rate, and we shall show, among other things, that p_r is the parameter estimated by the proband method under ideal conditions.

Notation

p	Disease or trait prevalence in the population and among twins
p_r	Disease prevalence among cotwins of affected twins
m	Rate of primary ascertainment in the population and among twins
m_r	Rate of primary ascertainment among cotwins of primarily ascertained twins
m'	Rate of secondary ascertainment among cotwins of primarily ascertained twins
c	Concordance rate; subscripts denote basis of computation: pw, pairs; cs, cases; pb, probands
N	Number of twin pairs, including those with and those without the disease or trait
C	Number of concordant pairs
C_1, C_2	Numbers of concordant pairs with one and two probands, respectively
D	Number of discordant pairs
Pb	Number of probands

		TWIN A		All Twins B
		Disease	No Disease	
TWIN B	Disease	pp_r	$p - pp_r$	p
	No Disease	$p - pp_r$	$1 - 2p + pp_r$	$1 - p$
All Twins A		p	$1 - p$	1.0

Fig. 1. Distribution of disease in twin pairs expressed in terms of prevalence rate (p) and concordance rate (p_r).

In the first stage of ascertainment a certain proportion, m , of affected twins in the population are detected. However, ascertainment like prevalence 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 assumptions, excess ascertainment in some pairs is compensated by deficient ascertainment in others. Primary ascertainment is independent for partners in the sense that both are screened without reference to their partner, and all cases detected in this phase are therefore properly defined as probands.

Probands' partners who are affected but not probands are detected in the second stage with frequency m' . This parameter will be zero when partners of probands are not subjected to secondary study. If partners of all probands are located and examined, m' will be virtually unity for a disease that is easily diagnosed. For diseases not so easily diagnosed, secondary cases detected by more focused examination are likely to include early or atypical cases, so that concordance rates may not be freely compared with population prevalence. This will be discussed below.

The superposition of ascertainment classification on twins already subdivided by the disease or trait is illustrated in Figure 2. In viewing these figures it should be remembered that labels apply to a twin, not a pair of twins. Thus, in Figure 2 the second column has the implied heading, "Proportion of Twins A affected with disease but not ascertained until the second stage".

The two main dividing lines of Figure 1 are represented in Figure 2 with the broken lines, as may be verified by observing the value, $1 - 2p + pp_r$ in the lowest right square. A twin

		TWIN A			
		Disease			No Disease
		Ascertained First Stage	Ascertained Second Stage	Missed 1st & 2nd Stages	
TWIN B	Disease				
	Ascertained First Stage	$m m_r p p_r$	$m m' (1-m_r) p p_r$	$m (1-m_r)(1-m') p p_r$	$m p (1-p_r)$
	Ascertained Second Stage	$m m' (1-m_r) p p_r$	0	0	0
No Disease	Missed 1st & 2nd Stages	$m (1-m_r) (1-m') p p_r$	0	$(1-2m + mm_r) p p_r$	$(1-m) p (1-p_r)$
	No Disease	$m p (1-p_r)$	0	$(1-m) p (1-p_r)$	$1 - 2p + pp_r$
Totals		$m p$	$mm'(1-m_r)pp_r$	$(1-m)p - mm'(1-m_r)pp_r$	$1 - p$
		$mp + mm'(1-m_r)pp_r$		$1 - mp - mm'(1-m_r)pp_r$	

Fig. 2. Cross-classification of twin pairs by status of each partner according to presence or absence of disease and stage in which he was ascertained, if at all.

with the disease or trait may fall in one of three categories with respect to ascertainment, as indicated in the margins: Proband, secondary case, or not ascertained. The marginal totals are shown only at the bottom under Twin A, but apply similarly to Twin B. Of particular interest is the small square in the upper left, containing pairs with two probands and exceeding chance expectation in the ratio $m_r p_r / mp$.

As in Figure 1, formulas in the various cells cannot be derived by multiplication of the marginal values, as would be the case if distributions of disease and primary ascertain-

		TWIN A	
		Ascertained	Not Ascertained or No Disease
TWIN B	Ascertained	$m m_r p p_r$ (C ₂) + $2 m m' (1-m_r) p p_r$ (C ₁) "Concordant" C	$m p (1-p_r)$ + $m (1-m_r) (1-m') p p_r$ "Discordant" ½ D
	Not Ascertained or No Disease	$m p (1-p_r)$ + $m(1-m_r) (1-m') p p_r$ "Discordant" ½ D	$1 - 2mp + m m_r p p_r$ "Unaffected" U

Fig. 3. Cross-classification of twin pairs by ascertainment status of each partner.

ment were random with respect to twin partners. The model assumes correlation in both processes.

A heavy line in the diagram (Fig. 2) indicates the observed distinction between concordant, discordant, and disease-free pairs, and in Figure 3 only those divisions are shown. Figure 3 simulates the simplicity of Figure 1, but each cell in Figure 3 is composed of two or more types of pairs, as shown in Figure 2.

FORMULAS FOR MEASURES OF CONCORDANCE

If C, D and U represent numbers of concordant, discordant and unaffected pairs, and C₂ and C₁, respectively, two-proband and one-proband concordant pairs, the usual measures of concordance are calculated as follows:

$$\text{Pairwise concordance} = c_{pw} = \frac{C}{C + D} \tag{1}$$

$$\text{Casewise concordance} = c_{cs} = \frac{2C}{2C + D} \tag{2}$$

$$\text{Proband concordance} = c_{pb} = \frac{2C_2 + C_1}{2C_2 + C_1 + D} \tag{3}$$

The casewise rate [6, 7, 8] differs from the proband rate in that all concordant pairs are counted twice without attempting to distinguish probands from secondary cases. It resembles Weinberg's [17] "Geschwistermethode" of analyzing complete sibships and Fisher's [4] "proband method". It is appropriate when there is no secondary ascertainment or when primary ascertainment is virtually complete, either in a twin population or in a sample of twin pairs assembled without reference to the disease or trait.

When these three measures are expressed in terms of the model parameters, it will be noted that both p , the prevalence, and m , the primary ascertainment rate in singletons, drops out of all the formulas:

$$c_{pw} = \frac{m_r + 2m' - 2m_r m'}{2 - m_r p_r} \cdot p_r \quad (4)$$

$$c_{cs} = \frac{m_r + 2m' - 2m_r m'}{1 + m'(1 - m_r)} p_r \quad (5)$$

$$c_{pb} = (m_r + m' - m_r m') \cdot p_r \quad (6)$$

The quantity p_r has been taken out of the numerator so as to express all three measures as nearly as possible as the product of p_r and a coefficient which represents effects of incomplete ascertainment. The pairwise and casewise rates both retain p_r intractably in the denominator, but the proband rate breaks clearly into the desired factors. Expression 6 shows that if either primary or secondary ascertainment is complete, the ascertainment factor becomes unity and the resulting value for proband concordance is simply p_r .

The expressions for pairwise and casewise concordance cannot be simplified much by bringing secondary ascertainment, m' , to unity. Complete primary ascertainment is more helpful. When primary ascertainment becomes complete, with $m = m_r = 1$, either in a defined, complete population or in a random sample of twin pairs, $1 - m_r$ becomes 0 and the parameter m' , irrelevant. The casewise rate is then identical with the proband as we should expect, because all cases are probands. Under such ascertainment, Formula 4 gives the "true" pairwise rate [14].

$$c_{pw} = \frac{p_r}{2 - p_r} \quad (7)$$

Some other relations and formulas are useful in making use of twin data. First, the total number of probands, P_b , is known:

$$P_b = 2C_2 + C_1 + D \quad (8)$$

This quantity can then be expressed in terms of the total number of twins, $2N$, together with the parameters applicable to the singleton population, provided that the conditions of our model are met. Prevalence of the trait or disease must be the same in twins and singletons and equally expressed at all ages that are included in the population. From Figure 2, the proportion of Twins "A" ascertained in the first stage is mp . A similar

proportion of Twins "B" become probands, making the total $2mp$ or, in numbers of individuals.

$$P_b = 2mpN \quad (9)$$

The total number of twins, $2N$, is presumably known, and if either m or p is known the other can be obtained from Formula 9. The value of p , prevalence, should be known for both twins and singletons to assure that twins are not more or less susceptible, but sometimes the prevalence is uncertain and ascertainment rate, m , can be determined, for example, from the frequency with which secondary cases are ascertained. Prevalence is then given as

$$p = \frac{P_b}{2mN} \quad (10)$$

Even if p and m are not known, a value can be obtained for m_r , the probability of primary ascertainment in affected cotwins. The proportion of double-proband pairs is given in the upper left cell in Figure 2:

$$C_2 = mm_r pp_r N$$

Substituting from Equation 10,

$$C_2 = \frac{1}{2} m_r p_r P_b$$

From Equations 3 and 6, when $m' = 1$ so that $p_r = C_{pb}$,

$$C_2 = \frac{1}{2} m_r \cdot \frac{2C_2 + C_1}{2C_2 + C_1 + D} \cdot P_b$$

Finally, substituting from Equation 8 and solving for m_r ,

$$m_r = \frac{2C_2}{2C_2 + C_1} \quad (11)$$

DISCUSSION

As stated above, complete primary ascertainment of a disease or trait yields all desired estimates without ascertainment bias, but it is usually impractical. Nearly the same end can be achieved by complete secondary ascertainment. If $m' = 1$, the coefficient of p_r in Expression 6 reduces to unity and one has a direct estimate of the proband concordance rate. This circumvents the problems of correlated, incomplete, primary ascertainment. There are, however, three difficulties. First, one may never be sure that secondary ascertainment is quite complete, but a similar doubt attaches to all population data. Second, as Weinberg [17] emphasized, the method requires accurate discrimination between true probands and conditionally identified, secondary cases. This is possible in a well differentiated, two-stage ascertainment process, where the first stage defines all probands. The validity of the definition in a given study can be assessed from the frequency of probands among all twins provided that (1) the probands are detected in the same case-finding procedure that defined the prevalence, and (2) twins have no differential susceptibility or resistance. Under these conditions, the frequency of probands among twins ought to be the same as the ascertained prevalence of the disease or trait in the population.

The third difficulty is that, if the prevalence of probands among twins, ie, the primary ascertainment rate, is the same as that of the disease in the population, any secondary ascertainment at all raises the prevalence rate in twins above that in singletons and implies a different definition of the disease [E. Kringlen, personal communication]. The general prevalence rate is needed together with the concordance rate to yield the correlation between twins when, for example, the disease is analyzed as a threshold trait. The primary prevalence rate would no longer serve this function. If secondary cases fall clearly within the diagnostic standards, the concordance rate is not in error; the problem is to obtain a comparable prevalence rate in the general population. This can be done by carrying out on a sample of the singleton population a secondary ascertainment procedure similar to that applied to cotwins of ascertained cases.

If secondary cases include more atypical or incomplete forms of the disease than do primary cases, as is usually true with psychiatric diagnoses, the disease may be regarded as varying quantitatively, and any use of a dichotomous classification is statistically inefficient and possibly erroneous. In more general terms, heterogeneity of risk or of etiology among affected families tends to invalidate any estimate of concordance rate. It has been suggested [C. Smith, personal communication citing N. Morton], that when risk is heterogeneous, the proband method estimates the sum of the concordance rate and its variance.

ILLUSTRATION

The use of the formulas and some of the problems to be expected in practice may be illustrated from the very thorough study of schizophrenic twins by Kringlen [9, 10]. Kringlen set up a twin registry in Norway that can be considered complete for the birth years 1901 through 1930. It included 25,000 pairs. Schizophrenia and other mental illnesses were ascertained primarily from a national registry of hospitalized mental patients. From his investigation of the families of affected twins he estimated that 80% of all persons with a functional psychosis in Norway are hospitalized and registered. This is probably an overestimate because of correlated ascertainment in sibs. If we accept his figure, we have a value of 0.80 for m .

Of 422 twin pairs with functional psychosis in one or both, Kringlen excluded 80, or 19%, because of early death of the cotwin and for other similar reasons, leaving 342 pairs. Of these, 126 were of opposite sex and accordingly DZ. Of the 216 same-sexed pairs, 75 were diagnosed as MZ and 131 as DZ. In 10 pairs zygosity remained unknown.

The above statistics are based, by choice, on Kringlen's entire sample of psychotics, but we shall discuss concordance only with respect to the patients who had schizophrenia (including schizophreniform psychosis). Sixty-nine schizophrenic probands were found in 55 MZ pairs, while among 90 DZ pairs there were 96 schizophrenic probands. Schizophrenia occurred in 6 of the undetermined pairs, all discordant. On examining the cotwins, Kringlen found 3 secondary cases among the MZ and 2 among the DZ pairs. It is not unlikely that he achieved complete secondary ascertainment: $m' = 1.0$.

Information is lacking on the frequency of schizophrenia in the whole population, but Kringlen states that he found it not to differ significantly from that in twins. In fact, some of his figures indicate an ascertained frequency lower in twins than in singletons, which supports the assumption that primary ascertainment was independent of presence of disease in the partner. We can best estimate the general prevalence by extrapolating from the frequency in twins. The total number of twin pairs in the registry is given as 25,000. If we

assume that the same proportion was broken by death as in the psychotic cases, this figure should be reduced by 19%, leaving 20,250 pairs or 40,500 individuals at risk. Of these, when we include 90 in opposite-sex pairs not further studied, 261 were schizophrenic, or 0.644%. Under Kringlen's estimated ascertainment rate of 80%, these 261 probands represent 326 cases altogether, and a total prevalence of 0.805%, our best estimate of p .

The parameters describing distribution of disease and ascertainment in the twins can be estimated separately for the MZ and DZ pairs. The following table gives the numbers needed to evaluate equations 3 and 11. It also gives the resulting estimates of the two model parameters, p_r and m_r , representing, respectively, the proband concordance rate, or prevalence in cotwins of affected persons, and the primary ascertainment rate in cotwins of ascertained persons:

	MZ	DZ
C_2	28	12
C_1	3	2
D	38	82
p_r	0.449	0.146
m_r	0.903	0.857

As expected, cotwin ascertainment, m_r , is higher in MZ twins than in DZ twins, but no significance can be attributed to the difference because it is based on such small numbers. Both values are higher than the figure of 80% estimated by Kringlen for m , the primary ascertainment rate, owing to the expected correlation within pairs. Since these cotwins were included in the estimate of m , their exclusion would have given an estimate below 80% and it probably would have been more accurate.

Despite detailed presentation of data in Kringlen's monograph, several additional items of information could have been used to advantage in the above analysis and ought to be given with twin data when available:

- (1) Prevalence of the disease in the general population;
- (2) Completeness of primary ascertainment, either among random individuals or, failing that, among sibs of probands, not including the cotwins;
- (3) The actual number of twins in the register. In this instance 25,000 appears to be a rounded figure, but it is accurate within 2.5% [Kringlen, personal communication];
- (4) Sex concordance figures for the entire twin register and for those pairs broken by early death; the first figure is available from the birth certificates usually employed in assembling a twin register.

Acknowledgments. This work was supported in part by contract NO1-HV-53010 with the National Heart, Lung, and Blood Institute through which the Twin Registry of the Medical Follow-up Agency, National Academy of Sciences-National Research Council is funded. Robert C. Elston and Charles Smith made helpful comments on an earlier draft.

REFERENCES

1. Allen G (1955): Comments on the analysis of twin samples. *Acta Genet Med Gemellol* 4:143-160.
2. Edwards JH (1960): The simulation of Mendelism. *Acta Genet Stat Med* 10:63-70.
3. Falconer DS (1965): The inheritance of liability to certain diseases, estimated from the incidence among relatives. *Ann Hum Genet* 29:51-76.
4. Fisher RA (1934): The effect of method of ascertainment upon the estimation of frequencies. *Ann Eugen* 6:13-25.
5. Gedda L, Brenci G (1966): Theoretical models in twin research. *Acta Genet Med Gemellol* 15:219-223.
6. Gottesman II, Shields J (1972): "Schizophrenia and Genetics." New York: Academic Press.

7. Hrubec Z (1973): The effect of diagnostic ascertainment in twins on the assessment of the genetic factor in disease etiology. *Am J Hum Genet* 25:15–28.
8. Hrubec Z, Allen G (1975): Methods and interpretation of twin concordance data (Letter to the editor). *Am J Hum Genet* 27:808–809.
9. Kringlen E (1966): Schizophrenia in twins. An epidemiological-clinical study. *Psychiatry* 29:172–184.
10. Kringlen E (1967): “Heredity and Environment in the Functional Psychoses. An Epidemiological-Clinical Twin Study.” Oslo, Universitetsforlaget; London: William Heinemann Medical Books.
11. Morton NE (1959): Genetic tests under incomplete ascertainment. *Am J Hum Genet* 11:1–16.
12. Selvin S (1970): Concordance in a twin population model. *Acta Genet Med Gemellol* 19:584–590.
13. Smith C (1972): Correlation in liability among relatives and concordance in twins. *Hum Hered* 22:97–101.
14. Smith C (1974): Concordance in twins: Methods and interpretation. *Am J Hum Genet* 26:454–466.
15. Stern C (1958): The ratio of MZ to DZ affected twins and the frequencies of affected twins in unselected data. *Acta Genet Med Gemellol* 7:313–320.
16. Vollmer RT (1972): Twin concordance: A set theoretic and probability approach. *J Theoret Biol* 36:367–378.
17. Weinberg W (1928): *Mathematische Grundlage der Probandenmethode*. *Z Induct Abstamm* 48:179–228.

Correspondence: Dr. Gordon Allen, NIMH, 5600 Fishers Lane, Rockville, MD 20857, USA.