Department of Biostatistics, University of California, Berkeley, California

Concordance in a Twin Population Model

Steve Selvin

SUMMARY

A twin population model is proposed, and two different definitions of the rate of concordance are investigated within the context of this model. It is found that the statistic formed from the ratio of the number of affected twins that come from concordant pairs to all affected twins has certain advantages over the traditional definition of the rate of concordance. This twin population model is extended to nontruncated sample of twins. The parameters of this situation are estimated. Finally, it is shown how these parameters relate to the two usual estimates of the proportion of DZ twins in a twin population, and the efficiency of these methods is computed.

1. Introduction

When a group of twins is used to study the relative importance of genetic and nongenetic factors of a particular trait (twin-study method), the statistics employed usually attempt to take advantage of the unique biology of twin pairs. The most widely used of these statistics is the rate of concordance. The rate of concordance is usually defined as the direct count of the proportion of twin pairs with two affected partners in a completely ascertained sample (Allen, 1955). This rate has been traditionally computed from twin data with little thought as to whether it is the best possible statistic. A theoretical structure for a twin population will be given here that yields an opportunity to investigate the properties of this statistic.

The use of the twin-study method implies that a twin pair will be included only if at least one member of the pair is affected. This ascertainment of a sample through affected individuals introduces certain complications since, in some cases, the probability that a twin pair is included in the sample depends on the number of affected twins in that pair. This type of problem was first recognized by geneticists investigating recessive genes in man. A number of methods have been proposed to compensate for this bias in the sample and yield unbiased estimates of population parameters (e.g., Haldane, 1938). The problem of ascertainment in twin samples has been discussed in a few papers such as Allen et al (1967) and Steinberg (1962). In the following development it will be assumed that the sample has been completely ascertained, which means that all affected twins are probands (Bailey, 1952).

Rec. 29. XI. 1969

Acta Genet. Med. Gemellol. (1970), **19**: 584-590

2. Population and Sample Structure

Two parameters associated with the twin-study method are of interest: the probability that a twin inherits the potential to express a set of genes, and the probability that the potential is expressed. Let the first probability be symbolized by p and the second by s. The probability that a set of genes is expressed can be interpreted in several fashions depending on the trait under investigation. The probability of expressivity of many traits is one or nearly one. An example of a trait where s is nearly equal to one, which will be used later, is the sex of the offspring. In this case p represents the probability of a male offspring. Clearly, there are numerous cases for which s is not equal to one. Since MZ twins have the same genes, any case in which the pair is discordant must be attributed to factors affecting the expression of those genes (i.e., $s<1$). In the context of a nature/nurture problem, p can be thought of as a measure of the genetic influence and s as a measure of the environmental influence (i.e., the lack of fit of the purely genetic model). If the trait is controlled by a simple dominant gene, then p is the relative frequency of the dominant phenotype and s measures any deviation from the expected relative frequency. These deviations may be caused by such things as observer error, observer bias, clerical error, and so forth. The twin population model is shown in Tab. I.

Tab. I

No. of affected	MZ	DZ
Two affected	ps^2M	p^2s^2D
One affected	$2ps\ (1\text{-}s)\ M$	$2ps\ (1\text{-}ps)D$
None affected	$[1\text{-}ps\ (2\text{-}s)]M$	$[1\text{-}ps\ (2\text{-}ps)]D$
Total	M	D

$D = 1 - M$ = proportion of DZ twins in the population

This model makes the same assumptions about p and s that are made in employing the usual twin-study method to actual data. Incorporated in the mathematical expressions are the assumptions that MZ twins have the same genes, MZ and DZ twin pairs experience the same nongenetic environment, and that the environment affects each twin independently. The last assumption is probably the most unrealistic. As mentioned before, the twin-study sample is obtained through affected individuals. A mathematical expression of this fact is given by a conditional probability. The population probabilities are conditioned by the fact that unaffected twins are not observed and these probabilities yield the sample structure shown in Tab. II.

Tab. II

	N_{mz}	MZ	N_{dz}	DZ
Concordant	a	$s/(2-s)$	c	$ps/(2-ps)$
Discordant	b	$2\ (1-s)/(2-s)$	d	$2\ (1-ps)/(2-ps)$
Total	N_{mz}	1.0	N_{dz}	1.0

$N_{mz} + N_{dz} = N$ = number of twin pairs in the sample

Note that the proportions of concordant and discordant MZ twins in the sample do not depend on p. This role of p is expected since MZ twins are always genetically concordant and only caused to be discordant by nongenetic factors.

3. Rate of Concordance

Stern (1958) observed that the ratio of MZ to DZ twins in a sample of twins where at least one is affected is not equal to the ratio of MZ to DZ twins in the twin population as a whole. Using the above population structure, this bias is expressed as (Steinberg, 1962):

$$\frac{ps^2M + 2ps(1-s)M}{p^2s^2D + 2p^2s(1-s)D + 2pqsD} = \frac{2-s}{2-ps} \times \frac{M}{D}.$$

The ratio M/D is the ratio of MZ twins to DZ twins in the twin population.

The calculated rate of concordance is easily expressed for the MZ and DZ samples as:

$$C_{mz} = a/N_{mz} \text{ and } C_{dz} = c/N_{dz}. \qquad [3.1]$$

The expectations of these quantities are functions of p and s. They are:

$$E(C_{mz}) = s/(2-s) \text{ and } E(C_{dz}) = ps/(2-ps). \qquad [3.2]$$

A complete interpretation of these twin concordance rates depends on knowledge of p and s, which is not given in the usual approach. An alternative definition of rate of concordance that does not have this problem of interpretation is achieved by doubling the number of concordant pairs in the MZ and DZ twin groups. This definition of concordance in terms of symbols is:

$$C^*_{mz} = 2a/(2a + b) \text{ and } C^*_{dz} = 2c/(2c + d). \qquad [3.3]$$

The expectation of these quantities is:

$$E(C^*_{mz}) \doteq s \text{ and } E(C^*_{dz}) \doteq ps. \qquad [3.4]$$

The approximate variance of these estimates is:

$$\text{and} \quad \begin{aligned} \text{Var}(C^*_{mz}) &= s(1-s)\ (2-s)^2/2N_{mz} \\ \text{Var}(C^*_{dz}) &= ps(1-ps)\ (2-ps)^2/2N_{dz}. \end{aligned} \qquad [3\cdot5]$$

Also the bias in the MZ/DZ ratio no longer exists:

$$\frac{2ps^2M + 2ps(1-s)M}{2p^2s^2D + 2p^2s(1-s)D + 2pqsD} = \frac{M}{D}.$$

These expressions for the rate of concordance are much easier to interpret because they relate in a simple way to the parameters under investigation in a typical twin-study. The following observations result directly from this new definition of concordance:

1) C^*_{mz} and C^*_{dz}/C^*_{mz} are, respectively, estimates of s and p;
2) $E(C^*_{mz}) > E(C^*_{dz})$: the magnitude of this difference measures the effect of p;
3) $E(C^*_{dz}) \doteq ps$ is the expected frequency of the trait in the single birth population.

4. Twin Population with no Diagnosis of Zygosity

The preceding model can be employed in twin populations where nonaffected pairs are included and no variables have been recorded which enable a researcher to separate the individual twins in the sample into different zygosity groups. This is sometimes the case with sources of data such as vital statistics, twin registries and most twin series that were not collected through a particular variable. In this case the theoretical population structure is derived by adding together the MZ and DZ twin proportions in the previous model, as shown in Tab. III.

Tab. III

No. of affected	Proportion of pairs observed	Probability
Two affected	x	$ps^2\ (M + pD)$
One affected	y	$2ps\ [1\text{-}s\ (M + pD)]$
None affected	z	$1\text{-}ps\ [2\text{-}s\ (M + pD)]$
Total	1.0	1.0

Based on this model, one gets estimates of p and s when $D = 1 - M$ is known.

The likelihood function is:

$$L = K[ps^2\pi]^x[2ps(1 - s\pi)]^y[1 - ps(2 - s\pi)]^z,$$

where $\pi = M + pD$ and K is a constant that does not depend on p and s. The derivatives of L with respect to p and s yield the following normal equations:

$$\frac{dL}{ds} = \frac{2x}{s} + \frac{(1 - 2s\pi)y}{s(1 - s\pi)} - \frac{2p(1 - s\pi)z}{1 - sp(2 - s\pi)} = 0 \quad \text{and}$$

$$\frac{dL}{dp} = \frac{(\pi + pD)x}{\pi p} + \frac{(1 - s[\pi + pD])y}{p(1 - s\pi)} - \frac{(2s - s^2(\pi + pD))z}{1 - sp(2 - s\pi)} = 0.$$

The maximum likelihood estimates of p and s are:

$$\hat{p} = \frac{(2x + y)^2 M}{4x - (2x + y)^2 D} \quad \text{and} \quad \hat{s} = \frac{2x + y}{2\hat{p}}. \qquad [4.1]$$

These estimates of p and s make it possible to approximate easily the rates of concordance $C^*_{mz} = \hat{s}$ and $C^*_{dz} = \hat{p}\hat{s}$ for almost all twin samples that include non-affected pairs. In actual fact, these samples in many cases will not be collected in a careful, unbiased fashion, so that valid inferences can be made from these estimates.

The expressions given in [4.1] can be used also in the following manner. If s is set equal to one and the expressions solved for D, one gets the maximum likelihood estimates for the proportion of DZ twins in a twin population:

$$\hat{p} = \frac{2x + y}{2} \quad \text{and} \quad \hat{D} = \frac{y}{2\hat{p}(1 - \hat{p})}. \qquad [4.2]$$

The trait that is usually observed with the frequency p and expressivity s assumed equal to one is the sex of the offspring [i.e., $p = 1 - q = P(\text{male})$]. Expression [4.2] was first derived in a much more straightforward manner by Weinberg (1901). He felt that $p = 1/2$ was a good approximation of the proportion of males in any twin population and employed $2y$ as an estimate of D which is called the Weinberg differential method. Since then, the maximum likelihood estimate of D has also been derived by Waterhouse (1950) and Gittlesohn (1964). The variances associated with the maximum likelihood and Weinberg estimates are, respectively:

$$V(\hat{D}_{ml}) = \frac{D(1 - pD)\ (1 - qD)}{Npq(2 - D)} \quad \text{and} \quad V(\hat{D}_w) = \frac{8Dpq(1 - 2pqD)}{N}, \text{ when } p = 1 - q$$

is known.

The efficiency of these two estimates for various values of p and D is listed in Tab. IV. This table shows that p must be greater than 0.54, before there is any appreciable difference in these two estimates of D. It should be noted that $M = 1 - D$

Tab. IV. The efficiency of the Weinberg differential estimation procedure

	$P = 0.50$	$P = 0.52$	$P = 0.54$	$P = 0.56$	$P = 0.58$	$P = 0.60$
$D = 0.50$	1.0	0.998	0.990	0.977	0.960	0.938
$D = 0.60$	1.0	0.998	0.991	0.980	0.964	0.944
$D = 0.70$	1.0	0.998	0.992	0.983	0.970	0.952
$D = 0.80$	1.0	0.998	0.994	0.987	0.976	0.963
$D = 0.90$	1.0	0.999	0.997	0.992	0.986	0.978

is an estimate of the correlation between the sex within a twin pair. The estimation of this type of correlation between bivariate discrete variables has been well investigated because of its application to the study of inbreeding (e.g., Wright, 1921). The quantity M is usually denoted as F and is called the coefficient of inbreeding. A complete review of its estimation is given by Li and Horwitz (1953).

References

Allen G. (1955). Comments on the analysis of twin samples. Acta Genet. Med. Gemellol, **4:** 143-160.

— Harvald B., Shields J. (1967). Measuring of twin concordance. Acta Genet. (Basel), **17:** 475-481.

Bailey N. J. (1952). A classification of methods of ascertainment and analysis in estimating the frequency of recessives in man. Ann. Hum. Genet., **16:** 223-225.

Gittlesohn A., Milham S. (1965). Observations on twinning in New York State. Brit. J. Prev. Soc. Med., **19:** 8-17.

Haldane J. B. S. (1938). The estimation of the frequency of recessive conditions in man. Ann. Eng., **28:** 251-255.

Li C. C., Horwitz G. (1953). Some methods of estimating the inbreeding coefficient. Amer. J. Hum. Genet., **5:** 107-117.

Steinberg A. G. (1962). Methodology in Human Genetics. Burdette W. J. Ed., Holden-Day.

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.

Waterhouse J. A. H. (1950). Twinning in pedigrees. Brit. J. Med., **4:** 193-216.

Weinberg W. (1901). Beitrage zur Physiologie und Pathologie der Mehrlingsgeburten beim Menschen. Pfluger. Arch., **88:** 346-430.

Wright S. (1921). Systems of inbreeding I-IV. Genetics, **6:** 111-178.

Riassunto

Viene proposto un modello di popolazione gemellare, nel contesto del quale vengono esaminate due diverse definizioni del tasso di concordanza. Risulta che l'uso del tasso del numero di gemelli affetti derivati da coppie concordanti sull'intero numero di gemelli affetti presenta alcuni vantaggi rispetto alla definizione tradizionale di tasso di concordanza. Questo modello di popolazione gemellare viene esteso a campioni gemellari non tronchi, e si stimano i parametri di tale situazione, i quali risultano correlati alle due stime usuali della proporzione di gemelli DZ in una popolazione gemellare. Viene, infine, calcolata l'efficienza di tali metodi.

Résumé

Un modèle de population gémellaire est proposé, dans le contexte duquel deux différentes définitions du taux de concordance sont examinées. Il se trouve que le taux du nombre de jumeaux atteints dérivés de couples concordants sur le total de jumeaux atteints présente quelques avantages sur le taux de concordance dans sa définition traditionnelle. Ce modèle de population gémellaire est aussi appliqué aux échantillons gémellaires non tronqués et les paramètres de cette situation sont estimés. Ces derniers résultent corrélés aux deux estimes usuelles de la proportion de jumeaux DZ dans une population gémellaire. L'éfficacité de ces méthodes est enfin évaluée.

Zusammenfassung

Es wird ein Modell für eine Zwillingsbevölkerung vorgeschlagen, im Zusammenhang mit dem zwei verschiedene Definitionen für den Konkordanzsatz untersucht werden. Daraus ergibt sich, dass eine Gegenüberstellung zwischen der aus konkordanten Zwillingen gewonnenen Zahl der von einer Krankheit befallen Paarlinge und der Gesamtzahl der befallenen Zwillinge einige Vorteile gegenüber der traditionellen Definition des Konkordanzsatzes aufweist.

Dieses Modell für Zwillingsbevölkerungen wird auf nicht unvollständige Zwillingsmuster ausgedehnt und die Parameter einer solchen Situation geschätzt, für die sich eine Korrelation zu den beiden üblichen Schätzungen des ZZ-Anteils einer Zwillingsbevölkerung ergibt.

Prof. Steve Selvin, New York State Department of Health, Birth Defects Institute, 84 Holland Ave., Albany, N. Y. 12208, USA.