SCHIZOPHRENIA: EVIDENCE FOR THE MAJOR GENE HYPOTHESIS*

R. C. Elston and M. A. Campbell

Department of Biostatistics and the Genetics Curriculum
University of North Carolina
Chapel Hill, North Carolina

ABSTRACT—A reanalysis of extensive data collected by Kallman (1938, 1946), shows that a major gene hypothesis satisfactorily accounts for the genetic component in schizophrenia; the results are consistent with the biochemical evidence to date. It is therefore unnecessary to postulate a polygenic theory for schizophrenia.

INTRODUCTION

THE possibility of a single major gene acting in the etiology of schizophrenia has been explored by several authors. Böök (1953) considered three forms of the major gene model, in each case invoking incomplete penetrance: the recessive hypothesis, the dominant hypothesis and the hypothesis that the heterozygote alone displays partial penetrance; i.e., that all homozygotes and a certain small proportion of the heterozygotes with the abnormal allele are affected. He selected this hypothesis of a partially penetrant heterozygote as providing the best fit to his data, in which there was a high morbid risk (3%). Such a hypothesis requires that, given the population incidence of schizophrenia, the gene frequency and the probability that the heterozygote be affected are functionally related.

This hypothesis of Böök was later applied by Slater (1958) to populations in which the incidence of schizophrenia is nearer the usual value of about 1%. He calculated algebraically the expected frequency of affected individuals among different relatives of schizophrenics. Then, considering the population incidence to be fixed at 0.8%, he was able to find parameter estimates (for the gene frequency and the proportion of heterozygotes affected) such that a reasonable fit was obtained to the observed frequencies in the various classes of relatives. Unforunately Slater selected the observed frequencies from three different investigations, thus greatly weakening his argument that the hypothesis provides a satisfactory fit to the data.

Since the classical Mendelian segregation ratios for a major gene are not found, and yet twin studies indicate a high heritable component for schizophrenia, a polygenic model has recently been proposed to explain the familial aggregation. It is therefore timely to point out that there is at present no published evidence for a polygenic hypothesis that cannot be equally well construed as evidence for

*This investigation was supported by a Public Health Service Research Career Development Award (1-K3-GM-31, 732), training grant (GM 00685) and research grant (GM HD 16697) from the National Institute of General Medical Sciences, and by a research grant (NGR 34-003-040) from the National Aeronautics and Space Administration.

a major gene hypothesis. The lines of evidence put forward by Gottesman and Shields (1967) for a polygenic hypothesis are: (1) the fact that classical ratios are not found among the relatives of schizophrenic probands; (2) the appearance of segregants in the offspring of normal parents; (3) the increased risk of schizophrenia in families that have several members affected; and (4) the relationship found between the severity of the disorder in twin probands and the level of concordance in their co-twins. We shall show through rigorous mathematical treatment and by restricting ourselves to the data of one investigator (Kallman, 1938, 1946) that a model involving a single major gene can account for all of these facts.

GENERAL MAJOR GENE MODEL

For a one locus model, assume there is a "normal" allele A with frequency p and a mutant allele a with frequency $q = 1-p$, the latter giving rise to some disorder in the functioning of a metabolic pathway. Then individuals of the three genotypes AA, Aa and aa will differ with respect to the extent to which this metabolic pathway functions, and hence, for any disorder, with respect to the probability that they will be classified as affected. Let the probability that a person with genotype i ($i = AA, Aa$ or aa) be classified as affected be f_i; and let these three probabilities be given by the vector $\underset{\sim}{\mathrm{f}}$ whose transpose is thus $\underset{\sim}{\mathrm{f}}' = (f_{AA}\ f_{Aa}\ f_{aa})$. We should expect f_{AA} to be near zero, but it need not be exactly zero, given the model as we have stated it.

To estimate p and $\underset{\sim}{\mathrm{f}}$ from a knowledge of the empirical risk of affected individuals in relatives of probands, we assume random mating with respect to the locus involved. In most cases any departure from random mating will be small enough to have a negligible effect. Let $\underset{\sim}{\mathrm{R}}$ be a matrix whose (i, j)-th element r_{ij} ($i, j = AA, Aa$, or aa) is the probability that a person should have genotype j given that he is a relative of a person with genotype i. Then under random mating (Li and Sacks, 1954) $\underset{\sim}{\mathrm{R}} = c_I\underset{\sim}{\mathrm{I}} + c_T\underset{\sim}{\mathrm{T}} + c_O\underset{\sim}{\mathrm{O}}$, where

$$\underset{\sim}{\mathrm{I}} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \qquad \underset{\sim}{\mathrm{T}} = \begin{bmatrix} p & q & 0 \\ \frac{1}{2}p & \frac{1}{2} & \frac{1}{2}q \\ 0 & p & q \end{bmatrix},$$

$$\underset{\sim}{\mathrm{O}} = \begin{bmatrix} p^2 & 2\mathrm{pq} & q^2 \\ p^2 & 2\mathrm{pq} & q^2 \\ p^2 & 2\mathrm{pq} & q^2 \end{bmatrix}$$

and the scalars c_I, c_T, and c_O are respectively the probabilities that the two relatives should have both, one or none of the genes at a locus identical by descent. Values of these scalars for various classes of relatives are given in Table 1. Thus R depends upon the class of relatives being considered; e.g., for full sibs

$$\underset{\sim}{\mathrm{R}} = \frac{1}{4}\underset{\sim}{\mathrm{I}} + \frac{1}{2}\underset{\sim}{\mathrm{T}} + \frac{1}{4}\underset{\sim}{\mathrm{O}}, \text{ but for first cousins } \underset{\sim}{\mathrm{R}} = \frac{1}{4}\underset{\sim}{\mathrm{T}} + \frac{3}{4}\underset{\sim}{\mathrm{O}}.$$

The probability that an individual should be classified as affected, given he is a relative of a person with genotype i, is the ith element of $\underset{\sim}{\mathrm{R}}\underset{\sim}{\mathrm{f}}$. Now let $s = p^2 f_{AA} + 2pq f_{Aa} + q^2 f_{aa}$ be the probability in the whole population of being classified as affected. Then the probability that an individual so classified should have genotype i is simply the ith element of the vector $\underset{\sim}{\mathrm{a}}$ whose elements are defined by $a_1 = p^2 f_{AA}/s$, $a_2 = 2pq f_{Aa}/s$ and $a_3 = q^2 f_{aa}/s$. From this it follows that the probability that the relative of a proband should be found to be affected is simply $\underset{\sim}{\mathrm{a}}'\underset{\sim}{\mathrm{R}}\underset{\sim}{\mathrm{f}}$.

The extension of this model to any number of unlinked loci is easily accomplished through the use of direct, or Kronecker, products denoted $\otimes$. A general discussion of this type of matrix multiplication is available in e.g. Searle (1967). For the particular case of two unlinked loci $\underset{\sim}{\mathrm{f}}$ becomes a vector with nine elements; i.e., the "penetrances" of the nine genotypes *AABB, AABb, AAbb, AaBB, AaBb, Aabb, aaBB, aaBb, aabb*. $\underset{\sim}{\mathrm{R}}$ becomes a 9 x 9 matrix, being a linear function of the nine matrix elements of $(\underset{\sim}{\mathrm{I}}_A\ \underset{\sim}{\mathrm{T}}_A\ \underset{\sim}{\mathrm{O}}_A) \otimes (\underset{\sim}{\mathrm{I}}_B\ \underset{\sim}{\mathrm{T}}_B\ \underset{\sim}{\mathrm{O}}_B)$, where the subscripts A and B indicate that the matrix is appropriate for the A, a and B, b loci respectively. The coefficients of these nine matrices are given by the corresponding elements of $(c_I\ c_T\ c_O) \otimes (c_I\ c_T\ c_O)$. Analogously, $\underset{\sim}{\mathrm{a}}$ becomes a vector with nine elements.

ESTIMATION OF PARAMETERS

Suppose we have data on m classes of relatives, a distinct $\underset{\sim}{\mathrm{R}}$ corresponding to each class. Since there are three or four parameters that can be estimated (depending on whether or not we fix $f_{AA} = 0$), we can always find estimates of these parameters that will fit the data perfectly, in terms of the empiric risk for each class, if $m<4$ (these estimates may not, however, lie in the interval $[0,1]$). Gottesman and Shields (1967) quote data from several studies but in each case $m<4$. Also, in view of the investigator differences in the diagnosis of schizophrenia, $\underset{\sim}{\mathrm{f}}$ should be estimated separately for each investigator and not over pooled data. For these reasons we have selected for analysis the most extensive set of available data on schizophrenia gathered by one person (Kallmann, 1938, 1946). The data are shown in Table 1; Kallmann used the abridged and proband methods of Weinberg to obtain the age corrected incidence of schizophrenia in the relatives of probands. On any genetic hypothesis the incidence in children of probands should equal the incidence in parents. (Kallmann 1938) quotes a much higher incidence in children (16.4%) than in parents (9.1%); but he does note that in families in which one parent is schizophrenic and the other normal (i.e. not belonging to the category which included "all manifestations of feeblemindedness, psychopathy and endogenous psychoses") the incidence of schizophrenia in the children is 11.9%. Thus the high incidence in children of all his families could be due to sampling bias, and presumably the correct figure lies somewhere between 11.9% and 16.4%.

Let $\underset{\sim}{\mathrm{R}}_k$, r_k and n_k be the $\underset{\sim}{\mathrm{R}}$-matrix, number of schizophrenic relatives and the sample size respectively for the kth class ($k = 1, 2, 3, 4$); r_k is taken to be

TABLE 1

Summary of Data Analyzed and Results

Relationship to Proband	Coefficients			Sample Size	Observed Incidence (%)	Expected Incidence (%)	
	c_I	c_T	c_O			I	II
Monozygotic twins[2]	1	0	0	174	69.0	32.5	29.8
Dizygotic twins[2]	¼	½	¼	517	10.3	12.7	12.8
Full sibs[3]				6453	10.9	12.7	12.8
Children[1]	0	1	0	2000	11.9, 16.4	8.5	9.9
Parents[2]				1191	9.1	8.5	9.9
Half sibs[3]				199	5.3	4.8	5.7
Grandchildren[1]	0	½	½	1016	4.3	4.8	5.7
Nephews, neices[1]				2170	3.9	4.8	5.7

[1]Kallmann, 1938.

[2]Kallmann, 1946.

[3]Pooled from Kallmann, 1938 and 1946 (weighted averages).

equal to the stated observed percentage of the sample size. Then, letting $P_k = \underset{\sim}{a}' = \underset{\sim}{R}_k\underset{\sim}{f}$, the estimates of the parameters can be found by maximizing the "log likelihood"

$$\sum_{k=1}^{4} \{r_k \ln P_k + (n_k - r_k) \ln(1 - P_k)\} \quad .$$

These estimates are consistent, and would in fact be the maximum likelihood estimates if the data were restricted to one relative per proband. The maximization was performed using a search alogrithm, and the standard errors of the parameter estimates were obtained, assuming one relative per proband, by numerical double differentiation of the log likelihood surface. (Elston and Kaplan, 1970). Since the individuals who make up the data in Table 1 are not independently sampled, the standard errors thus found are too small.

The maximization was performed subject to the constraint that all estimates should lie in the interval [0,1]. A first set of estimates was obtained using the incidence of 11.9% in children, the result being $f_{AA} = 0$, $f_{aa} = 1$, $f_{Aa} = 0.081 \pm 0.009$, and $q = 0.055 \pm 0.002$. These lead to an incidence of schizophrenia in the population, s, of 1.1%, and to the expected values given in column I of Table 1. The expected incidence in first cousins of probands is obtained by setting $c_I = 0$, $c_T = ¼$, and $c_O = ¾$, and found to be 3.0%; this compares favorably with the 2.6% quoted by Kallmann (1946) as being observed by others, and with the 2.9% quoted by Slater (1968). A second set of estimates, obtained using 16.4% for the incidence in children, is $f_{AA} = 0$, $f_{aa} = 1$, $f_{Aa} = 0.111 \pm 0.012$, and $q = 0.056 \pm 0.004$. These estimates lead to a population incidence of 1.5%, the expected frequencies given in column II of Table 1, and an expected frequency in first cousins of 3.6%. The two sets of estimates do not differ considerably and it is reasonable to assume that the correct estimates lie somewhere between these two.

The expected values in Table 1 do not take into account the different environmental covariances that might be expected to exist for the eight types of relatives, and so the fit can be considered adequate to allow for the possibility that a single major gene accounts for the genetic component. Only monozygotic twins show a very poor fit, but this is the class for which concordance rates in the literature are extremely variable. In the discussion we speculate as to the reason why the fit for monozygotic twins is poor.

The model of two major unlinked autosomal loci was proposed for the transmission of schizophrenia by Karlsson (1966). The estimates obtained when this model is fitted lead to a much better fit to the observed values in Table 1, but to an expected incidence in the population of 4.5×10^{-5}. Such a model is clearly unrealistic.

DISCUSSION

Our analysis leads to the same genetic hypothesis as that put forward by Slater (1958), but with somewhat different estimates of q and f_{Aa}. Whereas, however, Slater assumed $f_{AA} = 1$ and $f_{aa} = 0$, we have estimated them to be so from one investigator's data, assigning appropriate weights to the different classes of relatives. This materially strengthens the support for a major gene model.

Using the parameter estimates we have found here, we can make two predictions. Firstly, if there is random mating among persons classified as schizophrenic, the genotype distribution of their children will be $\left(a_1 + \frac{1}{2} a_2\right)^2$ AA, $2\left(a_1 + \frac{1}{2} a_2\right)\left(\frac{1}{2} a_2 + a_3\right) Aa$, $\left(\frac{1}{2} a_2 + a_3\right)^2$ aa; and hence the probability that a child of two such parents should himself be classified as schizophrenic is $2\left(a_1 + \frac{1}{2} a_2\right) \cdot \left(\frac{1}{2} a_2 + a_3\right) f_{Aa} + \left(\frac{1}{2} a_2 + a_3\right)^2 f_{aa}$. Using the estimates obtained from the data this is 42% or 44%, depending on whether 16.4% or 11.9% is used as the incidence of schizophrenia in children of probands. Kallmann (1946) calculated the age-corrected incidence to be 68% for this class of individuals, but Slater (1968), recalculating the age correction for Kallmann's data, arrived at 45%; and Rosenthal (1966), pooling four other sets of data, reports 35%. Here again, the predicted values ignore any environmental covariance, and so we can only expect the observed and expected values to be approximately similar.

Secondly, we can predict that $2pq(1 - f_{Aa})/(1 - s)$ of the normal individuals (i.e. persons who will never be classified as schizophrenic) carry the mutant allele a. Using either the figure 16.4% or 11.9%, our estimate of this quantity is approximately 10%. Thus any biochemical abnormality found in all schizophrenics might also be reasonably expected in 10% of non-schizophrenics.

Now Smith and Sines (1960) reported finding a peculiar odor in the sweat of 59 out of 85 (69%) schizophrenics but in the sweat of none, or possibly one, out of 13 non-schizophrenics. The substance trans–3–methyl–2–hexonic acid has recently been identified as the chemical responsible for this particular odor (Smith, *et al*, 1969). If the odor is detectable in 69% of schizophrenics, our model would suggest that it should also be detectable in 69% of 10%, i.e., about 7%, of non-schizophrenics; this could account for the fact that perhaps the sweat of one of the 13 non-schizophrenics examined did in fact have the peculiar odor noted in schizophrenics. Another biochemical trait of similar interest is the excretion of dimethoxyphenethylamine in urine (Friedhoff, 1967), found in 67% of schizophrenic patients and in 2 out of 27 normals.

If Smith and Sines are correct in believing that their finding is a step in the identification of a metabolic disorder responsible for schizophrenia, this substance should be found to be present in all persons who either have had or are liable to have schizophrenic attacks—though it may not always be in sufficient concentration to cause a noticeable odor. Such a finding would provide strong evidence that the polygenic theory of schizophrenia is incorrect. For if it is the presence or absence of a particular metabolic pathway that differentiates persons liable to attacks of schizophrenia from persons who are not, then it is reasonable to suppose that it is also the possession or non-possession of one or a particular set of enzymes, and hence genes, that differentiates these two classes. On the other hand it is not reasonable, in such a situation, to suppose that it is the *number* of such genes that differentiates the two classes, as is required by a polygenic hypothesis.

Several genetic arguments against a monogenetic model have been suggested by investigators favoring a polygenic model. On the proposed hypothesis we need only have $0<f_{Aa}<1$, $0<f_{aa}\leqslant 1$ and $0<q<1$ to obtain immediately the findings (1), (2), and (3) stated above in the introduction. The finding (4), that severity is correlated with concordance in twins, is explicable as being due to the similar environmental influences that twins share. As suggested by Shields (1968), if the role of the environment is not insignificant in comparison with the role of genetics, then it is not unreasonable to assume that a severely affected proband will have been exposed to a relatively severe environmental situation, and likewise his co-twin.

The incomplete penetrance in the heterozygote is sometimes considered to be a major stumbling-block for the major gene model, and so it is appropriate to discuss different mechanisms that can cause it. The first possibility is that purely environmental influences determine whether a heterozygote will become affected. This would explain the high concordance rate for monozygotic twins, since they share many of the same environmental influences. However, by the same token we should expect the concordance rate for dizygotic twins to be higher than that for sibs, whereas in fact it is slightly lower. Furthermore there is some evidence that the postnatal environmental covariance for monozygotic twins is similar to that for dizygotic twins (Scarr, 1968). If environmental influences are to be invoked these should be prenatal ones, since monozygotic twins are unique in that they often share a common chorion (Steiner, 1935). The second possibility is that genetic modifiers are responsible for the variable expression of the heterozygote. This would account for the high monozygotic twin concordance rate, though it would be necessary to assume the modifiers have no effect on homozygous individuals, to be consistent with the finding $f_{AA}=0$ and $f_{aa}=1$.

A third possibility, so far overlooked by the proponents of a major gene for schizophrenia, is that of autosomal gene inactivation. Beutler (1964) has listed some tentative examples of this phenomena in humans, and it has been confirmed for the Gm system in rabbits and mice (Herzenberg, 1967; David, 1969).

Analogous to the Lyon hypothesis, it is assumed that only one allele, determined at random, is active in each clone of cells. If the number of such clones is small, i.e., the determination of which allele shall be active in each cell line occurs early in embryonic life, the proportion of *a* alleles that are active in an individual will be quite variable over the population of heterozygotes. It is then reasonable to suppose that a certain threshold proportion of the active alleles in an individual must be the *a* allele for schizophrenia to be diagnosed. This hypothesis could lead to an elevated concordance for monozygotic twins if the twins sometimes separate at a later embryonic state than the time of clonal determination. Furthermore the concordance would be almost complete if the threshold proportion of *a* alleles is very low, such a situation representing a nearly completely penetrant dominant allele. It is relevant to note that Heston (1970) has proposed such a dominant gene hypothesis for "schizoid-schizophrenic disease," i.e., a person is considered affected if he has either schizoidia or schizophrenia. His suggestion is completely compatible with the hypothesis of heterozygote incomplete penetrance for schizophrenia.

Finally, we must stress that neither the analysis presented here, nor any other analysis so far presented, proves the existence of a major gene. Such proof will require either biochemical evidence, or evidence from rigorous segregation analyses of family or pedigree data. Such a method of analyzing pedigree data has recently been presented (Elston, 1969), and we intend applying it to the extensive pedigree data collected by Karlsson (1966). The present analysis differs from the cited earlier analyses in that it gives appropriate weights to the various observed incidences, and it shows that all the types of data reported so far are consistent with a major gene hypothesis. It is therefore unnecessary, at this stage, to postulate a polygenic hypothesis with the attendant difficulty of biochemical interpretation.

ACKNOWLEDGEMENT

We are grateful for the programming assistance of Mrs. K. C. Yelverton. We are also grateful to Dr. I. I. Gottesman for his criticisms of an early version of this paper, some of which were needed in the preparation of this revised version.

REFERENCES

Beutler, E. (1964). Gene Inactivation: the distribution of gene products among populations of cells in heterozygous humans. *Cold Spring Harbor Symposium on Quantitative Biology,* **29,** 261–71.

Böök, J. A. (1953). A Genetic and Neuropsychiatric Investigation of a North-Swedish population. *Acta genet. (Basel),* **4,** 1–100.

David, G. S. and Todd, C. W. (1969). Supression of Heavy and Light Chain Allotypic Expression in Homozygous Rabbits through Embryo Transfer. *Proc. nat. Acad. Sci.,* **62,** 860–6.

Elston, R. C. (1969). A probability model for the analysis of pedigree data. Paper presented before the Statistics Section, American Public Health Association meeting, Philadelphia, 11 November.

Elston, R. C. and Kaplan, E. B. Paper in preparation.

Friedhoff, A. J. (1967). Metabolism of dimethoxyphenethylamine and its possible relationship to schizophrenia. In *The Origins of Schizophrenia,* Proc. 1st Int. Conf. on Schizophrenia, 27, ed. J. Romano. Excerpta Medica Foundation, Amsterdam.

Gottesman, I. I. and Shields, J. (1967). A polygenic theory of schizophrenia. *Proc. nat. Acad. Sci.* **58,** 199–205.

Herzenberg, L. A., Herzenberg, L. A., Goodlin, R. C., and Rivera, E. C. (1967). Immunoglobulin Synthesis in Mice. *J. exp. Med.*, **125**, 701–13.

Heston, L. L. The Genetics of Schizophrenia and Schizoid Disease: The Evidence and a Dominant Gene Theory. *Science*, **167**, 249–56.

Kallmann, F. J. (1938). *The Genetics of Schizophrenia.* J. J. Augustin, New York.

Kallmann, F. J. (1946). The genetic theory of schizophrenia: an analysis of 691 schizophrenia twin index families. *Amer. J. Psychiat.*, **103**, 309–22.

Kallmann, F. J. and Roth, B. (1956). Genetics aspects of preadolescent schizophrenia. *Amer. J. Psychiat.*, **112**, 599–606.

Karlsson, J. L. (1966). *The Biological Basis of Schizophrenia.* C. C. Thomas, Springfield.

Li, C. C. and Sacks, L. (1954). The derivation of Joint Distribution and Correlation between Relatives by the use of Stochastic Matrices. *Biometrics*, **10**, 347–60.

Rosenthal, D. (1966). The Offspring of Schizophrenic Couples. *J. psychiat. Res.*, **4**, 169–88.

Scarr, S. (1968). Environmental Bias in Twin Studies in *Progress in Human Behavior Genetics*, ed. S. Vandenberg, Johns Hopkins Press, Baltimore.

Searle, S. R. (1967). *Matrix Algebra for the Biological Sciences.* John Wiley, New York.

Shields, J. (1962). Summary of the Genetic Evidence in *The Transmission of Schizophrenia,* ed. D. Rosenthal and S. Kety. Pergamon Press, New York.

Slater, E. (1958). The Monogenic Theory of Schizophrenia. *Acta. genet. (Basel)*, **8**, 50–6.

Slater, E. (1968). A review of earlier evidence on genetic factors in schizophrenia in *The Transmission of Schizophrenia,* Pergamon Press, New York.

Smith, K. and Sines, J. (1960). Demonstration of a Peculiar Odor in the Sweat of Schizophrenic Patients. *Arch. gen. Psychiat*, **2**, 184–8.

Smith, K., Thompson, G. F. and Koster, H. D. Sweat in Schizophrenic Patients: Identification of the Odorous Substance. *Science*, **166**, 398–9.

Steiner, F. (1935). Nachgeburtsbefunde bei Merrslingen und Ahnlichkeitsdiagnose. *Arch Gynäk.* **159**, 509–23.

Tienari, P. (1963). A Psychiatric Twin Study. *Acta. psychiat. scand.* 39, Suppl., **169**, 393–7.