Estimating Familial Effects on Age at Onset and Liability to Schizophrenia II. Adjustment for Censored Data

Charles J. MacLean, Michael C. Neale, Joanne M. Meyer, and Kenneth S. Kendler

Departments of Psychiatry (C.J.M., K.S.K.) and Human Genetics (C.J.M., M.C.N., J.M.M., K.S.K.), Medical College of Virginia, Richmond, Virginia

Genetic studies of disorders with adult onset often contain individuals who have not completed their age at risk when last observed. Without correction for such censoring, correlation in ages at onset among relatives is substantially underestimated. Moreover, without correction for the effect of correlated ages at onset, the relationship between age at onset in the proband and liability in relatives is substantially overestimated. The present paper describes methods for correcting the effects of censoring on these estimates. In a companion paper [Kendler and MacLean, *Genet Epidemiol* 7:409–417, 1990] these methods are applied to a large family study of schizophrenia.

Key words: age at onset, schizophrenia, familial correlation, liability

INTRODUCTION

In many diseases, both liability and age at onset vary among individuals, depending upon genetic and environmental factors [Falconer, 1967]. Correlation of familial liability with age at onset as well as correlation among ages at onset have both been observed [Heimbuch et al., 1980]. Liability and onset may be partly under control of the same genes or environmental effects, and partly under separate influences [Neale et al., 1989]. Their distributions and their relationship to each other comprise important genetic information.

Liability cannot be observed directly; it is usually inferred from risk among relatives [Falconer, 1965]. In diseases where onset occurs in adulthood, individuals may often be lost to follow-up before the end of their risk period, either by death, migration, refusal of information, or simply because they were young at the time of study.

Received for publication July 23, 1990; revision accepted August 30, 1990.

Address reprint requests to C.J. MacLean, Department of Psychiatry, Box 710, Medical College of Virginia, Richmond, VA 23298.

© 1990 Wiley-Liss, Inc.
Under such "censoring" of information, the observed proportion of relatives affected may misrepresent the true liability. Moreover, censoring may serve to confound estimation of the correlation between liability and age at onset [Risch et al., 1985]. Censored at a given age, families with an earlier age at onset distribution would appear to have a higher rate of disease than those with later onset, even if their uncensored, lifetime rates were equal. Under censoring, the distribution of ages at onset may also be distorted. Censoring tends to remove later onset cases disproportionately, shifting the mean to younger values. In addition, the correlation of ages at onset among relatives is artifactualy reduced by censoring. The present paper describes methods for estimating and adjusting the effects of censoring in the estimation of liability and age at onset as well as their correlation among relatives.

DATA

The methods described in the present paper have been applied to data gathered by Lindelius [1970] in a sample of 270 schizophrenic individuals and their families. Probands for the study consisted of all cases of schizophrenia admitted between the years 1900 and 1910 to a rural hospital in southwest Sweden. Relatives, consisting of parents, siblings, offspring, nieces, and nephews of the probands, were identified and traced by parish registers. Information is available on 3,997 relatives of whom 134 were affected. Details of sampling are given in the companion paper [Kendler and MacLean, 1990] as well as in Lindelius [1970].

Figure 1, showing the survival function in juxtaposition with the hazard function for schizophrenia in relatives, indicates that not all relatives had passed entirely through their risk period for schizophrenia when last observed. The value of the survival curve at a given age is the proportion of relatives still under observation at that age, the remainder being censored although not necessarily dead. The hazard rate for schizophrenia is the average annual proportion of surviving relatives to be newly diagnosed in 5 year intervals. Calculation of hazard rate is based upon survival at every age by the life table method [Kalbfleisch and Prentice, 1980].

![Figure 1](image_url)

Fig. 1. Survival function for relatives of probands, measured in total sample size, and annual schizophrenia hazard rate per 1,000 relatives for 5 years.
CORRELATION OF AGES AT ONSET

While standard life table methods can be used to adjust the age at onset distribution for individuals, in family data the correlation of ages at onset is also affected by censoring. The effect is similar to the reduction in correlation sustained in truncated variables [Martin and Wilson, 1982]. Therefore, we must adjust correlation of ages at onset among relatives for the underrepresentation of late onset cases. The correlation of ages at onset $x$ and $y$ is determined by univariate moments together with the expected cross product

$$ E(xy) = \int \int xy f(x,y) \, dx \, dy $$

where $f(x,y)$ is the joint probability distribution of onset. When values have been censored, calculation of the ordinary Pearson correlation leads to an estimate not of $E(xy)$, but rather of

$$ E^*(xy) = \int \int xy f(x,y) \, S(x,y) \, dx \, dy $$

where $S(x,y)$ is the joint survival function under which the data were recorded. Transformation from $E^*(xy)$ to $E(xy)$ may be accomplished by calculation under a weighting function $1/S(x,y)$. The idea is that since older pairs are underrepresented in the data, those that appear should be inflated to compensate. The survival function provides an estimate of the underrepresentation, and therefore its inverse provides an adjustment to account for the portion censored.

The present technique, analogous to the method of moments, employs only pairs of which both are affected. Although the maximum likelihood procedure [Meyer et al., in press] employs all observations, most of the information relevant to the correlation in ages at onset resides in the affected pairs. The information obtained from unaffected relatives mainly concerns other parameters of the distribution of onset. Furthermore, the maximum likelihood procedure has the disadvantage that the mathematical form of the onset distribution is not known. Bivariate normality or log-normality would be the usual assumption for $f(x,y)$, but a glance at Figure 1 does not inspire confidence.

The present technique relieves us of assumptions about the distribution of onset. However, distributional assumptions are replaced in this method by assumptions about the survival function. Thus, adjustment is based on the usual assumption that censoring is independent of both age at onset and liability to schizophrenia. Although it is difficult to prove absolute independence, preliminary analysis indicates no substantial correlations. Correlation between age at onset in probands and longevity in relatives is .09 (95% confidence interval .04-.15). Correlation between age at onset in probands and study follow up of relatives, often terminated by lost contact other than death, is only .05 (not statistically significant).

In the present analysis, each pair consisting of a proband and an affected relative must be treated as an independent observation despite actual clustering within certain families. This is necessary because the survival functions required for adjustment apply only to specific pairs, not to family averages. Of 270 probands, three had more than one affected sibling, three others had more than one affected offspring, 11 had more than one affected second degree relative, and none had multiple affected parents. Several probands had affected relatives of more than one type, but analysis is performed
on each type separately. To estimate the effect of familial clustering, we have calculated correlations without consideration of censoring by both the pairwise and by the more proper “within family variance” method used in nested analysis of variance. The results are quite similar, providing some confidence that the pairwise treatment of censoring is acceptable also.

Modeling Survival Functions

We represent the bivariate survival function as a product,

$$S(x,y) = S_1(x) S_2(y|x)$$

where $x$ is the age of proband and $y$ is the age of a corresponding relative. The survival function for probands has a somewhat different meaning from that of relatives. For probands we need a function representing the probability of appearing in the present sample from the appropriate population of schizophrenias. To appear in the sample, an individual had to be alive in the right place during 1900–1910 and have onset of schizophrenia before both death and 1910. This indicates that the demographic age distribution for the appropriate time and location would provide an estimate of $S_1(x)$. However, schizophrenics are not drawn at random from the population [Eaton, 1985]. Therefore, census data have serious potential bias.

If population size were stable, the cumulative age structure would be described by the survival distribution. This is not quite accurate in our case, because the population of Sweden grew during the nineteenth and early twentieth centuries at an average rate of about 0.84% per year [McKeown, 1976]. Therefore, the population had a higher proportion of younger ages than the survival distribution during this period. Nonetheless, a life table from the present data seems far more representative than demographic data which include an unknown mixture of the social factors affecting schizophrenia.

Although age at death is recorded for most probands, a life table from these data would be strongly biased, because all probands survived to appear in the sample. Therefore, as a substitute, we have chosen the life table for siblings of probands, presumably well matched for generation, social class, and other demographic factors. This representation fails to the extent that schizophrenia affects longevity [e.g., Tsuang and Woolson, 1977]. However, two factors mitigate the effect of a possible morbidity bias. First, for our purpose survival after onset is irrelevant. Adjustment is required only for ages at onset of both relatives. Although bias in the prodromal period is possible, it is apparently small [Eaton, 1985]. Second, if the survival curve for probands were simply shifted by a constant amount to earlier ages, the result would not affect correlation, which is standardized on average values. Only bias greater at one end of the age distribution would be transmitted to the correlation of ages at onset among relatives. There is no evidence of this.

The estimate of survival is adjusted to account for year of birth and sex by life table regression [Cox, 1972]. An example of $S_1(x)$ is shown in Figure 2 adjusted for a male proband, born in 1880. Since death is the event of importance, other censoring is treated as “withdrawal” by the life table method.

Survival function $S_2(y|x)$ is the probability that a relative survives to age $y$ given that the associated proband survived at least to age $x$. Unlike that of probands, the survival function for relatives concerns loss to follow up in the present study by any means. Since longevity, migration, and other censoring events are correlated within
families, we employ a life table for relatives conditional upon proband survival as well as type of relative, year of birth, and sex. Type of relative is statistically important to follow up in the present data, and is presumably related to the sampling procedure. Because of demographic trends in the population, year of birth also is an important factor in the present data, whereas sex appears to have a smaller effect on longevity. An example of $S_2(y|x)$ is shown in Figure 2 for a particular kind of relative: a male born in 1900, nephew of a female proband who survived to age 60. Note that survival curves are estimated from all the sample data and adjusted by regression coefficients for particular classes. Therefore, $S_2$ in Figure 2 is shown for individuals more than 60 years old even though final observation was in 1960.

**Application to Correlation**

Survival functions for probands and their affected relatives, evaluated for ages at onset of schizophrenia rather than at death, are employed in calculation of adjusted correlations, shown in Table I together with the corresponding unadjusted values. Since only four parents of probands were affected, correlations are not calculated for them. Even though ages at onset of second degree relatives are naturally less correlated than those of first degree relatives, the increase of adjusted over unadjusted correlation appears to be about the same proportion for all types of relatives: 30–50%.

Since the calculations providing adjustment for the correlation of ages at onset are rather complex, the sampling distribution of adjusted correlations is difficult to

<table>
<thead>
<tr>
<th>Type of relative</th>
<th>Unadjusted correlation</th>
<th>95% confidence interval</th>
<th>Adjusted correlation</th>
<th>95% confidence interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Offspring</td>
<td>.28</td>
<td>-.20 -.63</td>
<td>.42</td>
<td>-.17 -.74</td>
</tr>
<tr>
<td>Siblings</td>
<td>.30</td>
<td>.02 -.54</td>
<td>.44</td>
<td>.03 -.70</td>
</tr>
<tr>
<td>Nephews and nieces</td>
<td>.10</td>
<td>-.17 -.36</td>
<td>.13</td>
<td>-.23 -.47</td>
</tr>
</tbody>
</table>
derive theoretically. Therefore, we have instead recorded the variation of a simulated resampling of the data. The resampling method, called "bootstrap," is described by Efron [1981]. In the present case the population is taken to consist of pairs of probands and their relatives with all characteristics in the proportions found in the actual data. Observations are drawn from this "empirical distribution" by a computerized random process in samples the same size as the true sample. Each of the simulated samples is then subjected to the estimation procedure described above. The variation of correlations over 100 replicated samples is used under Fisher's transform to estimate the empirical standard deviation, which determines the confidence intervals shown in Table I. For unadjusted correlations, theoretical estimates of standard deviation can be derived [Kendall and Stuart, 1961]. These values are very close to the empirical values recorded for raw correlations from the same resampled data. Note that the confidence interval is increased for adjusted correlation compared to the unadjusted estimate, enough so that the lower limit remains about the same despite the shift in mean away from this value. Clearly, we should not gain statistical significance by this adjustment.

LIABILITY AND AGE AT ONSET

The relationship between liability and age at onset is derived from regression of risk among relatives onto age at onset in probands. Because of censoring, life table regression is employed. In this case, the event of importance is onset of schizophrenia and other censoring is treated as withdrawal. However, a special adjustment must be made because ages at onset are correlated among relatives so that individuals of a given age related to a proband with early onset have passed through a greater portion of their risk period than individuals of the same age who are related to late onset probands.

Adjustment employing the correlation of ages at onset is somewhat disturbing, because we suspect that some of the correlation in ages at onset we observe has been induced by the very correlation with liability that we attempt to estimate. However, the apparent contamination is illusory. The censoring effect would apply to correlated ages at onset whether liability and onset were correlated or entirely uncorrelated. If each proband were associated with enough relatives to provide a dependable life table for onset of schizophrenia, adjustment for censoring could be accomplished without regard for other probands and their relatives. In this case, the degree of correlation between age at onset and liability, or its genetic or environmental sources would clearly not be relevant to the procedure. Unfortunately, there are not enough data for such an independent adjustment. Therefore, we are forced to employ the correlation of ages at onset between probands and relatives, among other factors, to make up a single life table for all relatives.

Nonproportional Hazards

It seems unsafe to employ standard methods [Cox, 1972] for the adjustment of censoring in the present case, because the usual proportional hazards assumption is violated. For example, the schizophrenia hazard function for females is shifted to older ages and also appears somewhat altered in shape from that of males. Thus, separate life tables for the relatives of each proband would differ in shape as well as being shifted overall. Therefore, to remove the effect of correlation in ages at onset from liability estimates, we calculate a regression of age at onset in affected relatives onto
that of their probands, accounting for sex and type of relative simultaneously. Note that this regression is not attempted as an estimate of the true relationship between ages at onset in relatives; correlation in ages at onset was described in the preceding section. The regression line includes "regression to the mean" in addition to other effects on the relationship among relatives, and is of no interest. However, the residuals from this regression are uncorrelated with the proband age at onset, by algebraic design, at least among the affected individuals employed for regression. We hypothesize that the same adjustment applies also to as yet unaffected individuals.

Being uncorrelated is important, but several other qualities are also required for the residuals of age at onset in relatives to be equally distributed with respect to proband age at onset. First, the regression curve must fit the mean of the dependent variable at all values of the independent variable. Examination of residuals discloses no appreciable curvature either in the original data or under a logarithm transform of the age scale. The second quality required for proper regression analysis is homoscedasticity, i.e., the dependent variable must have the same variance at all values of the independent variable. Examination of residuals discloses greater variance of ages at onset among relatives of late onset probands than among those of early onset probands. This means that the onset life table for relatives of early onset probands has a steeper slope than that of late onset, as well as the overall shift to younger ages. To eliminate this bias we employ a transform of the age scale. The criterion employed to determine the best transform consists of an auxiliary regression equation for the squared residuals themselves regressed onto proband age at onset. In the original data, with greater variance at late onset, the regression coefficient is positive. Logarithms of age at onset in relatives also produces a positive coefficient, whereas log(age - 10) produces a negative regression of square residuals onto proband age at onset. From among the several scale transformations attempted, the homoscedasticity criterion is optimized by log(age - 5).

Regression of Liability Onto Age at Onset

Using residual age as "time to failure," the relationship of liability and onset may next be estimated by life table regression of schizophrenia among relatives onto age at onset of their probands. Assumptions concerning the distribution of age at onset or liability are not required. We only assume, as in the previous section, that the censoring function for relatives is independent of age at onset in probands. However, unlike the correlation method above, in the regression method occurrence of multiple relatives for each proband is no drawback. We wish precisely to measure the proportion of schizophrenics among all relatives of each proband.

Averaged over all classes, risk in relatives decreases 2.3% per year of age at onset in probands. The regression slope is significantly different from zero at $P = .05$. The 95% confidence limits are 0.4–4.2% decrease per year.

If ages at onset are positively correlated in relatives, application of the same age correction to relatives of probands with different ages at onset would make the relationship between age at onset and liability appear more negative than it really is. To measure the bias that would occur, another regression has been calculated without accounting for correlated ages at onset. In this case, the relationship between risk and onset is exaggerated to an apparent 2.8% decrease per year. Thus, in the present data, adjusting for censoring without properly accounting for the correlation of ages at onset produces an artifactual increase in the age–liability relationship of approximately 25%. 
As noted by Risch [1983], adjustment of risk calculations for censoring entails an additional estimate and a concomitant increase in estimation error. The sampling distribution resulting from this complex procedure is not known but can be estimated numerically in a resampling experiment similar to that described above for correlation. The confidence interval cited above was calculated by resampling.

ACKNOWLEDGMENTS

This work was supported in part by a grant from the John D. and Catherine T. MacArthur Foundation to Dr. Irving I. Gottesman, Department of Psychology, University of Virginia, and NIMH Grant MH41953 to K. S. K. The figures were prepared using computer graphics by Greg Porter, Department of Human Genetics, Medical College of Virginia. Data for this analysis were generously provided by Dr. Rolf Lindelius, Psychiatric Clinic, Soderjukhuset, Stockholm, and transferred to us with the kind assistance of Dr. Gottesman.

REFERENCES


Edited by D.C. Rao