Testing for familial correlation in age-at-onset

DANIEL RABINOWITZ*
Department of Statistics, MC4403, Columbia University, New York, NY 10027, USA
dan@stat.columbia.edu

REBECCA A. BETENSKY
Department of Biostatistics, Harvard School of Public Health, Boston, MA 02115, USA

SUMMARY

The analysis of family-study data sometimes focuses on whether a dichotomous trait tends to cluster in families. For traits with variable age-at-onset, it may be of interest to investigate whether age-at-onset itself also exhibits familial clustering. A complication in such investigations is that censoring by age-at-ascertainment can induce artifactual familial correlation in the age-at-onset of affected members. A further complication can be that sample inclusion criteria involve the affection status of family members. The purpose here is to present an approach to testing for correlation that is not confounded by censoring by age-at-ascertainment and may be applied with a broad range of inclusion criteria. The approach involves regression statistics in which subjects’ covariate terms are chosen to reflect age-at-onset information from the subjects’s affected family members. The results of analyses of data from a family-study of panic disorder illustrate the approach.

Keywords: Ascertainment; Cluster; Familial aggregation; Non-parametric; Panic disorder; Proportional hazards; Segregation analysis; Survival analysis; Truncation.

1. INTRODUCTION

For a disease whose occurrence has a genetic influence, familial association of age-at-onset can indicate that age-at-onset itself, as opposed to simply the occurrence of the disease, has a genetic influence. A genetic influence on age-at-onset can be relevant to screening, family counselling, and the design and conduct of genetic linkage and association studies. In the ongoing study of panic disorder, Knowles et al. (1998), for example, exploring whether age-at-onset showed familial aggregation was motivated by the decision of whether to subject age-at-onset to a linkage analysis and how to incorporate age-at-onset in a linkage analysis of affection status.

Exploring correlation in age-at-onset in affected family members can be complicated, however, by family-based sampling designs, as such designs can create spurious patterns of association. In the panic disorder study, for example, families came to the attention of researchers through affected individuals, and inclusion criteria were designed to over-represent families with several affected individuals. This sampling design promotes an excess of families in which several siblings experience onset prior to their current age. And, as siblings’s ages tend to aggregate in families, this promotes artifactual correlation in the age-at-onset of the siblings in the study. See, for example, Li et al. (1998); MacLean et al. (1990)

*To whom correspondence should be addressed

© Oxford University Press (2002)
and Langholz et al. (1999) for more general discussions of the confounding effects of censoring by age-at-ascertainment. See Fyer and Weissman (1999) for a comprehensive discussion of the data collection and the criteria used in defining affection status and onset.

In this paper, we develop an approach to testing for correlation that is robust to the effects of censoring by age-at-ascertainment and that is valid regardless of whether inclusion criteria involve the affection status of family members and regardless of whether age-at-onset influences the probability of becoming a proband. (Here, and in what follows, the term proband is used to denote the individual who brings a family to the attention of researchers when that family is ultimately included in the study sample.) We seek a methodology that is sensitive to familial correlation in age-at-onset, but only if that correlation is distinct from familial aggregation of affection status. That is, the goal is not adjusting for age-at-onset and age-at-ascertainment when examining familial aggregation of disease, but rather adjusting for age-at-ascertainment and inclusion criteria when examining familial aggregation in age-at-onset itself.

Our approach is not appropriate with every sampling design. It is predicated on the assumption that given the affection status of an individual’s relatives and given the age-at-onset and affection status of the individual, the probability that the individual becomes a proband is conditionally independent of the age-at-onset in the individual’s affected relatives. This assumption does not preclude, for example, a design in which families come to the attention of researchers through an affected family member, and are then included in the sample only if there are also additional affected family members. Nor does it preclude situations in which, for example, individuals come to the attention of researchers as incident cases. However, it does preclude designs where individuals seek inclusion in the study because their age-at-onset coincides with the age-at-onset of their affected relatives.

In the next section, we describe our approach to testing for correlation. The approach involves restricting attention to affected subjects, and viewing their experience backwards in age beginning at their age-at-ascertainment. The affected subjects’ data are used to compute a test statistic analogous to the efficient score in a proportional hazards regression model for truncated data. See, for example, Alioum and Commenges (1996). Each individual’s covariates in the regression model are derived from the age-at-onset data in the individual’s affected family members. This strategy is convenient for avoiding bias induced by familial correlation in age-at-ascertainment and for adjusting for sampling schemes designed to oversample affected individuals. In the subsequent section, the approach is illustrated through an analysis of family data from a study of panic disorder.

2. TESTING FOR CORRELATION

This section describes an approach to testing the null hypothesis that in randomly chosen affected individuals, age-at-onset is conditionally independent of both the disease occurrence and age-at-onset in the individuals’ family members.

Begin with some notation and assumptions. Let i index families in the study sample, and let j index the individuals within families who are affected prior to their ascertainment. Adopt the convention that \( j = 1 \) corresponds to the proband in each family. Let \( n \) denote the number of families in the sample, and let \( n_i \) denote the number of affected individuals in the ith family. Let \( T_{ij} \) denote the age-at-onset of the jth affected member in the ith family and let \( C_{ij} \) denote the corresponding age at ascertainment. We define \( C_{ij} \) to be the minimum of age at death or loss-to-followup if either death or loss-to-followup occurs prior to the age-at-ascertainment. Let \( Y_{ij}(t) \) denote the indicator that \( T_{ij} \leq t \leq C_{ij} \), and let \( N_{ij}(t) \) denote the indicator that \( t \leq T_{ij} \). These are, respectively, the indicator of at-risk status at time \( t \) for making a transition from having experienced the disease to not yet having experienced the disease when followed backwards in age, and the indicator that such a transition has occurred at an age greater than \( t \). Let \( R(t) \)
Testing for familial correlation in age-at-onset

79

denote the cardinality of the set of subjects at risk, excluding the probands, at age \( t \); that is

\[
R(t) = \sum_{i=1}^{n} \sum_{j=2}^{n_i} Y_{ij}(t).
\]

Let \( Z_{ij}(t) \) denote a (possibly time-varying) term that reflects a relevant aspect of the age-at-onset in the relatives of the \( j \)th subject in the \( i \)th family. For the data analyses described in the next section, \( Z_{ij}(t) \) is taken to reflect the age difference between the age \( t \) and the average age-at-onset of the \( j \)th subject’s affected siblings. Finally, let \( \bar{Z}(t) \) denote the average of \( Z_{ij}(t) \) in the subjects other than the probands at risk at age \( t \),

\[
\bar{Z}(t) = \frac{\sum_{i=1}^{n} \sum_{j=2}^{n_i} Z_{ij}(t)Y_{ij}(t)}{R(t)}.
\]

The test statistic proposed here is based on the quantity,

\[
\sum_{i=1}^{n} \sum_{j=2}^{n_i} [Z_{ij}(T_{ij}) - \bar{Z}(T_{ij})].
\]

This quantity is designed to test the influence of the covariates \( Z_{ij}(t) \) on the hazard, for subjects followed backwards in age from their age at ascertainment, of transitioning from having experienced onset to having not yet experienced onset—that is, from \( Y_{ij}(t) = 1 \) to \( N_{ij}(t) = 1 \). The statistic is patterned on a log-rank statistic for truncated data (followed backwards in time); at each transition age represented in the non-proband subjects, the covariate of the subject who transitions is compared to the average covariates in the subjects who are at risk for transitioning.

Note that the probands are excluded from the indices in the double sum. This is to avoid confounding by probands coming to the attention of researchers through their age-at-onset. The proband’s age-at-onset information is incorporated in the analysis, however, through its inclusion in the the regressors, the \( Z_{ij}(t) \).

A detailed derivation that the conditions on the ascertainment scheme described in the previous section are sufficient to ensure that, under the null hypothesis, and, under reasonable regularity conditions, the statistic is asymptotic to a mean zero variable, is available upon request to the corresponding author. The crux of the derivation lies in that under the null hypothesis and under the assumptions on the sampling scheme, for affected subjects other than the probands, the probability that onset occurs at an age in the interval \( (t - dt, t] \), given that onset is before or at age \( t \), given the affection status and age-at-onset in the family members, and given the proband’s age-at-onset, depends, under the null hypothesis, only on \( t \). To verify this result, note that, because inclusion in the sample is conditionally independent of the age-at-onset of family members given the family members’s affection status and given the proband’s age-at-onset, if it may further be assumed that the probability that any particular individual becomes a proband is negligibly small and the conditional hazard for relatives of probands is only negligibly different from the conditional hazard for randomly selected individuals, given the affection status and age-at-onset in family members. Since under the null hypothesis, age-at-onset in a family member, given that it occurs, is independent of the individual’s age-at-onset and of the affection status in the other family members, the conditional hazard in question is therefore the same as the conditional hazard in randomly selected individuals.

The approach presented here is flexible in that arbitrary codings of age-at-onset information may be used. This flexibility comes at the cost, however, of precluding clearly defined model parameters. It might be argued that the regression coefficients in the proportional hazards model corresponding to the test statistic are of intrinsic interest. However, the expectation of the estimates depends on interactions.
between the pattern of truncation, the choice of the covariates and the kinship relationships represented in
the data set; the estimates, therefore, are difficult to interpret.

There are approaches that do provide estimates of interpretable parameters. For example, non-
parametric estimators of the joint distribution of censored survival data have been developed. See, for
example, Burke (1988); Campbell (1981); Dabrowska (1988); Prentice and Cai (1992), and Tsai et al.
(1986). Extensions to truncated data are available as well. See, for example, Gurler (1996). Parametric
models and semi-parametric frailty models might also be considered. See, for example, Hougaard (1984).
See also Kent and O’Quigley (1988) and Lin and Ying (1993). When applicable, these approaches provide
interpretable estimates. These approaches are often not, however, readily available in standard computer
packages (especially for right-truncated outcomes), and extensions to data from multi-generational
families and selective sampling would seem to require some innovation. By restricting the focus to
estimation rather than testing, the approach proposed here is non-parametric in character, but may still
be implemented in a fairly straightforward fashion.

If the covariates in the test statistic were not functions of family members’s age-at-onset data, then
the test statistic would correspond to the efficient score from the partial likelihood in a proportional
hazards regression model for independent right-truncated data, and the canonical standard error based
on the observed information from the partial likelihood could be used to normalize the score. However,
the regressors associated with any one subject’s age-at-onset are functions of the age-at-onset in the
subjects’s relatives. The subject-specific terms in the counting process martingale representation of the
score, therefore, even under the null hypothesis, have familial correlation, and this correlation must be
accounted for in standard error computations. See, for example, Andersen et al. (1992) or Fleming and

A jack-knife estimator of the variance of the score that reflects the familial correlation in a manner
similar to the approach taken in Wei et al. (1989) and Liang and Zeger (1986) is given by

\[ \sum_{i=1}^{n} (S_{(i)} - S)^2, \]

where \( S_{(i)} \) is the score statistic that results from removing the \( i \)th family from the data set. The basis for
this choice is that the differences approximate the family-specific contributions to the score. Each term
\( S_{(i)} - S \) is a term that, up to lower-order terms, has expectation zero (under the null hypothesis), and
depends only on data from the \( i \)th family. That is, the variance estimator is simply a sum of squares of
(asymptotically) independent mean zero variables, and is therefore an unbiased estimate the variance of
the sum. A detailed derivation of this result is available upon request from the corresponding author. The
issues in the derivation are technical, the major issue being that the terms \( Z_{ij} - \bar{Z}(t) \) are not predictable
with respect to the filtration induced by observing onset in the affected subjects backwards from their age
at ascertainment.

Not all statistical packages provide the score \( S \), and, to our knowledge, those packages (for example,
SAS) that do provide the score, do not provide the score for analyses with time-varying covariates.
However, in order to compute the jack-knife estimate using standard software, the proportional hazards
regressions corresponding to the score statistics for right-truncated data may be fit, and the ratios of the
estimates of the regression coefficient to their corresponding variance estimates may be used in place of
the scores. (Justification for this substitution is the usual Taylor series approximation to the log partial
likelihood.)

We illustrate the approach (and the jack-knife variance estimator) through an application to data from
the panic disorder data. Computations were carried out in the SAS statistical package using primarily the
PHREG procedure. A SAS macro is available upon request to the corresponding author.
The covariate for the analysis was chosen to be

\[ Z_{ij}(t) = \exp(-0.5(t - \bar{T}_{ij})^2/\sigma^2), \]

where \( \bar{T}_{ij} \) is the average of the age-at-onset in all except the \( j \)th member of the \( j \)th family member’s siblings, and \( \sigma^2 \) was chosen to be approximately half the inter-quartile range of the non-parametric maximum likelihood estimate (NPMLE) of the conditional distribution of age-at-onset given that onset occurs. Computation of the NPMLE is described in Lynden-Bell (1971). This choice of covariate reflects familial aggregation in age-at-onset, as it is large when a subject’s age is close to the average age-at-onset in the subject’s affected relatives, but tends to zero when the subject’s age is far from the average. The parameter \( \sigma \) controls how quickly the covariate decays as the subject’s age becomes farther from the average.

There were 64 families in the data set. These families contained 81 sibships with more than one member. Of these sibships, 36, 32, nine, three and one, respectively, had two, three, four, five, and seven, affected siblings. Age-at-onset varied from five to sixty-two in the sibships, and age-at-ascertainment varied from 13 to 74. The non-parametric maximum likelihood estimator of the conditional distribution of age-at-onset, given that onset occurs, is depicted in Figure 1. Half the inter-quartile range was approximately 12.

The regression coefficient estimate from the proportional hazards regression corresponding to the score statistic was 0.71, and the variance estimate was 0.12. The corresponding nominal \( Z \)-statistic was 2.07, with nominal \( p \)-value less than 0.02. The ratio of the regression coefficient estimate to the covariance estimate was 6.0. The approximation to the jack-knife variance estimate was 8.8. The corresponding \( Z \)-statistic was 6.0/\( \sqrt{8.8} = 2.04 \) with \( p \)-value 0.02; the adjusted variance calculation results in a significant \( Z \)-statistic. The adjusted \( Z \)-statistic was stable over a fairly wide range of values of the smoothing parameter \( \sigma \) as well, as seen in Figure 2. The results suggest that there is significant correlation in age-at-onset of panic disorder that might be exploited in linkage and association analyses of genetic data from the family members.

Although the results of this analysis are statistically significant, some caution should be exercised in their interpretation. In particular, the methods developed here are not adjusted for secular trends that might induce correlation in the identification of panic disorder in affected siblings. Additionally, it is possible that the diagnosis of panic disorder in an individual might contemporaneously influence awareness of the disease in the individual’s family members; observed correlation might not represent a genetic etiology, but rather a sociological phenomenon. Finally, Goldstein et al. (1994) did not find correlation between age-at-onset of panic disorder in probands and first-degree relatives, and the application of the methods developed here to the data examined in Goldstein et al. (1994) also gave negative results.

3. DISCUSSION

The test statistic proposed here is designed to reflect relationships between subjects’ age-at-onset and the age-at-onset in their family members. Several authors have used similar strategies with dichotomous outcomes. See, for example, Betensky and Whittemore (1996) and Bonney (1987) or Blangero and Elston (1989). Similar strategies have also been applied with right-censored time-to-event survival outcomes. See, for example, Prentice and Hsu (1997); Clayton (1978) and Li and Thompson (1997). In the applied literature, it is common for age-at-onset in family members to be used in proportional hazards regression models in which proband affection status is used as a covariate. See, for example, Kloury et al. (1993). Such an approach is not applicable to the hypothesis of interest here, as it does not separate out the affects of familial aggregation of the occurrences of disease.
Fig. 1. Estimated conditional distribution function of age-at-onset in the panic disorder data. The horizontal axis corresponds to subjects’ age-at-onset. The vertical axis is the estimated conditional probability that onset occurs, given that it does occur. The points correspond to the values of the NPMLE at the ages at which the NPMLE makes a jump. Note that the conditional distribution does not extend completely to one; this is because when the panic data are viewed backwards in time, the risk set is temporarily zero about age 65.

Throughout, it has been assumed that each family is ascertained through a proband. The situation is not always so simple, however. Families may be brought to the attention of researchers because of, for example, the presence of several affected individuals. Conversely, families might be obtained as experimental controls through unaffected individuals, or families might be obtained through random sampling. Proband data are excluded from the indices of the double sum in the test statistic in order to account for the possibility that, conditional on the presence of disease, age-at-onset might further influence the probability of coming to the attention of researcher. In the less simple settings, all subjects (and only such subjects) whose inclusion in the study could have depended on their age-at-onset through more than simply their affection status should not contribute to the indices of the double sum. This modification, however, is not clearly applicable if different family members independently come to the attention of researchers, and it is only subsequently recognized that each is included in the family brought into the study by the other (see Vieland and Hodge, 1995). In this case, such families might be included in the calculation of the score statistic once for each proband, with all of each family’s several contributions removed simultaneously from the score statistic when computing the jack-knife variance estimator.

The assumptions on the ascertainment scheme included that age-at-onset should be conditionally independent of age-at-ascertainment. It was further noted that for subjects who are lost-to-followup or
who die prior to ascertainment, their age at loss-to-followup or death should be substituted for their age-at-ascertainment. This raises a difficulty, however, when death or loss-to-followup is influenced by onset of disease. In such cases, for example in an examination of correlation in age-at-onset of stroke, the methods proposed here are not appropriate.

The exposition here has been limited to a very simple choice of $Z_{ij}(t)$. However, other than that $Z_{ij}(t)$ should not be a function of $T_{ij}$, $Z_{ij}(t)$ may be chosen arbitrarily. For example, closer relatives could play a more prominent role than more distant relatives. Similarly, in order to incorporate additional covariates into the analysis, the test statistic could be replaced by the profile score in a proportional hazards regression model. Finally, the scores might be weighted to reflect family size; the covariates for subjects in larger families might be less variable, and so those subjects might be allowed a greater influence on the test statistic.

**ACKNOWLEDGEMENTS**

This work is supported by NIH grants GM5597 and CA75971 and the Taub Institute for Research on Alzheimer’s Disease and the Aging Brain. The authors are grateful to Abby Fyer and Myrna Weissman for the panic disorder data. Collection of the data was supported by NIH grants MH28274 and MH37592. The authors are grateful to Ruth Ottman, Priya Wickramaratne and Susan Hodge for thoughtful comments on preliminary drafts. We especially appreciate the efforts of Victoria Haghighi in assembling the database.
REFERENCES


[Received 17 July, 2000; revised 31 January, 2001; accepted for publication 26 February, 2001]