Some statistical issues in psychiatric epidemiology research

KUNG YEE LIANG and CLAYTON BROWN

INTRODUCTION

Like many other fields, psychiatric researchers rely heavily on the utilization of conventional epidemiologic designs and principles to pursue their scientific objectives. However, some of the issues, which to a large extent are specific to psychiatric disorders, have demanded that new designs and hence new statistical methods be developed to meet the challenges. These issues include, among others, uncertainty of diagnoses due to the lack of well-established biologic markers, high prevalence of comorbidity and lastly, etiologic heterogeneity.

In this note, we will elaborate on some problematic aspects of psychiatric epidemiology that pertain to the above three issues. We will then briefly discuss the shortcomings of current statistical approaches and provide suggestions for possible alternatives, some of which may require further development.

DIAGNOSTIC UNCERTAINTY

In the absence of well-accepted biologic markers, psychiatrists rely on clinical signs and symptoms as the main basis for their diagnoses. In epidemiological field studies it is critical to have accurate and valid measurement (Dohrenwend, 1990). As the criteria for many disorders overlap with regard to the presence of some clinical signs and symptoms, the question of «discrete versus continuous spectrum» is subject to constant debate. For example, Crow (1986) suggested that affective psychoses and schizophrenia are related to each other on a continuum of psychosis in which there is an increase in severity of illness from unipolar through bipolar affective disorder to schizoaffective disorder and schizophrenia. On the other hand, it has been argued that affective psychoses and schizophrenia represent discrete and distinct disorders as they are characterized by distinct psychopathological features, outcomes and unrelated predisposing genetic factors (Gershon & Rieder, 1980; Reich *et al.*, 1982). There is a similar debate within the context of anxiety disorders (Tyrer, 1985) as well as between generalized anxiety disorders and major depression (Kendler *et al.*, 1992).

Another issue, as a result of diagnostic uncertainty, is the use of discrete phenotypes to identify genetic loci that are thought primarily responsible for the occurrence of psychiatric disorders. The simplest discrete classification is dichotomously classifying subjects as «affected» versus «unaffected», for a particular disorder. A classification of a sample of subjects with schizophrenic spectrum disorders might be trichotomous with schizophrenia, schizoaffective, and affective psychosis diagnostic categories. The conventional genetic linkage studies have been hampered by the usage of such discrete phenotypes. This leads to the following question: Is it possible to refine the phenotypic definition in such a way that allows one to identify individuals who have inherited the susceptible gene(s), but have not manifested the illness in its typical form(s)? Some candidates for such refinement, known as «endo-phenotypes, have been suggested for schizophrenia including the measurement of eye tracking (Iacono *et al.*, 1988). An important question is then: how can one verify that the proposed candidates are

Indirizzo per la corrispondenza: Dr. K.Y. Liang, Department of Biostatistics, Johns Hopkins University, 615 North Wolfe Street, Baltimore, MD 21205-2179 (USA).
Fax +1 - 410-955.0958.
E-mail: Kyliang@yhsph.edu

https://doi.org/10.1017/S1121189X00004978 Published online by Cambridge University Press

indeed valid? One intuitive criterion is that these variables are specific to the targeted disorder.

The above issues of «discrete versus continuous spectrum» and «phenotype refinement», among others, demand innovative epidemiologic designs. One approach is to modify the conventional case-control design by using the family as the sampling unit. For example, to test the discrete versus continuous spectrum hypothesis for affective psychoses, one can compare the familial risk for schizophrenia among probands diagnosed with schizophrenia, schizoaffective disorder and affective psychoses. Here the familial risk is defined as the risk for relatives of probands, the probands being individuals whose pedigrees were sampled. Similarity in familial risk across the different diagnostic groups would refute the discrete hypothesis and be in favor of the continuum hypothesis.

This case-control/family study design (Liang & Pulver, 1996) might be useful in testing the endo-phenotype hypothesis as well. Specifically, one might compare the non-psychotic relatives of familial schizophrenics to non-psychotic relatives of healthy controls on the proposed phenotypes. Phenotypes which are distinguished between these two groups of individuals might be considered as «markers» of a vulnerability to schizophrenia.

A statistical complication of this design, which is not shared by the conventional case-control design, is that variables measured from individuals of the same family are likely to be correlated with each other. Inferences which ignore such a complication might lead to inaccurate conclusions. Liang & Pulver (1996) suggested the use of the Generalized Estimating Equations (GEEs) method (Liang & Zeger, 1986; 1993) to analyze the data derived from the case-control/family study design. This GEE method was designed to handle correlated data, which are common in family studies and in longitudinal studies for which repeated observations of the response variable from the same individual are collected. An illustration on the use of case-control/family design and the GEE method to address the discrete versus continuous spectrum hypothesis for schizophrenia was given in Liang & Pulver (1996).

ETIOLOGIC HETEROGENEITY

It is recognized that patients diagnosed with the same psychiatric disorder such as schizophrenia vary considerably on clinical expressions, course and response to treatments. One objective of laboratory, clinical, and epidemiological research is to determine relationships between the heterogeneous patterns of clinical signs and symptoms and heterogenous etiological causes and mechanism of psychopathology. Psychiatric epidemiologists study both genetic and environmental causes and mechanisms. Several hypotheses have been postulated for schizophrenia including genetic heterogeneity. The term «genetic heterogeneity» is descriptive of traits and diseases for which different genes or different genetic mechanisms are involved in different families or pedigrees. In the absence of additional clinical knowledge, genetic heterogeneity is a nightmare for psychiatric epidemiologists and geneticists who are searching for loci linked to a disorder.

Many attempts have been made to identify, based on clinical variables, «subtyping variables» which serve to subdivide the studied population into subpopulations which are more homogeneous etiologically. For example, loci have been identified for Alzheimer's disease by analyzing separately those with early age-of-onset from those with late age-of-onset. In the situation where the investigators have [a priori] hypothesized subtyping variables such as age-of- onset, one approach is to compare patients with different levels of the hypothesized variable on the clinical course over time. Age-of-onset, for example, may be considered as a legitimate subtyping variable if the longitudinal pattern of early onset patients is different from that of late onset subjects. To test such an hypothesis, one needs to adopt the longitudinal design in which the relevant clinical symptoms of each subject are measured repeatedly. Analytically, one faces two complications that require special attention. One is that more than one symptom is measured simultaneously for each subject at each occasion, all of which might be considered as response variables. Secondly, these clinical variables are likely to be a mixture of discrete and continuous measurements. Statistical methods for analyzing longitudinal data with one response variable have been well developed; see, for example, Diggle *et al.* (1994). Methods for longitudinal data with multiple discrete and continuous responses are less developed. While creating an artificial single response by summing the «scores» of multiple responses represents a simple approach, this simplified approach is far from being desirable. This is especially the case when these clinical variables cover different «domains» of the disorder. One example of this is the

https://doi.org/10.1017/S1121189X00004978 Published online by Cambridge University Press

positive and negative syndromes of schizophrenia. It is of considerable interest to examine how variables representing different domains interact with each other both concurrently and non-concurrently. More research is needed to develop innovative statistical methods for handling multidimensional longitudinal data.

Another approach that might be appropriate for genetic heterogeneity hypotheses is to compare characteristics (potential subtyping variables) of affected subjects on the familial risk of the studied disorder. For example, age of onset may be established as a subtyping variable, especially for genetic linkage studies, if probands with early onset have a higher familial risk than probands with late onset. This approach falls into the case-control/family study design and the GEE method mentioned previously.

When there are no overriding candidates for subtyping, the latent class models and grade of membership models (Woodbury & Manton, 1989) have been suggested as possible approaches. Both approaches assume that the clinical variables are manifested through a categorical variable which is not observed. This is known as the latent class variable or as «pure types». The main difference between these two approaches is that the latent class model assumes that each individual belongs to one class only; whereas the grade of membership model allows each individual to have partial membership in more than one class (pure type) where the extent of membership in each class is quantified by a «grade» between zero and one. Both approaches seem sensible in addressing the heterogeneity issue. There are, however, some outstanding statistical issues that have not been thoroughly resolved. For example, to determine the number of classes or pure types, it is typical that one would perform the likelihood ratio test to examine whether, for instance, having four classes would provide much better fit to the data than three classes. Unfortunately, this test statistic does not follow the conventional chi-squares distribution even when the number of subjects is large (Davies, 1977). Consequently, one needs to rely heavily on substantive knowledge to determine what constitutes the sensible number of classes. Another issue with these models has to do with the conventional and implicit assumption of «conditional independence». That is, assuming that one knows the class membership of an individual, the clinical variables of the same individual are statistically independent of each other («statistically independent» meaning knowledge of the value of one clinical variable provides no information about the value of another clinical variable). While sensible in some situations, there are circumstances that this conditional independence assumption might be violated. More work seems warranted to modify this key assumption and to thus assess how sensitive conclusions regarding subtyping are to this assumption.

PSYCHIATRIC COMORBIDITY

Psychiatric comorbidity is highly prevalent as shown in both clinical samples (Wolf *et al.*, 1988) and community samples (Robins *et al.*, 1991). For example, Wolf *et al.* (1988) reported that half of psychiatric patients in treatment have two or more diagnoses. Furthermore, the Epidemiologic Catchment Area (ECA) Study found that about half of all lifetime psychiatric disorders in the United States occur in individuals with a prior history of another psychiatric disorder (Robins *et al.*, 1991). Because multiple disorders complicate standard treatments and many types of comorbidity are associated with severe illness course, these results suggest that the prevention of comorbidity would be extremely valuable from both the clinical and public health viewpoint (Kessler & Price, 1993). Usually, before a prevention trial is carried out, one needs to understand, among other issues, which psychiatric disorders are clustered together, what are the causal processes regarding the occurrence of these disorders and what are the risk factors associated with the risk of the second and subsequent disorders after the onset of the first disorder.

In the absence of comorbidity, one can address the issue of risk factor identification through the conventional survival techniques such as the proportional hazards model (Cox, 1972). Here the response variable would be age-of-onset for the studied disorder and the hazard function would have the age-specific incidence rate interpretation. However, in the presence of comorbidity, which is indeed the main scientific objective, this commonly used technique would not be adequate. Instead, one needs a statistical model to describe the joint occurrence of several psychiatric disorders, presumably diagnosed at different ages. Furthermore, if the postulated causal process for a particular comorbidity suggests that the disorders share common genetic causes, then one needs to model the joint occurrence of the comorbidity in

https://doi.org/10.1017/S1121189X00004978 Published online by Cambridge University Press

individuals that are related to each other, such as siblings.

Statistically, the models briefly described above are known as multivariate survival models. There has been a renewed interest during the past decade in developing statistical models and methods that are useful for multivariate survival data; see, for example, a recent review on this topic by Liang *et al.* (1995). Some of the models reviewed therein are directly relevant to some of the issues discussed above, including the clustering of psychiatric disorders. However, more work is needed to address the issue of genetic determinants of comorbidity.

CONCLUSIONS

In this note, we provided a brief review of the statistical considerations for some substantive issues commonly encountered in psychiatric epidemiologic research. The three issues we have focused on here consist of diagnostic uncertainty, etiologic heterogeneity and comorbidity. For each of these three issues, we have pointed out the shortcomings of previously existing statistical methods when applicable, referred the readers to some newly developed methods that are relevant, and lastly identified some statistical areas of interest that warrant further research. Equally important, but not addressed in this note, is the availability of user-friendly statistical software based on the developed methods.

Another objective of this note is to bring to the attention of the reader the importance of constant mutual dialogue between the primary investigators, psychiatrists and psychiatric epidemiologists on the one hand, and statisticians on the other. For statistical methods to be useful, they must reflect scientific objectives (e.g. the parameters specified in the statistical model must be interpretable and address the question of interest). This can only be achieved if there is a constant dialogue between investigators and statisticians to ensure that the latter understand the substantive issues that are faced by the former.

REFERENCES

Cox D.R. (1972). Regression models and life tables (with Discussion). *Journal of the Royal Statistical Society*, Series B 34, 187-220.

Crow T.J. (1986). The continuum of psychosis and its implication for the structure of the gene. *British Journal of Psychiatry* 149, 419-429.

Davies R.B. (1977). Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika* 64, 247-254.

Diggle P., Liang K.Y. & Zeger S.L. (1994). *Analysis of Longitudinal Data*. Oxford University Press: Oxford.

Dohrenwend B.P. (1990). «The problem of validity in field studies of psychological disorders» revisited. *Psychological Medicine* 20, 195-208.

Gershon E.S. & Rieder R.O. (1980). Are mania and schizophrenia genetically distinct? In *Mania. An Evolving Concept* (ed. R.H. Belinaker and H.M. VanPraag). Spectrum: New York.

Iacono W.F., Bassett A.S. & Jones B.D. (1988). Eye tracking dysfunction is associated with partial trisomy of chromosome 5 and schizophrenia. *Archives of General Psychiatry* 45, 11-40-1141.

Kendler K.S., Neale M.C., Kessler R.C, Health A.C. & Eaves L.J. (1992). Major depression and generalized anxiety disorder. Same genes, (partly) different environments? *Archives of General Psychiatry* 49, 716-722.

Kessler R.C. & Price R.H. (1993). Primary prevention of secondary disorders: a proposal and agenda. *American Journal of Community Psychology* 21, 607-633.

Liang K.Y. & Pulver A.E. (1996). Analysis of case-control/family sampling design. *Genetic Epidemiology* 13, 253-270.

Liang K.Y. & Zeger S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* 73, 13-22.

Liang K.Y. & Zeger S.L. (1993). Regression analysis for correlated data. *Annual Review of Public Health* 14, 43-68.

Liang K.Y., Self S.G., Bandeen-Roche K. & Zeger S.L. (1995). Some recent developments for regression analysis of multivariate failure time data. *Lifetime Data Analysis* 1, 403-405.

Reich T., Cloninger C.R., Suarez B. & Rice J. (1982). Genetics of the affective disorder. In *Handbook of Psychiatry 3, Psychoses of Uncertain Aetiology* (ed. J.K. Wing and L. Wing). Cambridge University Press: Cambridge.

Robins L.N, Loche B.Z. & Regier D.A. (1991). An overview of psychiatric disorders in America. In *Psychiatric Disorders in America* (ed. L.N. Robins and D.A. Regier), pp. 328-366. Free Press: New York.

Tyrer P. (1985). Neurosis divisible. *Lancet* I, 685-688.

Wolf A.W., Schubert D.S.P., Patterson M.B., Grande T.P., Brocco K.J. & Pendelton L. (1988). Associations among major psychiatric diagnoses. *Journal of Consulting and Clinical Psychology* 56, 292-294.

Woodbury M.A. & Manton K.G. (1989). Grade of membership analysis of depression-related psychiatric disorders. *Sociological Methods and Research* 18, 126-163.

https://doi.org/10.1017/S1121189X00004978 Published online by Cambridge University Press