Primary Prevention of Secondary Disorders: A Proposal and Agenda

Ronald C. Kessler and Richard H. Price
The University of Michigan

This paper calls for consideration of a new class of preventive interventions designed explicitly to prevent comorbidity of psychiatric disorders. Epidemiologic data show that successful interventions of this type could be extremely valuable, as up to half of lifetime psychiatric disorders and an even larger percent of chronic and seriously impairing disorders occur to people with a prior history of some other disorder. Furthermore, a review of etiologic hypotheses concerning the causes of comorbidity suggests that interventions aimed at primary prevention of secondary disorders might be feasible. However, more basic risk factor research is needed on the causes of comorbidity before we can make a clear assessment of feasibility and discover promising intervention targets. A number of methodological problems arise in carrying out this type of formative research. These problems are reviewed and suggestions are offered for solutions involving innovations in measurement, design, and data analysis.

KEY WORDS: psychiatric disorder; comorbidity; risk factors; interventions; prevention.

INTRODUCTION

Comorbidity is of profound importance to clinicians, since half of psychiatric patients carry more than one diagnosis (Wolf et al., 1988). Such multiple disorders complicate ongoing treatment (Kranzler & Liebowitz, 1988) and precipitate relapses (Jaffee & Ciarulo, 1986). Similar patterns are observed among community samples. The Epidemiologic Catchment Area (ECA) Study found that half of all lifetime psychiatric disorders in the U.S. occur to people with a prior history of some other psychiatric disorder.

1All correspondence should be addressed to Ronald C. Kessler at the Institute for Social Research, The University of Michigan, Box 1248, Ann Arbor, Michigan 48106-1248.
(Robins et al., 1991). More strikingly, the recently completed National Comorbidity Survey (NCS) identified more than 80% of all severe current psychiatric disorders in the U.S. among the 13% of the population who have a lifetime history of three or more disorders (Kessler et al., 1994). These results suggest that the prevention of comorbidity (i.e., the prevention of the first onset of a second disorder) would reduce a substantial proportion of all lifetime psychiatric disorders and possibly an even greater proportion of ongoing disorders.

Yet despite such evidence, comorbidity has been largely ignored in risk factor research, in theorizing about the causes and consequences of psychopathology, and in the design and implementation of targeted preventive interventions. Interventions aimed specifically at the issue of comorbidity and targeted at the primary prevention of secondary disorders are long overdue. This paper discusses methodological issues involved in evaluating the feasibility of such interventions.

POSSIBLE CAUSAL PROCESSES

Epidemiologic data show that some disorders are especially likely to form comorbid clusters (Regier et al., 1990) and that one disorder in most such clusters typically occurs at an earlier age than the others (Kessler et al., in press a). Certain of these clusters are now considered distinct disorders in their own right, such as panic with agoraphobia (Eaton, Dryman, Weissman, 1991) and anxious depression (Akiskal, 1991). Others, such as comorbid phobia and substance use disorder, are even more common (Regier et al., 1990). People with a history of a primary disorder associated with a comorbid cluster are selectively more at risk for secondary disorders because of age or some other relevant risk factor should be eligible for interventions aimed at preventing those secondary disorders.

There would be several practical advantages of conducting preventive trials in populations at risk for comorbid disorders. First, the ease and reliability of identifying persons at risk for a secondary disorder is greater when they already meet criteria for a primary disorder. Second, already diagnosed groups may be at much higher risk for the development of the secondary disorder, increasing the efficiency and power of preventive trials. Third, primary prevention of secondary disorders may allow research to use already developed treatment technologies including pharmacological and behavioral interventions as part of the available technology of preventive intervention strategies. The longer history of intervention technology development in the treatment field can yield benefits for preventive intervention and may suggest new preventive uses for treatment techniques. Fourth, conducting preventive trials with diagnosed clinical populations at risk for the development of secondary disorders may increase the social warrant for preventive intervention. Various community constituencies and professional groups may more readily acknowledge the appropriateness of interventions with persons already suffering from a disorder, and recipients themselves may already be experiencing sufficient distress to make intervention more acceptable. Fifth, the prevention of secondary disorders might eliminate the exacerbation of primary disorders which is known to accompany the onset of secondary disorders, thereby addressing the issue of secondary prevention of primary disorders as a side benefit of the primary prevention of secondary disorders.

Despite these advantages, it is not clear whether such interventions are feasible. A number of practical problems need to be considered in this regard. For example, it is likely that more intensive effort is needed to prevent secondary disorders because primary disorders will complicate prevention efforts in the same way they complicate treatment efforts. Furthermore, if we rely on existing diagnoses for identification of subjects as having a primary disorder we will restrict the interventions to people who are already in the health care system. Yet if we go beyond treatment samples to make diagnoses in the general population, we risk stigmatization. These are complex issues that need to be addressed before we can seriously consider launching interventions for primary prevention of secondary disorder.

It might also be the case that it is more difficult to prevent the onset of secondary disorders than to prevent the onset of primary disorders for reasons that involve differences in the causal processes underlying these different kinds of onset. Several causal processes are worth considering in this regard. The first occurs when secondary disorders are directly caused by primary disorders. An example is cocaine-induced panic attack (Aronson & Craig, 1986; Liebowitz et al., 1984), thought to occur when prolonged cocaine use increases limbic-neuronal excitability (Charney, Woods, Goodman, & Heninger, 1987) to the point that pharmacological "kindling" of the brain permanently alters brain function in such a way that the threshold for seizure activity is lowered (Goddard, McIntyre, & Leech, 1969). Once this kindling occurs, panic attacks become independent of further cocaine use and persist despite discontinuation of the drug. There is evidence that similar long-term patterns of panic disorder can also be created by kindling processes linked to the use of psychostimulants (Abraham, 1986). Success-

1 A reviewer of an earlier draft noted that random association can bring about some comorbidity and remarked that this should be added to our list of causal processes. We have not done so because even though the co-occurrence of two or more disorders in a particular individual can be due to chance, the types of comorbidity of interest to us are those which have been shown to form significant clusters in aggregate analysis.
ful prevention of secondary disorders in cases of this sort is likely to be contingent on early treatment of the primary disorder (Rounsaville et al., 1991). In light of the difficulties associated with locating and successfully treating people with primary disorders of this type, interventions to prevent these direct secondary disorders seem unfeasible.

Secondary disorders may also occur as indirect consequences of primary disorders. One important example is substance abuse occurring as an unintended consequence of self-medicating a mental disorder. In one of the most dramatic demonstrations of this process, Bibb and Chambless (1986) screened a large sample of patients with agoraphobia for alcoholism and found that over 90% of the patients comorbid for alcoholism reported using alcohol to self-medicate their anxiety, compared to 43% among patients without alcoholism. Another example is agoraphobia secondary to panic disorder, which frequently develops when panic disorder leads to a disabling fear of attacks in situations where help is unavailable and escape is impossible. Such fears can severely restrict lifestyle and, in the extreme, provoke complete seclusion (Klein, Ross, & Cohen, 1987).

Interventions for such indirect secondary disorders seem feasible. An obvious first approach would be to develop a program of early detection and intervention to treat the primary disorder prior to the onset of the secondary disorder. In the case of agoraphobia secondary to panic, the evidence is clear that when panic is successfully treated, secondary agoraphobia often improves (Kaplan & Sadock, 1988). Another approach, not widely used but likely to be indicated when the primary disorder is more intractable but the secondary disorder has not yet occurred, is to develop an intervention aimed at teaching more adaptive and productive ways to cope with the primary disorder than those that lead to the secondary disorder. A program of this type might, for example, attempt to prevent the onset of substance use disorders secondary to social phobia by targeting socially phobic elementary school children and using a social skills training program (e.g., Botvin, 1986; Rotheram-Borus, 1988) to provide resources for managing fears without the abuse of alcohol or other drugs.

A third possibility is that primary disorders either create or are associated with contexts that potentiate the effects of other risk factors for secondary disorders. Although we are unaware of any systematic research that has documented such effects, there has been a good deal of speculation along these lines. For example, Bukstein, Brent, and Kaminer (1989) suggested that the family aggregation of either alcoholism or depression might increase the likelihood of experiencing a broad range of social problems which, in turn, could predispose to either disorder and so lead to comorbidity between alcoholism and depression. A related possibility is that primary disorders promote nonrandom assortment into social networks which are associated with risk of secondary disorders, such as the effects of deviant peer groups on risk of substance use secondary to conduct disorder (Meyer, 1986). Another example is related to the well-known finding that chronic depression leads to the erosion of social networks (Coyne, 1976), a process that could increase risk of secondary anxiety disorders by potentiating the effects of stressful life events which would otherwise be buffered by social support (Finlay-Jones, 1989). The challenge for preventive interventions in cases such as these is to target the contextual features associated with primary disorders that increase risk of secondary disorders so as to replace lost resources and reduce new vulnerabilities. These goals are quite similar to the goals of more conventional tertiary prevention programs, such as those developed to bolster the family resources of schizophrenics in order to prevent relapse (Fallow, Boyd, & McGill, 1984).

A fourth possibility is that the comorbid conditions share common causes. Recent twin research by Kendler, Neale, Kessler, Heath, & Eaves (1992b), for example, has documented common genetic determinants that largely explain the well-known association between generalized anxiety disorder and major depression (Maser & Cloninger, 1990), while data from the National Comorbidity Survey have shown that much of the comorbidity between post-traumatic stress disorder (PTSD) and other anxiety disorders is due to the fact that the traumatic experiences which lead to PTSD also cause panic, phobia, and generalized anxiety disorders. Another important example concerns the effects of community context on risk of multiple disorders (National Institute of Mental Health, 1993). The distinction between primary and secondary disorders is confused in situations of this sort, as no one disorder can be said to have causal priority over others in the cluster. Indeed, one might dispute the claim that there are multiple disorders in some cases of this sort and to argue, instead, that misdiagnosis has merely created the appearance of there being multiple disorders (as in the case of comorbidity between PTSD and other anxiety disorders, where there is considerable overlap of symptoms). It must also be recognized, though, that the existence of common causes does not in all cases ensure the eventual onset of the second disorder (or second phase of the single disorder) after the first one has occurred. This is true even in cases where the common causes are under genetic control, as genetic influences often have modifiable psychosocial mediators. To the extent that onset of the

3One of the two disorders might have an earlier age of onset. Long-term prospective studies, for example, show that early anxiety is considerably more predictive of later depression than early depression is of later anxiety (Angst, Vollrath, Merikangas, Ernst, 1990; Hagnell & Grabeck, 1990), even though recent evidence suggests that the comorbidity between anxiety and depression is probably due to common genetic causes (Kendler et al., 1992b).
secondary disorders is not inevitable, primary prevention of secondary disorders is feasible even when the comorbid conditions share common causes.

It should be clear from this review of possible causal mechanisms that our optimism about the possibility of developing successful preventive interventions for secondary disorders differs depending on the type of causal process creating the comorbidity. Our ability to evaluate this issue in particular circumstances depends on our success in assessing the causal processes that are involved in creating the comorbidity which we seek to prevent. There are formidable methodological problems involved in making this determination and in analyzing the causal processes in sufficient detail to select intervention targets. The remainder of this paper reviews the most important of these methodological problems and recommends strategies to surmount them. More substantive discussions of particular types of comorbidity and specific intervention targets must await the results of the studies proposed below.

EVALUATION OF THE PRIMARY-SECONDARY DISTINCTION

The first step in examining the causal processes leading to comorbidity should be to evaluate which disorder in a comorbid cluster is the primary disorder. The most common way of doing this is to use retrospective information about which disorder began first to define one as causally “primary” (i.e., typically having an earlier age of onset) and the others in the cluster as causally “secondary.” This approach confuses temporal priority with causal priority, though, and can lead to serious errors of inference. The inadequacy of this approach for inferring that the temporally prior disorder causes the later disorder can be seen in a simple example. Consider the fictitious Disorders A and B, each having a 20% lifetime prevalence and an odds-ratio of 2.3 between their lifetime prevalences. The first onset of A occurs a full decade earlier than the onset of B in exactly 50% of the people comorbid for the two disorders, while the first onset of B occurs five years earlier than the onset of A in the other 50% of comorbid cases, suggesting that neither disorder is “primary” relative to the other. Yet these patterns can be reproduced from an underlying causal process where Disorder A has absolutely no effect on risk of Disorder B, while Disorder B has a very powerful effect on risk of Disorder A. One set of underlying processes that could generate these observed patterns are as follows: 80% of the lifetime cases of Disorder A have onsets at birth; 100% of the lifetime cases of Disorder B have onsets at age 10 for reasons that are totally unrelated to prior history of Disorder A; and the remaining 20% of lifetime cases of Disorder A have onsets at age 20 that are powerfully influenced (odds-ratio of 20) by prior history of Disorder B. Minor calculations will show that these processes will yield an association between lifetime history of Disorders A and B that has an odds-ratio of 2.3. This example demonstrates that a simple examination of age-of-onset differences does not provide reliable information about causal priorities. Some additional approach is needed.

Survival Models with Time-Varying Covariates

When we are working with individual-level epidemiologic data on lifetime histories of comorbid conditions, we recommend using survival analysis models in which first onset of each disorder is treated as a time-varying covariate that can be used to predict the first onset of subsequent disorders. Models of this sort are capable of detecting asymmetric causal influences of the sort described above for the fictitious Disorders A and B. An illustration of this approach applied to the study of comorbidity between PTSD and other disorders is presented by Kessler, Sonnega, Bromet, & Nelson (submitted for publication). Models of this sort, when applied across a range of outcomes, are also capable of investigating whether primary disorders or combinations of primary disorders are associated not only with a single secondary disorder but with multiple secondary disorders.

The logic of this approach is easily grasped by considering the special case of a discrete-time survival model. The use of this model requires the researcher to create a separate observational record for each year of each respondent’s life. Each observational record would include a dichotomous outcome defining whether the respondent had a first onset of a target disorder in that year plus an array of predictor variables. Respondents who have the outcome disorder contribute one record for each year of their life up to and including the age of onset, with each record prior to the onset year coded 0 on the outcome variable and the final record coded 1. No record is included for years after the age of first onset. Respondents who never had the target outcome contribute one record for each year of their life and are coded 0 on the outcome for all records. Each person-year is also coded on a series of time-varying predictors, each of which defines the prior occurrence of other disorders in the comorbid cluster with a code of 0 for all years prior to the onset of a particular disorder and 1 for all subsequent years beginning with the year of first onset. More elaborate models sometimes include count variables for time since first onset or dummy variables defining a high-risk period of several years after first onset. Observational records for all respondents are combined into a single file with the person-year as the unit of analysis. This data array is analyzed
by means of logistic regression analysis with the target disorder as the outcome and the variables describing prior onset of the other disorders as predictors. Controls for both age at the time of the observational record and age at the time of the interview are included in the model. This analysis yields logistic regression coefficients analogous to those produced through the Cox proportional hazards model (Cox, 1972), but with the advantage of allowing the researcher flexibility in examining the effects of predictor variables which change values over time. Significant coefficients for comorbid disorders can be interpreted as showing that earlier onset of these disorders is associated with a significantly increased risk of the subsequent onset of the target disorder.

It is important to recognize that models of this type will not help the researcher detect the existence of common causes. If there are common causes, Disorder A will appear to be a risk factor for Disorder B and/or Disorder B a risk factor for Disorder A in standard survival models. The only way to evaluate the existence of this bias in the context of individual-level survival models is to introduce measures of common causes as control variables. In cases where these measures are not available, more complex designs are required.

**Family Aggregation Models**

One particularly attractive way of investigating the possibility that common causes account for comorbidity is to work with data from related individuals and study which disorders cluster within families (Smith, 1976). This method can sometimes provide information about causal priorities among comorbid conditions as well as about the contribution of common causes, even when the researcher has no specific hypotheses about what the common causes might be. An illustration of this method applied to the investigation of comorbid alcoholism, anxiety, and depression, is reported by Merikangas, Risch, and Weissman (in press).

The logic of this approach is based on the assumption that if one disorder causes the others in a comorbid cluster, we would expect the relatives of subjects with this primary disorder to have a high risk of this disorder, either singly or in conjunction with other disorders thought to be secondary to it. Relatives would not be expected to manifest pure forms of the putatively secondary disorders in higher proportions than the general population. Similarly, if other disorders are primary, the relatives of respondents with these disorders should have elevated prevalences of these disorders but not of disorders other than these except in the presence of the primary disorders. If common causes account for comorbidity, on the other hand, we would expect that the relatives of respondents with a pure disorder of one type would have elevated prevalences not only of that pure disorder but also of other pure disorders in the same comorbid cluster.

**Genetically Informative Models**

Although family aggregation models are capable of distinguishing causal effects of one comorbid condition on others from the effects of common causes, they cannot tell us whether the common causes are due to environmental or genetic influences. Needless to say, there may be important implications of whether the common causes are environmental or genetic for the feasibility of preventive interventions. This information can be obtained by using genetically informative designs, mainly adoption and twin designs. An illustration of the use of the adoption design to study comorbidity between anxiety and depression is presented by Cadoret and Winokur (1974), while an illustration of the use of the twin design to study comorbidity between generalized anxiety disorder and depression is presented by Kendler et al. (1992b).

The logic of the adoption design is a simple extension of the logic of the family aggregation design, with the important added benefit that any association between the disorders of parents and adopted away offspring can be attributed to genetic rather than to environmental influences. If, in comparison, comorbidity has no genetic basis, there will be no relationship between parental psychopathology and the disorders of adopted away offspring. The elegance of this logic, unfortunately, is compromised by the formidable practical problems associated with assembling a representative sample of adoptees and their biological parents.

The logic of the twin design is a good deal more complex. A detailed technical discussion can be found in Neale and Cardon (1992). Only a brief overview is presented here, focused on the simple case where there is comorbidity between only two disorders. A more complete development of this bivariate model can be found in Heath et al. (1993).

Estimation of genetic and environmental effects in a twin design requires separate representative samples of identical (MZ) and non-identical (DZ) twin-pairs, for each of which we obtain a 4 x 4 matrix of tetrachoric correlations for the liability to onset of Disorders A and B for twins 1 and 2. We would then fit a model like the one in Fig. 1 to these data by means of maximum-likelihood (Joreskog & Sorbom, 1989), asymptotic weighted least-squares (Browne, 1984), or some other fitting function. The figure represents the covariance between Disorders A and B for a single individual. Identification, however, requires that the model be estimated simulta-
neously for pairs of individuals and that coefficients be constrained across MZ and DZ pairs. Identification is achieved by imposing the following constraints on the covariance structure: (a) genetic influences $A_1$ and $A_2$ are assumed to be correlated perfectly within MZ pairs and correlated 0.5 in DZ pairs; (b) common environmental influences $C_1$ and $C_2$ are assumed to be correlated perfectly within all pairs, both MZ and DZ; (c) the effects of $A_1$, $C_1$, and $E_1$ are assumed to be mutually uncorrelated, as are those of $A_2$, $C_2$, and $E_2$, within twin-pairs; and (d) the values of the coefficients linking $A$, $C$, and $E$ to Disorders A and B are assumed to be the same for all respondents regardless of twin order or zygosity. When these assumptions hold, an evaluation of the coefficients $a'$, $c'$, and $e'$ can be used to make inferences about the extent to which genetic (A), common environmental (C), and unique environmental (E) influences account for the comorbidity between Disorders A and B.

This model can also be identified by working with samples of relatives other than twins, such as adoptees and their adoptive and biological parents, or full-sibs and half-sibs, so long as the samples are genetically informative. The model can also be elaborated to include nonadditive genetic effects as well as to estimate direct reciprocal effects between Disorders A and B. These extensions are discussed by Heath et al. (1993).

Other Complex Nonexperimental Models

There are other special-purpose complex designs that could profitably be used to evaluate specific hypotheses about the causes of comorbidity. For example, there is controversy over whether hyperactivity in childhood predicts adult substance abuse and antisocial personality disorder. Some researchers believe that hyperactivity is important in itself while others believe that hyperactivity only predicts adult problems if it occurs in conjunction with conduct disorder (Gillelman, Mannuzza, Shenker, & Bonagura, 1985; Gomez, Janowsky, Zetin, Huey, & Clopton, 1981; Loney, 1980). There is also some thinking that a common genetic basis might exist with hyperactivity linked to genetic risk of adult alcohol problems. The possibility of retrospective recall bias makes it difficult to carry out a definitive evaluation of this possibility with a retrospective design. Long-term longitudinal designs are difficult to implement because of high rates of dropout among antisocial and alcoholic young men. To address these problems, Tarter, Hegedus, and Gavaler (1985) used an elegant twogenerational design that compared delinquent sons of alcoholics with delinquent sons of nonalcoholics. As documented in their analysis, rates of hyperactivity did not differ in the two delinquent groups, providing compelling evidence free of retrospective recall bias that risk for alcoholism (as indexed by parental alcoholism) is not associated with hyperactivity when conduct disorder is held constant.

Experimental Models

The models discussed so far are all nonexperimental and, as a result, inherently equivocal concerning causal order among comorbid disorders. One unique opportunity to go beyond this sort of nonexperimental information is to use the results of randomized clinical trials. This could be done by modifying the outcomes assessed to include possible secondary disorders and to follow-up patients over a longer time interval to see if the experimental intervention reduced risk of first onset of secondary disorders which are known to cluster with the primary disorder. Evidence that the trial reduces risk of these other disorders could provide important information about causal pathways leading to comorbidity. Evidence that duration of treatment or treatment modality modified these effects would be even more illuminating. Furthermore, nonexperimental analysis could be
carried out to determine if the effect of the experimental intervention on reduced risk of these other disorders is mediated by remission of the disorder which was the focus of the clinical trial. If so, it could reasonably be concluded that either (a) this focal disorder is primary with respect to the other disorders or (b) the trial is effective in modifying common pathways to the different disorders. If the trial is effective in treating the focal disorder but does not reduce risk of first onset of other disorders known to be comorbid with the focal disorder, it could reasonably be concluded that either (a) the focal disorder is not primary with respect to the other disorders or that (b) the causal influences mediating between the focal disorder and later onset of other disorders are not mediated by course of the focal disorder. While still leaving room for uncertainty about the exact causal processes involved, information of this sort can be valuable in narrowing the range of plausible causal processes to consider in subsequent research.

**EVALUATION OF RISK FACTORS**

Once information is obtained to distinguish primary and secondary disorders, the analysis task turns to evaluating risk factors. It is important to appreciate that the analysis of risk factors for comorbidity is much more complex than the more conventional analysis of risk factors for morbidity. This is true because an analysis of the causes of comorbidity needs to focus on the determinants of covariance among two or more variables rather than on the determinants of variance in a single variable. A critical implication of this fact is that we cannot study risk factors for comorbidity in treatment follow-up studies unless we have external information that the risk factors have different effects in the absence of the primary disorder. This limits the extent to which we can understand the causal pathways leading to secondary disorders from studies of this type.

This limitation does not take away from the fact that treatment follow-up studies can be very useful in documenting significant effects of baseline predictors of the onset of secondary disorders among persons who have sought treatment for primary disorders. We can also investigate whether variation in treatment duration, modality, or effectiveness interacts with other risk factors to influence onset of secondary disorders. We cannot, however, determine from treatment follow-up studies whether these same processes exist independent of the primary disorder, an issue that is quite important in sorting out causal pathways.

General population data, in comparison, can be used to assess the extent to which risk factors have different effects depending on whether the person has a prior history of another disorder. In order to do this, prior history of the other comorbid disorder(s) must be included among the risk factors in each equation. This is easiest to do, as a practical matter, by working with discrete-time survival models that conceptualize the prior disorders as time-varying covariates. Models of this type were discussed earlier in the paper. Hypotheses about the determinants of comorbidity can be tested in these models by estimating the extent to which control variables explain the effects of the temporally prior disorders on subsequent first onset of the later disorders.

Two variants on the control variable approach are worth distinguishing. In the first, we look for causes of the primary disorder which also have direct effects on the secondary disorders. In the second, we look for consequences of the primary disorder which mediate the relationship between the primary and secondary disorders. These two classes of explanatory variables can be distinguished temporally. The first type must be in existence prior to the onset of the primary disorder, while the second type cannot be in existence until after the onset of the primary disorder. They also differ in their implications for intervention. If comorbidity is due to common causes, then intervention must focus on either changing these variables directly or on blocking the causal pathways between them and the secondary disorders. If comorbidity is due to mediating variables, there are a greater number of potential intervention targets, including treating the primary disorder, blocking the causal pathways between the primary disorder and the mediators, compensating for the effects of primary disorders on the mediators by providing direct intervention on the mediators, and blocking the causal pathways between the mediators and the secondary disorders.

It is also important to consider the possibility that the existence of temporally prior disorders might be associated with the exacerbation of the effects of other risk factors—either common causes or mediators—on the subsequent onset of secondary disorders. For example, the impact of exposure to a traumatic life event on risk of PTSD might be magnified among people with a prior history of panic or phobia (Kessler et al., submitted for publication). We are unaware of any epidemiologic research that has broadly evaluated the existence of such modifying effects. Yet this is central to the investigation of comorbidity, as a risk factor which has no association with the primary disorder can still be a risk factor for comorbidity if its effects on secondary disorders differ depending on the prior existence of the primary disorder. Systematic investigations of such interactive effects should become a routine part of future research on risk factors for comorbidity. The use of asymmetric survival models with time-varying covariates can facilitate this type of analysis, because hypotheses about modifying ef-
Effects can easily be integrated into such models (e.g., Kessler & Magee, in press).

Problems of distinguishing causal priority between presumed mediators and outcomes arise in models of this sort, as do problems of evaluating the true causal impact of presumed mediators in the absence of controls for possible confounding variables. Two types of nonexperimental model elaboration are available to address these problems. In addition, these problems can be addressed by embedding the risk factor analysis in an experimental intervention. These three possibilities are considered next.

Instrumental Variable Models

It sometimes occurs that presumed mediators represent processes which occur so close in time to the onset of the outcome that it is difficult to sort out which is the cause and which the effect. This might be the case, for example, in an analysis of the extent to which effective coping can protect against the onset of secondary disorders in the face of acute life stress. This problem can be resolved if baseline measures of coping dispositions are available and these, rather than the subsequent measures of actual coping strategies used in the face of stress, prove to be the important predictors, in which case the researcher would normally assume that dispositional competence in coping protects against onset of the disorder. It is less clear how to resolve the problem, though, in the case where coping behaviors are more important than coping dispositions in predicting the outcome. When this is the case, the researcher is left with the nagging question whether onset of the disorder led to maladaptive coping behaviors rather than these behaviors leading to onset of the disorder.

In cases of this sort it is possible to use simultaneous equation methods to sort out reciprocal effects between the presumed mediator and the outcome. This method requires that the researcher make an assumption that one or more “instrumental variables” has a direct effect on the mediator but not on the disorder. When variables of this sort exist and the identification assumptions are plausible, it is possible to obtain direct estimates of reciprocal effects between the mediator and the disorder (Kessler, 1987). In the above example, for example, the baseline measures of coping dispositions could be treated as instrumental variables, the assumption being that the effects of these dispositions in predicting onset of the disorder are entirely mediated by the actual coping behaviors exhibited by the person once he or she is confronted with a coping challenge.

A simplified model that uses the method of instrumental variables to distinguish the reciprocal effects of coping behaviors (CB) and disorder (D) is presented in Fig. 2. Assuming that there is only one measure of coping dispositions (CD), this model is just-identified. That is, there are three covariances among the variables and three structural parameters in the model, so that each parameter can be uniquely identified. The model is overidentified (i.e., there are more covariances than model parameters) if coping dispositions are assessed with multiple measures. If the model is just-identified, it can be estimated by the method of indirect least-squares (Duncan, 1975). If it is over-identified, it can be estimated by maximum-likelihood or asymptotic weighted least-squares and the fit of the model can be evaluated with a chi-square test. An example of using this model to distinguish reciprocal effects between a presumed mediator and onset of depression can be found in Kessler et al. (in press a).

Models of Disaggregated Genetic and Environmental Effects

We noted above that it can be useful to work with genetically informative family data in order to disaggregate the components contributing to comorbidity between pairs of disorders. This approach can also be useful in evaluating presumed risk factors to help determine if their effects are genuine rather than due to some unmeasured common cause. Two strategies are available for doing this. One is to work with the same kind of covariance structure models as those described earlier in the discussion of
decomposing the comorbidity between two disorders into genetic and environmental components. A model of this sort, applied to the somewhat different situation where we are studying the effect of a risk factor (R) on an outcome disorder (D), is presented in Fig. 3. Identification is achieved by imposing the same constraints as those discussed earlier in the context of the model in Fig. 1.

This model is over-identified with two degrees of freedom. Additional paths could be included for direct effects of any two of the three determinants of R on D (indicated in the figure by dotted lines). If these direct effects exist, the coefficient b in Fig. 3 is a biased estimate of the effect of R on D. If there are no direct effects of A, C, or E on D, however, b is unbiased. This means that it is possible to obtain a global test for spuriousness of the estimated effect of R on D, even in the absence of direct measures of the variables assumed to cause the spuriousness, by evaluating the fit of this model. A detailed example of using this model to evaluate spuriousness in the stress-buffering effect of social support can be found in Kessler et al. (in press a).

A second strategy for evaluating the same type of model can be embedded in a more general analysis of risk factors, including an analysis making use of discrete-time survival models with person-years as the unit of analysis, by analyzing data on twin-pairs rather than unrelated individuals. This method uses a measure of the risk factor from the respondent's co-twin to predict the respondent's outcome. If the effect of the risk factor is due solely to genetic influences, then in MZ pairs the co-twin's risk factor score should have the same effect as the respondent's own score on the respondent's outcome (due to the fact that MZ twins share all their genes). In the case of DZ twins, an exclusive genetic effect would lead to the respondent's score on the risk factor having twice as large an effect as his or her co-twin's score due to the fact that DZ co-twins share, on average, half their genes. In the more general case where genetic effects exist in conjunction with environmental effects, we would expect to find associations between co-twin scores on the risk factor and respondent outcomes that are significantly larger for MZ than DZ twins. Failure to find differences of this sort would strongly argue against the existence of a genetic influence on the outcome. A detailed example of using this model can be found in Kessler et al. (in press b).

Experiments

It is important to appreciate that nonexperimental analyses such as those discussed above can never do more than argue for the rejection of otherwise plausible hypotheses. Experimental interventions provide a much more persuasive way to evaluate causal hypotheses once a small number of plausible alternatives have been pinpointed. As recently noted by an NIMH report on prevention research, preventive interventions have been underutilized for this sort of basic risk factor research (National Institute of Mental Health, 1993). Future work on etiologic factors in comorbidity would profit from breaking this mold. A particularly powerful design would be one in which an experimental intervention is embedded in a longitudinal study of risk factors for comorbidity in a sample of people who have already had a first disorder and are in the age range of risk for secondary disorders. In the ideal case, this would be a genetically informative sample. For example, a quasi-cross-fostering design might be embedded in a longitudinal risk factor survey-intervention to prevent alcohol abuse secondary to depression by basing the study on an oversample of depressed children from broken homes who have not had contact with their biological fathers. This design would require the researchers to know whether these fathers have a history of alcoholism (from either mother reports or from direct interviews with the fathers). The paternal history of alcoholism could then be used as an indirect indicator of genetic risk for alcoholism in the same way as in an adoption design (e.g., Bohman, Cloninger, Sigvardsson, & van Knorring, 1981; Cadoret & Gath, 1978; Cloninger, Bohman, & Sigvardsson, 1981), with appropriate controls for confounding covariates (e.g., evaluating the effect of parental divorce and controlling for the possibility of maternal depression). If the intervention is successful in preventing onset of alcoholism, nonexperimental analysis
of intervening processes could be used to help interpret the causal pathways involved in this effect in an effort to provide insights concerning future modifications of the intervention.

**ADDITIONAL MEASUREMENT AND SAMPLING ISSUES**

The discussion of methodological issues up to now has focused on the problems involved in distinguishing primary from secondary disorders and in analyzing the causal processes linking primary disorders to secondary disorders in enough detail to select intervention targets. There are also a number of important measurement and sampling issues which need to be considered in studying comorbidity. These are the focus of this section of the paper.

**Measurement Issues**

*Systematic Diagnostic Assessment.* Any attempt to study comorbidity must begin by using a valid and reliable diagnostic instrument. Critics have noted that this has not always been done in previous research on comorbidity (Hesselbrock, Hesselbrock, Tennen, Meyer, & Workman, 1983; Kessler, Taylor, & Miller, 1979). The development of standardized diagnostic rules and interview procedures now make it possible to do this in community samples as well as in treatment samples. However, there has not been enough methodological work done to determine whether special procedures are needed to obtain reliable and valid diagnoses of comorbid conditions with standard research diagnostic interview schedules. (For an exception, see Schuckit, Irwin, Howard, & Smith 1988). There have been discussions of special difficulties in obtaining valid diagnoses posed by various aspects of comorbidity, such as problems in distinguishing true depression from the short-term effects of withdrawal in alcoholics undergoing detoxification (e.g., Ries & Ellingson, 1990) or biases in patient recall of substance abuse problems in their family of origin in using FHRDC rather than in-person methods to assess family history (e.g., Hesselbrock, 1986). There have been no serious efforts, however, to develop a general scheme of diagnostic interviewing that is sensitive to these difficulties.

One example of an important problem that currently exists in a widely used research diagnostic interview schedule is seen in the Composite International Diagnostic Interview (CIDI), an instrument developed by WHO and ADAMHA to provide a structured diagnostic interview schedule for purposes of general population epidemiologic research (Robins et al., 1988). Generalized Anxiety Disorder is assessed in the CIDI beginning with a question about whether the respondent ever had periods of feeling tense, worried or anxious “that were not due to use of alcohol or drugs.” This inclusion of a hierarchy qualification in the stem question for GAD introduces a clear bias in the assessment of anxiety secondary to substance abuse or dependence. There is a great need to develop diagnostic instruments without intrinsic biases of this sort for purposes of obtaining accurate information about comorbidity.

*Expanding the Assessment of Primary and Secondary Disorders.* As noted above, the distinction between primary and secondary disorders is most commonly induced based on which disorder met criteria at the earlier age. However, one can also find other related distinctions in the literature. For example, the primary disorder is sometimes defined as the more severe one. Disorders that develop independent of episodes of the second disorder are sometimes called primary, even if they begin at a later age than the other disorder. Results about which disorder is primary are, of course, importantly affected by which definition is used. This is a critical issue, so much so that future investigations should evaluate a number of different definitions and their implications for predicting comorbidity.

*Expanding the Assessment of Age of Onset.* In a similar way, there is disagreement concerning how to define age of onset. Parrella and Filstead (1987) reviewed five different definitions in common use in the literature (e.g., age of first symptom, age of first meeting full criteria for the disorder, etc.). Needless to say, conclusions about which disorder is temporally prior will be affected importantly by which of these definitions is used. This is such a critical issue to the study of comorbidity that we recommend that researchers collect data capable of dating onset in a variety of different ways and investigate the implications of these differences for results.

*Expanding the Assessment of Other Aspects of Comorbidity.* We also recommend that future studies incorporate explicit survey questions designed to obtain more detailed information about episode comorbidity (whether pairs of disorders occur at the same time) as well as about lifetime co-occurrence and timing of onset. Questions about episode comorbidity can be useful in resolving a number of special problems that arise in studying some of the most complex comorbid conditions. For example, there is considerable uncertainty involving the temporal priority between substance abuse disorder and antisocial personality disorder. This uncertainty is due, in part, to the fact that both of these disorders typically begin early in life, thus requiring retrospective reporting. Furthermore, there is considerable overlap in the symptoms of the two disorders. As noted by Bukstein et al. (1989), substance abuse by adolescents is almost always illicit and this creates a tautological association between substance abuse and delinquency
as well as secondary associations created by societal responses to adolescent substance use. Great care is consequently needed in evaluating the temporal and causal priorities between antisocial behavior problems and substance abuse. In particular, special insights might be obtained by probing explicitly to determine whether certain symptoms of conduct disorder during the adolescent years were caused directly by the use of illicit substances (e.g., fighting only when intoxicated) or indirectly by reactions to use of illicit substances (e.g., being arrested because of using or selling drugs, involvement in criminal activity in an effort to obtain money for drugs). It was shown by Bukstein and associates that a substantial percent of the adolescents who appear to be comorbid for conduct disorder and substance use disorder actually meet criteria for conduct disorder only because of symptoms that are caused by their use of alcohol or drugs.

Given that much of the work involved in obtaining these data must rely on retrospective self-reports or reports by family members, it is important that methods be devised to improve the reliability of lifetime recall of illness onset and course. There has been a good deal of recent methodological work by cognitive social psychologists and survey research methodologists aimed at designing methods to improve the accuracy of recall of health events (Loftus, Smith, Klinger, & Fiedler, 1992). There is a need to adapt, refine, and evaluate the extent to which these methods are able to improve accuracy of retrospective reports about onset and comorbidity of psychiatric disorders.

**Sampling Issues**

**Population Samples Versus Treatment Samples.** Most existing studies of comorbidity are based on treatment samples rather than on samples of the general population. There is clear evidence that individuals in treatment are more likely to have multiple disorders than those in the community (Helzer & Przybeck, 1988). This means that estimates of the prevalence of comorbidity based on treatment samples will be biased reflections of the true prevalence in the population (Berkson, 1946). It also means that studies of risk factors for comorbidity might be biased. The only reliable solution to these problems is to use general population surveys to study the prevalence and causes of comorbidity. Comparisons of general population data with treatment data can then be used to study barriers to help-seeking and unmet need. Treatment samples can also be used to study the consequences of comorbidity and to provide baseline samples of patients who have only primary disorders for prospective studies of the causes of comorbidity.

**Prevention of Secondary Disorders**

*Sex and Age Representativeness: Another sampling issue involves the representativeness of the sample with respect to sex and age. It is important to appreciate that much of the research on psychiatric comorbidity that has been published to date has been based on samples that over-represent older males (e.g., samples of patients in Veterans Administration hospitals). It is rare to find results concerning psychiatric comorbidity reported separately by sex, age (cohort), or age-of-onset. This is a serious limitation because the available evidence suggests that patterns of comorbidity differ by both sex and by age-of-onset. As noted above, there may also be important cohort differences in comorbidity. In an illustration of the importance of sex, Roy et al. (1991), recently reported on a consecutive series of 339 patients who met Research Diagnostic Criteria for alcoholism in treatment at the Laboratory of Clinical Studies at NIAAA. Lifetime comorbidity of alcoholism with major depression, intermittent depression, and phobia were all significantly higher among women than men. Higher comorbidity of alcoholism with anxiety and depression among treatment samples of women has also been reported in other studies (e.g., Woodruff, Guze, Clayton, & Carr, 1979), and the same is true of alcoholism with antisocial personality disorder among men (e.g., Hesselbrock, Hesselbrock, & Workman-Daniels, 1986).

Failure to distinguish between early-onset and late-onset cases is an equally serious problem in the literature on comorbidity. Clearly, given the fact that lifetime risk of disorders increase with age, lifetime comorbidity will increase with age within any birth cohort. This means that lifetime comorbidity will be higher in older samples than in younger samples, all other factors being equal. Yet, despite the fact that methods are available to adjust estimates of lifetime comorbidity for the fact that most respondents have not completed their lifetime risk periods at the time of study, it is rare to find published reports that present adjusted estimates of this sort. Furthermore, little attention had been paid to the observation of several researchers that patterns of comorbidity vary depending on the age of onset of the primary disorder. (For an exception, see Robins & Pryzbeck, 1985). Strong and consistent evidence has been found, for example, that depressed patients with an early onset have a stronger family history of both depression and alcoholism than those with a late onset (e.g., Mendlewicz & Baron, 1981). It is also known that patients with such family histories are at high risk of comorbidity for substance use disorders (e.g., Lewis, Rice, Andreasen, Endicott, & Hartman, 1986). Unfortunately very few investigations of comorbidity between affective disorders and substance use disorders have distinguished between early-onset and late-onset patients. This issue demands attention in future studies.

**Short-Term Historical Trends.** Another sampling issue involves the time frame of sample selection. Cohort analysis suggests that patterns of
psychopathology are changing in our society (Cross-National Collaborative Group, 1992). These changes can affect the base rates on which to evaluate the sensitivity of primary disorders as predictors of secondary disorders. This problem is illustrated by the work of Weiss, Mirin, and Griffin (1988), who assessed a sample of hospitalized cocaine abusers 1980-1982 and found high rates of primary affective disorder. In a replication between 1982 and 1988, however, much weaker evidence of primary affective disorder was found. The authors concluded that this change reflects the fact that cocaine use is becoming more widespread and, at least in some segments of society, normative.

While secular changes of this sort can complicate interpretation about the importance of mental disorders as causes of substance use and vice versa, they also present special opportunities. For example, the finding of Weiss and his colleagues that the impact of cocaine abuse on course of primary depression became smaller over time as the prevalence of cocaine use increased in the population suggests that the strong initial effect of cocaine in the early 1980s was due more to the social meanings of cocaine use during that period than to the direct effects of the substance itself. Analyses of comorbidity that use short-term historic changes of this sort in creative ways could provide important insights that have been overlooked in prior investigations.

Retrospective Versus Prospective Samples. A final issue concerns the fact that most of the evidence currently available on psychiatric comorbidity is based on retrospective samples. While retrospective data provide a quick and valuable overview of longitudinal patterns, they suffer from recall bias. Much more persuasive data could be obtained by carrying out true prospective studies. In the latter, results of retrospective studies would be used to target persons with a history of a first disorder who are thought to be at risk for the onset of inter disorders and these people would be followed over time. Retrospective studies suggest, for example, that elementary school girls with phobias are at very high risk of prescription drug abuse during late adolescence. A prospective study that recruited a sample of these girls during their elementary school years and followed them over a decade into late adolescence would be valuable in evaluating the importance of a wide variety of potential risk factors for onset of secondary substance abuse. The fact that risk of substance abuse is high in this segment of the population means that the sample would not have to be large. Furthermore, the time period over which the sample would be followed could be carefully selected on the basis of results obtained in earlier retrospective studies to keep the length of longitudinal assessment at a minimum.

Prevention of Secondary Disorders

OVERVIEW

We began this paper with an argument for the potential utility of a class of preventive interventions focused on primary prevention of secondary disorders. We recognize that such interventions could be seen as special types of indicated interventions (Gordon, 1987) aimed at either high-risk primary prevention or tertiary prevention (Leavell & Clark, 1953) depending on the intervention targets. Nonetheless, we believe that it is useful to think of them as distinct from these broader categories of intervention because of the special opportunities they create for prevention. As discussed earlier, the latter include ease and reliability of identifying target subjects who are at high risk of future disorders, the ability to use already developed treatment technologies as part of the interventions, the increased social warrant for and acceptance of the interventions by recipients, and the potential for reducing the exacerbation of primary disorders which is known to accompany the onset of secondary disorders. We recognize that there might also be disadvantages compared to other interventions. For example, if primary disorders complicate prevention efforts they will lead to more intensive effort being needed to prevent secondary disorders. There are also complex issues of recruitment and stigmatization to consider. These issues need to be addressed before we can seriously consider launching interventions for primary prevention of secondary disorder.

In addition to addressing these practical issues, we need to assess whether interventions of this type are feasible from an etiologic perspective. This requires formative research aimed both at clarifying the nature of the causal mechanisms that create comorbidity and at identifying the most promising points of intervention. A number of methodological problems arise in doing this. These problems were the major focus of this paper. We attempted, whenever possible, not only to describe the problems but also to offer suggestions for how they might be resolved. We recognize that these suggestions were sometimes inadequate, but this is an inherent limitation of formative research. The best we can hope for is to narrow the range of uncertainty involved in this part of the research process in order to help the intervention specialist make more informed judgments about the feasibility of preventive interventions and promising intervention targets, recognizing that the only way to resolve these uncertainties is to carry out the intervention and use experimental methods to evaluate its effects. This formative research has not yet been done in the area of comorbidity. Our hope is that the methodological considerations presented here will help motivate researchers to carry out this research and, in this way, help determine whether primary prevention of secondary disorders is feasible.
ACKNOWLEDGMENTS

Work on this paper was supported by Research Grant MH-41135, Training Grant MH-16806, Prevention Center Grant MH-38330, and Research Scientist Development Award MH-00507, all from the National Institute of Mental Health. We appreciate the very helpful comments of the journal reviewers.

REFERENCES


Prevention of Secondary Disorders


