Exploratory Data Mining Analysis Identifying Subgroups of Patients With Depression Who Are at High Risk for Suicide

Mark A. Ilgen, PhD; Karen Downing, MSI; Kara Zivin, PhD; Katherine J. Hoggatt, PhD; H. Myra Kim, ScD; Dara Ganoczy, MPH; Karen L. Austin, MPH; John F. McCarthy, PhD; Jignesh M. Patel, PhD; and Marcia Valenstein, MD

Objective: Although prior research has identified a number of separate risk factors for suicide among patients with depression, little is known about how these factors may interact to modify suicide risk. Using an empirically based decision tree analysis for a large national sample of Veterans Affairs (VA) health system patients treated for depression, we identified subgroups with particularly high or low rates of suicide.

Method: We identified 887,859 VA patients treated for depression between April 1, 1999, and September 30, 2004. Randomly splitting the data into 2 samples (primary and replication samples), we developed a decision tree for the primary sample using recursive partitioning. We then tested whether the groups developed within the primary sample were associated with increased suicide risk in the replication sample.

Results: The exploratory data analysis produced a decision tree with subgroups of patients at differing levels of risk for suicide. These were identified by a combination of factors including a co-occurring substance use disorder diagnosis, male sex, African American race, and psychiatric hospitalization in the past year. The groups developed as part of the decision tree accurately discriminated between those with and without suicide in the replication sample. The patients at highest risk for suicide were those with a substance use disorder who were non–African American and had an inpatient psychiatric stay within the past 12 months.

Conclusions: Study findings suggest that the identification of depressed patients at increased risk for suicide is improved through the examination of higher order interactions between potential risk factors.

J Clin Psychiatry 2009;70(11):1495–1500

Submitted: *October 13, 2008; accepted January 2, 2009 (doi:10.4088/JCP.08m04795).*

Corresponding author: *Mark A. Ilgen, PhD, 4250 Plymouth Rd, Ann Arbor, MI 48109-5763 (marki@umich.edu).*

Suicide is a major and potentially preventable public health problem.[1] In the United States, over 30,000 people die by suicide every year, and suicide accounts for 4% of all deaths among adults aged 18–65 years.[2] Depression is a major risk factor for suicide.[3,4] Two comprehensive meta-analyses[5,6] estimated that 2%–6% of individuals with affective disorders will die by suicide.

However, the vast majority of patients with depression do not die by suicide, and more research is needed to better understand profiles of risk among patients with depression.[7,8] Typically, studies of suicide risk in individuals with depression have examined the independent effects of demographic factors, such as age and race, and of comorbid psychiatric conditions, such as substance use or personality disorders. Prior research[9–11] has identified a number of risk factors, including a diagnosis of an alcohol or drug use disorder, male sex, white race/ethnicity, single marital status, unemployment, prior psychiatric hospitalization, and poor physical health. For example, Zivin and colleagues[12] found that among Veterans Affairs (VA) health system patients with depression, inpatient psychiatric hospitalizations in the year prior to their qualifying depression diagnosis and comorbid substance use were independent risk factors for suicide. Also, patients living in the southern or western United States were more likely to die by suicide. This approach is helpful for identifying general indicators of risk, but it does not provide information about how factors work together to influence this risk.

Suicide is a rare event, and risk assessment requires the use of very large samples. It is therefore not surprising that few studies have evaluated interactions among risk factors. One report[13] observed that problem drinking may increase risks differentially in whites compared to blacks. Also, few studies have examined relationships between more than 2 risk factors. Standard logistic regression analyses typically examine interactions only between pairs of risk factors, and their results are often difficult to apply in clinical practice.

An alternative approach uses exploratory techniques to identify subgroups of patients, characterized by clusters of specific risk factors, who are either at particularly high or low risk for suicide.[14,15] Data mining approaches identify the single strongest indicator of risk within the overall sample

For Clinical Use

- In veterans with depression, a substance use disorder is the strongest risk factor for suicide.
- Knowledge of combinations of risk factors could potentially help clinicians identify veterans at particularly high risk for suicide.
- Clinicians treating patients with depression should regularly assess for suicidal thoughts, plans, and recent attempts throughout the course of treatment.

and certain subgroups in each sequential assessment point, taking into account the presence or absence of stronger predictors.[16] Baca-García and colleagues[14] reported that a series of higher-order interactions among 3 or more variables developed through data mining could better predict rehospitalization among individuals with a prior suicide attempt than could traditional multivariate analyses. However, to the best of our knowledge, this approach has not been directly applied to examine suicide in individuals with depression.

In the present study, we conducted an exploratory analysis designed to derive an empirically based set of interactions related to rates of suicide in a national sample of VA patients diagnosed with depression. This analysis was conducted with a randomly selected portion of our sample (ie, two-thirds of the original sample), and a confirmatory analysis of the utility of these groups was conducted on the remaining one-third of the sample.

METHOD

Procedures and Participants

This study utilized data mining techniques to reanalyze data from a retrospective cohort study of suicide in VA patients with depression.[12] Participants were drawn from the VA's National Registry for Depression (NARDEP), based at the VA Serious Mental Illness Treatment Research and Evaluation Center in Ann Arbor, Michigan.[17] The NARDEP includes data on demographic factors, diagnostic characteristics, pharmacy data, and service utilization for VA patients with depression. The present study uses data from April 1, 1999, through September 30, 2004. Participants were included if they either (1) had been diagnosed with a depressive disorder and had been prescribed an antidepressant or (2) were diagnosed with depression during 2 separate medical visits in the study period. Patients were excluded if they had bipolar I disorder, schizophrenia, or schizoaffective diagnoses recorded during the observation period or the year prior to entry into the cohort. The use of these inclusion and exclusion criteria resulted in a sample of 887,859 patients.

To examine suicide risk in these patients, we linked treatment record data with information from the National Death Index (NDI). The NDI queries were submitted for the cohort of patients with a date of death in the VA Beneficiary Identification and Records Locator System (BIRLS) Death File during the study period. Because BIRLS data may not detect all individuals who died within a given year, we also initiated NDI searches for patients who did not use VA services in the year following the study period, even if there was no date of death for these patients in VA data. This comprehensive process identified 1,892 patients in the study cohort as having died of suicide using *ICD-9* or *ICD-10* codes specifically identifying suicide in NDI data (*ICD-9* codes E950–E959 or *ICD-10* codes X60–X84, Y87.0).

The sample was divided into primary and replication samples. A total of 589,825 (66.4% of the original sample) were randomly selected to be part of the primary sample, and 298,034 (33.6% of the original sample) were selected to be in the replication sample. On average, participants in the primary sample were 58.63 (± 14.38) years of age; other characteristics of participants in the primary sample are presented in Table 1. The primary and replication samples did not significantly differ on any of these characteristics.

Measures

All analyses focused on predicting suicide as the primary outcome. As noted, suicide deaths were identified using NDI searches. The NDI is a national repository of data collected from state death certificates, and it is considered the "gold standard" of US mortality databases.[18]

The following candidate risk factors were derived from patient treatment records: age, race (white, African American, other, Asian, American Native, Native Hawaiian or Pacific Islander, multiracial, or unknown), sex (male or female), substance use disorder diagnosis (no or yes; ie, a clinical diagnosis of any alcohol or drug use disorder), diagnosis of posttraumatic stress disorder (PTSD; no or yes), prior VA hospitalization (no or yes; for any psychiatric diagnosis in the year prior to cohort entry), service connection (no or yes; indicating some VA-recognized disability stemming from injuries or conditions that occurred or were exacerbated during military service), and baseline medical comorbidity reflecting the sum of all conditions in the past 12 months included in the Charlson Comorbidity Index.[19]

Analyses

The main set of data mining analyses was conducted on the primary sample. The replication sample was used to

Table 1. Patient Characteristics of the Primary Sample (N = 589,825)

Characteristic	n	%
Sex		
Male	542,109	91.9
Female	47,716	8.1
Race and ethnicity		
White	450,704	76.4
Black	73,126	12.4
Other	12,803	2.2
Unknown	53,192	9.0
Hispanic	27,522	4.7
Substance abuse	130,228	22.1
PTSD	130,124	22.1
Inpatient psychiatric stay in past 12 mo	21,047	3.6
Service connection	186,261	31.6

Abbreviation: PTSD = posttraumatic stress disorder.

verify the original set of results using more traditional statistical methods. The data mining procedures used Bayesian Dirichlet Equivalent (BDE) methods to identify the strongest predictor. This data mining process examined possible higher-order relationships between predictors.[20] For continuous variables, these determine the optimal cut points for splitting continuous predictors, such as age, into levels that maximally differentiate between levels of suicide risk.[21] This method has the advantage of using prior information to allow for the application of these methods to rare outcomes, such as suicide. In order to increase the potential utility of the results, we limited the results to a maximum of 3 cut points (similar to a 3-way interaction). This process resulted in an empirically derived decision tree with patients grouped at each level based on similarities in level of suicide risk. Additionally, for all cut points on the figures, we present rates of suicide per 100,000 person-years (PYs). This information is provided for descriptive purposes and is not directly part of the decision tree analysis.

Following the development of the decision tree in the primary sample, we examined the extent to which the factors identified from the data mining approach could accurately discriminate between those with and without a suicide in the replication sample using χ^2 analyses and odds ratios. A cut-off of $P < .01$ was used as an indicator of statistical significance.

RESULTS

The overall rate of suicide was 89.55/100,000 PYs in the primary sample of depressed veterans. The data mining analysis identified a diagnosis of a substance use disorder as the strongest single risk factor for suicide (Figure 1). The suicide rate for those with a substance use disorder was 128.27/100,000 PYs relative to 76.48/100,000 PYs for those without a substance use disorder. The data mining analysis split the sample into those with or without a substance use disorder and identified the most important remaining risk factors in these 2 separate groups.

Individuals With a Substance Use Disorder

Depressed VA patients with a substance use disorder were subsequently divided on the basis of their race. Specifically, African Americans were significantly less likely to die by suicide than individuals of any other race/ethnicity group (28.54/100,000 PYs vs 157.11/100,000 PYs). In African Americans with a substance use disorder, no other variables reliably distinguished those who did versus those who did not die by suicide. In non–African Americans (who were predominately white), having been admitted to inpatient psychiatric treatment at least once in the prior 12 months conferred an additional risk for suicide. In the subgroup of depressed, non–African American veterans with a substance use disorder, the suicide rate was 247.21/100,000 PYs among those with a prior inpatient psychiatric stay, whereas it was 143.05/100,000 PYs among those with no psychiatric admissions in the prior year.

Individuals Without a Substance Use Disorder

In those without a substance use disorder, sex was the strongest indicator of suicide risk. The suicide rate for men in this group was 82.08/100,000 PYs versus 26.22/100,000 PYs for women. In women, no other available risk factors reliably identified individuals at differing levels of risk. In men without a substance use disorder, race again was related to suicide risk. Specifically, African American men without a substance use disorder were less likely to die by suicide, with a rate of 33.30/100,000 PYs compared to 86.78/100,000 PYs for non–African American men.

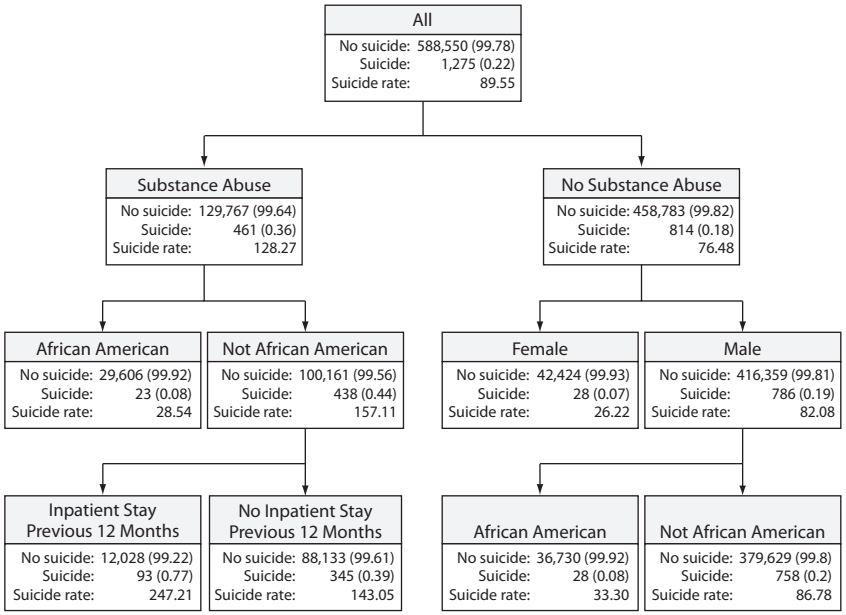
Results From the Replication Sample

A series of comparisons were conducted in the replication sample to examine the extent to which more traditional analytic techniques would yield similar results. For these analyses, we reconstructed the prior steps in the decision tree in an a priori manner and examined the extent to which each additional step significantly distinguished between the groups. Table 2 presents the results of χ^2 testing of the groups generated by the data mining analyses in both the primary and replication samples. In all cases, the groups produced by the initial exploratory data mining analyses significantly discriminated those who died of suicide from those who did not.

DISCUSSION

Using a large national database of VA patients with depression, we found that a combination of factors, including demographic characteristics, psychiatric disorders, and prior psychiatric hospitalizations, discriminated between individuals with and without a suicide. Consistent with prior research on this sample,[12] a substance use disorder was the single strongest indicator of suicide risk in depressed veterans. However, the sample could be further divided into smaller groups of high- and low-risk individuals. Depending

Figure 1. Suicide Rates in Depressed Veteran Affairs Patients[a]



[a]Suicide rates are presented as estimated number of suicides per 100,000 person-years and No suicides and Suicides are measured as n (%).

Table 2. Associations Between Risk Factors and Suicide

Step	Group/Subgroup	Predictor	χ^2	*P*	Odds Ratio (95% CI)
Primary sample					
1	Primary sample	Any substance use disorder (yes vs no)	147.2	<.0001	2.0 (1.8–2.2)
2a	Those with a substance use disorder	Non–African American (yes vs no)	83.1	<.0001	5.6 (3.7–8.6)
3a	Those with a substance use disorder who are not African American	Any inpatient treatment in the prior 12 mo (yes vs no)	35.0	<.0001	2.0 (1.6–2.5)
2b	Those without a substance use disorder	Sex (male vs female)	32.7	<.0001	2.9 (2.0–4.2)
3b	Those without a substance use disorder who are male	Non–African American (yes vs no)	27.0	<.0001	2.6 (1.8–3.8)
Replication sample					
1	Full replication sample	Any substance use disorder (yes vs no)	47.6	<.0001	1.8 (1.5–2.1)
2a	Those with a substance use disorder	Non–African American (yes vs no)	44.7	<.0001	8.5 (4.0–18.1)
3a	Those with a substance use disorder who are not African American	Any inpatient treatment in the prior 12 mo (yes vs no)	25.6	<.0001	2.3 (1.6–3.2)
2b	Those without a substance use disorder	Sex (male vs female)	16.7	<.0001	2.9 (1.7–4.9)
3b	Those without a substance use disorder who are male	Non–African American (yes vs no)	13.6	<.0001	2.6 (1.5–4.5)

on the constellation of risk factors, the risk of suicide in veterans with depression ranged from a low of 26.22/100,000 PYs among women without a substance use disorder to a high of 247.21/100,000 PYs among non–African American patients with a substance use disorder and an episode of inpatient psychiatric treatment within the past year. This substantial variability in level of risk highlights the need to better understand how risk factors interact and combine to influence risk for suicide.

The finding that a diagnosis of a substance use disorder was the strongest single predictor of suicide in depressed veterans is consistent with prior research[22] in nondepressed samples indicating that presence of an alcohol or drug use disorder is one of the most potent risk factors for suicide mortality. According to an empirical review of cohort studies,[6] individuals with alcohol dependence are almost 10 times more likely to die by suicide compared to the general population. However, this research has not generally focused on additional risks conveyed by substance use disorder in other psychiatric populations, including depressed individuals. The strength of the association in the present sample of depressed veterans reinforces the importance of substance use disorder as an additional risk factor even among depressed patients who are already at an elevated risk for suicide.

Examining interactions among risk factors yields other potentially important findings, including the implication that sex is less immediately relevant in determining suicide

risk once substance use is determined to be present. Sex is one of the most potent risk factors for suicide, with men being 4 times more likely to complete suicide than women in the general population.[2] The odds ratios for suicide associated with male compared to female sex are still high among depressed populations, although somewhat smaller than in the general population. The ordering of risk factors in this study emphasizes the need to remain alert for suicide among depressed women as well as men once substance use is identified. It is also important to note that among depressed patients, once substance use is determined to be present, the most important factor to consider next is race. Sex is not among the top 3 risk factors for suicide among patients with both depression and substance use. This finding has implications for clinicians completing suicide risk assessments of dual-diagnosis patients. Among patients who do not have substance use, sex again clearly emerges as one of the strongest predictors of suicide, even greater than other potentially important indicators of illness severity such as prior inpatient stays. However, these findings also most likely reflect the composition of the present sample. It is important to note that the sample is predominantly male and that we may not have been able to detect the impact of sex on suicide risk in certain subgroups.

This was an exploratory data analysis, and the decision tree was derived purely on the basis of the ability to discriminate between participants on completed suicide. This decision process does not adhere to established theory of suicide risk, and the associations are not necessarily causal. Instead, these factors are markers of risk. Thus, they may reflect a direct influence on suicide risk or the influence of other unmeasured variables.[23] Also, even in our highest risk groups, the vast majority of participants did not die by suicide. These results highlight the difficulty in developing specific models of low base-rate outcomes like suicide. Similarly, the risk of suicide most likely fluctuates over time, and we do not have data on proximal risk factors for suicide. Identifying individuals at immediate risk for suicide requires a combination of knowledge of risk factors, attention to recent life events (losses, conflicts), and clinical judgment.

Several additional limitations of this study should be noted. The sample is composed of depressed veterans receiving health services from the Department of Veterans Affairs. Thus, it is important to be cautious in generalizing these findings to other depression treatment populations. All predictors in these analyses were based on data derived from VA treatment records. The measurement of diagnostic features, including PTSD and substance use disorders, reflects the real-world utilization of these diagnoses and most likely differs from gold-standard measurements of these disorders with structured clinical instruments. In this sample, no reliable data were available on prior suicide attempts. Future research in other samples would benefit from the inclusion of information related to prior suicidal behaviors. Also, designations of suicide mortality were obtained from NDI data and are therefore necessarily limited by the accuracy of information in death certificates.

Despite these limitations, this is the first study of which we are aware to use an empirically derived decision tree to segment the depression treatment population by suicide risk. Our results indicate that the association between a known risk factor and suicide in veterans with depression is clearly not independent of another risk factor. Unlike traditional logistic regression analyses that include only main effects or interaction terms specified a priori by researchers, these analyses allowed for the examination of more complicated interactions that may not have been hypothesized to be important based on the existing research. Additionally, the decision tree allows for the presentation of complicated interactions, such as the one found between substance use disorder status, race/ethnicity, and 12-month hospitalization among suicide risk factors in a manner that is likely to be readily understandable by clinicians and policy makers.

Patients at highest risk for suicide among depressed patients are those with a substance use disorder, those of non–African American race, and those with an inpatient stay in the past 12 months. Among patients without substance use disorder, those who are male and non–African American are at the greatest risk of suicide. Providers treating patients with these characteristics should be aware of these risks and consider target strategies to screen for current suicidal thoughts and plans.

Disclosure of off-label usage: The authors have determined that, to the best of their knowledge, no investigational information about pharmaceutical agents that is outside US Food and Drug Administration–approved labeling has been presented in this article.
Author affiliations: Veterans Affairs Serious Mental Illness Treatment Research and Evaluation Center, Ann Arbor, Michigan (Drs Ilgen, Zivin, Hoggatt, Kim, McCarthy, Valenstein, and Mss Downing, Ganoczy, and Austin); and Department of Psychiatry (Drs Ilgen, Zivin, McCarthy, Valenstein, and Ms Downing), Department of Epidemiology (Dr Hoggatt), Center for Statistical Consultation and Research (Dr Kim), and Department of Electrical Engineering and Computer Science (Dr Patel), University of Michigan, Ann Arbor.
Financial disclosure: **Drs Ilgen**, **Zivin**, **Hoggatt**, **Kim**, **McCarthy**, **Patel**, and **Valenstein**, and **Mss Downing**, **Ganoczy**, and **Austin** have no personal affiliations or financial relationships with any commercial interest to disclose relative to the article.
Funding/support: This work was supported by US Department of Veterans Affairs Health Services Research and Development Service grants MRP 05-137 and IIR 04-104-2 and by National Institute of Mental Health grant R01-MH078698-01.
Disclaimer: The views expressed in this report are those of the authors and do not necessarily represent those of the US Department of Veterans Affairs or the National Institute of Mental Health.

REFERENCES

1. Institute of Medicine. *Reducing Suicide: A National Imperative*. Washington, DC: National Academies Press; 2002.
2. Centers for Disease Control and Prevention. *Web-based Injury Statistics Query and Reporting System: Leading Causes of Death Reports*. Atlanta, GA: Centers for Disease Control and Prevention; 2007.
3. Bertolote JM, Fleischmann A, De Leo D, et al. Psychiatric diagnoses and suicide: revisiting the evidence. *Crisis*. 2004;25(4):147–155.
4. Henriksson MM, Aro HM, Marttunen MJ, et al. Mental disorders and

comorbidity in suicide. *Am J Psychiatry*. 1993;150:935–940.
5. Bostwick JM, Pankratz VS. Affective disorders and suicide risk: a reexamination. *Am J Psychiatry*. 2000;157(12):1925–1932.
6. Inskip HM, Harris E, Barraclough B. Lifetime risk of suicide for affective disorder, alcoholism and schizophrenia. *Br J Psychiatry*. 1998;172:35–37.
7. Dumais A, Lesage AD, Alda M, et al. Risk factors for suicide completion in major depression: a case-control study of impulsive and aggressive behaviors in men. *Am J Psychiatry*. 2005;162(11):2116–2124.
8. McGirr A, Renaud J, Seguin M, et al. An examination of DSM-IV depressive symptoms and risk for suicide completion in major depressive disorder: a psychological autopsy study. *J Affect Disord*. 2007;97(1–3):203–209.
9. Conwell Y, Duberstein PR, Cox C, et al. Relationship of age and Axis I diagnoses in victims of completed suicide: a psychological autopsy study. *Am J Psychiatry*. 1996;153(8):1001–1008.
10. Qin P, Agerbo E, Mortensen PB. Suicide risk in relation to socioeconomic, demographic, psychiatric, and familial factors: a national register-based study of all suicides in denmark, 1981–1997. *Am J Psychiatry*. 2003;160(4):765–772.
11. Waern M, Runeson BS, Allebeck P, et al. Mental disorder in elderly suicides: a case-control study. *Am J Psychiatry*. 2002;159(3):450–455.
12. Zivin K, Kim HM, McCarthy JF, et al. Suicide mortality among individuals receiving treatment for depression in the Veterans Affairs Health System: associations with patient and treatment setting characteristics. *Am J Public Health*. 2007;97(12):2193–2198.
13. Castle K, Duberstein PR, Conner KR, et al. Risk factors for suicide in blacks and whites: an analysis of data from the 1993 National Mortality Followback Survey. *Am J Psychiatry*. 2004;161(3):452–458.
14. Baca-García E, Perez-Rodriguez M, Basurte-Villamor I, et al. Using data mining to explore complex clinical decisions: a study of hospitalization after a suicide attempt. *J Clin Psychiatry*. 2006;67(7):1124–1132.
15. Tiet QQ, Ilgen MA, Byrnes HF, et al. Suicide attempts among substance use disorder patients: an initial step toward a decision tree for suicide management. *Alcohol Clin Exp Res*. 2006;30(6):998–1005.
16. Obenshain MK. Application of data mining techniques to healthcare data. *Infect Control Hosp Epidemiol*. 2004;25(8):690–695.
17. Blow FC, Owen RE, Valenstein M, et al. *Specialty Care for Veterans With Depression in the VHA 2002 National Depression Registry Report*. Ann Arbor, MI: Department of Veterans Affairs National Serious Mental Illness Treatment Research & Evaluation Center; 2003.
18. Cowper DC, Kubal JD, Maynard C, et al. A primer and comparative review of major U.S. mortality databases. *Ann Epidemiol*. 2002;12(7):462–468.
19. Charlson ME, Pompei P, Ales KL, et al. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *J Chronic Dis*. 1987;40(5):373–383.
20. Heckerman D, Geiger D, Chickering M. Learning Bayesian networks: the combination of knowledge and statistical data. Technical Report MSR-TR-94-09. Redmond, WA: Microsoft Research; 1994
21. Heckerman D, Geiger D, Chickering DM. Learning Bayesian networks: the combination of knowledge and statistical data. *J Mach Learn Res*. 1995;20:197–243.
22. Wilcox HC, Conner KR, Caine ED. Association of alcohol and drug use disorders and completed suicide: an empirical review of cohort studies. *Drug Alcohol Depend*. 2004;76(suppl):S11–S19.
23. Kraemer HC, Lowe KK, Kupfer DJ. *To Your Health: How to Understand What Research Tells Us About Risk*. New York, NY: Oxford University Press; 2005.

For the CME Posttest for this article, see pages 1614–1615.