Interrater Agreement Among Psychiatrists in Psychiatric Emergency Assessments

Bruce B. Way, Ph.D., Michael H. Allen, M.D., Jeryl L. Mumpower, Ph.D., Thomas R. Stewart, Ph.D., and Steven M. Banks, Ph.D.

Objective: The authors' purpose in this study was to investigate the interrater agreement among psychiatrists in psychiatric emergency service settings. The interrater reliability of many of the key concepts in psychiatric emergency service settings has not been studied. **Method:** Videotapes of 30 psychiatric emergency service patient assessment interviews conducted by psychiatrists were shown to eight experienced psychiatric emergency service psychiatrists. The eight psychiatrists rated each videotape on dimensions such as severity of depression and psychosis and recommended a disposition for each patient. Interrater reliability was then explored. **Results:** The level of agreement (intraclass correlation coefficient) among the reviewing psychiatrists was higher for psychosis and substance abuse but lower for psychopathology, impulse control problems, danger to self, and disposition. The reviewers' disposition recommendations did not match well with the assessing psychiatrist's actual disposition, but comparisons with actual practice should be considered only suggestive. **Conclusions:** Psychiatric emergency service assessments need improvement. This may be accomplished by exploring the underlying structure of psychiatric emergency service concepts, the creation and validation of structured assessment tools, and the creation of practice guidelines.

(Am J Psychiatry 1998; 155:1423–1428)

The psychiatric assessment conducted in a psychiatric emergency service and the resulting disposition have major physical, psychological, and fiscal effects on the patient, his or her family, the community, and insurance carriers. Inappropriate release may lead to violence against another community member (1) and, as a result, negative media attention. Inappropriate release may also increase an individual's risk of suicide and other sources of mortality, may burden his or her support system, and may result in further deterioration of the underlying condition.

On the other hand, inappropriate admissions may be disruptive and stigmatizing to the individual (2) and may "determine the choice of subsequent treatment plans and often influences the course of the problem or the illness" (3, p.1). Inappropriate admissions may lead to the loss of jobs, housing, and child custody and may have negative financial implications for families that depend on the hospitalized individuals.

There are recent changes that may increase both the difficulty and consequences of the psychiatric emergency service assessment task. Psychiatric emergency rooms have become a main entry point for patients seeking mental health treatment (4, 5), and the clinical profiles of patients have changed to include a much higher proportion who are chronically ill and difficult to engage in treatment (4, 5). At the same time, access to inpatient care is increasingly "managed," which limits the psychiatrist's ability to "lean" toward hospitalization for borderline cases. The psychiatric emergency service task, therefore, may require greater thoroughness and precision.

Furthermore, the psychiatric emergency service assessment, unlike most other services where patient and evaluator are beginning a treatment relationship, is not static. It may be directed in at least five different ways under different circumstances, and the task may change dynamically as time elapses, or as new infor-

Presented in part at the 150th annual meeting of the American Psychiatric Association, San Diego, May 17–22, 1997. Received July 2, 1997; revision received Jan. 16, 1998; accepted Feb. 26, 1998. From the Center for the Study of Issues in Public Mental Health, Albany, N.Y.; the New York State Office of Mental Health; the Comprehensive Psychiatric Emergency Program, Bellevue Hospital and New York University, New York City; and the Center for Policy Research, Rockefeller College of Public Affairs and Policy, University at Albany, State University of New York, Albany. Address reprint requests to Dr. Way, New York State Office of Mental Health, 44 Holland Ave., Albany, NY 12229; Way@iris.rfmh.org (e-mail).

Supported in part by NIMH grant MH-51359 to the Center for the Study of Issues in Public Mental Health.

mation or events develop, or the effects of emergency interventions appear. Evaluation is first directed at strategies to contain serious behavioral disturbances and facilitate further evaluation. Next, it is necessary to separate consequential medical problems with psychiatric symptoms from primary psychiatric syndromes. Evaluation may then be directed at identification of the major problems contributing to the presentation and a determination of the setting in which these will be further evaluated and treated. A diagnosis of some sort will then be rendered.

Finally, there is no consensus about the scope or the role of diagnosis in psychiatric emergency service evaluations (6). Specific diagnosis, which requires a longitudinal rather than cross-sectional approach and is labor intensive, has not traditionally been highly valued in the psychiatric emergency service. For a variety of reasons, emergency assessment is usually directed narrowly at the current circumstances, focusing on dangerousness, severity of illness, and the possible benefits of treatment, particularly hospitalization. Gerson and Bassuk (3) described the dominant psychiatric emergency service practice as "rapid evaluation, containment, and referral" and advocated a focus on "the patient's and the community's adaptive resources and competence and minimizing subtle diagnostic considerations" (p. 9). In their model, patients are triaged and little or no treatment is provided in the psychiatric emergency service. However, as gatekeeping increases and alternatives in the community proliferate, specific diagnosis and immediate treatment in the psychiatric emergency service are becoming more important (7).

Despite the importance and difficulty of emergency assessments, few studies have reported on the interrater reliability of psychiatric assessments conducted in a psychiatric emergency service. No one has investigated whether there is agreement among psychiatric emergency service evaluators on disposition or on the measurement of issues such as danger to self.

The reliability of prescribing psychiatric medication, judging depression, and making psychiatric diagnoses has been studied in nonpsychiatric emergency service settings, but in the psychiatric emergency service, only diagnosis has been studied. Several studies have investigated the reliability of psychiatric diagnosis (8–14); many of these reported low interrater reliability. As an example, Hjortso et al. (9) created 24 vignettes based on clinical records and asked seven experienced psychiatrists to make diagnostic judgments. Agreement coefficients were 0.55 for psychiatric syndromes such as schizophrenia, 0.52 for personality disorders, 0.66 for psychosocial stressors, and 0.47 for global functioning.

Studies have compared psychiatric emergency service diagnoses with subsequent diagnoses. Lieberman and Baker (10) compared psychiatric emergency service diagnoses with the subsequent discharge diagnoses and reported the following kappas: psychosis (kappa=0.64), depression (kappa=0.62), alcoholism (kappa=0.77), schizophrenia (kappa=0.41), and bipolar disorder (kappa=0.55). A more recent study (14), however, found higher levels of agreement between psychiatric emergency service and discharge diagnoses for schizophrenia and schizoaffective disorder (kappa=0.82), bipolar disorder (kappa=0.72), substance use disorders (kappa=0.87), and major depression (kappa=0.64).

Gillis and colleagues (15, 16) investigated psychiatric medication decisions and found that interrater reliability among psychiatrists for class of medication was 0.6, which was no better than chance. Fisch and colleagues (17, 18) investigated severity of depression ratings made by psychiatrists and general medicine physicians and found an average interrater agreement of 0.52. Several studies have found that use of structured assessment instruments and explicit criteria are associated with improvements in interrater reliability (7, 11, 19, 20).

METHOD

Videotapes of 30 psychiatric emergency service assessment interviews between the psychiatrist and the patient were reviewed and rated by eight psychiatrists, most of whom were senior psychiatrists. The 30 interviews, from a total of 97 collected in 1994–1995 at four urban psychiatric emergency services, were selected to cover the range of psychiatric disorders, to represent each hospital approximately equally, and to represent the dispositional categories. Tapes were not selected if they had audio or video problems, if they were conducted by one of the reviewing psychiatrists (because the reviewer might remember information not included on the tape), or if they were brief. Personal and geographical identifiers as well as the disposition were edited from the tapes. An attending psychiatric emergency service psychiatrist at Albany Medical Center, N.Y., reviewed most of the selected tapes and pilot-tested the rating instrument used. All protocols were approved by the appropriate institutional review committees. No data were collected without the patient's previous written consent.

Two psychiatrists from each psychiatric emergency service were asked to participate. Six were the most senior attending psychiatrists regularly seeing patients in their psychiatric emergency service, one was a fifth-year fellow in psychiatry, and another was a chief resident. Fellows and chief residents usually act as attending psychiatrists by making dispositional decisions without consultation. The reviewers' psychiatric emergency service experience varied from 11 to 200 months with a median of 38 months.

The reviewers were asked to rate the videotaped patient on danger to self, danger to others, psychopathology, depression, psychosis, impulse control problems, substance abuse, social support, ability to care for self, benefit of inpatient treatment, and patient cooperation. These concepts or cues have been identified in the literature as important issues related to disposition (21). The psychiatrists also rated need for information from collateral sources (e.g., family), recommended disposition, rated confidence in their dispositions, and rated the quality of the videotaped assessment interview. All of the questions used an 8-point response scale (0=none, 1=low, and 7=high, except as follows: 0=definitely discharge and 7=definitely admit on the recommended disposition scale, 0=very unconfident and 7=very confident on the confidence scale, and 0=poor and 7=excellent on the quality of the interview scale). Eight reviewers of 30 interviews yielded a total of 240 observations for analysis.

Use of videotapes is an improvement over most investigations of interrater agreement. Typically, case vignettes ("paper cases") with varying values on the concept dimensions are presented to experts, who are asked to form judgments (22, 23). Videotape methods, on the other hand, more closely approximate real practice by permitting

TABLE 1. Intraclass Correlations and Variance Decomposition for Emergency Service Concepts Rated by Eight Emergency Service Psychiatrists After Viewing 30 Videotapted Assessments

Concept	Intraclass Correlation Coefficient	Mean Sum of Squares			Estimate of Variability	
		Due to Patient (df=29)	Due to Physician (df=7)	Error (df=203)	Due to Patient	Due to Physician
Psychopathology	0.28	6.52[a]	15.50[a]	1.58	0.62	0.46
Ability to care for self	0.28	6.51[a]	12.10[a]	1.56	0.62	0.35
Impulse control problems	0.30	9.75[a]	32.31[a]	2.17	0.95	1.00
Benefit of inpatient treatment	0.30	12.19[a]	17.80[a]	2.70	1.19	0.50
Danger to self	0.32	12.00[a]	26.95[a]	2.56	1.18	0.81
Patient cooperation	0.42	9.17[a]	14.50[a]	1.34	0.98	0.44
Danger to others	0.44	13.05[a]	9.08[a]	1.77	1.41	0.24
Depression	0.48	13.28[a]	23.81[a]	1.59	1.46	0.74
Social supports	0.51	10.48[a]	4.04[a]	1.14	1.16	0.10
Psychosis	0.64	25.91[a]	2.16	1.72	3.02	0.01
Substance abuse	0.65	23.01[a]	11.44[a]	1.44	2.70	0.33
Recommended disposition	0.33	14.19[a]	8.06[a]	2.84	1.42	0.17
Confidence in disposition rating	0.18	5.45[a]	6.47[a]	1.98	0.43	0.15
Need for collateral information	0.28	7.28[a]	23.50[a]	1.80	0.68	0.72
Quality of interview	0.30	6.50[a]	6.74[a]	1.48	0.63	0.18

[a] $p<0.05$, F test.

the psychiatrist to see and hear the patient and by providing data sequentially instead of all at once (23).

Videotape methods are also an improvement over studies that have two psychiatrists interview the same patient (8). In these studies, each evaluator would undoubtedly ask different questions, and this itself could lead to interrater differences (24). Further, the process itself may distort patient's responses from the first to the second interview. With videotape methods, all evaluators are reaching a judgment based on the same information, eliminating one source of variability. Videotape methods also permit the inclusion of several psychiatrists from different hospitals.

The current methods, however, do vary from psychiatric emergency service practice. For example, besides the interview, the psychiatric emergency service psychiatrist could have other information available such as previous records and interviews with family members. There is no way to know whether such additional information would increase or decrease interrater agreement, but there is some evidence to suggest that having more information available may actually decrease agreement (25).

The main questions addressed by this paper are the following: 1) To what extent did the reviewing psychiatrists agree with each other in the assessment of key psychiatric emergency concepts (cues) that the literature suggests are related to disposition? 2) To what extent did reviewers agree on disposition? 3) Did the reviewing psychiatrists agree with the disposition given by the assessing psychiatrist?

The interviews, after editing, varied in length from 9.5 minutes to 57 minutes, with a median of 24.6 minutes. Twenty-seven of the 30 tapes were longer than 15 minutes. The mean rating of the quality of the tapes was 3.9, which was higher than the 3.5 midpoint of the scale. Twenty-four of the interviews were rated higher than 3.0 in quality.

The demographic and diagnostic characteristics of the 30 patients in the interviews matched well with a larger study group of 465 patients collected from all four emergency rooms. Twenty-three (77%) of the patients in the videotapes were men, compared with 64% of the larger group; 15 (50%) of our videotaped patients were white, compared with 51% of the larger group; and the average age of the videotaped patients was 36 years, compared with 37 in the larger group. Twelve (40%) of the videotaped patients had a diagnosis of major mental illness (schizophrenia, bipolar disorder, or psychosis not otherwise specified), compared with 42% in the larger group. Nine (30%) of the videotaped patients were released from the hospital, compared with 48% of the larger group, but this difference was deliberate. Patients admitted to special 72-hour beds (extended observation beds) located in the psychiatric emergency service (N=6) were oversampled to permit future study.

RESULTS

Agreement Among Psychiatrists on Concept Ratings and Recommended Disposition

An intraclass coefficient (26, 27) was calculated for ratings of each of the 15 concepts in emergency psychiatry; these are presented in table 1. Low intraclass correlations suggest a lack of agreement among psychiatrists (1.0 indicates perfect agreement). The intraclass correlation was relatively lower for recommended disposition, psychopathology, impulse control problems, ability to care for self, danger to self, and quality of the interview conducted by the assessing psychiatrist. Judgments with relatively higher agreement were psychosis and substance abuse.

The low levels of agreement found here might be attributed to the quality of the interviews. To investigate this issue, Pearson correlations were calculated between the reviewers' mean ratings of interview quality and level of agreement. Interview quality was not positively related to interjudge agreement on any of the dimensions tested (danger to self, depression, psychosis, impulse control problems, and recommended disposition).

The level of agreement on recommended disposition could also be affected by the oversampling of patients admitted to extended observation beds because this is a fairly new disposition option. To examine this issue, the intraclass coefficient was recalculated with the six extended observation bed cases deleted; the coefficient increased only slightly, to 0.34.

Analysis of variance techniques can also measure the differences attributable to patients, psychiatrists, and the unexplained or error variance. For a perfectly reliable measure, all the variance would be accounted for by differences among patients, and unexplained variance and variance due to differences between physicians would be zero. Variance explained by differences

TABLE 2. Relation of Recommended Patient Disposition (Divided at Midpoint of Scale) to Actual Disposition for Eight Emergency Service Psychiatrists Who Rated 30 Videotaped Assessments

	Percentage Correct			
Physician or Item	Release (N=9)	Admit (N=21)	Overall (N =30)	Cohen's Kappa
Physician 1	77.8	47.6	56.6	0.23
Physician 2	66.7	60.0	62.1	0.23
Physician 3	66.7	42.9	50.0	0.07
Physician 4	77.8	47.6	56.6	0.20
Physician 5	55.6	52.4	53.3	0.07
Physician 6	55.6	52.4	53.3	0.07
Physician 7	77.8	30.0	44.8	0.07
Physician 8	55.6	71.4	66.6	0.26
Overall	66.7	50.6	55.5	0.16
Physicians confident in disposition recommendation[a]			57.4	
Physicians judged interview to be of high quality[a]			56.0	
Extended-observation cases removed			59.7	
Optimal cutoff points for recommended disposition scale			60.9	

[a] Gave variable a rating of 4 or higher on a 0–7 scale.

between physicians is a measure of how consistently the psychiatrists varied in their judgments.

Mean-square statistics and estimates of variability (expected mean squares) were calculated by using standard variance decomposition procedures (26); these are displayed in table 1. As seen in table 1, differences between physicians explained statistically significant variations in all the judgments except psychosis. In terms of the relative size of the three variability estimates, the judgments regarding psychosis and substance abuse have the most desirable characteristics of all the judgments in the study. The variability due to differences among patients is larger than, almost double, the amount due to error, and the amount due to differences among physicians is low. However, there is still a good deal of error that could be reduced in judgments of psychosis and substance abuse. Recommended disposition also had a low amount of consistent physician variability (estimate=0.17), but the error (2.84) was about twice the size of the amount explained by patient differences (estimate=1.42). For impulse control problems, the difference among physicians was even larger than patient variability, and the error term was large. Other judgments with relatively high levels of variance associated with psychiatrists were danger to self and depression.

Actual Versus Recommended Disposition

Pearson correlation coefficients were calculated between the reviewers' recommended dispositions and the actual disposition (admit or release) given by the assessing psychiatrist. The coefficients varied from 0.11 to 0.31, and none was statistically significant.

Although Pearson correlations provided an overall agreement statistic, questions remained concerning whether the recommended disposition matched the actual disposition. Estimation of the percent correct required that the recommended disposition scale be dichotomized, and a logical starting point was to divide the scale in the middle with values 0, 1, 2, and 3 classified as release, and values 4, 5, 6, and 7 classified as admission. Table 2 displays recommended dispositions and actual dispositions. Overall, the physicians matched 55.5% of the time, and, as measured by Cohen's kappa, this was a 16% improvement over chance.

The cases with a mean recommended disposition confidence rating of less than 4 on the 0–7 scale (N=43 of 240) were eliminated and the analysis was rerun. The percent correct improved only slightly, to 57.4% from 55.5%. Similar analyses were conducted after deleting the interviews judged to be of lower quality and the six extended observation cases. As can be seen in at the bottom of table 2, neither procedure greatly improved the percentage correct.

Since no precise admission threshold was specified for the recommended disposition scale, an analysis was conducted to determine the maximum level of agreement with actual disposition. The optimal scale-dividing point for two of the physicians remained in the middle, but for three it was shifted higher—to between 4 and 5—and for three physicians it was shifted lower—one between 2 and 3, and two between 1 and 2. Choosing the optimal scale cutoff points improved the overall percent correct to 60.9% from 55.5%.

DISCUSSION

These results indicate that there was considerable disagreement among psychiatrists concerning the disposition a psychiatric emergency service patient should receive. Segal et al. (24) suggested that psychiatrists would agree on disposition if they all had the same data, but here, when all had exactly the same data, they did not. There was also considerable disagreement on many of the important emergency psychiatric concepts. These are important findings because, to our knowledge, the reliability of these concepts has never before been studied in psychiatric emergency service settings, nor have studies compared recommended dispositions with actual dispositions. Also, we are not aware of any previous study in which psychiatrists have reviewed videotaped psychiatric emergency service assessment interviews.

Apsler and Bassuk (28) found large variability among psychiatric emergency service psychiatrists in the information they use to admit patients; these authors recommended that admission standards be created that specify the important variables. This is a worthwhile goal, but even if most evaluators used the same variables, the results of the current study suggest

that there would still be large variability in practice until the variables were reliably measured.

The levels of interrater agreement found here are similar to those found in studies of other psychiatric settings. Fisch et al. (18) reported a 0.52 agreement coefficient in judgments of severity of depression, which is similar to the 0.48 coefficient found here.

Reasons for Disagreement

A potential cause of low agreement could be that different psychiatrists have different "mental models" (22, 29, 30) about how each concept should be measured. They may disagree on what objective pieces of information available in the interview should be selected and how they should be weighted and combined with other information to form a judgment. This type of disagreement has been found in many areas of expert judgment (22, 23, 29), including psychiatry , 28).

An alternative explanation of the low interrater agreement could be that the current psychiatric emergency service assessment interview does not provide adequate information. Psychiatrists may agree on what objective bits of information are important to measure each concept, but often psychiatric emergency service interviews do not contain the information. In the current study, this alternative explanation was not supported; no relationship was found between the reviewer's ratings of interview quality and levels of agreement. The ratings of quality, however, were for the whole interview, and it would have been useful to have ratings on the adequacy of the assessment for each concept.

Possible Solutions—Improved Definition of Concepts, Measurement, Practice Guidelines, and Emphasis on Diagnosis

To increase reliability in assessment of concepts, we recommend that the field of emergency psychiatry begin to focus on the definitions and underlying structure of the less reliable concepts, such as impulse control problems. Expert panels could be convened to build a consensus on the meaning of these key concepts (31).

We also recommend that the psychiatric emergency service interview be improved by the development or inclusion of assessment scales that target the important psychiatric emergency service judgments. Use of structured instruments has been shown to lead to high interrater reliability (12, 13, 19). Variability due to different "mental models" would be eliminated because the instrument standardizes the combining and weighting of information. To prevent such an emergency assessment from becoming impractical, however, it would be necessary to develop a screening tool consisting of a small number of questions for each dimension. These questions would be used to determine if the complete assessment for that dimension should be administered.

Some consideration should be given to the combining and weighting of the key dimensions to produce a recommended disposition score. Such a score could reduce interrater disagreement on disposition. Although the clinical decision must remain the responsibility of the psychiatrist, the recommended disposition score could be a useful additional piece of information. Attempts have been made in this regard (20, 31).

Further, it may be useful to increase the importance of diagnosis in emergency assessment. This would presumably have salutary effects, such as permitting earlier and better treatment, and diagnosis may be among the better predictors of some outcomes, such as suicide (32).

The scope of emergency assessment, definitions of key concepts, and improved instruments could be incorporated into practice guidelines. Practice guidelines are rapidly being developed in psychiatry and general medicine and are heralded as facilitating more consistent, effective, and efficient medical care (6, 33); however, compliance by physicians is low (34). Research is beginning to focus on techniques to enhance adoption of guidelines (35), and these efforts should continue and include psychiatry.

Even if reliability is improved, psychiatry needs to explore the validity of as many of the psychiatric emergency service judgments as possible. With a "gold standard," the predictive accuracy of judgments can be explored. Often a state's mental health law suggests at least three outcomes that could be measured. They are intentionally engaging in behavior dangerous to self after hospital release (danger to self), unintentionally engaging in behavior dangerous to self (ability to care for self), and engaging in behavior dangerous to others (danger to others). Information regarding these questions is hard to collect because the behavior will occur in the future and outside the hospital, but such studies can and should be done. Patients and significant others could be contacted in the community to determine if the target behavior has occurred, and this could be cross-referenced with criminal justice and coroner's records.

Issues Concerning Interrater Agreement

Assessing the values of key concepts and recommending a disposition with videotaped patient assessments conducted by other psychiatrists are different in some important ways from actual practice. For example, psychiatrists in the real world have access to additional information, such as previous records and interviews with policemen, families, and other clinicians. Further, psychiatrists in real practice often discuss the case with a nurse or social worker or another physician who has had contact with the patient during the episode of care. These resources were unavailable to the reviewers in this study. Also, in real practice, psychiatrists must consider the availability and quality of treatment services in their hospitals and the community. Finally, the reviewers were making abstract disposition decisions; a real patient was not in the psychiatric emergency service waiting for service. Despite these

shortcomings, the present research more closely approximates real practice than previous designs.

The results presented here were based on videotapes of interviews conducted in four urban public psychiatric emergency services and were based on the reviews of eight predominantly senior psychiatrists who worked in these psychiatric emergency services. The assessment of key issues such as psychosis, depression, and impulse control problems is important for disposition decisions in all settings. There is no obvious reason to believe, therefore, that the results would not apply to other urban public hospitals, to private hospitals who may receive only insured patients, or to rural emergency rooms.

REFERENCES

1. Man with sword kills 2 and injures 9 on SI ferry. New York Times, July 8, 1986, pp 1, B-16
2. Stroul B: Residential crisis services: a review. Hosp Community Psychiatry 1988; 39:1095–1099
3. Gerson S, Bassuk E: Psychiatric emergencies: an overview. Am J Psychiatry 1980; 137:1–11
4. Bassuk EL: Psychiatric emergency services: can they cope as last-resort facilities? in Emergency Psychiatry at the Crossroads: New Directions for Mental Health Services 28. Edited by Lipton FR, Goldfinger SM. San Francisco, Jossey-Bass, 1985, pp 11–20
5. Hughes DH: Trends and treatment models in emergency psychiatry. Hosp Community Psychiatry 1993; 44:927–928
6. American Psychiatric Association: Practice guideline for psychiatric evaluation of adults. Am J Psychiatry 1995; 152(Nov suppl):63–80
7. Allen MH: Definitive treatment in the psychiatric emergency service. Psychiatr Q 1996; 67:247–262
8. Fernando T, Mellsop G, Nelson K, Peace K, Wilson J: The reliability of axis V of DSM-III. Am J Psychiatry 1986; 143:752–755
9. Hjortso S, Butler B, Clemmesen L, Jepsen PW, Kastrup M, Vilmar T, Bech P: The use of case vignettes in studies of interrater reliability of psychiatric target syndromes and diagnoses: a comparison of ICD-8, ICD-10 and DSM-III. Acta Psychiatr Scand 1989; 80:632–638
10. Lieberman PB, Baker FM: The reliability of psychiatric diagnosis in the emergency room. Hosp Community Psychiatry 1985; 36:291–293
11. Riskind JH, Beck AT, Berchick RJ, Brown G, Steer RA: Reliability of DSM-III diagnoses of major depression and generalized anxiety disorder using the structured clinical interview. Arch Gen Psychiatry 1987; 44:817–820
12. Skodol AE: Axis IV: a reliable and valid measure of psychosocial stressors? Compr Psychiatry 1991; 32:503–515
13. Zimmerman M, Coryell W: The reliability of personality disorder diagnoses in a non-patient sample. J Personality Disorders 1989; 3:53–56
14. Warren MD, Peabody CA: Reliability of diagnoses made by psychiatric residents in a general emergency department. Psychiatr Serv 1995; 46:1284–1286
15. Gillis JS, Moran TJ: An analysis of drug decisions in a state psychiatric hospital. J Clin Psychol 1981; 37:32–42
16. Gillis JS, Lipkin JO, Moran TJ: Drug therapy decisions. J Nerv Ment Dis 1981; 169:439–447
17. Fisch HU, Hammond KR, Joyce CRB, O'Reilly M: An experimental study of the clinical judgment of general physicians in evaluating and prescribing for depression. Br J Psychiatry 1981; 138:100–109
18. Fisch HU, Hammond KR, Joyce CRB: On evaluating the severity of depression: an experimental study of psychiatrists. Br J Psychiatry 1982; 140:378–383
19. Malone KM, Szanto K, Corbitt EM, Mann JJ: Clinical assessment versus research methods in the assessment of suicidal behavior. Am J Psychiatry 1995; 152:1601–1607
20. Bengelsdorf H, Levy LE, Emerson RL, Barile FA: A Crisis Triage Rating Scale: brief dispositional assessment of patients at risk for hospitalization. J Nerv Ment Dis 1984; 172:424–430
21. Marson DC, McGovern MP, Pomp HC: Psychiatric decision making in the emergency room: a research overview. Am J Psychiatry 1988; 145:918–925
22. Cooksey RW: Judgment Analysis. San Diego, Academic Press, 1996
23. Wigton RS: Social judgement theory and medical judgement. Thinking and Reasoning 1996; 2:175–190
24. Segal SP, Watson MA, Nelson S: Consistency in the application of civil commitment standards in psychiatric emergency rooms. J Psychiatry Law 1986; 14:125–147
25. Stewart TR, Moninger WR, Heideman KF, Reagan-Cirincione P: Effects of improved information on the components of skill in weather forecasting. Organizational Behavior and Human Decision Processes 1992; 53:107–134
26. Winer BJ: Statistical Principals in Experimental Design. New York, McGraw-Hill, 1962
27. Bartko JJ: Intra-class correlation coefficient. Psychol Rep 1966; 19:3–11
28. Apsler R, Bassuk E: Differences among clinicians in the decision to admit. Arch Gen Psychiatry 1983; 40:1133–1137
29. Hammond KR, Stewart TR, Brehmer B, Steinmann DO: Social judgment theory, in Human Judgment and Decision Processes. Edited by Kaplan MF, Schwartz S. New York, Academic Press, 1975, pp 271–312
30. Mumpower JL, Stewart TR: Expert judgement and expert disagreement. Thinking and Reasoning 1996; 2:191–211
31. Dawes SS, Bloniarz PA, Mumpower JL, Shern D, Stewart TR, Way BB: Supporting Psychiatric Assessment in Emergency Rooms. Albany, NY, University at Albany, Center for Technology in Government, 1995
32. Fawcett J, Clark D, Busch KA: Assessing and treating the patient at risk for suicide. Psychiatr Annals 1993; 23:244–255
33. Hayward RS, Laupacis A: Initiating, conducting and maintaining guidelines development programs. Can Med Assoc J 1993; 148:507–512
34. Lomas J, Anderson GM, Domnick-Pierre K, Vayda E, Enkin MW, Hannah WJ: Do practice guidelines guide practice? N Engl J Med 1989; 321:1306–1311
35. Karuza J, Calkins E, Feather J, Hershey CO, Katz L, Majeroni B: Enhancing physician adoption of practice guidelines. Arch Intern Med 1995; 155:625–632