Clinicin Reliability and Accuracy in Judging Appropriate Level of Care

Leonard Bickman, Marc S. Karver, and L. James A. Schut
Center for Mental Health Policy, Vanderbilt University

Accurately assigning children to the most appropriate level of care is widely recognized as important. Managed care companies conduct utilization reviews in which they monitor the level of care to which clients are assigned using written placement criteria. However, no research has examined the ability of clinicians to perform this task. In the present study, 47 child and adolescent clinical profiles consisting of 48 variables were developed. Eighteen clinicians, trained to use their agency’s level-of-care guidelines, made level-of-care decisions on these profiles. Their interjudge reliability in assigning a child to an appropriate level of care was close to zero (κ = .07). There was a small, statistically significant correlation between client placement and actual placement, but chance-corrected agreement between client placement and actual placement was very low (κ = .09). Implications of these findings for clinical research, practice, policy, and training are discussed.

Demonstration, who were trained to make level-of-care decisions based on preestablished placement criteria. Appropriateness was an important dimension addressed in the Fort Bragg Evaluation Project (FBEP; see Bickman et al., 1995), an evaluation of the continuum-of-care model advanced by advocates of the Child and Adolescent Service System Program (CASSP; Stoul & Friedman, 1986). The program model developed to describe services at the demonstration site (Fort Bragg, North Carolina) assumed that because a full continuum-of-care was available, children would be placed in the most appropriate level of care (i.e., the level of care that best matched their needs), thereby achieving more improvement than children at the comparison sites in which only outpatient and residential services were available. The system of care for the demonstration also used a comprehensive, uniform assessment process, written standards for making level-of-care placements (Cardinal Mental Health Group, 1993), and interdisciplinary treatment teams and case managers to ensure that placements were appropriate. Despite these systemwide efforts, the main hypothesis that clinical outcomes would favor the demonstration was not supported. Instead, children and adolescents in the two types of systems displayed statistically equivalent mean improvement on several outcome variables.

Much of the literature on clinical decision making has been concerned with the assignment of a diagnostic category. Whereas empirical research has not been supportive of the reliability and accuracy of diagnosis (Dawes, 1994), level-of-care decisions are less complex. For instance, in this study, level-of-care decisions required using a 13-page criteria list, as opposed to diagnostic decisions, which require using a diagnostic and statistical manual consisting of hundreds of pages of criteria.

The present study is the only known study to evaluate clinicians' ability to apply preestablished level-of-care criteria. Previous work has been limited to studying groups of experts as they attempt to establish level-of-care criteria (e.g., Strauss, Chassin, & Lock, 1995). However, no study has validated placement by demonstrating that clients who are placed using level-of-care criteria experience better outcomes. Before such a validation study can be conducted, however, an important first step

Leonard Bickman, Marc S. Karver, and L. James A. Schut
Center for Mental Health Policy, Vanderbilt University

This research was funded by the U.S. Army Health Services Command (DA-DA10-89-C-0013) as a subcontract from the North Carolina Department of Human Resources/Division of Mental Health, Developmental Disabilities, and Substance Abuse Services, and by Grant RO1MH-46136 from the National Institute of Mental Health. This research is also supported by Training Grant T32MH-19544-02 from the National Institute of Mental Health.

We thank Ken Berry and Warren Lambert for their considerable assistance with statistical methods.

Correspondence concerning this article should be addressed to Leonard Bickman, Center for Mental Health Policy, Vanderbilt University, 1207 18th Avenue, South, Nashville, Tennessee 37212. Electronic mail may be sent via Internet to bickman@uansv5.vanderbilt.edu.
to examining validity is to determine if preestablished criteria can be reliably applied (Bickman & Salzer, in press).

Over a decade ago, Friedman and Street (1985) noted that the state of knowledge then did not allow for the identification of interventions that would best match children’s needs. Accurate admission and discharge criteria were simply not in place to support appropriate placement procedures. Since the mid-1980s, mental health systems have undergone widespread change in the implementation of systems of care supported by federal policy and funds. However, the greatest emphasis in children’s mental health has been the development of continua or systems of care, mostly ignoring the appropriateness issue. This study explores the possibility that one reason for the failure of the Fort Bragg continuum-of-care model may be the lack of knowledge about appropriate treatment choices. Can clinicians reliably and accurately assign children to the most appropriate level of care? Without this ability, one cannot expect a continuum of care to produce better outcomes than care provided in a fragmented system.

Method

The method selected to study clinical decision making was to have clinicians make placement decisions based on clinical case profiles (see Falvey & Hebert, 1992; Newman, Pitt, & Heverly, 1987). Profiles were modeled after Rumbaugh Clinic case summaries, and consisted of 48 variables (17 domains) that have been found to distinguish children treated in various levels of care (see Table 1).1

Eighteen clinicians were recruited from the Rumbaugh Mental Health Clinic. Five clinicians were clinical supervisors, six were case managers, and three were treatment team leaders. Five clinicians had a PhD degree, seven had either a terminal MA or MS degree, and four had an MSW degree. Clinical experience ranged from 3 to 28 years (M = 12.1, SD = 6.7).

Pilot testing revealed that clinicians could be expected to evaluate a maximum number of 50 clinical profiles. Profiles were first dichotomized by site because the original intent of the study was to test the hypothesis that the initial levels of care to which children were assigned were more appropriate at the demonstration than at the comparison. Furthermore, to assess a wide range of mental health settings, we categorized the most restrictive initial (i.e., within the first 21 days of entering treatment) level of care in which children were actually treated into five groups, ranging from the least to the most restrictive settings. Given these groups, 10 cases could be selected from each of five levels: (a) outpatient therapy, (b) intermediate nonresidential care, (c) intermediate residential care, (d) residential treatment center (RTC), and (e) inpatient care.

The lack of a continuum of care at the comparison site posed a challenge to our sampling of cases because intermediate services were not available. It became necessary to match those children at the demonstration who received intermediate services with children at the comparison (see Table 1).2

We selected cases from the 987 participants in the FBEP using randomized cluster sampling. However, only 47 cases were selected because fewer than 10 clients were initially placed in Level 4. Of these 47 children, 14 were placed in Level 1, 5 in Level 2, 5 in Level 3, 6 in Level 4, and 17 in Level 5.

Participants received a background sheet, 47 clinical profiles (randomly ordered), 47 copies of the Appropriateness Questionnaire,2 a comment sheet, and supplementary documents (level-of-care descriptions and a key to clinical measures) designed to aid the clinicians in making the best informed decisions. The Appropriateness Questionnaire requested the clinician to (a) rate client severity on a scale ranging from 1 to 5, (b) rate the appropriateness of assigning the client to each respective level of care on a scale ranging from 1 to 5, (c) select the most ideal and appropriate initial placement, (d) select the most appropriate initial placement taking cost into account, and (e) indicate which factors were most influential in making the placement decision. The clinicians were asked to read each profile and complete the corresponding questionnaire only on the basis of their clinical training and experience without using any additional aids. They were given 1 week to complete this task and were paid $100.

Results and Discussion

A multivariate kappa statistic (Berry & Mielke, 1988) was used to examine interjudge reliabilities of assigning children to (a) the most ideal and appropriate initial placements, (b) the most appropriate initial placements taking cost into account, and (c) the severity of each client: All kappas were very low (KS = .07, .03, and .15, respectively). Agreement was also very low for ratings of the relative appropriateness of assigning clients to each of the five levels of care (range of KS, .02–.11). In addition, analyses that were limited to placement decisions of clinicians who agreed that they had adequate information still resulted in agreement close to zero (K = .03), although agreement was better for ratings of severity (K = .27, p < .001).6 Collapsing level-of-care placements into residential and nonresidential services still resulted in poor agreement (K = .07).

Did clinicians with similar degrees, positions, or experience have higher levels of agreement? All groupings by background similarity had only slight agreement on level-of-care placements. However, three groupings had fair agreement (range of KS, .21–.40) on severity ratings: (a) masters-level clinicians, (b) clinicians with 9–10 years of experience, and (c) clinicians with 11–16 years of experience. In general, clinicians appeared unable to reliably use level-of-care criteria, regardless of their discipline, position, or experience.

Clinical accuracy was examined by comparing clinicians’ assignments to levels of care actually received by the children. The actual most restrictive initial level of care served as the comparison because most children in the FBEP study received an outpatient assessment before being assigned to a level of care. There was a small but significant correlation between clinicians' placements and actual most restrictive initial placements (r = .07).7

References

.27, p < .0001). The clinicians ordered the clients in a similar manner in which they were actually placed, but they were inexact in their placement because of a very strong bias toward assigning children to lower levels of care. This bias translated into low percentage agreement (29%) and even lower chance-corrected agreement (κ = .09). The significant correlation suggests that the clinicians' decisions were not random. Those children who received the most restrictive care in reality were assigned the most restrictive care in the study but to a lesser degree of restrictiveness.

To better understand the nature of the clinicians' lack of accuracy, we examined sensitivity and specificity with respect to the actual most restrictive initial placements (see Figure 1). Sensitivity, the percentage of the levels of care actually received by the clients that the clinicians correctly identified, was highest for Level 1 (67%) because the clinicians assigned most children to outpatient services. Had the 18 clinicians placed every client in Level 1, sensitivity would have been 100%. Sensitivity scores for Levels 2-5 (3%-26%) indicate that rarely were children assigned to the levels of care they actually received.

As evident, sensitivity was an insufficient indicator of clinical accuracy. These calculations were greatly affected by the clinicians' biases in assigning Level 1. Therefore, the specificity of the clinicians' level-of-care assignments was calculated. Specificity, the percentage of actual cases correctly identified as not getting a specific level of care, was lowest for Level 1 (51%) because of the high rate of false positives. On the other hand, specificity for Level 5 was very high (99%). When clinicians assigned children to inpatient hospital care, although a rare occurrence, they were usually correct. If the clinicians had not assigned any of the children to Level 5 care, specificity for Level 5 would have been 100%.

Although clinicians were asked to assign cases to the most ideal initial levels of care that they thought were appropriate, they may have made assignments based on initially conservative less restrictive placements. Thus, clinicians' placements were compared with actual placements 1 week postreferral (see Table 2). This increased the possibility that a child's level of care would be determined as outpatient because of the initial outpatient assessment. The clinicians did have a similar placement rate as compared with actual initial placements for outpatient care, but they missed the large percentage of clients who were initially hospitalized.

Many of the clinicians commented that they wanted additional information from the profiles. However, Fallon, Adnopoz, Solnit, and Cohen (1992) conducted a study in which clinicians had unlimited access to information on a case, and still clinicians had poor agreement (r = .13). Only after a discussion to achieve a consensus did agreement improve. Even so, the correlation was only .52. Another possible limitation to these findings is that the present study did not use a group of clinicians (i.e., a treatment team approach to decision making). However, in a separate study, Bickman, Karver, and Schut (1995) found low agreement among six clinicians after a meeting to attempt reaching a consensus on level-of-care placements. Furthermore, several studies that directly compared groups and individuals on decision-making tasks found that groups fall prey to the same errors and biases as do individual decision makers and that groups are subject to additional characteristic problems (Carrol & Johnson, 1990).

The lack of agreement among the 18 clinicians and the inaccuracy relative to actual placements in this study is consistent with other studies. Wrobel (1993) and Newman et al. (1987) also conducted studies suggesting clinician disagreement on level-of-care decisions. Leff, Mulker, Lieberman, and Raab (1994) found extremely low agreement between services prescribed by consensus panels and actual service use. Additionally, research examining clinician agreement and accuracy has not been posi-

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Domains and Variables of Clinical Profiles</th>
</tr>
</thead>
<tbody>
<tr>
<td>Domain</td>
<td>Variables</td>
</tr>
<tr>
<td>Identifying information</td>
<td>Age, race, gender</td>
</tr>
<tr>
<td>Presenting problem</td>
<td>Up to four presenting problems reported by parent</td>
</tr>
<tr>
<td>Psychiatric history--treatment</td>
<td>Past levels of care and when child received that care</td>
</tr>
<tr>
<td>Family structure--resources</td>
<td>Who child lives with, parent education, family income</td>
</tr>
<tr>
<td>Parent symptomatology</td>
<td>Brief Symptom Inventory (Detergatis &amp; Spencer, 1981), alcohol and drug history of parents</td>
</tr>
<tr>
<td>Family functioning</td>
<td>Family Assessment Device (Miller, Epstein, Bishop, &amp; Keizer, 1985), Burden of Care Questionnaire (Braman, Heffinger, &amp; Bickman, 1994)</td>
</tr>
<tr>
<td>Health and substance use</td>
<td>Physical health rating, alcohol and drug use (parent and child report)</td>
</tr>
<tr>
<td>Abuse history</td>
<td>Sexual, emotional, and physical abuse (parent and child report)</td>
</tr>
<tr>
<td>Developmental history</td>
<td>History of mental retardation or developmental delay</td>
</tr>
<tr>
<td>Contact with law enforcement--misconduct</td>
<td>Law enforcement contact, use of weapons, cruelty to animals, fighting with others, running away from home (parent and child report)</td>
</tr>
<tr>
<td>Social status</td>
<td>PCAS friends content area</td>
</tr>
<tr>
<td>Recreation</td>
<td>PCAS activities content area</td>
</tr>
<tr>
<td>Problem behavior</td>
<td>Child Behavior Checklist (Achenbach, 1991) and Youth Self-Report (Achenbach &amp; Edelbrock, 1987) Total Internalizing, Total Externalizing, and Total Problem scores</td>
</tr>
<tr>
<td>Self-harm Impression</td>
<td>Suicidal ideation or behavior (parent and child report)</td>
</tr>
<tr>
<td>Client--family participation</td>
<td>PCAS and CAS generated diagnoses, Global Level of Functioning Scale (Hodges, 1990)</td>
</tr>
<tr>
<td></td>
<td>Parent involvement in development of treatment plan, child report of willingness to participate in treatment</td>
</tr>
</tbody>
</table>
There are several implications of this study. First, the clinic’s written standards for level-of-care assignments may have been too general to be prescriptive or may have been applied inconsistently. Either way, children may not have received the level of services that were best matched to their needs. Second, it appears that clinical judgment alone provides little assurance that a child will be assigned to the most appropriate level of care. Clinicians may need better training in following written guidelines. Third, standards of care may need clarification so clinicians can use them more proficiently.

A more basic implication of this and other studies is that preestablished criteria may not be valid representations of appropriate care. In studies in which mental health appropriateness has been investigated, researchers have measured appropriateness of care with expert clinician ratings and consensus panels (Leff et al., 1994; Newman et al., 1987; Newman, Smukler, Griffin, & Fishman, 1992). Whereas expert judgment and panel approaches may be important sources for the development of mental health appropriateness standards, these standards are still prone to significant errors in clinical judgment. For instance, even after a 3-day meeting to determine criteria for the appropriateness of hospitalization, agreement of a consensus panel peaked at only 55% (Strauss et al., 1995). Several other researchers have questioned the validity of consensus panel approaches for determining appropriateness (Black, 1994; Hayward, McMahon, & Bernard, 1993; Phelps, 1993).

Furthermore, standards have not been linked to clinical and functional outcomes. Outcomes should be the ultimate criteria for determining appropriateness of care (Salzer, Nixon, Schut, Karver, & Bickman, in press). In the health care field, it is believed that appropriate care results in increased life expectancy, relief of pain, reduced anxiety, and improved functional capacity (Paccaud & Guillain, 1994). In a similar manner, appropriate mental health care should result in enhanced child development, increased functioning, reduced symptomatology, more consumer satisfaction, and better observance of client rights. Thus, even if clinicians in this study had been reliable and accurate in their judgments of level of care, these judgments would require validation against clinical outcomes.

Future research should focus on developing standards of appropriate care based on linkages between settings and client outcomes as opposed to relying just on clinical judgment. Clinical tools, such as actuarial measures of severity, need to be

Table 2
Percentages of Clinicians’ Assignments to Levels of Care and Levels of Care Actually Received by Children

<table>
<thead>
<tr>
<th>Variable</th>
<th>No care needed</th>
<th>Outpatient</th>
<th>Intermediate nonresidential</th>
<th>Intermediate residential</th>
<th>Residential treatment center</th>
<th>Inpatient hospital</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clinicians’ assignments</td>
<td>3</td>
<td>54</td>
<td>26</td>
<td>7</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Most restrictive level of care received</td>
<td>0</td>
<td>30</td>
<td>11</td>
<td>11</td>
<td>13</td>
<td>36</td>
</tr>
<tr>
<td>Initial level of care received</td>
<td>0</td>
<td>66</td>
<td>4</td>
<td>0</td>
<td>9</td>
<td>21</td>
</tr>
</tbody>
</table>
developed to aid clinicians in assigning children to the most appropriate levels of care. In this study, the clinicians' severity ratings were compared with an actuarial severity rating that was found to be highly predictive of hospitalization within the full sample of children in the FBEP (Bickman et al., 1996). A correlation of .46 between the two ratings of severity was found. It appears that the clinicians were able to differentiate the most severe cases from the least severe cases to some extent. However, clinician-rated severity information did not appear to help clinicians in making reliable or accurate placement decisions. This finding suggests that clinicians may have adequate information to make appropriate level-of-care judgments but that they do not know how to consistently apply the information. Perhaps clinicians can be trained how to apply actuarial severity measures in their clinical practice, thereby improving their ability to make more reliable level-of-care assignments.

Finally, more research needs to go into understanding why clinicians disagree in their decision making. Perhaps the study clinicians were uncomfortable with the number of research instruments used to make the clinical profiles or with the clinical profile format. However, for 62% of the cases, the clinician felt they had adequate information to make a decision, and they were not more accurate for those cases. Conceivably, the clinical profile format used in this study changed the decision-making process that the clinicians normally use, thus making the format unrepresentative. However, several studies have provided evidence of adequate correlations between judgments on written profiles of people and judgments on the actual people (Brehmer & Brehmer, 1988). Future studies should (a) determine what information clinicians believe is most important for making level-of-care and other treatment decisions, (b) determine what profile format is best, (c) examine the underlying thought processes guiding clinical decisions by comparing the factors that clinicians believe they use in making treatment decisions to the factors they actually use, and (d) determine whether treatment decision-making processes using clinical profiles parallel decision-making processes in actual settings.

The results of this study raise questions not only about the reliability of clinical judgment but also about the belief that guidelines will help make those judgments more valid and reliable. As managed care becomes even more widespread, practitioners will continue to experience significant increases in requirements to use written guidelines and protocols. There will be continued pressure to develop methods that will help ensure that services will not be underutilized, such as evaluating the appropriateness of care (Sharpe & Faden, 1996). However, these apparently positive innovations should not be taken at face value and need to be validated by research if there is going to be any true advantage to the consumer.

References


Received June 5, 1996
Revision received August 29, 1996
Accepted November 25, 1996

---

**Low Publication Prices for APA Members and Affiliates**

**Keeping you up-to-date.** All APA Fellows, Members, Associates, and Student Affiliates receive—as part of their annual dues—subscriptions to the *American Psychologist* and *APA Monitor*. High School Teacher and International Affiliates receive subscriptions to the *APA Monitor*, and they may subscribe to the *American Psychologist* at a significantly reduced rate. In addition, all Members and Student Affiliates are eligible for savings of up to 60% (plus a journal credit) on all other APA journals, as well as significant discounts on subscriptions from cooperating societies and publishers (e.g., the American Association for Counseling and Development, Academic Press, and Human Sciences Press).

**Essential resources.** APA members and affiliates receive special rates for purchases of APA books, including the *Publication Manual of the American Psychological Association*, and on dozens of new topical books each year.

**Other benefits of membership.** Membership in APA also provides eligibility for competitive insurance plans, continuing education programs, reduced APA convention fees, and specialty divisions.

**More Information.** Write to American Psychological Association, Membership Services, 750 First Street, NE, Washington, DC 20002-4242.