Agreement Between Clinician and Patient Ratings of Adaptive Functioning and Developmental History

Jared A. DeFife, Ph.D.
Rebecca Drill, Ph.D.
Ora Nakash, Ph.D.
Drew Westen, Ph.D.

Objective: Psychiatric researchers rely heavily on patient report data for clinical research. However, patient reports are prone to defensive and self-presentation biases. Recent research using practice networks has relied on clinician reports, and both forensic and personality disorder researchers have recently turned to quantified data from clinically expert observers as well. However, critics have raised legitimate concerns about the reliability and validity of data from clinician informants. The aim of this study was to assess the validity and diagnostic efficiency of clinician reports of their patients' adaptive functioning and developmental histories, using patient reports as the comparative standard traditionally used in psychiatric research.

Method: Eighty-four clinicians and their patients completed a clinical data form designed to assess a range of patient functioning, clinical history, and developmental relationship variables used in multiple clinician report studies. The authors correlated clinician and patient reports across a number of clinically relevant adaptive functioning variables and calculated diagnostic efficiency statistics for a range of clinical history variables, including suicide attempts, hospitalizations, arrests, interpersonal conflicts affecting employment, and childhood physical and sexual abuse.

Results: Across variables, patient-therapist correlations (0.40–0.66) and overall correct classification statistics (0.74–0.96) were high.

Conclusions: The data demonstrate that clinicians’ judgments about their patients’ functioning and histories agree with patients’ self-reports and that in areas of discrepancy, clinicians tend to make appropriately conservative judgments in the absence of clear data. These findings suggest that quantified clinical judgment provides a vast untapped potential for large-sample research on psychopathology and treatment.

(Am J Psychiatry DeFife et al.; AiA:1–7)
Given the increasing use of clinician report methods in psychiatric research, an important question is whether clinicians can assess reliably and validly dimensions such as functional impairment, personality health-sickness, and clinical and developmental history. Perhaps the major objection to the use of clinicians as informants is the large body of research on limitations and biases in clinical judgment (19, 20). Whether these biases are more pervasive than those of other observers, such as patients with personality disorders self-reporting their pathology or their experiences with significant others, is unclear. Two objections frequently raised regarding the use of clinician informants are a lack of comprehensiveness of standard clinical interviewing practices and a bias toward over-pathologizing. Wood et al. (21) and others suggest that clinical information should be obtained almost exclusively from highly structured and standardized research interviews that tend to rely exclusively or nearly exclusively on patient reports.

In contrast to the arguments against clinician reporting, recent data suggest that clinicians can actually make highly reliable and valid observations at low levels of clinical inference (17, 22, 23) or when they are provided with psychometrically sound instruments to quantify their clinical observations (24). For example, clinicians are able to make highly reliable judgments across functional domains using the Global Assessment of Functioning Scale (GAF), the Global Assessment of Relational Functioning, and the Social and Occupational Functioning scales provided in DSM-IV-TR; in one study, all three measures showed high interrater reliability, with intraclass correlation coefficients ranging from 0.85 to 0.89 (25). In another study, licensed clinicians’ ratings of general intelligence were highly correlated with full-scale IQ scores (r=0.70) obtained from administration of the WAIS (26).

Westen and colleagues (17, 18, 27) have argued that aggregating clinical data quantitatively into the same kinds of scales typically developed for self-reports is one of the best ways to maximize the reliability and validity of clinical data.

The goal of the present study was to examine empirically the validity and diagnostic efficiency of clinician report data by assessing their agreement with patient reports on a number of clinically relevant variables. To capture a broad spectrum of information of interest to clinical practice, we obtained clinician and patient reports of current level of adaptive functioning, clinical history, and quality of early developmental experiences—precisely the kinds of judgments routinely made in clinical practice.

Method

Participants

Two groups of participants were studied: 1) patients receiving treatment in multiple outpatient sites with the Departments of Psychiatry and Psychology at Emory University (including Grady Memorial Hospital, an urban public hospital associated with Em-
history variables rated in both adolescents and adults were correlated in expected ways with measures of psychopathology and attachment status (30–33), although to date they have never been examined in relation to patient reports of the same variables.

Because aggregated variables tend to be more reliable and hence of greater use in research, and in order to test the validity of scales used in numerous research reports relying exclusively on clinician report data, we standardized clinical data form items so that no item held greater weight, and we calculated composite variables of overall functioning (all variables), relational functioning (quality of friendships, quality of romantic relationships, and number of close confidants), work functioning (employment functioning, loss of job in the past 5 years), psychiatric status (clinician GAF score, suicide history, psychiatric hospitalization history), developmental relationships (quality of relationships with mother and father, family warmth, family stability), and abuse history (physical and sexual abuse).

Results

Table 1 provides Pearson correlation coefficients for each of the patient- and therapist-rated composite functioning variables. All correlations were significant, with large effect sizes (34). Table 1 also provides correlations for each of the dimensionally rated individual clinical data form items. Interestingly, most of the variables related to early developmental history (quality of relationship with father, family stability, and family warmth) had slightly larger correlations (r values ranged from 0.53 to 0.66) than items related to current social and occupational functioning (r values ranged from 0.40 to 0.48), although both were large and statistically significant. To account for the effects of time in treatment on patient-therapist rating agreement, we ran partial correlations controlling for number of sessions as a secondary analysis. Controlling for time in treatment had negligible effects (Δr ranged from +0.01 to +0.07); all correlations remained large and statistically significant.

Finally, we calculated diagnostic efficiency statistics for each of the dichotomous historical event variables recorded (e.g., suicide attempts, childhood sexual abuse). The five statistics calculated were overall correct classification rate (the overall “hit rate” or proportion of patients and clinicians matching in their response), sensitivity (the ability of clinicians to identify correctly the occurrence of a historical event that a patient endorsed), specificity (the ability of clinicians to identify correctly the absence of an event the patient did not endorse), positive predictive power (the probability that a patient endorsed an event the clinician identified as having occurred), and negative predictive power (the probability that the patient did not endorse an event when the clinician did not endorse it either). Table 2 summarizes these statistics.

We also obtained published prevalence rate measurements from extensive surveys of the U.S. general population. Prevalence rates of suicide attempt history, prior psychiatric hospitalization history, and history of childhood physical and sexual abuse are available from the National Comorbidity Survey (35). We were unable to obtain reliable data on the prevalence of individuals who lost a job because of interpersonal problems in the past 5 years. Arrest information is available from the FBI (36). Because of unaccounted multiple and repeat offenses, we used prevalence data from 2005 only. Following recommendations by Streiner (37), Table 3 reports diagnostic efficiency statistics adjusted for U.S. prevalence rates.

As can be seen from Tables 2 and 3, overall correct classifications rates were high, with concordance rates of 0.70 and above. The patterns of higher versus lower diagnostic efficiency also suggest that clinicians followed the instructions we used in this and all prior studies using the clinical data form to make judgments conservatively, essentially sacrificing sensitivity for specificity and negative predictive power (i.e., not diagnosing any event unless they were certain, thereby maximizing false negatives but minimizing false positives). For example, if clinicians reported a history of physical or sexual abuse, patients virtually always reported it, although many patients reported abuse histories of which clinicians were either unaware or unsure. Adjusting for U.S. prevalence rates resulted in increased overall correct classification so that all variables rated above 0.90.

Discussion

These results support the validity of clinician reports for a number of clinically relevant variables related to adaptive functioning, developmental history, and occurrence of significant events in both childhood and adulthood. Correlations between patient and clinician reports across broad domains of functioning were greater than typically expected of cross-method correlation coefficients (1) and fell into the upper-quartile range of correlation coefficients seen across a wide sampling of psychological studies (38). Contrary to suggestions that clinicians are prone to an overpathologizing bias, clinicians’ ratings of adaptive functioning were quite consistent with patients’ own

<table>
<thead>
<tr>
<th>Table 1. Agreement of Patient- and Clinician-Rated Adaptive Functioning and Developmental History Variables (N=84)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clinical Data Form Ratings</td>
</tr>
<tr>
<td>Composite overall functioning</td>
</tr>
<tr>
<td>Composite psychiatric status</td>
</tr>
<tr>
<td>Composite relational functioning</td>
</tr>
<tr>
<td>Composite work functioning</td>
</tr>
<tr>
<td>Composite family relationship</td>
</tr>
<tr>
<td>Composite abuse history</td>
</tr>
<tr>
<td>Quality of friendships</td>
</tr>
<tr>
<td>Number of close confidants</td>
</tr>
<tr>
<td>Quality of romantic relationships</td>
</tr>
<tr>
<td>Current school/work quality</td>
</tr>
<tr>
<td>Relationship with mother</td>
</tr>
<tr>
<td>Relationship with father</td>
</tr>
<tr>
<td>Family stability</td>
</tr>
<tr>
<td>Family warmth</td>
</tr>
</tbody>
</table>

***p<0.001.
views of their lives and functioning. Clinicians also tended to see patients’ developmental histories (e.g., relationships with their parents and overall warmth and stability of their familial experiences) in ways that agreed with patients’ experience of their histories.

Overall, clinicians were highly accurate in reporting significant historical events in the same way as their patients (all overall correct classification coefficients except one were >0.80). The data were imperfect, however, which suggests the importance in all psychiatric research of collecting data from multiple informants. In general, sensitivity and negative predictive power were extremely high, suggesting that clinicians tended to be more conservative in reporting events than patients. This could reflect any of several factors: an appropriate level of caution on the part of clinicians in making assumptions about events that occurred in the past without documentation or convincing evidence from the patient; a reluctance on the part of patients to report to their clinicians events about which they felt ashamed; or clinicians’ adherence to our instructions to make ratings conservatively. Another factor potentially involved is that clinicians may at times fail to inquire about significant life history events, such as physical or sexual abuse or prior hospitalizations. Such history may not seem immediately relevant to the treatment work, clinicians may be overly sensitive in their approach to inquiring about painful events, or clinicians may be appropriately concerned about suggestion. For example, while the vast majority of clinicians consider a history of sexual abuse relevant to the therapeutic work, one study found that only half of therapists reported asking all or most of their patients about a sexual abuse history (39).

On the other hand, the positive predictive power statistics were imperfect as well, with clinicians at times rating events as present that patients did not report. (The low base rate of arrest history in our sample, with only two patients endorsing it, contributed to the particularly low positive predictive power of this event.) These discrepancies have several possible explanations. Clinicians may have simply been mistaken in their judgments or reporting, or patients may have failed to disclose certain events on the questionnaire because of forgetfulness, concerns about privacy, or a different interpretation of events than their therapist had (for example, a patient not considering an occurrence severe enough to be labeled abusive).

The area of greatest discrepancy in event reporting is reflected in the sensitivity statistic. As also found by Russ et al. (40), clinicians were more conservative in their reporting of events, endorsing “no” or “unsure” for each item with greater frequency than their patients. Clinicians clearly were not willing to identify the occurrence of significant life events without a high degree of certainty.

**Limitations**

These findings have four limitations. First, we examined only two main sets of variables—adaptive functioning and developmental history. It is possible (and indeed likely) that diagnostic judgments, particularly those made without the aid of instruments designed specifically to maximize accurate use of clinical judgment and minimize error, are far less reliable and valid. That is in fact why much of our laboratory’s research has focused on developing diagnostic methods and instruments for use by experienced clinicians that rely on the same kinds of psychometric principles typically used in more traditional psychiatric research (17, 18).

Second, many of the clinicians in this study were trainees. Thus, generalizability to the population of ex-

---

**TABLE 2. Diagnostic Efficiency Statistics for Agreement of Patient and Clinician Reports of Categorical Events (N=84)**

<table>
<thead>
<tr>
<th>Measure</th>
<th>Overall Correct Classification</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Positive Predictive Power</th>
<th>Negative Predictive Power</th>
<th>Sample Prevalence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Suicide history</td>
<td>0.85</td>
<td>0.44</td>
<td>0.96</td>
<td>0.73</td>
<td>0.86</td>
<td>0.21</td>
</tr>
<tr>
<td>Prior psychiatric hospitalization</td>
<td>0.91</td>
<td>0.71</td>
<td>0.98</td>
<td>0.94</td>
<td>0.89</td>
<td>0.29</td>
</tr>
<tr>
<td>Loss of job in past five years because of interpersonal conflicts</td>
<td>0.74</td>
<td>0.50</td>
<td>0.82</td>
<td>0.50</td>
<td>0.82</td>
<td>0.26</td>
</tr>
<tr>
<td>Arrest within the past five years</td>
<td>0.96</td>
<td>0.50</td>
<td>0.98</td>
<td>0.33</td>
<td>0.99</td>
<td>0.02</td>
</tr>
<tr>
<td>Childhood physical abuse</td>
<td>0.80</td>
<td>0.39</td>
<td>0.95</td>
<td>0.75</td>
<td>0.81</td>
<td>0.27</td>
</tr>
<tr>
<td>Childhood sexual abuse</td>
<td>0.81</td>
<td>0.46</td>
<td>0.93</td>
<td>0.71</td>
<td>0.83</td>
<td>0.27</td>
</tr>
</tbody>
</table>

**TABLE 3. Diagnostic Efficiency Statistics for Agreement of Patient and Clinician Reports of Categorical Events, Adjusted for U.S. Population Prevalence (N=84)**

<table>
<thead>
<tr>
<th>Measure</th>
<th>Overall Correct Classification</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Positive Predictive Power</th>
<th>Negative Predictive Power</th>
<th>Sample Prevalence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Suicide history</td>
<td>0.93</td>
<td>0.44</td>
<td>0.96</td>
<td>0.34</td>
<td>0.97</td>
<td>0.05</td>
</tr>
<tr>
<td>Prior psychiatric hospitalization</td>
<td>0.97</td>
<td>0.71</td>
<td>0.98</td>
<td>0.64</td>
<td>0.99</td>
<td>0.04</td>
</tr>
<tr>
<td>Loss of job in past five years because of interpersonal conflicts</td>
<td>N/A</td>
<td>0.50</td>
<td>0.82</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Arrest within the past five years</td>
<td>0.95</td>
<td>0.50</td>
<td>0.98</td>
<td>0.52</td>
<td>0.97</td>
<td>0.05</td>
</tr>
<tr>
<td>Childhood physical abuse</td>
<td>0.91</td>
<td>0.39</td>
<td>0.95</td>
<td>0.38</td>
<td>0.95</td>
<td>0.04</td>
</tr>
<tr>
<td>Childhood sexual abuse</td>
<td>0.92</td>
<td>0.46</td>
<td>0.93</td>
<td>0.22</td>
<td>0.98</td>
<td>0.07</td>
</tr>
</tbody>
</table>

* The U.S. prevalence of arrests is based on a 1-year period in 2005.
experienced clinicians is limited. However, if trainees are capable of making judgments about adaptive functioning, developmental history, and events of psychiatric significance that strongly agree with patient reports, it seems unlikely that more experienced clinicians would lose the ability over time. The data we present here are, in this respect, likely conservative, underestimating rather than overestimating the ability of seasoned clinicians to make reliable judgments of this sort.

Third, even with high cross-observer correlations and overall correct classification diagnostics, the relation between clinician and patient reports was far from perfect. As in most other studies, we lacked a gold standard from which to assess the validity of clinician report data, so we simply used patient reports as the standard, which at times could represent an over- or underestimation of the variables being assessed. Furthermore, the extent to which therapists and patients agree in their assessment of adaptive functioning and developmental history is not the same as measuring the external validity of such judgments. For example, a patient may report experiences of a hostile/cold family history whereas a sibling or friend of the family views the family dynamic as warm and stable. In standard clinical practice, however, the patient’s narrative and the therapeutic relationship often constitute the only available raw material. We would recommend the use of a greater range of informants (e.g., family, friends, and teachers) rather than the standard approach in psychiatric research—which is to rely on structured interviews and self-reports that all presume the accuracy of a single informant, the patient—as a way of obtaining more accurate data and minimizing informant effects.

Finally, because we collected data from a clinical sample, the prevalences of historical events such as suicide history, abuse, and psychiatric hospitalization were greater than those observed in the general U.S. population. Adjusting for these prevalence rates resulted in greater overall correct classification and negative predictive power, with reduced positive predictive power. Still, the diagnostic efficiency results seen in our sample may not generalize to populations with disproportionately higher or lower base rates (a forensic setting, for example). For interested researchers, a diagnostic efficiency statistics calculator with features to adjust for observed prevalence rate in a sample is available through our lab’s web site at www.psychsystems.net/manuals.

Implications

These data have two primary implications, one for practice and one for research. With respect to practice, a number of commentators, particularly psychologists, have criticized clinical judgment for years, arguing that clinical judgment is riddled with so many biases that it is essentially useless. Indeed, a monograph was recently published (41) that has drawn attention in the popular media suggesting that clinicians are so faulty in their thinking, so caught up in their own biases, and so unscientific in their outlook that they should be forced to practice only from detailed treatment manuals that restrict any use of informed clinical judgment. The data reported here suggest that clinical expertise in psychiatry is likely no different from expertise in any other medical field and that clinical observers, even those in training, are capable of making valid observations about their patients’ functioning and generalizing information about their developmental histories from the narratives they offer in treatment that closely resemble patients’ own views of their histories. Given that patients often have their own biases and that clinicians are likely to be more accurate some of the time than patients about, for example, their ability to form and maintain relationships with others, the fact that we obtained correlations for composite variables in the range of 0.50–0.70 suggests that clinicians are far from the unenlightened caricatures suggested by the psychological literature.

Second, from an empirical perspective, the data presented here provide further impetus to the development of practice networks and other novel methods of data collection that quantify the data of practicing clinicians to study the nature, classification, etiology, and treatment of psychopathology. With tens of thousands of doctoral-level clinicians in practice, each seeing multiple patients, we have access to data on an extraordinary number of patients drawn from samples that look precisely like the patients treated in clinical practice because they are, in fact, sampled from precisely that population. This makes possible, for example, treatment effectiveness research (stud-
ies of the effectiveness of psychotherapeutic and pharmacological treatment as practiced in the community) on hundreds or thousands of patients, which can complement clinical trials. The two methods offer very different trade-offs of experimental control versus external validity (i.e., applicability to real patients seeking treatment for the problems for which they seek treatment in practice, not the single-disorder presentations for which patients are recruited in clinical trials) and specialized samples (people willing to enter into a clinical trial, usually at a university setting) versus patients who present in everyday practice (42, 43). Neither approach alone is the Holy Grail for psychiatric research, but the absence of practice-based research networking has led to an imbalance in what is considered evidence-based practice that reduces real-world significance and drives a wedge between researchers and clinicians. The results of this study suggest that we need not do so because clinicians are capable of making judgments about, for example, important treatment outcomes such as adaptive functioning in ways that are not only, as prior research suggests, highly reliable but, as this study suggests, valid as well.

Received Oct. 16, 2009; revisions received Jan. 18 and April 12, 2010; accepted April 19, 2010 (doi: 10.1176/appi.ajp.2010.09101489). From the Departments of Psychiatry and Psychology, Emory University. Address correspondence and reprint requests to Dr. DeFife, Emory University, 36 Eagle Row, Ste. 270–Westen Lab, Atlanta, GA 30322; jdefife@emory.edu (e-mail).

All authors report no financial relationships with commercial interests. The authors gratefully acknowledge David Streiner, Ph.D., for his statistical consultation.

References


