Constraints on the Validity of Computer Diagnosis

Robert L. Spitzer, MD; Jean Endicott, PhD; Jacob Cohen, PhD; Joseph L. Fleiss, PhD

Issues in the constraints on the validity of computerized psychiatric diagnosis are illustrated by the analysis of diagnoses produced by treating clinicians, expert diagnosticians, and the DIAGNO III computer diagnostic system. The results indicate modest agreement between the computerized diagnoses and both clinicians and experts, and not much better agreement between the experts and between the treating clinicians. The main constraint on the validity of computerized diagnoses is not in any inherent limitation on computer processing but rather in the limitations of the current diagnostic system itself. Improvements in computer diagnosis await improvements in the diagnostic system, along the lines of simplification, explicit criteria, and limitation of the categories to those conditions for which validity evidence exists.

The Russian novelist Gogol wrote that when taken to see a bear dance, one does not comment about how well he dances. But we are past the stage of marveling that a computer can make a psychiatric diagnosis and should now pause and consider critically the constraints on its validity.

The term validity has many specific meanings. As applied to computerized psychiatric diagnosis, it can mean: agreement with a clinician (expert or otherwise), correlations with behavioral or physiological criteria, prediction of outcome, correctness of treatment assignment, etc. Most work thus far, ours as well as others', has been limited to demonstrating the first of these. Thus, the validity question becomes how well the computer can reproduce the diagnoses of clinicians. Limiting validity to this meaning is in itself a constraint, given the at best modest reliability of clinicians' diagnoses.

In this report, we will first compare the models that have been employed, and then offer some new data from our most recent program, DIAGNO III, that illustrate some of the constraints on the validity of computerized diagnosis.

Models

The most frequently employed computer diagnostic models are statistical: discriminant function and Bayes classification. Both are empirically derived from data on patients (usually symptoms, but possibly including demographic and historical material) for whom there are available clinical diagnoses. These methods develop optimal classification rules for the particular sample to which they were originally applied, the "developmental" sample. The derived rules are then applied to new cases. The major

Accepted for publication March 28, 1974.
From Biometrics Research, New York State Department of Mental Hygiene, New York State Psychiatric Institute; and the departments of psychiatry (Dr. Spitzer and Endicott) and biostatistics (Dr. Fleiss), Columbia University; and the Department of Psychology (Dr. Cohen), New York University.
Reprint requests to the New York State Psychiatric Institute, 722 W 168th St, New York, NY 10032 (Dr. Spitzer).
constraint in these systems is their dependence on a very large sample size. Even with several thousand diagnosed patients, no more than a score of diagnostic categories can be employed. A second constraint is the restriction of their applicability to samples that come from the same population as that of the developmental sample. We have shown that optimal rules for both statistical procedures developed on a heterogeneous psychiatric sample yielded diagnoses that were in good agreement with clinical diagnoses in another sample from the same population, but in poor agreement in a sample from a different population, one consisting of selected maternity patients.

The alternative model is the logical decision tree approach that is patterned after the clinical differential diagnostic procedure and consists of a sequence of questions, the answer to each of which is either "true" or "false." The answer to each question rules out one or more diagnoses and determines which question is to be examined next. Some questions may query the presence of a single sign or symptom, others whether a numeric score is in a certain range, and still others whether or not a given pattern of both signs and scores is present. This approach has the advantage of being independent of any specific body of data although, of course, not of accumulated clinical or research experience. Therefore, the method "travels well" from one psychiatric population to a quite different one, since it is not designed to be optimal for a particular population. This was demonstrated by us in the study noted above: the logical decision tree approach yielded diagnoses for the maternity cases that were in better agreement with clinical diagnoses than were those of the two statistical methods. A more detailed comparison of the statistical and the logical decision tree models is given in Fleiss et al. 6

**Development of DIAGNO III**

DIAGNO III is the most recent realization of the decision tree approach from our laboratory. The first two programs utilized data collected on rating scales, primarily designed for research studies. DIAGNO I was limited to current status information and made one of only 27 diagnoses. 7 DIAGNO II utilized current status, as well as some historical information, and made one of 46 diagnoses. 8 DIAGNO III was developed to assist the three main consumers of automated record keeping systems: the clinician who supplies the input data and wants help in the understanding, management, and treatment of his patient; the administrator who needs summary information on patient characteristics for planning, reporting, and evaluation; and the researcher who wants standardized procedures for selecting homogeneous or stratified subgroups for special studies. 9,10

The first version of DIAGNO III used data from the two major clinical intake forms of the Multi-State Information System: the Mental Status Examination Record (MSER) (Figure) and the Psychiatric Anamnestic Record (PAR). 11 The MSER is a four-page form designed to enable a rater to record the results of a mental status examination. Its coverage is divided into the following sections: attitude toward rater, reliability and completeness of information, appearance, motor behavior, general attitude and behavior, mood and affect, quality and content of speech and thought, somatic functioning and concern, perception, sensorium, cognitive functions, judgment, potential for suicide or violence, insight and attitude toward illness, overall severity of illness, and change in condition during the past week. No structure is imposed on the method of data collection. However, the rater is encouraged to consider all of the items on the MSER when conducting his examination. The evaluation covers behavior and symptoms that occurred during the week prior to evaluation, including the day of evaluation itself, whether the evaluation is done on admission or at a later time.

The PAR is a four-page form designed to cover the information that is generally included in a psychiatric case history. Its coverage includes the following: characteristics of current condition, psychosocial disturbance in family, previous treatment for psychosocial disturbance, childhood problems, adolescent and adult friendship patterns, education and intellectual capacity, occupational history, adolescent and adult heterosexual adjustment, marital history, physical health, personality traits, and history of arrests.

Because of the extensive coverage of these two documents, the DIAGNO III program is sufficiently complex to make as many as 75 of the discrete psychiatric diagnoses listed in the American Psychiatric Association's diagnostic and statistical manual. 12 The program is also capable of making multiple diagnoses on a given patient, when appropriate. The output lists the most likely diagnosis, as well as other diagnoses for which there is some evidence and which should also be considered. Within the "most likely" category, multiple diagnoses may appear with the following limitations: transient situational disturbance, conditions without manifest psychiatric disorder, nonspecific condition, and no mental illness; each excludes all other conditions. At most, only one functional psychosis, one personality disorder, or one neurosis is noted; when a
functional psychosis or a personality disorder is diagnosed, neurtic disorders are excluded (but still may appear in the category "some evidence ").

In the output of diagnoses, the conditions are listed in the following order that is roughly in decreasing degree of severity: organic brain syndromes, functional psychoses, sexual deviations, alcoholism and drug addictions, personality disorders (arranged in order of decreasing severity), neurotic disorders (arranged in order of decreasing severity), psychophysiological disorders, mental retardation, conditions without manifest psychiatric disorder, non-specific conditions, and no mental illness.

The DIAGNO III program is written in Fortran IV for the IBM 360 system. The initial version was developed on an ad hoc basis by expanding the logic of DIAGNO I and II. The program was modified successively by using data from over 300 newly admitted inpatients from the Washington Heights Community Service of the New York State Psychiatric Institute. For each case, the clinical diagnosis, the raw data of the MSER and PAR, and the DIAGNO III output were examined and changes made whenever the data in the MSER and PAR justified a modification in the logic. With each successive revision, we again looked at the data for the 300 developmental cases until we concluded that further changes would be of little use.

Because many of the facilities in the automated record keeping system where DIAGNO III is used do not use the history form (the PAR), the program was modified to form the DIAGNO III-M (for MSER), so that it would be possible to go through the decision tree with the mental status data collected on the MSER alone. However, many diagnoses, particularly the personality disorders, cannot be made without historical information. Also, brain syndromes could no longer be divided into acute vs chronic. And finally, some of the subtypes of schizophrenia had to be combined into schizophrenia, unspecified type" and some specific affective illnesses combined into "psychotic depressive mood disorder." The modified program thus makes only 42 diagnoses.

**Method**

Since the purpose of DIAGNO III is to simulate the diagnostic practices of expert diagnosticians, we designed a study that would permit a comparison of the degree to which DIAGNO III agreed with a pair of experts. To provide an appropriate context for such a comparison, we also determined the agreement between DIAGNO III and a pair of clinicians who were responsible for the care of the patients, and between the clinicians and the experts. Finally, the agreement within the pair of clinicians and within the pair of experts was also determined.

One hundred consecutive admissions to the Washington Heights Community Service of the New York State Psychiatric Institute (different from those patients used in the development of DIAGNO III) provided the material for diagnosis that was accomplished in the following manner:

For each patient, the clinicians consisted of the patient's therapist and the attending psychiatrist responsible for supervising his work. There were two attending psychiatrists and a total of 23 therapists, consisting of 17 psychiatric residents and six psychology interns. This was a live clinical situation in which no attempt was made to prevent the usual discussion between the therapist and the attending psychiatrist prior to the therapist's recording of his admission diagnosis. Our impression is that most of this discussion concerned management issues and only secondarily differential diagnosis. The diagnoses of the attending psychiatrists were made after all 100 cases had been admitted, resulting in an interval between the therapist and attending psychiatrist's diagnoses of from a few weeks to three months. To refresh his memory of the case, the attending psychiatrist was provided with a copy of the initial case history dictated by the therapist, from which his diagnosis was deleted. The case histories ranged from two- to five-typewritten pages and included a description of the present illness, mental status on admission, and a brief psychiatric history. The attending psychiatrist's diagnosis are thus based on his recall of personal contact with the patient, throughout the patient's hospitalization up to the time of his assessment, and the information in the initial case history. Both attending psychiatrists reported that they had little difficulty recalling the patients and making diagnoses.

The computer diagnosis was based on the information supplied by the therapist on the two automated recording forms, generally completed within the first two weeks of admission. For each case, there were two computer diagnoses: one based on both forms (DIAGNO III), and one based on the mental status form alone (DIAGNO III-M). The results reported in the tables used

### Table 1. Agreement (Mean Weighted Kappa) for Main Diagnosis Among Three Sources of Diagnosis (N = 100)

<table>
<thead>
<tr>
<th></th>
<th>Experts</th>
<th>Clinical Staff</th>
<th>DIAGNO III</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experts</td>
<td>0.59*</td>
<td>0.52</td>
<td>0.45</td>
</tr>
<tr>
<td>Clinical staff</td>
<td>0.52</td>
<td>0.50†</td>
<td>0.36</td>
</tr>
<tr>
<td>DIAGNO III</td>
<td>0.45</td>
<td>0.36</td>
<td>1.00</td>
</tr>
</tbody>
</table>

* Agreement between the two experts.
† Agreement between the therapist and attending psychiatrist.

### Table 2. Agreement (Frequency and Kappa) for Various Diagnostic Groups Among Three Sources of Diagnosis

<table>
<thead>
<tr>
<th>Average Frequency</th>
<th>Diagnostic Group</th>
<th>Experts</th>
<th>Clinical Staff</th>
<th>DIAGNO III</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experts</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>Organic brain syndrome</td>
<td>0.88†</td>
<td>0.67</td>
<td>0.64</td>
</tr>
<tr>
<td>64</td>
<td>Schizophrenia</td>
<td>0.78</td>
<td>0.68</td>
<td>0.62</td>
</tr>
<tr>
<td>12</td>
<td>Affective psychosis</td>
<td>0.62</td>
<td>0.59</td>
<td>0.15</td>
</tr>
<tr>
<td>4</td>
<td>Neurosis</td>
<td>0.17</td>
<td>0.32</td>
<td>0.08</td>
</tr>
<tr>
<td>5</td>
<td>Personality disorder</td>
<td>0.09</td>
<td>0.28</td>
<td>0.12</td>
</tr>
<tr>
<td>Clinical staff</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Organic brain syndrome</td>
<td>0.67</td>
<td>0.59‡</td>
<td>0.45</td>
</tr>
<tr>
<td>54</td>
<td>Schizophrenia</td>
<td>0.68</td>
<td>0.65</td>
<td>0.62</td>
</tr>
<tr>
<td>12</td>
<td>Affective psychosis</td>
<td>0.59</td>
<td>0.45</td>
<td>0.13</td>
</tr>
<tr>
<td>9</td>
<td>Neurosis</td>
<td>0.32</td>
<td>0.48</td>
<td>0.10</td>
</tr>
<tr>
<td>8</td>
<td>Personality disorder</td>
<td>0.28</td>
<td>0.29</td>
<td>0.26</td>
</tr>
<tr>
<td>DIAGNO III</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>Organic brain syndrome</td>
<td>0.64</td>
<td>0.45</td>
<td>...</td>
</tr>
<tr>
<td>60</td>
<td>Schizophrenia</td>
<td>0.62</td>
<td>0.62</td>
<td>...</td>
</tr>
<tr>
<td>9</td>
<td>Affective psychosis</td>
<td>0.15</td>
<td>0.13</td>
<td>...</td>
</tr>
<tr>
<td>7</td>
<td>Neurosis</td>
<td>0.08</td>
<td>0.10</td>
<td>...</td>
</tr>
<tr>
<td>2</td>
<td>Personality disorder</td>
<td>0.12</td>
<td>0.26</td>
<td>...</td>
</tr>
</tbody>
</table>

* For each source, balance from 100 is frequency of "Other."
† Between the two experts for set.
‡ Between therapist and attending psychiatrist for set.
DIAGNO III as the computer diagnosis.

One of the experts (R.L.S.) briefly interviewed each patient, generally within the first 48 hours of admission. He had access to the admission note and provisional diagnosis. At no time prior to the final analysis of the data did he see the two forms used by the computer in making its diagnosis. The length of the examination varied but typically was five to ten minutes. The purpose of this examination was to confirm the major findings noted in the admission note and to make a tentative diagnosis. Once the initial history was typed, it was reviewed and the diagnosis was sometimes changed in the light of additional (usually historical) information. The attempt here was not to have a completely independent examination, but rather to simulate the role of a consultant who, by and large, is heavily dependent on the therapist’s description of the patient’s current and past status.

The second expert was a research-oriented psychiatrist with 14 years of postresidency experience and an interest in diagnosis. For each case, he initially made a diagnosis after examining the two automated forms, the MSER and the PAR. He then read the initial case history from which the therapist’s diagnosis had been deleted, and made a second diagnosis, now taking into account the additional information. The results referred to in Tables 1 and 2 are based on this last diagnosis.

Data Analysis

Kappa statistics were used for measuring diagnostic agreement. Kappa is the proportion of agreement corrected for chance. It varies from negative values for less than chance-expected agreement, through zero for exactly chance-expected agreement, to plus one for perfect agreement. Unweighted kappa is used in this study to measure agreement on a single category, for example, schizophrenia vs nonschizophrenia. Weighted kappa is used for measuring agreement simultaneously over the entire set of diagnostic categories, taking into account the differing degrees of diagnostic disagreement. For example, schizoaffective schizophrenia is considered closer to manic-depressive manic than it is to any personality disorder.

So that the reader can have a basis for evaluating the values of kappa reported here, values obtained in previous studies of diagnostic agreement are summarized: well-trained clinicians who studied precoded research protocols, rather than observing interviews of patients, had weighted kappas, ranging from 0.25 to 0.80 with a mean of 0.45. Using audio tapes of interviews, three research psychiatrists on a project studying diagnostic practice obtained weighted and unweighted kappas of 0.50 to 0.60 on 23 cases. Sandifer conducted three studies, using several diagnosticians observing live interviews, and kappas calculated from his data ranged from 0.43 to 0.59.

Results

Table 1 presents agreement values (mean weighted kappa) for the main diagnosis among the three sources of diagnosis: experts (R.L.S. and second expert), clinical staff (attending psychiatrist and therapist), and DIAGNO III. Agreement between experts and clinical staff actually involves four discrete weighted kappa coefficients. Agreement between experts and DIAGNO III involves two coefficients, as does agreement between clinical staff and DIAGNO III. These sets of coefficients were averaged for presentation in Table 1. Thus, the value of 0.52 between the experts and the clinical staff is an average of the four clinician-expert kappas, 0.49, 0.50, 0.52, and 0.55. The first two diagonal values are between the therapist and the attending psychiatrist, and between the two experts, respectively. We note by 1.00 the obvious fact that computer diagnoses are perfectly reliable.

It is a fundamental principle of psychometric theory that unreliability constrains validity. In the present context, this simply means that because the experts agree with each other at the level of 0.59, one can hardly expect any other source of diagnosis to agree with the experts at a higher level (although it is mathematically possible). When we find that DIAGNO III’s kappa with the experts is 0.45, we must keep in mind that it is not relative to 1.00 that it must be appraised, but rather to 0.59. Further, the agreement of DIAGNO III with experts, 0.45, is somewhat lower than that of clinical staff with experts, 0.52. The agreement between DIAGNO III and clinical staff is even lower, 0.36, despite the fact that DIAGNO III’s diagnoses were based on ratings made by the clinical staff.

Table 2 presents the mean frequency and mean agreement values for various diagnostic groups, among the three sources of diagnosis. It offers a per-diagnosis look at the agreements that were summarized above. With regard to specific diagnostic categories, agreement between the experts ranges from virtually none (0.09 for personality disorder, and 0.17 for neurosis) to excellent (0.78 for schizophrenia, and 0.88 for organic brain syndrome). Agreements between clinical staff and experts on the one hand, and between DIAGNO III and the experts on the other are ordered similarly to agreement between experts: poor for neurosis and personality disorder and good for organic brain syndrome and schizophrenia. However, the clinical staff’s agreement with experts is never lower than DIAGNO III’s agreement with experts for any of the five diagnostic categories. For organic brain syndrome and schizophrenia, where the agreement is good, it is only slightly better for clinical staff, whereas, the differences are larger where the agreement is poor. The most discrepant degree of agreement with experts, by clinical staff as compared to DIAGNO III, occurs for affective psychosis, 0.59 vs 0.15.

We hypothesize that the reason for the above pattern of findings is that, for the diagnoses of organic brain syndrome and schizophrenia, there are a few critical signs or symptoms that, when present, make these diagnoses highly probable. The diagnosis of neurosis and personality disorder depends on more complex conditional judgments that seem inherently less reliable, not only between computers and clinicians, but even between the experts.

In the agreement between DIAGNO III and clinical staff, the same general pattern obtains, with, however, organic brain syndrome lower (0.45). This can be accounted for, at least in part, by the fact that the agreement between the two clinical staff members on organic brain syndrome is only modest (0.59).

In evaluating the agreement between DIAGNO III and the experts and clinical staff, as compared to the agreement among all four clinicians, three points should be kept in mind. The weighted kappa of 0.50 between the attending psychiatrist and the therapist is undoubtedly spuriously inflated, since no attempt was made to keep their
diagnoses independent. The second point is that the experts shared a common source of data, the initial case history, that contained information not available to the computer. The third point is that all four clinicians, both experts and clinical staff, shared the initial case history. The importance of this initial case history data in enhancing agreement among clinicians is evident when one compares the agreement between the second expert and the other clinicians, when he was limited to the MSER and the PAR data alone, with his agreement when he had the initial case history. These differences are shown in Table 3. This table also shows that the computer performed about as well as the second expert, when he was limited to the input information available to the computer.

Agreement with regard to affective psychosis has a distinct pattern in Table 2. Agreement is fairly good between the two experts (0.62), and only moderately good among the clinical staff (0.45). However, the agreement of DIAGNO III with each of these other sources (0.15 and 0.13) is poor. It is inescapable that DIAGNO III diagnoses affective psychosis poorly. This may well be a consequence of its specific diagnostic strategy and not an inherent limitation of the decision tree method. DIAGNO III treats certain schizophrenic-like symptomatology as contraindicating affective disorder. It was the belief of the authors of DIAGNO III that this is the way most clinicians in the country approach this diagnostic problem. However, the Renard Hospital group, in their concept of primary affective illness, permit a diagnosis of affective illness, even when a symptom that might otherwise suggest schizophrenia is present, such as tactile hallucinations. If this concept of the diagnostic criteria for affective illness were incorporated into DIAGNO III, greater agreement on affective psychosis would likely result, but perhaps at the expense of agreement on schizophrenia.

Tables 1 and 2 were based on the full DIAGNO III program that utilized both mental status (MSER) and historical (PAR) information. Similar analyses were made, using the output of the DIAGNO III-M program that is limited to mental status (MSER) information. The agreement, between the limited program and each of the other sources of diagnoses, was virtually identical to that of the full program. One would have expected that the additional historical information might have improved agreement of the DIAGNO III program with all of the clinical sources, since they all had historical information. Apparently, for newly admitted inpatients, most of the diagnostic variability is accounted for by the mental status findings, so that the historical information contained on the PAR and utilized by this program contributed as much error as useful information. Thus, all of the data reported above for DIAGNO III apply equally well to DIAGNO III-M.

When two sources of diagnoses make a given diagnosis at substantially different rates, the agreement with regard to individual cases cannot be high; the determining consideration is the ratio of the rates. Thus, we find in the frequency data (Table 2) that the rates at which schizophrenia is diagnosed by the three sources are comparable in ratio terms, and indeed, as already noted, the kappas are quite high. In contrast, the ratio of rates at which neurosis and personality disorders are diagnosed by the three sources are large (eg, 8/2 = 4 between clinical staff and DIAGNO III on personality disorder) and here, indeed, is where the kappas are poorest. Note, however, that although dissimilarity in the rates with which a diagnosis is made must yield poor agreement, similarity in rates does not assure good agreement.

**Comment**

Before one can discuss the validity of computerized diagnoses and its constraints, it is first necessary to discuss the criterion against which it is evaluated: in this case, the diagnosis produced by expert diagnosticians. We are distinguishing the diagnoses produced by experienced clinicians who have a special interest in diagnosis from that produced routinely by staff psychiatrists, many of whom, unfortunately, have little, if any, interest in psychiatric diagnosis. As already noted from the literature and as found in this study, the overall agreement (weighted kappa) among the experts when they have either direct contact with the patient or access to clinical case records, is typically in the 50s. If one takes into account the lack of independence between the diagnoses made by the therapist and the attending psychiatrist in this study, one can only assume that the weighted kappa of 0.50 would, at best, be in the low 40s, if the diagnoses had been made completely independently. Therefore, in assessing the validity of computerized diagnosis, it is against the more reliable diagnosis of the experts that it should be assessed, and, indeed, compared with the validity of routine clinical diagnoses assessed against the same criterion.

The relevant values from this study are the weighted kappa for DIAGNO III with experts of 0.45 vs the weighted kappa for clinical staff with experts of 0.52. The difference is modest but it is there and represents an estimate of the magnitude of the superiority of routinely made clinical diagnoses over diagnoses made by a particular computer system.

Using a more stringent basis for comparison, one can compare the agreement of 0.45 between DIAGNO III and the experts with their own agreement with each other, 0.59. (The reader waiting for statements concerning statistically significant differences will wait in vain. Unfortunately, the sampling distribution of differences between correlated values of kappa is unknown, thus precluding formal significance tests.) Weighed against this latter criterion, the validity of DIAGNO III falls substantially short. However, it should be noted that the experts in this study shared the same case record material and were thus not strictly independent.
In previous reports, we presented data supporting the contention that "the agreement between computer diagnoses and clinical diagnoses equals the diagnostic agreement between clinicians given the same information." In those studies, both the clinicians and the computer were limited to precoded information on the psychopathology of the patient. In contrast, in this study, in which the experts had access to additional information (direct contact with the patient or case history or both), the agreement between the experts exceeded the agreement between the computer and the experts. Does this mean that there are inherent limitations in the validity of computerized diagnosis? Not necessarily. We believe that all of the constraints on computerized diagnosis are of a practical nature and are inherent in neither the kinds of information that computers can process, nor in the nature of the algorithms available to them.

This study, unlike our previous studies in computerized diagnosis, made use of data collected routinely in an automated record-keeping system by clinicians with limited interest in diagnosis rather than of data gathered in a research context. Undoubtedly, this was a constraint on the quality of the data that was available to the computer in this study. In another article, we have examined the specific ratings made by the clinicians and determined that, in a consequential number of the cases, the disagreement between the computer and the experts was due to sheer blunders in the ratings made by the clinical staff. We do not know to what degree data of better quality would have improved the agreement between the computer and the experts. Obviously, therefore, one practical constraint on the validity of computerized diagnosis is the quality of the data supplied to the computer. Any system that relies on routinely collected data must institute training and administrative procedures, such as feedback to the clinicians, to insure high quality data.

Although there exist experimental computer systems that accept narrative text as input, the practical state of the art constrains the input of psychiatric information to precoded categories. Many clinicians believe that the more subtle and important clinical features that are critical for psychiatric diagnosis are not amenable to precoding. We contend, on the contrary, the any clinical feature that is capable of explicit verbal formulation can be precoded and, therefore, the only issue is the number of categories that are required for an adequate psychiatric description of all potential patients by all potential psychiatrists. This could be accomplished by an order of magnitude increase in the number of items on the rating form. This would be no strain on the computer's capacity, but an obvious one on the clinician's patience.

We believe that improvement in computerized diagnosis does not in fact require increasing the number of items checked and thus the tolerance of clinicians for making more detailed ratings, but rather requires improvements in the specific content of the item set. For example, none of the computerized psychiatric diagnosis programs has utilized items descriptive of the course of the illness and the sequence of symptom development. It is almost certain that such items would increase the validity of computerized diagnosis, since expert clinicians utilize this kind of information.

Another presumed limitation to computerized diagnosis is the inability of the computer to recognize patterns of symptoms. Clinicians are capable of recognizing different patterns by attaching to their elements probabilities or other weights for assessing the relative importance of each element of the profile. For example, apathy and lack of pleasure are only suggestive of schizophrenia to the extent that other evidence of depression is absent. Thus, the diagnosis of schizophrenia is a joint, not simply an additive, function of manifestations of apathy and absence of signs of depression. Computers are fully capable of applying such formulations, since they are readily reducible to logical and arithmetic operations. Thus far, they have been applied, if at all, in only the most rudimentary form.

In our judgment, therefore, the constraints on the validity of computer diagnosis are neither in the nature of the data that computers can process nor in the nature of the operation that computers can perform. The major constraint lies, rather, in the traditional diagnostic system itself. Recall that the two diagnostic experts achieved a degree of agreement indexed by a weighted kappa of only 0.59. The equivalence has been demonstrated of weighted kappa and the intraclass correlation coefficient, a measure of reliability for quantitative data that is interpretable as the percent of total variance that is in agreement. A value for weighted kappa of 0.59, therefore, implies, by analogy, that some 40% of the variance in even expert psychiatric diagnosis is in disagreement. No attempt to simulate by computer such an unreliable system can hope to attain adequate validity for broad clinical and research use.

Although improvements in the raw data and the computer algorithms will lead to some improvement in computer diagnosis, using contemporary criteria, a quantum jump in the validity of computer diagnosis awaits changes in the system itself in the direction of substantial improvement in its own validity. One attempt in this direction is being made by the Renard Hospital group. They propose a system, limited at present to 16 diagnoses, for which they believe strong validity evidence exists in the form of discriminability from other conditions, a typical specified course, and frequently elevated familial incidence. Whereas, in current diagnostic practice, the clinician determines to which of the standard diagnostic stereotypes the patient most closely conforms, in the Renard system, explicit criteria for any given diagnosis must be met by the patient before the diagnosis is made. For example, for a diagnosis of the depressive form of primary affective disorder the three requirements are dysphoric mood, a psychiatric illness lasting at least one month with no preexisting psychiatric condition, and at least five of the following eight symptoms: poor appetite or weight loss; sleep difficulty; loss of energy; agitation or retardation; loss of interest in usual activities or decrease in sexual drive; feeling of self-reproach or guilt; complaints of or actually diminished ability to think or concentrate; thoughts of death or suicide; and absence of a massive or peculiar alteration of perception or thinking as a major manifestation of the illness.

A consequence of the Renard approach is the necessity
for an "undiagnosed psychiatric disorder" category for those patients who do not meet any of the criteria for the specified diagnoses. In actual use, this category is applied to some 20% of newly admitted inpatients.

In the past, our efforts in computerized diagnosis were aimed at simulating the diagnostic practices of expert clinicians who used the current nosological system, with its exhaustive list of possible diagnoses and their variants, and its lack of explicit criteria. Our current efforts, in collaboration with the Renard group and others (National Institute of Mental Health Clinical Research Branch, collaborative studies of the psychobiology of depressive illness), is in the direction of a change in the diagnostic system itself with emphasis on simplification, explicit criteria, and limiting the categories to those conditions for which validity evidence exists. This work has led to the development of a new instrument for eliciting and recording all of the salient current and historical features necessary to characterize and diagnose a selected group of functional disorders: the Schedule for Affective Disorders and Schizophrenia (SADS). In addition, it has led to the development of the Research Diagnostic Criteria (RDC), a glossary containing the specific criteria for making these diagnoses. In this work, the emphasis is on improving selected areas of the diagnostic system itself. Initial work by clinicians, using the RDC, indicates far superior reliability than that obtained on the same cases when clinicians use the vague criteria of DSM II.

We will continue to explore the use of computerization because of the obvious advantages in mechanical efficiency and accuracy over clinical summarization, although its role in the diagnostic process will be incidental rather than central. However, if the RDC not only increases reliability but also validity, then some of the arbitrary features of the RDC may be improved on by exploiting the computer's unique capacities. For example, many of the all-or-none specifications of the criteria (e.g., "age of onset prior to age 40," "duration of illness more than two weeks") can be modified by tying them to a probability continuum more in keeping with the continuity that characterizes the real world.

This investigation was supported in part by National Institutes of Health research grants MH 65594 and MH 14994.

Noble A. Endicott, MD, was the second diagnostic expert.

References