Quantification of Agreement in Multiple Psychiatric Diagnosis

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

To deal with the problem of indexing diagnostic agreement when clinicians make multiple diagnoses on the same case, an algorithm for calculating a specific level of disagreement for an individual case was developed, and a summary statistic for measuring overall agreement is defined. The output of a computer program for quantifying agreement on multiple diagnoses is described, and an application is given.

PSYCHIATRIC diagnosis is used for such varied purposes as selecting subjects for the study of the effectiveness of treatment; for research into the etiology, course, and duration of mental disorders; and for the evaluation of automated diagnostic procedures; to say nothing of its clinical use in providing a framework for managing a given patient. For each of these purposes, one would desire assurance that the diagnosis given a patient is valid, i.e., actually serves the purpose intended. In the absence of ultimate criteria for validating psychiatric diagnosis, such as are usually provided by various laboratory tests in other branches of medicine, we are thrown back on determining its reliability, since the degree of agreement between diagnosticians necessarily represents the upper limit of validity."'}

To date, the evidence on the reliability of psychiatric diagnosis, measured by the degree of agreement among different diagnosticians, indicates that reliability may not be adequate for the many uses to which diagnosis is put. A problem which prevents an accurate assimilation of the evidence accumulated so far, and prevents a comparison of the degree of reliability attained in one setting with that attained elsewhere, is the lack of uniformity in measuring diagnostic agreement. This communication discusses...
some of the alternative measures which have been used and describes a procedure for quantifying agreement between two clinicians who assign multiple psychiatric diagnoses to a series of patients.

Some Measures of Agreement

The most popular measure of diagnostic reliability appears to be the simple proportion of all cases on whom there was perfect diagnostic agreement. Another popular measure is the chi square (or, equivalently, the contingency coefficient) for association between the diagnoses given by two clinicians. Detailed criticisms of these methods, and numerical examples of their weaknesses, have been presented elsewhere.\(^5\)\(^6\) Here we only review the more salient points.

Simple percent agreement suffers because it does not take into account the degree of agreement expected solely on the basis of chance. The failure to take into account the base rates with which the various diagnoses are made necessarily results in an overestimation of the actual degree of agreement.

Chi square and the contingency coefficient suffer because they measure association of any sort between the two clinicians' diagnoses, and not agreement per se. For example, if, with fair consistency, one clinician diagnoses schizophrenia when another diagnoses manic-depressive psychosis, chi square and the contingency coefficient will be large because of this association, but, patently, agreement is poor. Large values for chi square and the contingency coefficient are necessary indicators of good agreement, but are clearly not sufficient.

A weakness common to both simple percent agreement and chi square is their failure to discriminate between serious disagreements and disagreements involving only closely related diagnoses. For example, paranoid schizophrenia and paranoid personality are obviously closer to each other than paranoid schizophrenia and depressive neurosis, but percent agreement and chi square effectively view these two disagreements as equally serious.

A valid measure of agreement should therefore incorporate a correction for the degree of agreement expected by chance; it should be amenable to a test of the statistical significance of the degree of agreement and not just of the degree of association, and should distinguish among degrees of diagnostic disagreement. Such a measure is weighted kappa.\(^5\)\(^7\)

Weighted kappa has the following properties: (1) it gives partial credit for moderate disagreements; (2) it measures agreement corrected for, i.e., over and above, what is expected purely by chance; (3) it is scaled from -1 to +1 so that negative values indicate worse than chance agreement, 0 indicates exactly chance agreement, and positive values indicate better than chance agreement; and (4) it has a well-defined standard error which permits a statistical assessment of the significance of the observed degree of agreement.\(^8\)

Measuring Agreement When Diagnoses Are Made Using DSM-II

The major requirement of weighted kappa is the quantification of the degree of disagreement between each possible pair of diagnoses, whether single or multiple, that might be assigned to the same patient by two clinicians. Such a quantification was effected by two of us (R.L.S. and J.E.) for situations where each clinician made only one diagnosis per patient using the 1962 nomenclature of the American Psychiatric Association (DSM-I).\(^8\) Levels of disagreement between all possible pairs were scaled by clinical judgment from 0 (identical diagnoses, hence no disagreement) to 9 extreme disagreement, e.g., an organic brain syndrome versus not ill) with intermediate levels for moderate disagreement (e.g., 3 for involutional melancholia versus depressive neurosis). These levels were based on a disagreement scheme originally developed by Sandifer (unpublished data), and take into account similarity of symptomatology, severity, and probable nature of the underlying condition.

With the introduction of the 1968 nomenclature (DSM-II),\(^9\) modifications in the quantitative levels of disagreement between single diagnoses, necessitated by changes made in the official categories, were made by two of us (R.L.S. and J.E.), again using clinical judgment. To validate their judgments, we determined the extent to which the disagreement levels assigned by them agreed with clinical judgments of two experienced psychiatrists who were not familiar with our levels. They independently assigned disagreement levels (also using a 0 to 9 scale) to 184 pairs of single diagnoses that were selected to represent the full range of potential diagnostic disagreements. The correlation between the levels assigned by the two clinicians was 0.67; the correlations between each clinician's levels and those assigned by the authors were both 0.59. Our levels seemed to agree well with those of expert clinicians, and were therefore incorporated into a look-up table to be used in all analyses.

A more serious modification in the measurement of agreement was necessitated by DSM-II's encouragement, for the first time, of multiple diagnoses for a single patient. The use of multiple diagnoses adds a wholly new dimension to the problem of indexing the level of disagreement on a single case. For example, suppose that two clinicians diagnose a series of patients, using multiple diagnoses when deemed appropriate. If only the first diagnosis is relevant to some given purpose, agreement can be assessed by a straightforward application of weighted kappa to the series of first diagnoses only. If, however, one wishes to take into account all of the diagnoses given a patient, it is by no means clear how to assess the degree of agreement between the two sets of diagnoses.

Consider the situation shown in Table 1 where two clinicians, A and B, diagnose a single patient. Agreement between the initial diagnoses of paranoid schizophrenia by A and alcohol addiction by B is poor (found in our look-up table to be scaled 7 on our 0 to 9 scale). On the other hand, both clinicians agree that the patient has alcohol addiction, and only disagree as to whether it is primary. To complicate the matter, one of B's diagnoses, paranoid personality, is relatively close to one of A's diagnoses, paranoid schizophrenia (level = 3), while B's third diagnosis has no counterpart in A's list. Clearly a set of rules sufficient to the task of assessing agreement under these complex conditions must be formulated.
Table 1.—Diagnoses by Two Clinicians on the Same Patient

<table>
<thead>
<tr>
<th>Clinician A</th>
<th>Clinician B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Paranoid schizophrenia</td>
<td>Alcohol addiction</td>
</tr>
<tr>
<td>Alcohol addiction</td>
<td>Paranoid personality</td>
</tr>
<tr>
<td>Homosexuality</td>
<td></td>
</tr>
</tbody>
</table>

Indexing Disagreement on a Single Case

After trying out various procedures we developed the following algorithm for indexing disagreement on a single case, illustrated in Table 2 for the diagnoses of Table 1.

1. Determine from the look-up table the disagreement level between each of A's and each of B's diagnoses.

2. Determine, for each of A's diagnoses, its minimum disagreement level with B's diagnoses. Thus, A's first diagnosis, paranoid schizophrenia, has disagreement levels of 7, 3, and 9 with B's diagnoses, the minimum being 3.

3. Average the minimum disagreement levels for A's diagnoses. In this case, the mean for A is 1.50.

4. Apply steps 2 and 3 to B's diagnoses. The mean of B's minimum levels is 3.33.

5. Take an unweighted average of the mean for A and the mean for B. This yields a case-summary disagreement level of 2.42 for the illustrated patient.

Note that this algorithm ignores differences in the order assigned to the diagnoses. Thus, two clinicians who assign two identical diagnoses but in reverse order would be characterized by a case-summary disagreement level of 0. We considered introducing disagreements as to order into the algorithm but decided not to do so for a variety of reasons. The most important reason is that differences as to order are not fundamentally disagreements as to what conditions the patient has, but rather as to their relative importance for some particular purpose. Furthermore, the decision as to which condition should be listed first often involves contradictory principles, such as the most serious condition (for example, chronic undifferentiated schizophrenia) versus the one most urgently in need of treatment (for example, acute alcohol intoxication). Finally, when issues of order are present, they usually involve disagreements as to the diagnosis listed first. This can easily be handled by merely considering, as a separate issue, disagreements between the first diagnoses listed.

To validate this algorithm, we asked the same two expert clinicians reported on earlier to assign a disagreement level to a series of 54 cases, where each case had two sets of diagnoses, at least one being multiple. The clinicians were instructed to ignore the order in which the diagnoses were listed. The correlation between the two clinicians was 0.55, whereas the correlations between each clinician and the algorithm employing our look-up table were 0.55 and 0.66. Thus the algorithm seems to be a valid reflection of expert clinical judgment.

Summarizing Disagreement over Patients

We next had to solve the problem of summarizing disagreement over a series of patients. The simple mean across all patients of the case-summary disagreement levels is inadequate as an overall index by itself because it does not take into account the degree of disagreement expected by chance alone. The index we selected, analogous to weighted kappa, corrects the simple mean for the average degree of disagreement to be expected by chance and rescales it to limits of -1 to +1.

The correction has been programmed as follows. For each clinician, an array is accumulated containing his sets of diagnoses for all patients. After the diagnoses on all patients have been read in, the program calls for a comparison of each set of diagnoses made by the first clinician with each set made by the second. The case-summary disagreement level for each such pairing (there are N² of them, where N is the number of subjects) is determined according to the algorithm outlined above, and the mean across all N² pairings is calculated, giving the chance-expected mean disagreement level.

If D₀ denotes the mean of the observed case-summary disagreement levels, and D₁ the chance-expected mean, the final index is defined as

\[ K = \frac{D₀ - D₁}{D₁} \]

If agreement is perfect, i.e., D₀ = 0, then K = 1. If the observed mean disagreement level is equal to the expected mean, indicating no more than chance agreement, then K = 0. If the observed mean disagreement level is greater than expected, then K < 0. The smallest possible value for K is -1, indicating complete disagreement.

Since order disagreements are not taken into account by our algorithm, one might expect weighted kappa for multiple diagnoses to be greater than weighted kappa for first or single diagnoses. Whether this is true in any given situation is an empirical issue since there is no mathematically necessary relationship between the two measures.

Description of Computer Output

A Fortran IV program for the IBM 360 has been written for this procedure, and is available on request. The look-up table for disagreement levels between single diagnoses exists as a separate subroutine. (1) For each patient, the program prints out the two sets of diagnoses made on him, the disagreement level for the pair of first diagnoses, and the case-summary disagreement level. (2) After all data have been read in, the program prints out the frequency with which each clinician used each of the diagnoses actually made. There is a separate tabulation for the frequency with which the diagnoses were listed first, second and third. (3) A set of summary statistics is then printed out describing agreement on the pairs of first diagnoses only. First, weighted kappa and its standard error are printed out. Next, the value of chi square for assessing the statistical significance of weighted kappa is printed out. Further, the value of unweighted kappa, its standard error and chi square are printed indexing agreement on each of 14 broad diagnostic categories (e.g., organic brain syndrome, schizophrenia, and affective psychosis). (4) The same statistics described in (3) are printed...
Table 2.—Summarizing the Disagreement Between the Diagnoses of Table 1*

<table>
<thead>
<tr>
<th></th>
<th>Clinician A</th>
<th>Clinician B</th>
<th>Alcohol Addiction</th>
<th>Paranoid Personality</th>
<th>Homosexuality</th>
<th>Minimum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Paranoid schizophrenia</td>
<td>7</td>
<td>3</td>
<td>9</td>
<td>3</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Alcohol addiction</td>
<td>0</td>
<td>7</td>
<td>7</td>
<td>0</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Minimum</td>
<td>0</td>
<td>3</td>
<td>7</td>
<td>0</td>
<td>3</td>
<td></td>
</tr>
</tbody>
</table>

*Average of minima for A = 1.50; average of minima for B = 3.33; unweighted mean of two averages = 2.42.

Application

As part of a comparative study of psychiatric diagnosis in the United Kingdom and United States, a team of psychiatrists interviewed and diagnosed samples of American and British mental patients. Multiple diagnoses were used when the interviewer felt that a single diagnosis could not adequately describe a patient's condition. Two of the reports by this project11,12 describe uses to which the multiple diagnoses have been put.

As part of a reliability study, interviews of 23 American patients were recorded on audio tape. The tapes were rated independently by another team psychiatrist, and the diagnoses made by the interviewer were compared with those made by the rater. Measures of overall agreement and of agreement on the broad categories of schizophrenia and affective psychosis for the 23 pairs of first diagnoses and for the 23 pairs of multiple diagnoses are presented in Table 3. In each of these instances, agreement on multiple diagnoses was better than agreement on first diagnoses. For some other broad categories (eg, alcoholism), however, agreement actually declined.

This work was supported in part by grant No. MH 08534 from the National Institute of Mental Health.

Dr. N. Endicott and L. Sharpe helped validate the disagreement levels.

Table 3.—Agreement on First and Multiple Diagnoses Made by Interviewer and Rater on 23 Cases

<table>
<thead>
<tr>
<th></th>
<th>First Diagnoses</th>
<th>Multiple Diagnoses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kappa</td>
<td>Chi Square*</td>
<td>Kappa</td>
</tr>
<tr>
<td>Overall</td>
<td>0.51</td>
<td>27.00</td>
</tr>
<tr>
<td>Schizophrenia</td>
<td>0.59</td>
<td>8.44</td>
</tr>
<tr>
<td>Affective psychosis</td>
<td>0.59</td>
<td>6.07</td>
</tr>
</tbody>
</table>

*Each chi square has one degree of freedom.
Indexed by weighted kappa.
Indexed by unweighted kappa.

References