THE AGREEMENT CHART

by

Shrikant I. Bangdiwala

Department of Biostatistics
University of North Carolina at Chapel Hill

Institute of Statistics Mimeo Series No. 1859

September 1988

The Agreement Chart

Shrikant I. Bangdiwala

Department of Biostatistics

University of North Carolina

Chapel Hill, North Carolina 27514 USA

SUMMARY

Various measures of agreement between two raters independently categorizing N items into the same set of k nominal or ordinal categories are reviewed. A chart for displaying the agreement contingency table data is proposed. The construction of the agreement chart is described. An intuitive alternative statistic for observer agreement is derived from the chart. A permutation testing procedure for calculation of exact significance levels of the statistic in small samples is described. Large sample conditional and unconditional normal distribution tests are also provided. The agreement chart's utility in addressing the observer bias problem and assessing partial agreement is discussed and the techniques illustrated with various examples.

Key Words: Kappa statistic, p-p plot, pair chart, permutation test, reliability testing

1. INTRODUCTION

If two raters independently classify N items into the same set of k nominal or ordinal categories, one wishes to develop a measure of agreement or "reliability" between the raters. The data can be presented in the format of a two-way contingency table, where the cell entry, X_{ij}, $i,j,=1,\ldots,k$, denotes the number of items classified into the jth category by rater A that were classified into the ith category by rater B. The row and column totals, $X_{i.} = \sum_{j=1}^{k} X_{ij}$ and $X_{.j} = \sum_{i=1}^{k} X_{ij}$, denote the total number of items classified into the ith category by rater B and the total number of items classified into the jth category by rater A, respectively.

Reliability may be incorrectly assessed in terms of the magnitude of the two-way table's chi-squared statistic, which is inadequate since the chi-squared statistic measures association of any kind, and not necessarily agreement. It can be easily shown that a high degree of association (positive or negative) can exist with absolutely no agreement between the raters. Agreement can be regarded as a special kind of association: perfect association implies the ability to perfectly predict the category of one rater from knowledge of the category of the other rater, while perfect agreement implies that both raters classify the same items into the same categories. Intuitively, a high level of agreement implies that the diagonal cell entries X_{ii}, $i=1,\ldots,k$, would be larger than the off-diagonal cell entries X_{ij}, $i \neq j$, $i,j=1,\ldots,k$. However, the naive measure of agreement $p_o = \sum_{i=1}^{k} X_{ii}/N$, the overall proportion of items on which there is perfect agreement, can be misleading since it does not account for the information contained in the marginal totals (Fleiss 1981, sec. 13.1).

Various measures of agreement have been proposed. Galton (1892) developed but later discarded a measure to assess agreement between fingertip patterns on left and right hands beyond what would be expected on the basis of chance alone. Cohen (1960) proposed the kappa statistic, which takes into account the information contained in the table marginal totals. The kappa statistic,

$$\hat{\kappa} = (N \sum_{i=1}^{k} X_{ii} - n_t)/(N^2 - n_t), \qquad (1.1)$$

where $n_t = \sum_{i=1}^{k} X_{i.}X_{.i}$, measures agreement above what would be expected due to chance agreement. The statistic is normalized to take on values from $-n_t/(N^2-n_t)$ to +1. The range of values can be seen to depend on the observed marginal totals. The kappa statistic's values are intuitively interpreted as follows: if $\hat{\kappa} < 0$, there is agreement smaller than expected due to chance alone; if $\hat{\kappa} = 0$, there is agreement only as expected due to chance alone; if $\hat{\kappa} > 0$, there is agreement beyond what is expected due to chance alone, and if $\hat{\kappa} = 1$, there is perfect agreement. The kappa statistic is asymptotically normally distributed.

The kappa statistic has enjoyed widespread popularity and has been generalized in various ways. Cohen (1968) and others have extended the statistic to a "weighted kappa", which takes into account the off-diagonal terms as indicators of "partial agreement". Fleiss (1971), Light (1971), and later Kraemer (1980) and Davies and Fleiss (1982) extend the concept to more than two raters. When agreement on quantitative scales is assessed, the intraclass correlation coefficient can be utilized as a measure of reliability. Fleiss and Cohen (1973) and later in a series of papers, Landis and Koch (1977a, 1977b, 1977c) developed kappa-like measures of agreement from a contingency table point of view which have an intraclass correlation coefficient interpretation. Landis and Koch (1977a) also looked at the concept of observer bias. Focusing on the raters' marginal distributions, the hypothesis of marginal homogeneity or "no interobserver bias" is an extension of the test for correlated proportions due to McNemar (1947). More recently, O'Connell and Dobson (1984) developed a general class of kappa-like measures of observer agreement for individual items and for groups of items as means of the measures on individual items.

All of the above kappa-like statistics have as drawbacks their negative lower bound which depends on the observed marginal totals and their lack of a graphical representation. A graphical representation of the agreement contingency table data is proposed. An intuitive statistic obtainable from the agreement chart is proposed for measuring observer agreement. The statistic accounts for chance agreement by depending upon the constraint of the marginal totals. The development of the agreement chart and the proposed statistic is presented in Section 2 along with their advantages and disadvantages. In Section 3 a permutation test is developed for calculating exact significance levels for testing the hypothesis of observer agreement in small samples. Large sample tests of hypothesis are also discussed. Observer bias and partial agreement extensions are discussed in Section 4. Finally, the techniques are illustrated with various examples in Section 5.

2. DESCRIPTION OF THE AGREEMENT CHART

The motivation for developing the agreement chart was to provide a visual representation of the contingency table data. Unfortunately, few graphical representations are available to represent cross-classified data. Riedwyl and Schuepbach (1983) use colorful grids to denote cell frequencies in their graphical display of a contingency table. The "mosaic" of Hartigan and Kleiner (1981, 1984), in which the individual cell entry counts are represented by rectangles or tiles of area proportional to the count, makes it possible to visually compare sets of counts, suggest hypotheses, check assumptions and interpret results for up to a practical maximum of six cross-classifying variables. The concept of utilizing rectangles to represent counts is adapted for producing the agreement chart, for the specific purpose of assessing observer agreement and investigating other relevant hypotheses such as observer bias.

Figure 1 displays the information from a hypothetical table with $k=3$ classification categories. It is constructed as follows: (1) Draw an NxN square, the limitation imposed by the size of the sample. (2) Draw k rectangles of dimensions $(X_{i\cdot})(X_{\cdot i})$, $i=1,\ldots,k$, within the NxN square. The rectangles are positioned such that the first rectangle of dimension $(X_{1\cdot})(X_{\cdot 1})$ is in the lower left corner of the NxN square, the second rectangle's lower left corner touches the first rectangle on the first one's upper right corner, and so on until the kth rectangle is in the upper right corner of the NxN square. These rectangles contain the information on the marginal totals of the contingency table. The marginal totals thus determine the particular set of rectangles describing the observed contingency table from among all possible sets of rectangles realizable given the sample size N. (3) The $X_{\ell\ell}$ diagonal entry, $\ell=1,\ldots,k$, will determine a square of perfect agreement of dimension $X_{\ell\ell}$ x $X_{\ell\ell}$ which will be blackened and positioned within the ℓth marginal total rectangle such that, within the rectangle, $\sum_{j=1}^{\ell-1} X_{\ell j}$ spaces are to the left of it, $\sum_{j=\ell+1}^{k} X_{\ell j}$ spaces are to the right of it, $\sum_{i=1}^{\ell-1} X_{i\ell}$ spaces are below it, and $\sum_{i=\ell+1}^{k} X_{i\ell}$ spaces are above it. If the two observers agreed exactly, that is, if all the off-diagonal cells of the two-way table were zeroes, the agreement chart would consist of squares rather than rectangles, and they would all be blackened. If there is no agreement whatsoever, that is, if all the diagonal cells of the two-way table were zeroes, there would be no blackened area. The shape of the rectangles of the chart would reflect the marginal distribution of each observer, as usual.

Under independence, the expected value of the diagonal cell entry X_{ii} is the area $X_{i}.X._{i}$ of the corresponding rectangle, $i=1,\dots,k$,times N^{-1}. Thus, a descriptive measure for agreement between the observers that naturally arises from such a chart is the portion of the total area of the k rectangles that represents agreement between the two observers. This area is $\sum_{i=1}^{k} X_{ii}^{2}$. Dividing by the total area of the k rectangles, one obtains the statistic

$$B_N = \sum_{i=1}^{k} X_{ii}^{2} / \sum_{i=1}^{k} X_{i}.X._{i}, \tag{2.1}$$

a measure of agreement ranging from 0 to +1, with zero being "no agreement" and one "perfect agreement". Thus, the range of values of B_N does not depend on the cell frequency counts.

Utilizing the agreement chart as a means of providing a possible graphical interpretation for the kappa statistic, kappa can be shown to equal the following ratio:

$$\frac{\text{(The number of items classified)}\ \text{(The number of perfect agreements)} - \text{(Sum of the areas of the rectangles of marginal totals)}}{\text{(Total Area)} - \text{(Sum of the areas of the rectangles of marginal totals)}},$$

which cannot be easily visualized from Figure 1. On the other-hand, the proposed B_N statistic (2.1) can be interpreted as the ratio of the sum of areas of squares of perfect agreement to the sum of areas of rectangles of marginal totals, or

$$\frac{(\text{Sum of areas of squares of perfect agreement}) \;/\; (\text{The number of items classified})}{(\text{Expected number of perfect agreements based on chance alone})},$$

so while some of the same statistics are involved in the computation of the B_N and $\hat{\kappa}$ statistics, we have an altogether different measure of agreement. The B_N statistic does utilize the information contained in the marginal totals, and thus accounts for chance agreement. While the value of B_N under agreement due only to chance is not intuitively interpretable as the zero value is for kappa, the statistic T_N^* given in equation A.9 for large sample tests does have the value of zero when $X_{ii} = X_{i\cdot}X_{\cdot i}/N$, $i = 1,\ldots,k$.

The choice of ordering of the categories in the contingency table will affect the visual interpretation of the agreement chart. This is not a concern for ordinal categories or for nominal categories with a preferred ordering, but a reasonable choice of ordering categories, say, in descending order by the size of the diagonal cell entries, can eliminate any arbitrariness in visual impact encountered for nominal categories.

The square layout of the agreement chart recaptures all the information contained in the two-way contingency table. Each axis contains the vectorized two-way table, one row-wise and one column-wise, as in the discrete p-p plot [see Wilk and Gnanadesikan (1968)]. Displaying the information of the chart with k vertical bars of height $X_{i\cdot}X_{\cdot i}$, blackened as far as X_{ii}^2, $i=1,\ldots,k$ would lose the observer bias diagnostic features of the square layout and the visual information on the marginal distributions, discussed in Section 4. Finally, note that when $k = N$, the agreement chart provides a graphical display of the rank scores comparison between two raters. Kendall and Smith (1939) examine this problem of rankings and more recently, Iachan (1984) discusses measures of agreement for incompletely ranked data. In this situation, one essentially has continuous data and the agreement chart reduces to the pair chart of Quade (1973).

3. TESTS OF HYPOTHESES FOR OBSERVER AGREEMENT

3.1 Small Sample Permutation Test

Of interest would be to test the null hypothesis of no observer agreement versus the alternative of having agreement. Since the statistic B_N is based upon areas within the rectangles determined by the marginal totals, a test conditional on the fixed marginals can be utilized. The conditional probability of observing X_{ij} given the marginal totals $X_{i\cdot}$ and $X_{\cdot j}$, i,j=1,...,k, and under the condition of independence (hence, no agreement), is given by the multivariate hypergeometric distribution underlying the two-way contingency table (Rao 1973, eq. 6d.1.3),

$$(N!)^{-1} \prod_{i=1}^{k} X_{i\cdot}! \prod_{j=1}^{k} X_{\cdot j}! \prod_{i=1}^{k} \prod_{j=1}^{k} (X_{ij}!)^{-1} . \tag{3.1}$$

Thus, a permutation test can be constructed by an analogous extension of Fisher's Exact Test, as follows. All tables with a more extreme value of the test statistic are constructed and their probabilities as given by (3.1) are added to obtain an exact level of significance for the test of observer agreement. Under the null hypothesis of no observer agreement, the statistic B_N equals 0. Thus, larger values of the statistic B_N would indicate departures from the null. This exact test of significance, with potentially cumbersome computational effort because of the requirement to generate all kxk contingency tables with fixed marginals, can be feasibly computed using the computational techniques proposed by Mehta and Patel

(1983) and Baglivo et. al. (1985). However, for large values of k and N and for an observed table that is close to the null, the actual calculations of the significance levels themselves can be computationally cumbersome. The large sample conditional and unconditional distributions of the B_N statistic provide results for an asymptotic test.

3.2 Large Sample Tests

Under the conditional assumption of fixed row and column marginals, it is shown in Appendix A that under the null hypothesis of independence, the statistic T^*_N given in (A.9) has asymptotically a normal distribution with mean zero and variance γ^2, where γ^2 is given by equation (A.12).

If the null hypothesis is not true or if one is not willing to assume fixed row and column marginals, the unconditional multinominal distribution must be utilized. It is shown in Appendix B that the same statistic T^*_N has asymptotically a normal distribution with mean Γ given in (B.9) and variance given in (B.10). Larger values of B_N and thus of T^*_N indicate departures from the null in favor of the alternative hypothesis. The test is consistent against all alternatives.

The unconditional test is less convenient computationally than the conditional test since the unknown parameters Π_{ij}, $i,j = 1,\ldots,k$, need to be estimated. The conditional assumption of fixed row and column marginals seems a reasonable one, given the formula for B_N.

4. EXTENSIONS OF THE GRAPHICAL TECHNIQUE

Similar to the notions of a "weighted kappa" as discussed in Cohen (1968), the B_N statistic can be generalized to a "weighted B_N" statistic by meaningfully incorporating the information contained in the off-diagonal cell entries. In the agreement chart, "partial agreement" can be visualized as successively lighter shaded areas the further away one is from the perfect agreement diagonal entries. Thus, for $\ell=1,\ldots,k$, cell entries $X_{\ell-1,\ell}$, $X_{\ell+1,\ell}$, $X_{\ell,\ell-1}$, and $X_{\ell,\ell+1}$ would determine a shaded rectangle surrounding the $X_{\ell\ell}$ determined square of perfect agreement which would represent an area $A_{1\ell}$ of "close to perfect agreement". Successively further away cells from the diagonal cells will determine successively lighter shaded areas, $A_{2\ell}$ up to $A_{k-1,\ell}$, $\ell=1,\ldots,k$, as shown in Figure 2. These concepts are especially appropriate for ordinal categories, where appropriate weights w_i can be attached to the shaded areas and a "weighted B_N" statistic constructed as follows:

$$B_N^w = \frac{1 - \sum_{i=1}^{k} [X_{i\cdot}X_{\cdot i} - X_{ii}^2 - \sum_{s=1}^{q} w_s A_{si}]}{\sum_{i=1}^{k} X_{i\cdot}X_{\cdot i}} \qquad (4.1)$$

where w_s is the weight for A_{si}, the shaded area s levels away from the ith diagonal entry, $s=1,\ldots,q$, where q is the furthest level of "partial agreement" wishing to be considered, $q=1,\ldots,k-1$. The advantage of such a statistic would again be its graphical interpretation. Of course, the choice of weights can be appropriate for specific alternatives one is wishing to detect.

The concept of observer bias discussed in Landis and Koch (1977a) can be studied utilizing the agreement chart as well. Observer bias refers to the tendency of one rater to classify items more into certain categories than the other rater, and is thus a measure of the difference in the observers' marginal distributions. A common case of observer bias is either an increasing or decreasing preference for categories by one of the observers if the categories are ordinal. This concept can be easily visualized in the agreement chart (see Figure 2, for example) by noting the deviation of the path of marginal totals rectangles (say, formed by the positive slope diagonals of the rectangles) from the diagonal straight line of marginal homogeneity. The closer to the diagonal the path is, the closer in agreement the raters marginal distributions are. This concept is useful only for ordinal categories, since the rearrangement of the rectangles, possible for nominal categories, makes the path of rectangles an arbitrary criterion.

Analogous to the interpretation of the diagonal path in the pair chart of Quade (1973) for comparing two distributions, the agreement chart can be used to compare the raters marginal distributions as shown in Figure 3. Figure 3 portrays schematic representations of the path of rectangles under the situation where the number of rectangles k and the number of observations N approach infinity and a smooth curve is produced. Figure 3(a) has rater A classifying items into lower categories than rater B and vice versa in Figure 3(b). Figure 3(c) has rater A classifying items more towards the middle categories than rater B, who is more variable or classifies items more towards the extreme categories (vice versa for Figure 3(d)). A comparison of the areas above the path of rectangles with the areas below can, analogously to the Mann-Whitney U-statistic interpretation of the areas in a pair chart, provide a useful test of the hypothesis of observer bias. The examination of the deviation of the path of rectangles from the diagonal can be linked to the analogous examination of the proportion of the residual sums of squares over the total sums of squares when fitting a diagonal line with simple linear regression. A further communication by the author will examine the observer bias question.

Extensions to more than two raters, though possible in concept, would entail a generalization of the notions of areas to volumes and higher dimensions and thus would lose the attractive graphical representability of the test statistic.

5. ILLUSTRATIVE EXAMPLES

Table 1 displays the information on independent diagnostic classification by New Orleans and Winnipeg neurologists of multiple sclerosis patients from the two cities. The kappa statistic for the New Orleans patients [Table 2(a)] is .297, while for the Winnipeg patients [Table 2(b)] is .208. The B_N statistic is .285 and .272, respectively. Landis and Koch (1977a) examine specific patterns of disagreement between the neurologists on the diagnostic classification of individual subjects by considering a hierarchy of weights which by successively combining adjoining categories of diagnoses, create potentially less stringent measures of agreement. They also consider sets of weights which assign varying degrees of partial credit to the off-diagonal cells depending on the extent of the disagreement. This latter concept can be visually examined in Figures 2(a) and 2(b), where shading in the agreement chart denotes partial agreement as discussed in Section 4. By examining the path of rectangles in Figure 2, it seems that the Winnipeg neurologist tends to classify the New Orleans patients into worse diagnostic classes than the New Orleans neurologist.

A practical example of the technique is in comparing the reliability of classification of causes of death by a trained nosologist with classification by a panel of cardiologists, a costly and cumbersome process as described by Curb et al. (1983). In elderly cases, the presence of a multitude of factors at death may make the selection of the underlying cause of death especially difficult. Relying solely on the death certificate may not provide sufficient information to properly classify the death. Table 2(a) presents information on 155 nonelderly deaths among participants of the Lipid Research Clinics Prevalence Follow-Up Study. Table 2(b) is the 268 elderly deaths in the same cohort of subjects. The kappa statistic for the nonelderly (.558) does not differ from that for the elderly deaths (.580), while the B_N statistics are .720 and .614, respectively. Examining Figures 4(a) and 4(b) confirms the high level of reliability between the two different measures of classification. The asymptotic p-values for testing agreement with the B_N statistic are highly significant. However, the graphical technique provides an added feature beyond a statistical test, a visual test of the belief that classifying deaths in the elderly is especially difficult. Death certificates tend to emphasize cardiovascular-related causes of death. The panel of cardiologists examine additional information and is better able to rule out cardiovascular disease as an underlying cause while the nosologist, relying solely on the death certificate, may not. In Figure 4(b) for elderly deaths, the upper right-hand rectangle in the chart has a large area of disagreement indicating that a large number of elderly deaths are classified as "Other noncardiovascular" by the panel of physicians but not so (and thus as cardiovascular of some type) by the nosologist. This phenomenon is not seen to occur in Figure 4(a) for the nonelderly deaths.

*Shrikant I. Bangdiwala is Research Assistant Professor of Biostatistics, Department of Biostatistics, University of North Carolina at Chapel Hill, NC 27514. This work was partially supported by the National Heart, Lung, and Blood Institute of the U.S. National Institutes of Health. The author is greatly indebted to Professor P.K. Sen for helpful comments on the asymptotic distributional properties of the proposed statistic. Thanks also to Hope E. Bryan for the beautiful computer-drawn charts, and Ernestine Bland for careful typing of the manuscript.

APPENDIX A: THE CONDITIONAL NULL DISTRIBUTION OF THE B_N STATISTIC

Let P_N denote the conditional distribution of the contingency table cell entries X_{ij}, $i,j = 1,\ldots,k$, given fixed row and column marginals. Under the null hypothesis of independence, the probability of the $(i,j)^{th}$ cell is $\Pi_{ij} = \Pi_{i\cdot}\Pi_{\cdot j}$, $i,j = 1,\ldots,k$, and P_N is given by equation 6d.1.3 of Rao (1973) as in equation (3.1). Given (3.1), it is easily shown that for $i,j = 1,\ldots,k$,

$$E[X_{ij} \mid P_N] = Na_i b_j, \tag{A.1}$$

where $a_i = X_{i\cdot}/N$ and $b_j = X_{\cdot j}/N$, and for $i,i',j,j' = 1,\ldots,k$,

$$E[X_{ij}(X_{i'j'} - \delta_{ii'}\delta_{jj'}) \mid P_N] = X_{i\cdot}(X_{i'\cdot} - \delta_{ii'})\ X_{\cdot j}(X_{\cdot j'} - \delta_{jj'})/N(N-1), \tag{A.2}$$

where $\delta_{ij} = 1$ if $i=j$ and $=0$ otherwise. From (A.1) and (A.2) it follows that,

$$\mathrm{Cov}[X_{ij},\ X_{i'j'} \mid P_N] = [N^2/N-1]\ a_i b_j\ (\delta_{ii'} - a_{i'})(\delta_{jj'} - b_{j'}), \tag{A.3}$$

for $i,j,i',j' = 1,\ldots,k$. Now, let $\underset{\sim}{X}$ denote the k^2x1 vector of concatenated transposed rows of the contingency table cell entries. Haberman (1974) shows that the cell entries X_{ij}, $i,j = 1,\ldots,k$ can be taken as independent Poisson random variables with parameters λ_{ij}, $i,j = 1,\ldots,k$. Thus, under the convergence of the Poisson distribution to the Normal distribution, the asymptotic distribution of $\underset{\sim}{X}$ is the k^2-multivariate Normal with mean vector $\underset{\sim}{\lambda}$, the k^2x1 vector of Poisson parameters corresponding to the means of the random variables in $\underset{\sim}{X}$, and covariance matrix $\underset{\sim}{\lambda}\underset{\sim}{I}_{K}{}^{2}$, the k^2xk^2 diagonal matrix with $\underset{\sim}{\lambda}$ along the diagonal and zero elsewhere.

The conditional expectation of X_{ij} given in (A.1) equals the estimator of λ_{ij} under the null hypothesis of independence, $i,j = 1,\ldots,k$. The conditional variance of X_{ij}, obtainable as a special case from (A.3) with $(i,j) = (i',j')$, is for large N Na_ib_j, the estimator of λ_{ij} under the null hypothesis of independence, $i,j = 1,\ldots,k$. Furthermore, the conditional covariance of X_{ij} and $X_{i'j'}$, $i,i',j,j' = 1,\ldots,k$, given in (A.3) is asymptotically zero for $(i,j)\neq(i',j')$, again the same as the covariance under Haberman's (1974) model. Thus, under P_N of fixed row and column marginals, the conditional asymptotic distribution of $\underset{\sim}{X}$ under the null hypothesis of independence is the k^2-multivariate Normal with mean vector $\underset{\sim}{\mu}$ and covariance matrix $N^2/(N-1)\underset{\sim}{\Sigma}$, where for $p = 1,\ldots,k^2$,

$$\mu_p = Na_ib_j \tag{A.4}$$

and for $p,q = 1,\ldots,k^2$,

$$\Sigma_{pq} = a_ib_j(\delta_{ii'} - a_{i'})(\delta_{jj'} - b_{j'}), \tag{A.5}$$

where the p,q correspond to the concatenated transposed rows of the subscripts of the contingency table cell entries [e.g., for $k = 3$, $\underset{\sim}{X}' = (X_{11}, X_{12}, X_{13}, X_{21}, X_{22}, X_{23}, X_{31}, X_{32}, X_{33})$; if $p = 7$, $\mu_7 = E(X_{31} \mid P_N)$; if $(p,q) = (2,6)$, $\Sigma_{26} = \mathrm{Cov}(X_{12},X_{23} \mid P_N)$].

Define the k^2x1 vector

$$\underset{\sim}{Z}^* = N^{-1/2}\ (\underset{\sim}{X} - E[\underset{\sim}{X} \mid P_N]), \tag{A.6}$$

a transformation of the elements from $\underset{\sim}{X}$. From (A.4) and (A.5), it follows that $\underset{\sim}{Z}^*$ has the k^2-multivariate Normal distribution with mean vector $\underset{\sim}{0}$ and covariance matrix $\underset{\sim}{\Sigma}^* = [N/(N-1)]\underset{\sim}{\Sigma}$. Therefore, the subset $\underset{\sim}{Z} = \underset{\sim}{A}\underset{\sim}{Z}^*$ corresponding to the diagonal cell entries of the contingency table, where $\underset{\sim}{A}$ is a kxk^2 matrix such that the ith row has a one in the column corresponding to the X_{ii} cell entry and 0 elsewhere, $i = 1,\ldots,k$, has the k-multivariate Normal distribution with mean vector $\underset{\sim}{0}$ and covariance matrix $\underset{\sim}{A}\underset{\sim}{\Sigma}^*\underset{\sim}{A}'$. The (r,s) element of $\underset{\sim}{A}\underset{\sim}{\Sigma}^*\underset{\sim}{A}'$ is

$$a_r b_r(\delta_{rs} - a_s)(\delta_{rs} - b_s)\ N/(N-1), \tag{A.7}$$

$r,s = 1,\ldots,k.$

Now, B_N as given by (2.1) can be rewritten using (A.1) and (A.6) in terms of the Z's as

$$[\sum_{i=1}^{k} a_i^2 b_i^2] + 2N^{-1/2} \sum_{i=1}^{k} a_i b_i Z_i + N^{-1}\ \underset{\sim}{Z}'\underset{\sim}{Z}][\sum_{i=1}^{k} a_i b_i]^{-1}. \tag{A.8}$$

Define the statistic

$$T_N^* = 1/2\ N^{1/2}(B_N - A_N^*), \tag{A.9}$$

where $A_N^* = \sum_{i=1}^{k} a_i^2 b_i^2 \ / \sum_{i=1}^{k} a_i b_i$. From (A.8),

$$T_N^* = (\sum_{i=1}^{k} a_i b_i Z_i)\ /\ (\sum_{i=1}^{k} a_i b_i) + O_p(N^{-1/2}). \tag{A.10}$$

Under P_N, the a's and b's are given so that asymptotically as $N \to \infty$, T_N^* is a linear combination of the Normal variables $\underset{\sim}{Z}$. It is not necessary that all X_{ii} be large, since each cell entry can be characterized as Poisson variables under Haberman's (1974) model and the Central Limit Theorem applies. Thus, T_N^* is normally distributed with expectation zero and variance γ^2 equal to

$$[\sum_{i=1}^{k} a_i^2 b_i^2\ \mathrm{Var}(Z_i) + 2 \sum_{i>j}\sum a_i b_i a_j b_j\ \mathrm{Cov}(Z_i,Z_j)](\sum_{i=1}^{k} a_i b_i)^{-2}. \tag{A.11}$$

From (A.7), γ^2 can be expressed as

$$N(N-1)^{-1}\ (\sum_{i=1}^{k} a_i b_i)^{-2} \sum_{i=1}^{k} a_i^2 b_i^2\ [a_i b_i(1 - a_i - b_i) + \sum_{i=1}^{k} a_i^2 b_i^2]. \tag{A.12}$$

APPENDIX B: THE UNCONDITIONAL DISTRIBUTION OF THE B_N STATISTIC

The unconditional distribution of the contingency table cell entries X_{ij}, $i,j = 1,\dots,k$, is given by the multinomial distribution as in equation 6d.1.1 of Rao (1973):

$$N! \prod_{i=1}^{k} \prod_{j=1}^{k} \frac{(\Pi_{ij})^{X_{ij}}}{X_{ij}!} \qquad \text{(B.1)}$$

where Π_{ij} is the probability of the (i,j) cell, $\Pi_{i\cdot} = \sum_{j=1} \Pi_{ij}$ and $\Pi_{\cdot j} = \sum_{i=1}^{k} \Pi_{ij}$, $i,j = 1,\dots,k$. Define for $i,j = 1,\dots,k$,

$$w_{ij} = (X_{ij} - N\Pi_{ij})N^{-1/2}, \qquad \text{(B.2)}$$

and let $w_{i\cdot} = (X_{i\cdot} - N\Pi_{i\cdot})N^{-1/2}$ and $w_{\cdot j} = (X_{\cdot j} - N\Pi_{\cdot j})N^{-1/2}$, where $X_{i\cdot} = \sum_{j=1}^{k} X_{ij}$ and $X_{\cdot j} = \sum_{i=1}^{k} X_{ij}$. From (B.1) and (B.2), it is seen that for $i,j = 1,\dots,k$,

$$E(w_{ij}) = 0, \qquad \text{(B.3)}$$

and

$$\mathrm{Var}(w_{ij}) = \Pi_{ij}(1 - \Pi_{ij}), \qquad \text{(B.4)}$$

and that the covariances for $(i,j) \neq (i'j')$, $i,i',j,j' = 1,\dots,k$, are given by

$$\mathrm{Cov}(w_{ij}, w_{i',j'}) = -\Pi_{ij}\, \Pi_{i'j'}. \qquad \text{(B.5)}$$

Now, let $\underset{\sim}{X}$ denote the k^2x1 vector of concatenated transposed rows of the contingency table cell entries, and $\underset{\sim}{w}$ the corresponding transformation as in (B.2). With a straightforward extension of the result for the multinomial distribution of Bishop, Fienberg and Holland (1975, Example 14.3-3, pp. 469-470), $\underset{\sim}{X}$ and therefore $\underset{\sim}{w}$ will asymptotically have a k^2-multivariate Normal distribution. The mean vector for $\underset{\sim}{w}$ is $\underset{\sim}{0}$ and the (s,r) element of the covariance matrix $\underset{\sim}{\Sigma}_w$ corresponding to the (i,j) and (i',j') cell is $\Pi_{ij}(1 - \Pi_{ij})$ if s=r and $-\Pi_{ij}\,\Pi_{i'j'}$ if s≠r.

The statistic B_N as given by (2.1) can be rewritten in terms of the w's using (B.2) as

$$\frac{\sum_{i=1}^{k}\Pi_{ii}^{2} + 2N^{-1/2}\sum_{i=1}^{k}\Pi_{ii}w_{ii} + N^{-1}\sum_{i=1}^{k}w_{ii}^{2}}{\sum_{i=1}^{k}\Pi_{i\cdot}\Pi_{\cdot i} + N^{-1/2}\sum_{i=1}^{k}(\Pi_{i\cdot}w_{\cdot i} + \Pi_{\cdot i}w_{i\cdot}) + O_p(N^{-1})}. \qquad (B.6)$$

Thus, the statistic $T_N = 1/2\ N^{1/2}(B_N - \Delta)$, where $\Delta = \sum_{i=1}^{k}\Pi_{ii}^{2} \Big/ \sum_{i=1}^{k}\Pi_{i\cdot}\Pi_{\cdot i}$, is

$$\frac{(\sum_{i=1}^{k}\Pi_{ii}w_{ii})(\sum_{i=1}^{k}\Pi_{i\cdot}\Pi_{\cdot i}) - 1/2\ [\sum_{i=1}^{k}(\Pi_{i\cdot}w_{\cdot i} + \Pi_{\cdot i}w_{i\cdot})](\sum_{i=1}^{k}\Pi_{ii}^{2})}{(\sum_{i=1}^{k}\Pi_{i\cdot}\ \Pi_{\cdot i})^{2}} + O_p(N^{-1/2}), \qquad (B.7)$$

which is a (complicated) linear combination of the w's.

Now, Δ is an unknown parameter and so the statistic T^{*}_{N} of (A.9) is used instead. Now, $T^{*}_{N} = T_{N} + 1/2\ N^{1/2}(\Delta - A^{*}_{N})$. From (B.2), $1/2\ N^{1/2}(\Delta - A^{*}_{N})$ equals

$$\Gamma + \frac{1/2(\sum_{i=1}^{k}\Pi_{ii}^{2})\sum_{i=1}^{k}(w_{\cdot i}\Pi_{i\cdot} + w_{i\cdot}\Pi_{\cdot i}) - (\sum_{i=1}^{k}\Pi_{i\cdot}\Pi_{\cdot i})[\sum_{i=1}^{k}\Pi_{i\cdot}\Pi_{\cdot i}(w_{\cdot i}\Pi_{i\cdot}+w_{i\cdot}\Pi_{\cdot i})]}{(\sum_{i=1}^{k}\Pi_{i\cdot}\ \Pi_{\cdot i})^{2}} + O_p(N^{-1/} \qquad (B.8)$$

where

$$\Gamma = 1/2\ N^{1/2}\left[\frac{\sum_{i=1}^{k}\Pi_{ii}^{2} - \sum_{i=1}^{k}\Pi_{i\cdot}^{2}\Pi_{\cdot i}^{2}}{\sum_{i=1}^{k}\Pi_{i\cdot}\Pi_{\cdot i}}\right]. \qquad (B.9)$$

Thus, from (B.8) and (B.9), T_N^* is still a (more complicated) linear combination of the w's and thus asymptotically is normally distributed with mean Γ and variance obtainable from $\underset{\sim}{\Sigma}_w$ and the coefficients determining the linear combination. It is not necessary that all X_{ij} involved in (B.8) be large, since they can be characterized as sums of independent indicator variables and the same result is obtained with the Central Limit Theorem. If ℓ_{ij} is the coefficient for w_{ij} and $\ell_{i'j'}$ is the coefficient for $w_{i'j'}$, the variance is given by

$$\sum_{(i,j)} \sum_{(i',j')} \ell_{ij}\, \ell_{i'j'} \operatorname{Cov}(w_{ij}, w_{i'j'}). \tag{B.10}$$

If the k^2xl vector $\underset{\sim}{\ell}$ of coefficient ℓ_{ij} is proportional to the k^2xl vector of 1's, (B.10) will be zero and T_N^* will have a degenerate normal distribution. Under the null hypothesis of independence, $\Gamma = 0$. For any fixed alternative $|\Gamma| \to \infty$ as $N \to \infty$ so the test is consistent. For local alternatives of the form $\Pi_{ii} = \Pi_{i\cdot}\Pi_{\cdot i} + N^{-1/2}\,\delta_i$, $i = 1,\ldots,k$, the statistic T_N^* will have a normal distribution with mean $\Gamma = \left(\sum_{i=1}^{k} \delta_i \Pi_{i\cdot}\Pi_{\cdot i}\right) / \left(\sum_{i=1}^{k} \Pi_{i\cdot}\Pi_{\cdot i}\right)$ and variance as described above in (B.10), provided $\sum_{i=1}^{k} \delta_i \Pi_{i\cdot}\Pi_{\cdot i} \neq 0$.

REFERENCES

Baglivo, J., Oliver, D., and Pagano, M. (1985). Computing Fisher and likelihood ratio exact tail probabilities for contingency tables, *Proceedings of the Statistical Computation Section of the American Statistical Association*, Annual Meeting at Las Vegas, Nevada, 70-77.

Bishop, Y.M.M., Fienberg, S.E., and Holland, P.W. (1975). *Discrete Multivariate Analysis: Theory and Practice*, The MIT Press, Cambridge, Mass.

Cohen, J. (1960). A coefficient of agreement for nominal scales, *Educational and Psychological Measurement* *20*, 37-46.

Cohen, J. (1968). Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit, *Psychological Bulletin* *70*, 213-220.

Curb, J.D., Babcock, C., Pressel, S., Tung, B., Remington, R.D., and Hawkins, C.M. (1983). Nosological coding of cause of death, *American J. of Epidemiology* *118*, 122-128.

Davies, M. and Fleiss, J.L. (1982). Measuring agreement for multinomial data, *Biometrics* *38*, 1047-1051.

Fleiss, J.L. (1971). Measuring nominal scale agreement among many raters, *Psychological Bulletin* *76*, 378-383.

Fleiss, J.L. (1981). *Statistical Methods for Rates and Proportions*, 2nd Edition, Wiley, New York.

Fleiss, J.L. and Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability, *Educational and Psychological Measurement* *33*, 613-619.

Galton, F. (1892). *Finger Prints*, London, Macmillan.

Haberman, S.J. (1974). *The Analysis of Frequency Data*, University of Chicago Press, Chicago, pp. 6-9.

Hartigan, J.A. and Kleiner, B. (1981). Mosaics for contingency tables, *Proceedings of the 13th Symposium on the Interface Between Computer Science and Statistics*, ed. W.F. Eddy, New York, Springer-Verlag.

Hartigan, J.A. and Kleiner, B. (1984). A mosaic of television ratings, *The American Statistician* *38*, 32-35.

Iachan, R. (1984). Measures of agreement for incomplete ranked data, *Educational and Psychological Measurement* *44*(4), 823-830.

Kendall, M.G. and Smith, B.B. (1939). The problem of m rankings, *Annals of Mathematical Statistics* *10*, 275-287.

Kraemer, H.C. (1980). Extension of the kappa coefficient, Biometrics 36, 207-216.

Landis, J.R. and Koch, G.G. (1977a). The measurement of observer agreement for categorical data, Biometrics 33, 159-174.

Landis, J.R. and Koch, G.G. (1977b). An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers, Biometrics 33, 363-374.

Landis, J.R. and Koch, G.G. (1977c). One-way components of variance model for categorical data, Biometrics 33, 671-679.

Light, R.J. (1971). Measures of response agreement for qualitative data: some generalizations and alternatives, Psychological Bulletin 76, 365-377.

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages, Psychometrika 12, 153-157.

Mehta, C.R. and Patel, N.R. (1983). A network algorithm for performing Fisher's exact test in rxc contingency tables, J. American Statistical Association 78, 427-434.

O'Connell, D.L. and Dobson, A.J. (1984). General observer-agreement measures on individual subjects and groups of subjects, Biometrics 40, 973-984.

Quade, D. (1973). The pair chart, Statistica Neerlandica 27, 29-45.

Rao, C.R. (1973). Linear Statistical Inference and Its Applications, 2nd Edition, Wiley, New York.

Riedwyl, H. and Schuepbach, M. (1983). Siebdiagramme: graphische darstellung von kontingenztafeln", Technischer Bericht No. 12, Institut fur Mathematische Statistik und Versicherungslehre, Universitat Bern Switzerland.

Wilk, M.B. and Gnanadesikan, R. (1968). Probability plotting methods for the analysis of data. Biometrika 55, 1-17.

Table 1: Diagnostic Classification Regarding Multiple Sclerosis

(a) New Orleans Patients		Winnipeg Neurologist					
		Diagnostic Class	1	2	3	4	Total
New Orleans Neurologist	Certain MS	1	5	3	0	0	8
	Probable MS	2	3	11	4	0	18
	Possible MS	3	2	13	3	4	22
	Doubtful or No MS	4	1	2	4	14	21
	Total		11	29	11	18	69

(b) Winnipeg Patients		Winnipeg Neurologist					
		Diagnostic Class	1	2	3	4	Total
New Orleans Neurologist	Certain MS	1	38	5	0	1	44
	Probable MS	2	33	11	3	0	47
	Possible MS	3	10	14	5	6	35
	Doubtful or No MS	4	3	7	3	10	23
	Total		84	37	11	17	149

Source: Landis and Koch, 1977a.

TABLE 2: Nosological Coding* By Mortality Classification Panel (MCP) for Underlying Cause of Death. Lipid Research Clinics Prevalence Follow-Up Study

(a) Nonelderly (Age at Death Less than 65 Years)

Nosological Classification	MCP Classification						
	440.2,445.0	441-442	437	410-414	390-458	Other Non-CVD	Total
Atherosclerosis of peripheral arteries with gangrene (440.2, 445.0)	0	0	0	0	0	0	0
Atherosclerotic arterial aneurysm with rupture (441-442)	0	1	0	0	2	0	3
Atherosclerotic cerebrovascular disease (437)	0	0	6	1	6	1	14
Atherosclerotic coronary heart disease (410-414)	0	0	0	84	5	3	92
Other cardiovascular disease (390-458)	0	0	0	10	7	1	18
Other noncardiovascular disease	1	0	0	5	4	18	28
Total	1	1	6	100	24	23	155

(b) Elderly (Age at Death 65 Years and Over)

Nosological Classification	MCP Classification						
	440.2,445.0	441-442	437	410-414	390-458	Other Non-CVD	Total
Atherosclerosis of peripheral arteries with gangrene (440.2, 445.0)	0	0	0	0	0	0	0
Atherosclerotic arterial aneurysm with rupture (441-442)	0	4	0	0	2	0	6
Atherosclerotic cerebrovascular disease (437)	0	0	20	1	4	15	40
Atherosclerotic coronary heart disease (410-414)	0	1	5	100	12	10	128
Other cardiovascular disease (390-458)	2	0	1	5	15	10	33
Other noncardiovascular disease	0	0	4	1	6	50	61
Total	2	5	30	107	39	85	268

*Eighth Revision ICDA.

Figure 1. Illustration of an Agreement Chart for $k = 3$. The blackened squares of perfect agreement within the rectangles determined by the marginal totals provide a measure of agreement between two observers categorizing the same N objects into k categories.

Figure 2. The Agreement Charts of New Orleans Patients and Winnipeg Patients for the Diagnostic Classification Regarding Multiple Sclerosis (MS) data from Table 1. Source of data: Landis and Koch (1977a). The Winnipeg neurologist is more likely to classify patients as Certain or Probable MS than the New Orleans neurologist. This is markedly true for the Winnipeg patients.

Figure 3. Diagrams of the Path of Rectangles in an Agreement Chart. The path of rectangles can provide an intuitive indication of observer bias with rater A preferring lower categories (a), higher categories (b), middle categories (c) or extreme categories (d) as compared to rater B.

Figure 4. The Agreement Charts of Nonelderly (age at death < 65 years) and Elderly (age at death > 65 years) deaths in the Lipid Research Clinics Prevalence Follow-Up Study for the Classification of Underlying Cause of Death data from Table 2. The classification methods agree substantially for both groups, but some elderly deaths tend to be classified as cardiovascular by the nosologist while the panel of cardiologist classifies them as noncardiovascular.

FIGURE 1

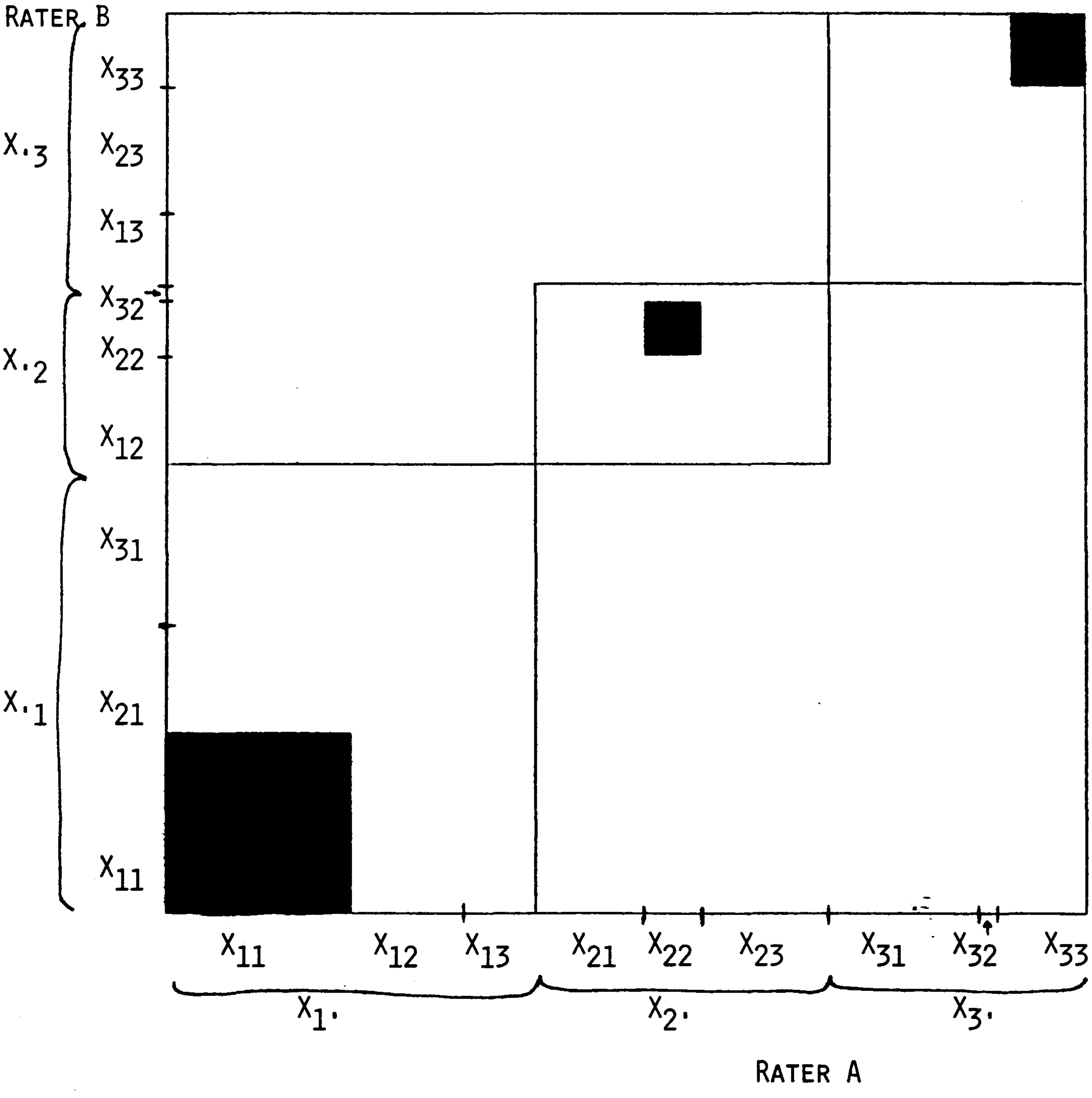
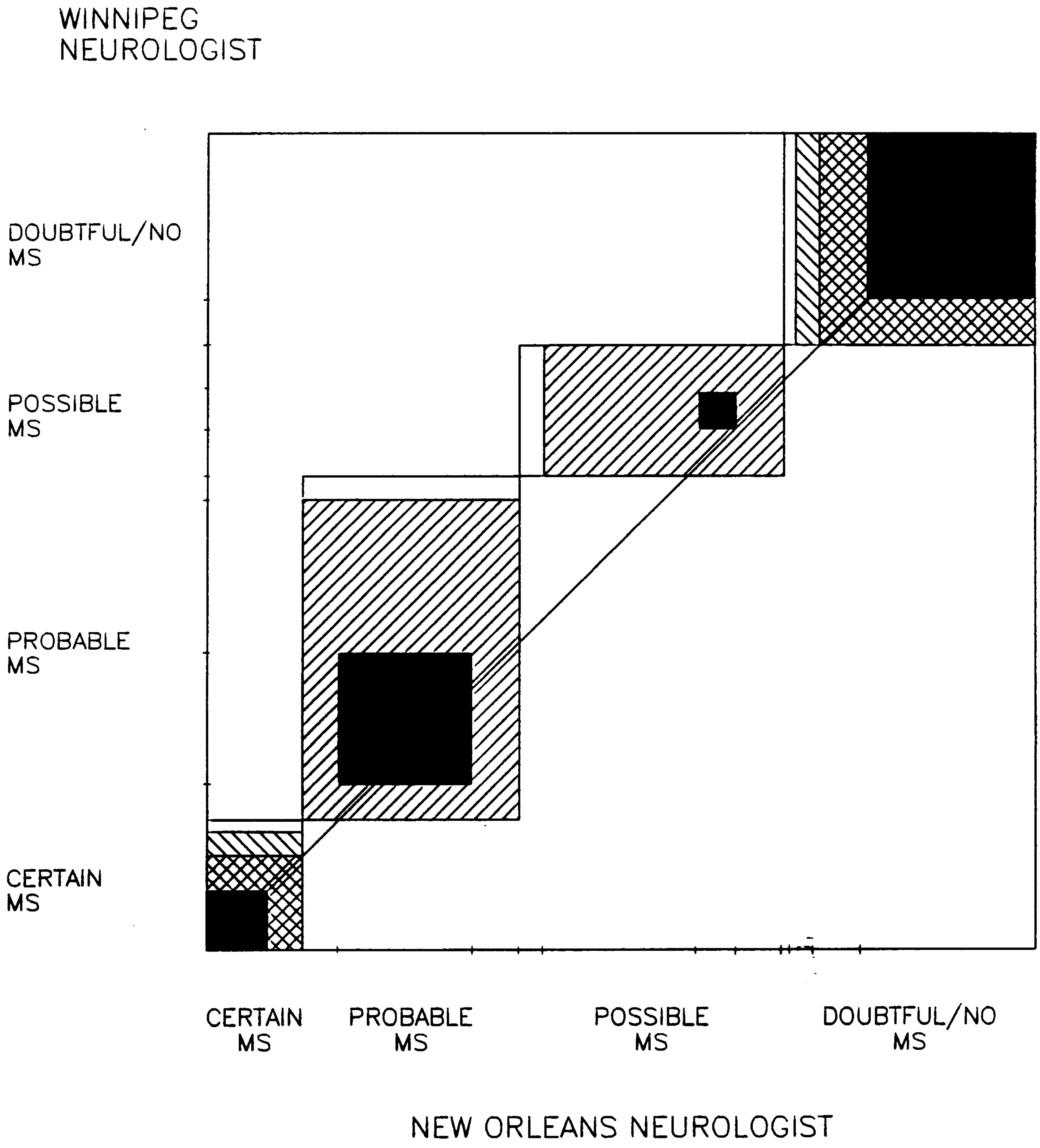

RATER B
X_{33}
X_{23}
X_{13}
X_{32}
X_{22}
X_{12}
X_{31}
X_{21}
X_{11}
$X_{.3}$
$X_{.2}$
$X_{.1}$
X_{11}
X_{12}
X_{13}
X_{21}
X_{22}
X_{23}
X_{31}
X_{32}
X_{33}
$X_{1.}$
$X_{2.}$
$X_{3.}$
RATER A

FIGURE 2 (A)


WINNIPEG
NEUROLOGIST
DOUBTFUL/NO
MS
POSSIBLE
MS
PROBABLE
MS
CERTAIN
MS
CERTAIN
MS
PROBABLE
MS
POSSIBLE
MS
DOUBTFUL/NO
MS
NEW ORLEANS NEUROLOGIST

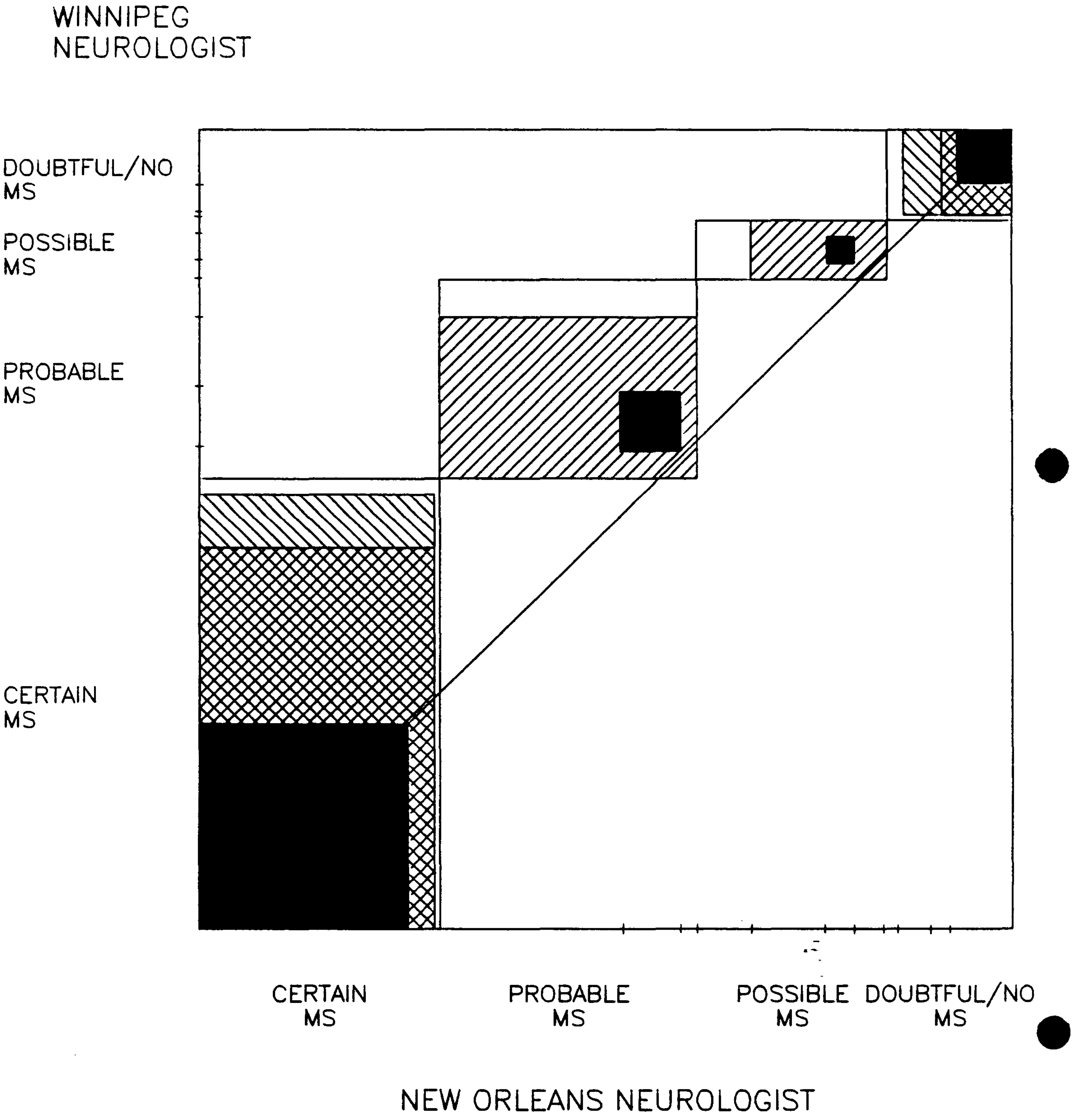

WINNIPEG
NEUROLOGIST
DOUBTFUL/NO
MS
POSSIBLE
MS
PROBABLE
MS
CERTAIN
MS
CERTAIN
MS
PROBABLE
MS
POSSIBLE
MS
DOUBTFUL/NO
MS
NEW ORLEANS NEUROLOGIST

Figure 3

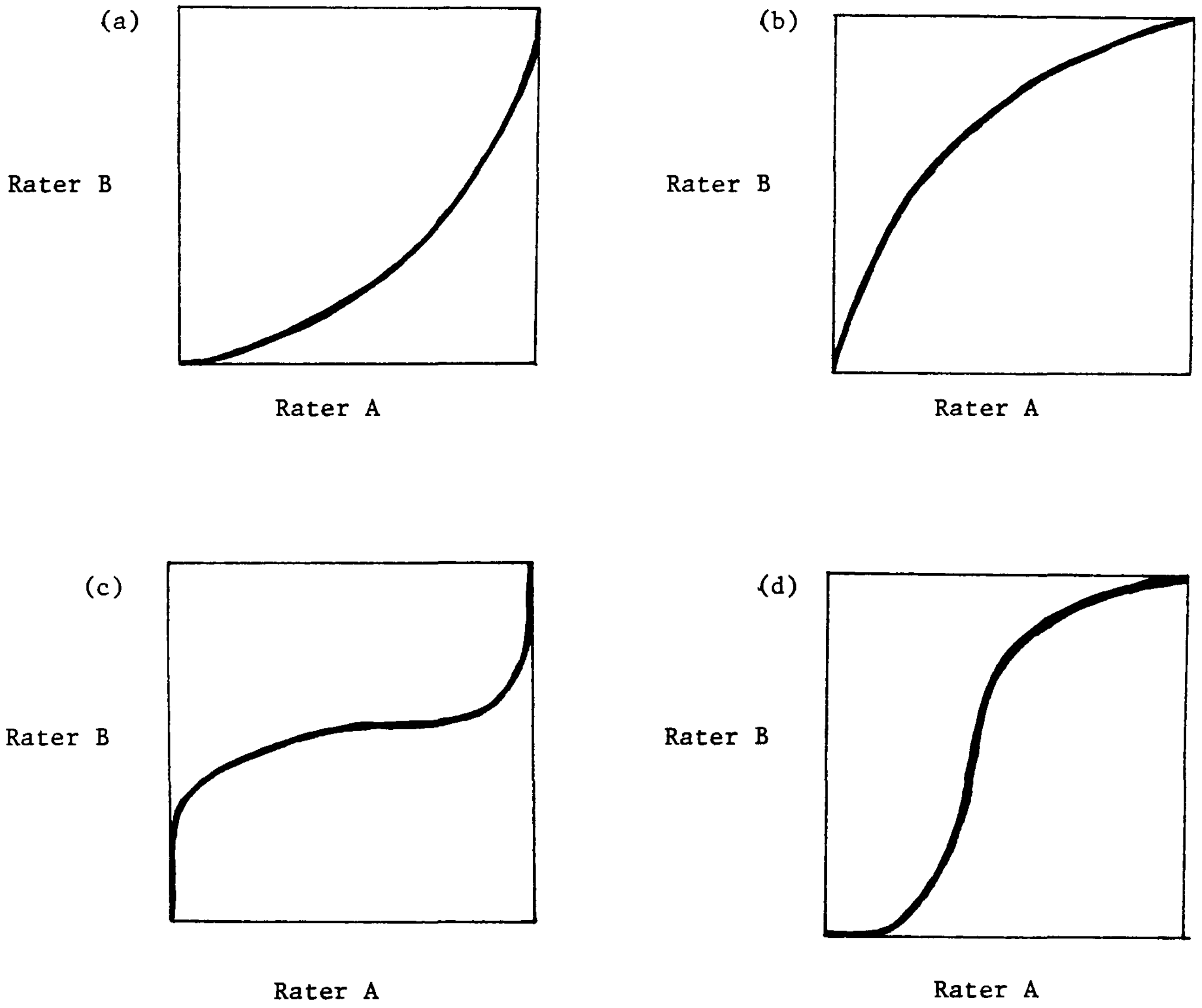

(a)
Rater B
Rater A
(b)
Rater B
Rater A
(c)
Rater B
Rater A
(d)
Rater B
Rater A

FIGURE 4 (A)

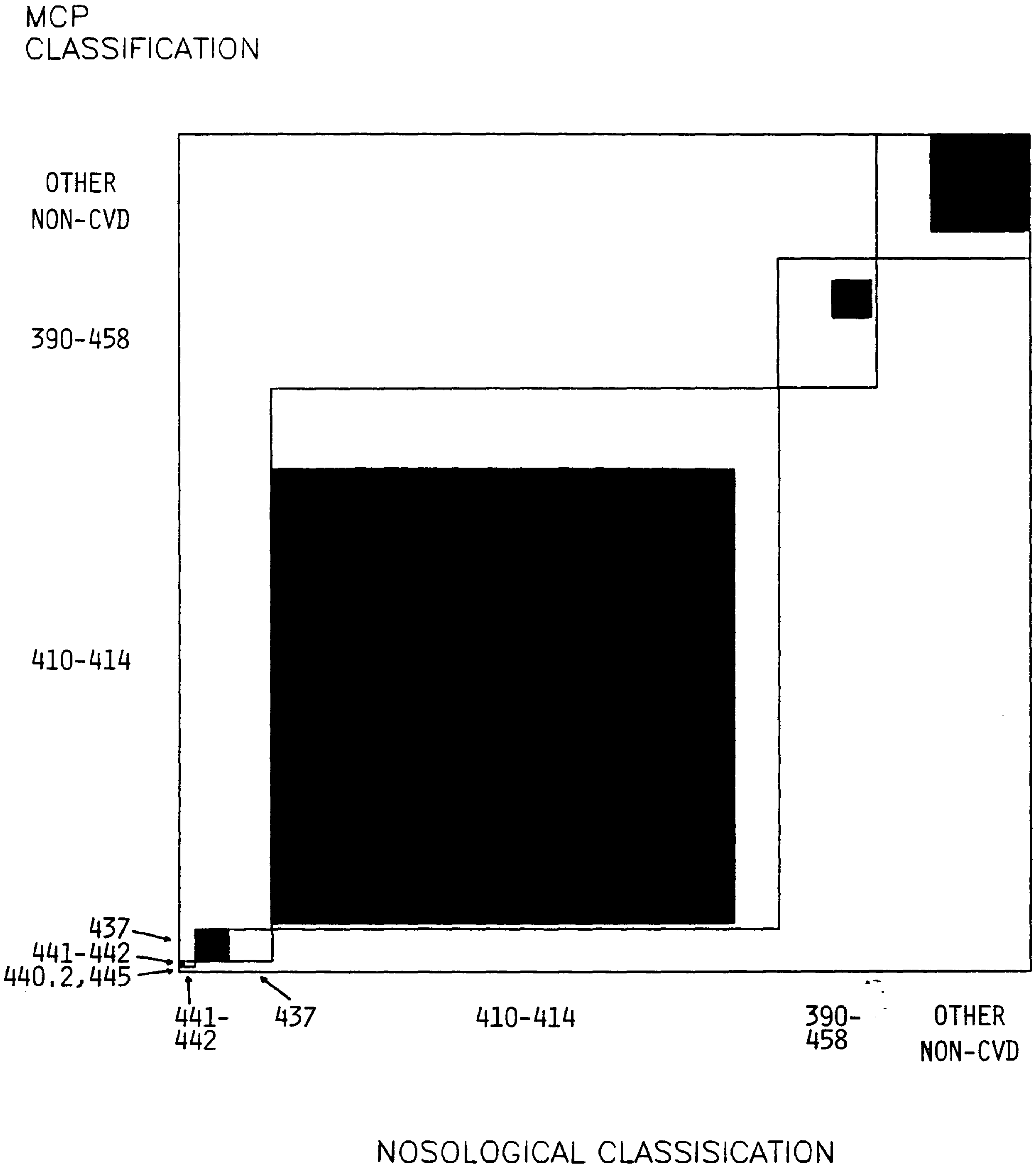
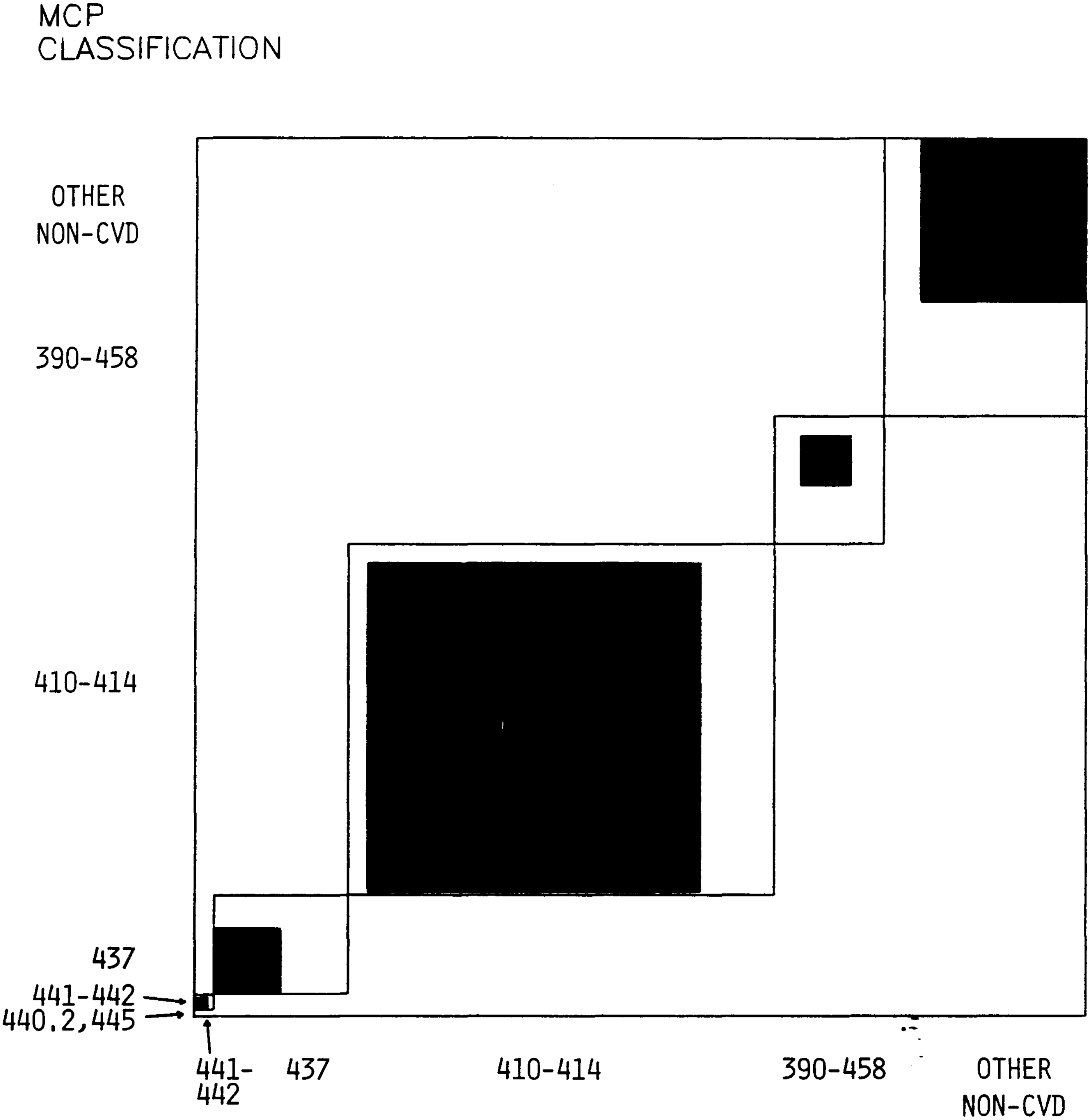

MCP
CLASSIFICATION
OTHER
NON-CVD
390-458
410-414
437
441-442
440.2,445
441-
442
437
410-414
390-
458
OTHER
NON-CVD
NOSOLOGICAL CLASSISICATION

FIGURE 4 (B)



NOSOLOGICAL CLASSISICATION