Intraobserver and interobserver reproducibility of the FAB classification was assessed for two independent observers whose decisions are acted on for treatment of patients with acute leukemia in the Hamilton region. Intraobserver reproducibility was assessed for Wright-stained preparations that were examined independently on two consecutive occasions at least 2 weeks apart. A third reading was performed with Wright stain and cytochemical data, and the fourth reading was done with addition of immunophenotype data. Concordance was calculated using a statistic that corrects for chance-expected agreement (k), and a weighted statistic that takes into account the seriousness of disagreements was used. Samples were available for morphological and cytochemical assessment on 105 patients, and immunophenotype data were available on 93 specimens. Intraobserver concordance was 64.8% and 70.5% for observers A and B, respectively, with kappa values of .56 and .62. There were 37 discordant readings for observer A and 31 for observer B, with each observer discordant between lymphocytic:nonlymphocytic phenotypes in ten cases. Concordance between observers was 63% (k = .54) and 72% (k = .65) for each of two separate readings for Wright-stained preparations only. Reproducibility improved to 89% (k = .86) when cytochemistry was added. When immunophenotype information was provided in addition to Wright-stained and cytochemical preparations, the agreement was 99%. Lymphocytic:nonlymphocytic discordance between observers occurred on nine occasions when Wright-stained preparations only were available and four times when cytochemistry was added; it did not occur with immunophenotyping. The study suggests that immunophenotyping, when added to morphological assessment of acute leukemia, may contribute substantially to agreement between observers.

© 1986 by Grune & Stratton, Inc.
Intraobserver concordance. From December 1983 to July 1985, specimens from 105 consecutive patients were submitted for evaluation. Morphological evaluation with cytochemistry was performed on all specimens. Immunophenotyping data were available for 93 samples.

Figure 1 shows the intraobserver agreement matrix for each observer for a six-category scale in which M3 and M6 were pooled and all lymphoid subtypes were pooled separately into single categories. The weightings for disagreement among categories, which were used to calculate the $k_w$ statistic, are inserted in the right upper corner of each cell in (A).

Fig 2. Agreement matrix within observers of a two-category scale (lymphocytic:nonlymphocytic) for separate readings. (A) Observer A; (B) observer B. The weightings used for partial agreement are indicated in the upper right corner of each cell in (A).

Results

Intraobserver concordance. From December 1983 to July 1985, specimens from 105 consecutive patients were
Each observer agreed 90% of the time as to whether the leukemia was lymphocytic or nonlymphocytic. For each observer, there were ten discordant readings. Reference to the agreement matrices of Fig 1 show that of the 20 lymphocytic:nonlymphocytic disagreements, 17 (85%) involved the FAB M-1 subtype.

Figure 3 shows intraobserver concordance for lymphoid subcategories (L-1 through L-3) of acute lymphocytic leukemia. For observer A, there was agreement on only 67.7% of cases \((k = .41)\), whereas for observer B, the results were similar (62.5% agreement, \(k = .43\)). However, of the ten discordant readings for observer A, all represented lymphocytic:nonlymphocytic disagreements, whereas for observer B, of the 12 discordant readings, ten represented lymphocytic:nonlymphocytic disagreements. Therefore, of the 22 disagreements, only two represented errors within lymphoid subtypes. There was no disagreement between L-3, and L-1, 2.

Table 1 summarizes concordance results for intraobserver ratings for both observers using the six-category scale. When morphology alone is examined, the overall percentage of agreement for the first and second examinations is 64.8% for observer A and 70.5% for observer B. When chance-expected agreement is accounted for, the kappa value is .56 for observer A and .62 for observer B. The \(k_w\) scores are higher than the \(k\) scores, since the weightings give credit for partial agreement.

The relationship between the \(k\) score and the strength of agreement will be discussed subsequently. There are no universally accepted guidelines to determine an "acceptable" level of agreement.

Interobserver concordance. The agreement between observers was assessed for each of two separate readings using Wright-stained slides only, and for a third reading in which cytochemical information was added. Figure 4 shows the agreement matrix for each of two separate readings of FAB classification using the six-category scale for which morphological information alone was available, and Fig 5 shows the agreement matrix when cytochemical data were also provided.

To determine how the observers would use the classification based on immunophenotype, this information was provided to each separately accompanied by cytochemical and Wright-stained preparations. This is not a true test of concordance for the immunophenotype method, since immunophenotype preparations themselves were not assessed. However, the study does provide information on how these data are used by the observers.

The data presented in Table 2 summarize the results of interobserver agreement for all conditions. When Wright-stained preparations only were available, there was 63% agreement (\(k = .54\)) on the first reading and 72% (\(k = .65\)) on the second reading. When weightings were used to credit partial agreements, \(k_w\) was .63 on the first reading and improved to .71 on the second reading. This suggests the possibility of a learning effect from the first to the second reading. Although the number of discordant readings declined from the first to the second (39 to 29), the proportion of lymphocytic:nonlymphocytic discordant pairs was identical (31%) among the total number of disagreements for the first (12/39) and second (9/29) readings.

Figure 5 shows the pattern of concordance between observers when cytochemical information was provided.
Interobserver concordance improved from $k = 0.65$ in the absence of cytochemistry to 0.86 when cytochemical data were added, and $k_\text{p}$ improved from 0.71 to 0.88. The availability of immunophenotyping was associated with almost perfect agreement ($k = 0.99$). The influence of cytochemistry and immunophenotyping on nonlymphocytic:lymphocytic discordance is shown in Table 2. When Wright-stained preparations alone were examined, there were nine occasions in which lymphocytic and nonlymphocytic subtypes were discordant. Addition of cytochemical information resulted in only four such disagreements, while immunophenotyping eliminated this problem. These data suggest that observers depended primarily on the immunophenotype pattern when making a final judgment.

Although the demonstration of a high degree of reproducibility does not necessarily establish the validity of a classification, we decided to use the immunophenotyping results as a “gold standard” to determine the final phenotype among discordant readings involving the lymphocytic:nonlymphocytic categories. These data are shown in Table 3. Among the ten discordant readings by observer A when Wright-stained preparations only were available, eight are classified as lymphocytic and two are of nonlymphocytic origin. For observer B, among ten discordant readings, six were lymphocytic and four, nonlymphocytic. Interobserver discordance among lymphocytic:nonlymphocytic subtypes in Wright-stained preparations occurred 12 times on the first reading, of which nine were immunophenotyped as lymphocytic and three as nonlymphocytic. On the second reading, eight of the nine discordant readings immunophenotyped as lymphocytic and one as nonlymphocytic. There were only four lymphocytic:nonlymphocytic discordant readings with cytochemistry, and all immunophenotyped as lymphocytic.

In view of the high degree of concordance with immunophenotyping, we have characterized all of the 93 cases for which this information was available. Fifty-five patients had acute nonlymphocytic leukemia, 25 had acute lymphocytic leukemia, and 13 had mixed phenotypes of which 12 were mixed lineage and one had dual markers on the same cells. The acute lymphocytic leukemias, including those with dual markers, could be subclassified into 15 CALLA positive, 12 T cell, 5 null cell, 2 B cell, and 1 pre-B cell. There was only one disagreement among 93 samples in which observer A classified a sample as M-1 and observer B classified the same sample as M-2. For all other samples, there was perfect agreement within the nine-category scale of the FAB classification.

The almost perfect agreement with immunophenotyping depends on agreed-to specific objective criteria for subclassification. On this basis we can conclude that observers who accept these objective criteria for immunophenotyping will agree in most cases. The objective criteria used by our observers is the subject of a separate report.8

**DISCUSSION**

The appropriate treatment of acute leukemia is influenced by the type of disease, which is usually characterized by morphological assessment. This is especially relevant for the distinction between acute lymphocytic and nonlymphocytic leukemias, in which induction regimens are quite different. Apart from a choice of induction therapies, the intensity of postremission consolidation chemotherapy may be influenced if an association between FAB subtype and remission duration is found in future studies. There is evidence that the FAB classification may distinguish between leukemias that are biologically different, since correlations between FAB subtype and specific nonrandom chromosomal abnormalities have been shown.1,2 In view of these considerations, it is important that the reproducibility of the FAB classification and of any additional diagnostic procedures be properly evaluated.

We have presented a formal evaluation of an agreement
study between observers whose decisions in almost all cases influence patient management in our region. We have used a statistical procedure that accounts for chance-expectation agreement. Furthermore, we have attempted to determine the contribution of cytochemistry and immunophenotyping to reproducibility of the FAB classification. Our data indicate that immunophenotyping may contribute substantially to the morphological classification of acute leukemia. In this study, biphenotypic leukemias were diagnosed only by the immunophenotype method.

The majority of studies that have examined the reproducibility of the FAB or a closely related classification in acute leukemia focused on the agreement between single observers and a panel of acknowledged experts (see Table 4). Only one study mentioned the extent of intraobserver concordance, although no details regarding the pattern of disagreement within observers was presented. In the majority of cases, agreement of panels of experts or of three or more observers was examined. As Table 4 indicates, some of these studies examined extent of agreement using morphological criteria alone, while others included cytochemical evaluation. The number of categories that were used to determine extent of agreement differed among the studies. In view of these differences, it is difficult to compare the results across studies. Furthermore, none of the previously reported studies used a statistical analysis that accounts for chance-expectation agreement.

There are no universally recognized guidelines for what represents an acceptable level of reproducibility for the morphological classification of acute leukemia. The results reported in this study are similar to those of other studies, but for the reasons given, the results are not directly comparable. It would be highly desirable if reproducibility were perfect. However, where subjective judgments play an important role in classifying disease, it is unrealistic to expect that all observers will act identically. Thus, classifications that depend on human interpretations are subject to wide variation, which can be considered part of the measurement error. The establishment of easily recognizable and interpretable objective criteria will contribute substantially to reducing this variability.

An estimate of the level of agreement that can be expected when judgments about clinical information are made by two observers has been reviewed by Koran. None of the studies reviewed in Koran's report dealt with hematologic/morphological evaluation. Moreover, most of the data reviewed dealt with simple two-category judgments, such as whether a disease or clinical sign (on physical examination, radiologic examination, EEG, EKG, etc) was present or absent, or normal or abnormal. For these evaluations, kappa scores ranged from .3 to .83.

Demonstration of reproducibility is not a sufficient condition for establishing validity. Two observers can agree on an event, but both may be wrong. In this case, reproducibility may be high, but the validity of the observations cannot be said to have been established. Although reproducibility is not a sufficient condition for establishing validity, the remarkable consistency of the immunophenotyping results in this study suggest that surface marker analysis of acute leukemia may be established among investigators. Proper evaluation of concordance for these methods will require further studies. In addition, further studies will be required to determine whether clinical decisions based on different methods of morphological classification of acute leukemia will improve treatment results.

REFERENCES


4. Bennett JM, Catovsky D, Daniel MT, Flandrin G, Galton

Table 4. Agreement Studies Using FAB Classification in Acute Leukemia

<table>
<thead>
<tr>
<th>First Author</th>
<th>Year</th>
<th>Ref. No.</th>
<th>Data Base</th>
<th>No. of Diagnostic Categories</th>
<th>No. of Cases</th>
<th>Agreement (%)</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Whitacker</td>
<td>1979</td>
<td>13</td>
<td>M,C</td>
<td>6</td>
<td>278</td>
<td>—</td>
<td>45.7</td>
</tr>
<tr>
<td>Viana</td>
<td>1980</td>
<td>14</td>
<td>M</td>
<td>4</td>
<td>170</td>
<td>—</td>
<td>74.8–86.8</td>
</tr>
<tr>
<td>Mertelsmann</td>
<td>1980</td>
<td>3</td>
<td>M</td>
<td>11</td>
<td>228</td>
<td>—</td>
<td>69.4</td>
</tr>
<tr>
<td>Bennett</td>
<td>1981</td>
<td>4</td>
<td>M</td>
<td>3</td>
<td>100</td>
<td>—</td>
<td>60</td>
</tr>
<tr>
<td>Head</td>
<td>1985</td>
<td>15</td>
<td>M,C</td>
<td>9</td>
<td>233</td>
<td>63.8,76,58.3–80</td>
<td>Individual classification</td>
</tr>
<tr>
<td>This study</td>
<td>M</td>
<td>6</td>
<td>105</td>
<td>64.8,70.5</td>
<td>72</td>
<td>86</td>
<td>agreement</td>
</tr>
</tbody>
</table>

Concordance results using consensus are not included, since consensus is not associated with independent blind assessment. ANLL, acute nonlymphocytic leukemia; M, morphology (Wright stain); C, cytochemistry; I, immunophenotype.
The contribution of cytochemistry and immunophenotyping to the reproducibility of the FAB classification in acute leukemia

GP Browman, PB Neame and P Soamboonsrup