Validation of Consensus Panel Diagnosis in Dementia

Matthew J. Gabel, PhD; Norman L. Foster, MD; Judith L. Heidebrink, MD; Roger Higdon, PhD; Howard J. Aizenstein, MD, PhD; Steven E. Arnold, MD; Nancy R. Barbas, MD, MSW; Bradley F. Boeve, MD; James R. Burke, MD, PhD; Christopher M. Clark, MD; Steven T. DeKosky, MD; Martin R. Farlow, MD; William J. Jagust, MD; Claudia H. Kawas, MD; Robert A. Koeppe, PhD; James B. Leverenz, MD; Anne M. Lipton, MD, PhD; Elaine R. Peskind, MD; R. Scott Turner, MD, PhD; Kyle B. Womack, MD; Edward Y. Zamrini, MD

Background: The clinical diagnosis of dementing diseases largely depends on the subjective interpretation of patient symptoms. Consensus panels are frequently used in research to determine diagnoses when definitive pathologic findings are unavailable. Nevertheless, research on group decision making indicates that many factors can adversely affect panel performance.

Objective: To determine conditions that improve consensus panel diagnosis.

Design: Comparison of neuropathologic diagnoses with individual and consensus panel diagnoses based on clinical scenarios only, fludeoxyglucose F 18 positron emission tomography images only, and scenarios plus images.

Setting: Expert and trainee individual and consensus panel deliberations using a modified Delphi method in a pilot research study of the diagnostic utility of fludeoxyglucose F 18 positron emission tomography.

Patients: Forty-five patients with pathologically confirmed Alzheimer disease or frontotemporal dementia.

Main Outcome Measures: Statistical measures of diagnostic accuracy, agreement, and confidence for individual raters and panelists before and after consensus deliberations.

Results: The consensus protocol using trainees and experts surpassed the accuracy of individual expert diagnoses when clinical information elicited diverse judgments. In these situations, consensus was 3.5 times more likely to produce positive rather than negative changes in the accuracy and diagnostic certainty of individual panelists. A rule that forced group consensus was at least as accurate as majority and unanimity rules.

Conclusions: Using a modified Delphi protocol to arrive at a consensus diagnosis is a reasonable substitute for pathologic information. This protocol improves diagnostic accuracy and certainty when panelist judgments differ and is easily adapted to other research and clinical settings while avoiding the potential pitfalls of group decision making.

Arch Neurol. 2010;67(12):1506-1512

Many dementing diseases lack distinctive physical findings or validated biomarkers, thus making accurate clinical diagnosis challenging. Clinicians often must reach a diagnosis based solely on their judgment of informant history of variable quality and the relative prominence of deficits in specific cognitive domains. Because these subjective judgments understandably differ among individual clinicians, the accuracy and confidence of diagnoses also vary. Diagnostic criteria have been developed to provide guidance for clinicians, but applying these criteria also requires interpretation and judgment. Consequently, neuropathologic examination findings continue to be the standard criterion for determining the cause of a dementing illness.

The validity of research results depends on accurate diagnosis. Recognizing the limitations of individual clinician diagnoses, research studies often use the consensus of a panel when histopathologic information is unavailable. It is hoped that a panel will achieve greater diagnostic reliability, accuracy, and certainty than even an individual expert. Despite this hope, there has been little examination of consensus panel performance in determining the cause of dementia. The limited empirical evidence available suggests that consensus panel results may be suspect. For example, similarly composed medical panels often reach varying conclusions about the same sets of questions, raising serious doubts about panel reliability. In addition, theoretical and empirical studies of group decision making indicate that depending on their composition and procedures, consensus panels may...
not achieve highly accurate decisions. Consequently, the absence of strong evidence regarding the efficacy of consensus panels is a potentially serious problem for dementia research.

Bringing empirical evidence to bear on this question is complicated by the variety of consensus panel goals, memberships, and procedures currently in use. Given this variety, we need to identify effective panels and cannot simply assume that any single panel will be as accurate as others. For example, consensus panels can have different goals. Some are designed to identify only patients for whom a diagnosis is likely to be highly accurate, whereas others seek the best diagnosis for all patients, recognizing that accuracy may be higher in some situations than in others. Consensus panels also vary in their composition and organization. Members may include only a single specialty or may be multidisciplinary. Some panels include individuals who have personally examined the patient with the intent of ensuring the most direct and detailed information. Other panels explicitly exclude individuals with “special knowledge” of the patient out of concern that such individuals would exert disproportionate influence on group judgments and suppress independent analysis, which is the theoretical advantage of panel diagnosis. Furthermore, panel rules for arriving at a group diagnosis also are variable. For some, majority agreement is sufficient. For others, unanimity is expected or required. Finally, the panel may follow a rigorous protocol or be quite informal. Some simply determine whether there are objections to the individual physician judgment, whereas others expect each panelist to arrive at a diagnosis independently. Social science research shows that these aspects of panel organization affect the accuracy of consensus judgments.

The Delphi method of consensus is a formal and rigorous procedure that incorporates organizational features that social science theory indicates promote accurate individual and group judgments. This method is commonly used to set professional priorities and establish guidelines, but the exact protocol can vary in panel size, the use of face-to-face discussion, and the number of iterations before a final decision is reached. The essential features of the Delphi method are (1) presentation of a uniform set of information to the panel (thus excluding individuals with unique special knowledge), (2) an initial independent decision of each panelist that is recorded and subsequently shared with others, (3) discussion of the recorded opinions of panelists, and (4) a final group decision. Votes are used to ensure independent judgments, and diversity of opinions is encouraged through panel membership and during discussions.

We took advantage of an opportunity to explore diagnostic performance of consensus panels provided by trials we conducted to examine the diagnostic utility of fludeoxyglucose F 18 positron emission tomography (FDG-PET). Consensus panels generally are convened only when there is no standard criterion available. In these trials, however, neuropathologic findings were available, and we undertook these studies to determine the extent to which consensus panel diagnosis might be a justifiable alternative to postmortem examination. In the United States, FDG-PET currently is reimbursed in dementia only when physicians find it difficult to distinguish Alzheimer disease (AD) from frontotemporal dementia (FTD). Thus, it was scientifically appropriate in these trials to restrict diagnostic options to these 2 possibilities. The requirement of a binary decision was fortuitous because it significantly simplified the analysis of panel performance. Diagnostic decisions inherently vary widely in difficulty, and repeated use of exactly the same decision in this study allowed us to evaluate key variables, including the diversity of diagnostic perspectives, the types of patient information reviewed, and the decision criteria for consensus. Although clinical diagnosis is complex and requires the consideration of multiple conditions, binary decisions are relevant to clinical practice. For example, after an extensive dementia evaluation, researchers often must make critical diagnostic judgments, choosing between only 2 of the most likely possibilities such as demented or nondemented, mild cognitive impairment or normal for age, AD or not AD, and AD or vascular dementia.

METHODS

Two consensus panels, each composed of 6 panelists, and 6 additional individual raters reviewed clinical data to arrive at a diagnosis of AD or FTD. None of the panelists or raters had direct interaction with the patients being considered. Although panelists and raters were aware that patients had only 1 of 2 possible diagnoses, they did not know the proportion with each diagnosis.

PANEL CHARACTERISTICS

A “trainee” panel met twice and consisted of 6 physician trainees in specialties involved in dementia care from a single institution: 2 neurology residents, 2 geriatric medicine fellows, 1 psychiatry resident, and 1 geriatric psychiatry fellow. One of these trainees was present for the review of only 28 of the 45 patients. A second “expert” panel met 3 times at least 6 months apart and was composed of 6 physicians (4 neurologists and 2 geriatric psychiatrists) involved in dementia care and research at 1 of 4 National Institute on Aging–funded Alzheimer centers.

RATER CHARACTERISTICS

Distinct from the members of the panels, this study also involved 6 “raters”: dementia specialist neurologists, each with 10 to 23 years of experience in dementia care, 2 from each of 3 National Institute on Aging–funded Alzheimer centers. Raters arrived at a diagnosis based solely on their private consideration of the same patient information provided to the panels. They did not convene as a panel for discussion or share information with each other about their diagnoses. These raters provided a set of decisions by individual experts to compare with panel diagnoses.

PATIENT DATA

Clinical scenarios and FDG-PET images were evaluated from 45 patients with a postmortem examination documenting a histopathologic diagnosis of AD (n=31) or FTD (n=14) uncomplicated by other abnormalities, such as a stroke or a significant number of cortical Lewy bodies. Foster et al provide a full description of the pathologic findings in these cases, scenario development, imaging methods, and training of raters and panelists in image interpretation. Neuropsychological data were not included. Three sets of data were prepared for each pa-
tient: clinical scenario alone, FDG-PET images alone, and scenarios plus PET images. Patient data were labeled using random number identifiers, with a different series of random numbers used in each data set.

**DIAGNOSTIC DELIBERATIONS**

Consensus panel deliberations uniformly followed the RAND–University of California at Los Angeles modified Delphi procedure. Each set of data was presented on a different day and in a different patient order to keep panelists blind to their previous diagnostic judgments. A panel leader organized the meeting and encouraged discussion but did not participate in discussion or voting. Panelists began by privately considering the information provided about each patient. They then marked a card indicating their diagnosis of AD or FTD and their level of confidence in that diagnosis (very confident, somewhat confident, or uncertain). The panel leader collected the cards and announced the “vote tally” (eg, 3 AD and 3 FTD) to the panel. At that point, the panelists were encouraged to discuss the case and their reasons for arriving at a specific diagnosis. During individual review and group deliberations of the clinical scenarios, we encouraged reference to published diagnostic criteria for AD and FTD, but we neither suggested nor imposed any rules regarding the interpretation of the criteria or individual patient information.

After discussion, panelists again marked a card in private indicating their diagnosis and diagnostic confidence. After these cards were collected, the group was asked to arrive at a final diagnosis. The panelists were not provided with a decision rule (eg, simple majority) but were told that they needed to return a decision for the panel. The leader then recorded the consensus decision, and the panel turned to the next patient and repeated the same procedure. There was no time limit for individual deliberation or group discussion. Research staff recorded the time taken for these deliberations and made qualitative observations.

Individual raters not involved in the panels reviewed the same 3 types of data as panelists and provided a diagnosis of AD or FTD and their level of confidence. In all, there were 810 diagnostic judgments by individual raters, 2126 judgments by individual panelists, and 180 consensus judgments by panels.

**STATISTICAL ANALYSIS**

Diagnostic judgments of raters, panelists, and the consensus panels were compared with the neuropathologic diagnoses (the reference standard). For each panel, we computed statistics for sensitivity, specificity, predictive value, and likelihood ratio. With only 2 diagnostic options, positive and negative predictive values were complementary, and sensitivity and specificity for FTD were reciprocal to those for AD. We used κ statistics to evaluate the reliability of consensus diagnoses across panels and the level of diagnostic agreement within panels. The degree of agreement was rated as fair (κ = 0.20–0.39), moderate (κ = 0.40–0.59), substantial (κ = 0.60–0.79), or almost perfect (κ = 0.80–1.0), according to convention. We analyzed consensus panel performance relative to that of raters and panelists by fitting logistic regression models to a binary variable representing correct diagnosis, with raters, panelists, and the consensus panel as covariates. This provides an estimate of the odds ratio that an expert was more accurate than the panel, which served as the reference category. The change in panelist diagnostic accuracy from before to after discussion in each panel was analyzed using logistic regression models fit to a binary response variable for whether the prediscussion or postdiscussion diagnosis was correct and included the timing of the diagnosis as a covariate (before or after the diagnosis). The change in diagnostic confidence from prediscussion to postdiscussion was evaluated in a similar manner, fitting the model to a binary variable for whether the panelist was “very confident.”

To determine the extent to which changes in panelists’ diagnoses were beneficial, we estimated logistic regression models for all the panelists who changed their confidence or diagnosis from prediscussion to postdiscussion. We fit the model to a binary variable indicating whether a change was beneficial, defined as a shift to the correct diagnosis, an increase in confidence in a correct diagnosis, or a decrease in confidence in an incorrect diagnosis. The intercept provides an estimate of the log odds ratio that the change was beneficial.

Because diagnoses of the same case by different panelists or of different cases by the same panelists are potentially correlated, estimates of standard errors were adjusted to account for violations of standard independence assumptions. Where relevant, standard errors were adjusted for the longitudinal nature of the prediscussion and postdiscussion data in some analyses. Specifically, the standard errors of the statistical tests were adjusted using a robust covariance estimator that incorporated estimates of correlation between panelists and between patients. We then used the adjusted variance estimate to generate corrected P values. Also, where relevant, P values were adjusted for multiple tests using the Hochberg correction. McNemar χ² tests were used to assess whether consensus diagnoses were more accurate than alternative methods of group diagnosis (eg, simple majority rule).

**RESULTS**

**RELIABILITY, ACCURACY,**
**AND CONFIDENCE OF DIAGNOSIS**

The accuracy of the consensus diagnoses of the trainee and expert panels was superior to that of the individual diagnoses of their own members when considering clinical scenarios alone (Figure 1A). In general, the accuracy of the consensus diagnoses were superior to those of expert raters making individual judgments (Figure 1B). The consensus diagnoses were more accurate than the diagnoses of 10 of the 11 individual panelists and 5 of the 6 individual expert raters, and these differences often reached statistical significance. On average, the 12 experts individually performed better than the 5 trainee panelists, although after deliberation, the trainee and expert panels had the same diagnostic accuracy. Indeed, the trainee panel was statistically significantly more accurate than the individual opinions of 3 of the 6 expert panelists (eFigure; http://www.archneurol.com).

Individual diagnostic accuracy and confidence were high with review of FDG-PET images with or without scenarios, and there was less individual variation in diagnoses. In these situations, panel accuracy was rarely superior to that of individual raters or expert panelists, and deliberations did not provide the same benefits seen with scenarios alone (Figure 2). Indeed, most individual experts had the same or higher diagnostic accuracy as the panel.

The consensus diagnoses ranged from 84% accurate when based exclusively on clinical scenarios to 89% when the diagnosis included review of FDG-PET images. The AD sensitivity and FTD specificity (89%–94%) were higher than AD specificity and FTD sensitivity (71%–86%) (eTable 1). As expected from previous experience, the
diagnostic accuracy of individuals and panels was less when considering FTD compared with AD. Despite the concerns of other researchers, the consensus judgments were highly reproducible across panels (2-way \( \kappa = 0.68-0.90 \)) despite differences in panel memberships and diagnostic information reviewed (Figure 3).

Panelists’ judgments tended to converge after discussion in all situations, as indicated by the increase in mean \( \kappa \) agreement scores within panels, and diagnostic confidence also increased (Table). This increase in agreement after deliberation was not uniformly associated with beneficial changes in diagnosis or confidence (eTable 2). Similar to the panel diagnoses, the salutary effect of the consensus process varied by type of diagnostic information. Panelists typically made beneficial changes when reviewing scenarios alone. These changes were predominantly due to panelists who were uncertain or only somewhat confident in their initial diagnoses (eTable 3).

Similarly, panelists who were not very confident in their initial diagnosis accounted for all diagnostic changes when reviewing images. However, compared with reviewing scenarios alone, these changes were fewer in number and were typically not beneficial (eTable 3).

**EFFECT OF PANEL CONSENSUS RULES ON DIAGNOSTIC ACCURACY**

After discussion and the second vote, the panel was asked to determine a single final consensus diagnosis. When 5 of 6 or 6 of 6 panelists agreed on a prediscussion diagnosis, this diagnosis was always adopted as the consensus diagnosis. The final diagnosis also never deviated from the majority diagnosis after discussion. As the threshold for consensus increases from 4 of 6 to unanimity, accuracy generally improves, although gains are small and at the expense of many patients going undiagnosed (eTable 4). Voting again after discussion allowed more patients to be diagnosed and by a larger majority of panelists. None of the alternative rules exhibited a statistically significantly higher accuracy than the forced consensus rule (eTable 4).

In general, discussion caused panelists to converge around the prediscussion majority diagnosis, regardless of whether that diagnosis was correct or incorrect. The only exceptions were 3 cases in the trainee panel where discussion led to a change from a simple majority incorrect diagnosis to a majority correct diagnosis. There were no instances of discussion changing a correct majority diagnosis.
PRACTICAL IMPLICATIONS

Review of the literature raises concerns about many of the consensus procedures currently in use in dementia research. Other consensus procedures may not provide similar positive results as does the modified Delphi protocol used in this study. The limitation of other consensus methods may not be readily apparent to investigators because there often is a high pretest probability of a single diagnosis. In this situation, diagnostic errors will change autopsy confirmation rates only slightly. On the other hand, in this study, pretest probability of FTD was unknown to the raters but was considerably higher than that in many AD research studies and, thus, provided an informative setting to assess consensus.

Properly constituted consensus panels are time consuming and are expensive, and require considerable effort...
Discussion among equals using identical patient data. This study demonstrates the value of open discussions involving a multidisciplinary perspective. Diversity of opinion is important for realizing the potential benefits of consensus panels, and panel membership should be multidisciplinary whenever feasible. It might be helpful for individuals with personal information about patients to present data for consideration and response to questions, but including them on the diagnostic panel is problematic because it could discourage diverse opinions voiced by those without “special knowledge.” This study demonstrates the value of open discussion among equals using identical patient data.

**Potential Limitations**

Given the variety of consensus panels, these findings may not generalize to other settings. The clinical scenarios reviewed in this study were based on a comprehensive longitudinal prospective study and varied considerably in the number of examinations, the detail and length of the medical record, and the quality of the medical history. Although this reflects many clinical situations, restricting data to an initial visit may provide more limited or ambiguous diagnostic information and would likely cause more practitioner error than observed in this study. In contrast, prospectively collected comprehensive longitudinal data would probably produce less error because diagnostic accuracy improves with longitudinal information.1,2 We can only speculate as to whether diagnostic accuracy would be affected by a change in the quantity or quality of patient information. Nevertheless, as long as panelists can independently review and interpret the patient information, we would expect a benefit from consensus panels. Although not a desirable setting, situations that provide limited and ambiguous information likely would cause more individual diagnostic errors and provide a greater opportunity for improvement using consensus methods. Likewise, an expanded set of diagnostic choices is likely to reduce the reliability of consensus diagnosis but could result in even stronger performance of panels relative to individuals than was found in this study.

Eventually, validated biomarkers may make interpretation of clinical data less subjective. Until that elusive goal is achieved for dementing diseases, consensus diagnosis following a carefully considered protocol that allows for diverse opinion and deliberations involving a multidisciplinary panel without special knowledge will be an appropriate approach to maximizing diagnostic accuracy.

**Accepted for Publication:** June 9, 2010.

**Author Affiliations:** Department of Political Science, Washington University, St Louis, Missouri (Dr Gabel); Center for Alzheimer’s Care, Imaging, and Research and Department of Neurology, University of Utah, Salt Lake City (Dr Foster); Department of Neurology, University of Michigan, Ann Arbor (Dr Heidebrink); Neurology Service, Department of Veterans Affairs Medical Center, Ann Arbor (Dr Heidebrink); and Seattle Children’s Research Institute, Seattle, Washington (Dr Higdon); Department of Psychiatry, University of Pittsburgh, Pittsburgh, Pennsylvania (Dr Aizenstein); Departments of Neurology (Drs Barbas and Turner) and Radiology (Dr Koepepe), University of Michigan, Ann Arbor; Departments of Psychiatry (Dr Arnold) and Neurology (Drs Arnold and Clark), University of Pennsylvania, Philadelphia; Department of Neurology, Mayo Clinic, Rochester, Minnesota (Dr Boeve); Department of Neurology, Duke University, Durham, North Carolina.

---

**Table. Individual Panelist Accuracy, Confidence, and Agreement Before and After Panel Deliberation**

<table>
<thead>
<tr>
<th>Panel</th>
<th>Diagnostic Information</th>
<th>Panelist Accuracy, Mean (Range), %</th>
<th>Panelist Confidence, Mean (SE)</th>
<th>Panelist Agreement, Mean (Range), %</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Prediscussion</td>
<td>Postdiscussion</td>
<td>Prediscussion</td>
</tr>
<tr>
<td>Trainee</td>
<td>Scenario</td>
<td>71 (64-78)</td>
<td>79 (73-92)</td>
<td>0.45 (0.08)</td>
</tr>
<tr>
<td>Expert</td>
<td>Scenario</td>
<td>79 (71-87)</td>
<td>82 (78-84)</td>
<td>0.61 (0.07)</td>
</tr>
<tr>
<td>Expert</td>
<td>Images</td>
<td>90 (84-96)</td>
<td>90 (87-93)</td>
<td>0.81 (0.03)</td>
</tr>
<tr>
<td>Expert</td>
<td>Scenario and images</td>
<td>90 (89-93)</td>
<td>89 (87-91)</td>
<td>0.84 (0.07)</td>
</tr>
</tbody>
</table>

*Panelists (trainees and experts) reviewing scenarios alone were more likely to increase than decrease their diagnostic accuracy after deliberation (odds ratio [OR]=1.56, P<.01 for trainees; OR=1.26, P=.03 for experts). They were also more likely to increase their diagnostic confidence after deliberation (OR=1.85, P<.01 for trainees; OR=1.61, P<.01 for experts). In contrast, panelists reviewing scenarios plus images or images alone were no more likely to increase than decrease their diagnostic accuracy after deliberation (OR=.87, P=.16 for scenarios plus images; OR=1.0, P>.99 for images). However, these panelists were more likely to increase their diagnostic confidence after deliberation (OR=1.25, P=.04 for scenarios plus images; OR=1.45, P>.01 for images only). The mean 2-way interpanelist κ value increased with deliberation on all panels (P<.01, difference in means test). Standard errors for κ were calculated using the bias-corrected bootstrap method.

*Adjusted for 17 cases not reviewed by 1 panelist.*
Additional Contributions: David E. Kuhl, MD; Sid Gilman, MD; Henry Buchtel, PhD; David Knesper, MD; R. Scott Turner, MD, PhD; and Kirk Frey, MD, PhD, made images from their research available for this study; Charles DeCarli, MD, contributed as a site investigator for grant AG22394 from the National Institutes of Health; Peijun Chen, Charles Davies, Shelley Herschner, and Joseph O. Nnodim served on the pilot trainee panel; and Jeff Gill, PhD; Ryan Moore, PhD; and Diana O’Brien, MA, provided valuable statistical advice.

REFERENCES


