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.

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 2 possible diagnoses, they did not know the proportion with each diagnosis.

**METHODS**

**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 sce-

narios plus PET images. Patient data were labeled using ran-
dom 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 proce-
dure. Each set of data was presented on a different day and in
a different patient order to keep panelists blinded to their pre-
vious diagnostic judgments. A panel leader organized the meet-
ing and encouraged discussion but did not participate in discus-

sion 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 confi-
dent, or uncertain). The panel leader collected the cards and an-
nounced 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 re-
garding the interpretation of the criteria or individual patient
information.

After discussion, panelists again marked a card in private in-
dicating 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 pro-
cedure. There was no time limit for individual deliberation or group
discussion. Research staff recorded the time taken for these de-

liberations 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 diag-
nostic judgments by individual raters, 2126 judgments by in-
dividual 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 predic-
tive values were complementary, and sensitivity and specific-
ity for FTD were reciprocal to those for AD. We used χ2 sta-
tistics 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), moder-
ate (κ=0.40-0.59), substantial (κ=0.60-0.79), or almost per-
fect (κ=0.80-1.0), according to convention. We analyzed con-
sensus panel performance relative to that of raters and panelists
by fitting logistic regression models to a binary variable repre-
senting correct diagnosis, with raters, panelists, and the con-
sensus 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 diag-
nostic accuracy from before to after discussion in each panel
was analyzed using logistic regression models fit to a binary re-
sponse variable for whether the prediscussion or postdiscus-
sion diagnosis was correct and included the timing of the di-
agnosis as a covariate (before or after the diagnosis). The change
in diagnostic confidence from prediscussion to postdiscus-
sion 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’ diag-
noses were beneficial, we estimated logistic regression models
for all the panelists who changed their confidence or diagno-
sis 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 con-
fidence 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 corre-
lated, estimates of standard errors were adjusted to account for
violations of standard independence assumptions. Where rele-
vant, standard errors were adjusted for the longitudinal na-
ture of the prediscussion and postdiscussion data in some analy-

ses. Specifically, the standard errors of the statistical tests were
adjusted using a robust covariance estimator that incorpo-
rated estimates of correlation between panelists and between
patients. We then used the adjusted variance estimate to gen-
erate corrected P values. Also, where relevant, P values were
adjusted for multiple tests using the Hochberg correction. Mc-
Nemar χ2 tests were used to assess whether consensus di-
agnoses 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 clinici-

(Reprinted) Arch Neurol/Vol 67 (No. 12), Dec 2010   www.archneurol.com

©2010 American Medical Association. All rights reserved.

Downloaded From:  by a Non-Human Traffic (NHT) User on 10/24/2018
diagnostic accuracy of individuals and panels was less when considering FTD compared with AD. Despite the concerns of other researchers,\textsuperscript{3} 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
Figure 3. Panel diagnostic accuracy by patient. Each horizontal line represents a single patient. Panel diagnoses in agreement with neuropathologic diagnoses are shown in gray. Panel diagnostic errors are shown in peach. Panels often were in error in the same patients. The pairwise \( \kappa \) agreement (SE) between the diagnoses of the trainee and expert panels for scenarios was 0.79 (0.15); for trainee (scenario) and expert (images) panels, 0.69 (0.15); for trainee (scenario) and expert (scenario + images) panels, 0.79 (0.15); for expert (scenario) and expert (images) panels, 0.69 (0.15); and for expert (images) (scenario + images) panels, 0.79 (0.15).

**COMMENT**

The modified Delphi protocol resulted in reliable consensus diagnoses across panels of varying expertise and diagnostic information. The expertise of individuals does not negate the benefit of consensus; consensus improved the accuracy of nonexpert and expert panelists alike. When reviewing only clinical scenarios, trainee and expert consensus panel diagnoses were typically as accurate or more accurate than individual expert diagnoses. In addition, the consensus process led panelists to improve the accuracy of their individual diagnoses. Thus, when reviewing scenarios, a modified Delphi protocol for consensus panels provided sufficiently accurate diagnoses to be considered ideal when histopathologic information is unavailable.

In contrast, consensus diagnoses when reviewing FDG-PET images, with or without scenarios, were rarely better than those of individual experts, and panelists typically made adverse diagnostic changes after deliberation. What accounts for this variation in performance? These results are consistent with social science research on group decision-making and the conditions under which consensus should be of value.\(^{20}\) A key determinant of the benefit of consensus is the level of diversity of individual panelist judgments. When reviewing clinical scenarios exclusively, the trainee and expert panelists were evaluating a type of information familiar to them and to which they could apply their own idiosyncratic diagnostic experience in reaching their judgments. In contrast, the interpretation of FDG-PET images offered relatively little room for variation in interpretation. As a result, the panelists demonstrated higher interrater agreement when reviewing images than when reviewing the clinical scenario alone (Table). This lower diversity led to lower panel performance in terms of relative accuracy of consensus diagnoses compared with individual diagnoses (Figure 2) and in terms of lower number and lesser quality of diagnostic changes by panelists (eTables 2 and 3).

Thus, a critical issue for application of this modified Delphi protocol is to ensure that the panels have sufficient diversity. The selection of an appropriate panel requires identifying panelists who are likely to make different errors in judgment.\(^{20}\) Sources of such diversity include variation in clinical training, medical specialty, and experience with particular socioeconomic, ethnic, and racial groups. These factors are particularly important when relying on the rich variety of information provided by a detailed clinical history.

**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.
to organize. In situations where resources are limited, these results suggest some steps that could increase efficiency without major loss of diagnostic accuracy. For panels designed to accurately diagnose all patients in a study, the best protocol involves forced consensus after deliberation. But when the panelists’ initial judgments are unanimous or nearly so, simply adopting that position as the consensus judgment provides similar accuracy. Indeed, if the costs to conduct deliberation are particularly high, one might also consider lower majority thresholds applied to the panelists’ initial diagnoses. To the extent that the panel seeks to identify patients with highly accurate diagnoses and has little regard for the share of patients diagnosed, a high-threshold rule without discussion is appropriate.

It is important to note that panels also confer professional legitimacy that typically does not accompany an individual judgment. Thus, to the extent that the legitimacy and accuracy of the judgment are to a particular question, the cost of the modified Delphi protocol may well be justified, even if the improvement in judgment accuracy is modest.

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 respond 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 not 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 panelist 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. 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 Koeppe), 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 Caro...
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 Hershner, 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