Published in final edited form as:
Arch Neurol. 2010 December ; 67(12): 1506–1512. doi:10.1001/archneurol.2010.301.

Validation of Consensus Panel Diagnosis in Dementia

Matthew J. Gabel, PhD, **Norman L. Foster, MD**, **Judith L. Heidebrink, MD**, and **Roger Higdon, PhD for the Pilot PET Study Group***
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, Utah (Drs Foster and Zamrini); Departments of Neurology (Drs Barbas, Heidebrink, and Turner) and Radiology (Dr Koeppe), University of Michigan, Ann Arbor, Michigan; Neurology Service (Drs Heidebrink and Barbas) and GRECC (Dr Turner), Department of Veterans Affairs Medical Center, Ann Arbor, Michigan; Seattle Children's Research Institute, Seattle, Washington (Dr Higdon); Department of Psychiatry, University of Pittsburgh, Pittsburgh, Pennsylvania (Dr Aizenstein and formerly Dr DeKosky); School of Medicine, University of Virginia, Charlottesville, Virginia (Dr DeKosky); Departments of Psychiatry (Dr Arnold) and Neurology (Drs Arnold and Clark), University of Pennsylvania, Philadelphia, Pennslvania; Department of Neurology, Mayo Clinic, Rochester, Minnesota (Dr Boeve); Department of Neurology, Duke University, Durham, North Carolina (Dr Burke); Department of Neurology, Indiana University, Indianapolis, Indiana (Dr Farlow); Departments of Neuroscience and Public Health, University of California at Berkeley, Berkeley, California (Dr Jagust); Departments of Neurology and Neurobiology and Behavior, University of California at Irvine, Irvine, California (Dr Kawas); Veterans Affairs, Puget Sound Health Care System (Drs Leverenz and Peskind); Departments of Neurology (Dr Leverenz) and Psychiatry & Behavioral Sciences (Drs Leverenz and Peskind), University of Washington, Seattle, Washington; Departments of Neurology (Dr Womack and formerly Dr Lipton) and Psychiatry (Dr Womack), University of Texas Southwestern, Dallas, Texas; Department of Neurology, Texas Health Presbyterian Hospital, Dallas, Texas (Dr Lipton).

Abstract

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

Objective—To determine conditions that improve consensus panel diagnosis.

Correspondence to: Matthew J. Gabel, PhD Department of Political Science, Washington University One Brookings Drive, St. Louis, MO 63130-4899 Tel: (314) 935-6613, Fax: (314) 935-5856 mgabel@artsci.wustl.edu .

*Members of the Pilot PET Study Group: 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; Norman L. Foster, MD; Matthew J. Gabel, PhD; Judith L. Heidebrink, MD; Roger Higdon, PhD; 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.

Author Contributions: Drs Gabel and Foster had full access to all of the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis. *Study concept and design:* Foster, Gabel, Heidebrink, and Lipton. *Acquisition of data:* Aizenstein, Arnold, Barbas, Burke, Clark, DeKosky, Farlow, Foster, Gabel, Heidebrink, Jagust, Kawas, Leverenz, Lipton, Peskind, Turner, Womack, and Zamrini. *Analysis and interpretation of data:* Arnold, Boeve, DeKosky, Farlow, Foster, Gabel, Heidebrink, Higdon, Koeppe, and Lipton. *Drafting of the manuscript:* Gabel and Foster. Critical revision of the manuscript for important intellectual content: Aizenstein, Arnold, Barbas, Boeve, Burke, Clark, DeKosky, Farlow, Foster, Gabel, Heidebrink, Higdon, Jagust, Kawas, Koeppe, Leverenz, Lipton, Peskind, Turner, Womack, and Zamrini. *Statistical analysis:* Gabel, Higdon, and Koeppe. *Obtaining funding:* Foster and Turner. *Administrative, technical, and material support:* Arnold, Boeve, Burke, DeKosky, Farlow, Foster, and Heidebrink. *Study supervision:* DeKosky, Gabel, and Foster. *Member of consensus panel:* Aizenstein.

Financial Disclosure: None reported.

Design—Comparison of neuropathological diagnoses with individual and consensus panel diagnoses based on clinical summaries, FDG-PET scans, and summaries with scans.

Setting—Expert and trainee individual and consensus panel deliberations using a modified Delphi method in a pilot research study of the diagnostic utility of FDG-PET imaging.

Patients and Methods—Forty-five patients with pathologically confirmed Alzheimer's disease or frontotemporal dementia. 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 potential pitfalls of group decision-making.

Many dementing diseases lack distinctive physical findings or validated biomarkers, thus making accurate clinical diagnosis challenging. Clinicians often must reach a diagnosis based solely upon their judgment of informant history of variable quality and the relative prominence of deficits in specific cognitive domains. Since 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, the findings on a neuropathological examination continue to be the gold standard for determining the cause of a dementing illness.

The validity of research results depends upon accurate diagnosis. Recognizing the limitations of individual clinician diagnoses, research studies often use the consensus of a panel when histopathological information is unavailable.[1, 2] 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 surprisingly 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.[3, 4] In addition, theoretical and empirical studies of group decision-making indicate that, depending upon their composition and procedures, consensus panels may fail to achieve highly accurate decisions.[5] 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 wide variety of consensus panel goals, membership, 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, while others seek the best diagnosis for all patients, recognizing accuracy may be higher in some situations than others. Consensus panels also vary in their composition and organization. Members may include only a single specialty or be multidisciplinary. Some panels include individuals who have personally examined the patient with the intent of assuring 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.[5] Furthermore, panel rules to arrive 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 clinician judgment, while 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. [5]

The Delphi method of consensus is a formal and rigorous procedure incorporating organizational features that social science theory indicates promote accurate individual and group judgments.[5,6,7] This method is commonly employed to set professional priorities and establish guidelines, but the exact protocol can vary in size of the panel, the use of face-to-face discussion, and the number of iterations before a final decision is reached.[8,9,10] 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 insure independent judgments and diversity of opinions is encouraged through panel membership and during discussions.

We took advantage of an extraordinary opportunity to explore diagnostic performance of consensus panels provided by trials we conducted to examine the diagnostic utility of FDG-PET imaging.[11] Consensus panels generally are convened only when there is no "gold standard" available. In these trials, however, neuropathological findings were available and we undertook these studies to determine the extent to which consensus panel diagnoses 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's disease (AD) and frontotemporal dementia (FTD). Thus, it was scientifically appropriate in these trials to restrict diagnostic options to these two possibilities. The requirement of a binary decision was fortuitous because it significantly simplified 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 very complex and requires the consideration of multiple conditions, binary decisions are very relevant to clinical practice. For example, after an extensive dementia evaluation, critical diagnostic judgments often must choose only between two of the most likely possibilities, such as demented or non-demented, mild cognitive impairment or normal for age, AD or not AD, and AD or vascular dementia.

Methods

Two consensus panels, each composed of six panelists, and six 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. While panelists and raters were aware that patients had only one of two possible diagnoses, they did not knowthe 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, a psychiatry resident, and a 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 one of 4 NIA-funded Alzheimer Centers.

Raters

Distinct from the members of the panels, our study also involved 6 "raters": dementia-specialist neurologists, each with 10 to 25 years of experience in dementia care, two from each of three NIA-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 of individual experts to compare with panel diagnoses.

Patient Data

Clinical scenarios and positron emission tomography brain scans with 18F-fluorodeoxyglucose (FDG-PET) were evaluated from 45 patients with a postmortem examination documenting a histopathological diagnosis of AD (31) or FTD (14) uncomplicated by other pathology such as a stroke or significant number of cortical Lewy bodies. Foster et al.[11] provides a full description of the pathological 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 patient: clinical scenario alone, FDG-PET images alone, and scenarios with 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.[12] Each set of data was presented on a different day and in a different patient order to keep panelists blinded to their prior 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 level of confidence in that diagnosis (very confident, somewhat confident, or uncertain). The panel leader collected the cards and announced the "vote tally" (e.g., 3AD, 3FTD) 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[13-16], but we neither suggested nor imposed any rulesregarding the interpretation of the criteria or individual patient information.

Following discussion, panelists again marked a card in private indicating diagnosis and diagnostic confidence. After these cards were collected, the group was asked to arrive at a final diagnosis. The panelists were not provided a decision rule (e.g. simple majority), but were told they needed to return a decision for the panel. The leader then recorded the consensus decision and the panel turned to the next subject 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 with their level of confidence. In all, there were a total of 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 to the neuropathological diagnoses, our reference standard. For each panel, we computed statistics for sensitivity, specificity, predictive value, and likelihood ratio. With only two diagnostic options, positive and negative predictive values were complementary, and sensitivity and specificity for FTD were reciprocal to that for AD. We used Kappa 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 (0.2-0.39), moderate (0.4-0.59), substantial (0.6-0.79), or almost perfect (0.8-1.0), according to convention.[17] We analyzed consensus panel performance relative to 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 pre- to post-discussion in each panel was analyzed using logistic regression models fit to a binary response variable for whether the pre- or post-diagnosis was correct and included the timing (pre or post) of the diagnosis as a covariate. The change in diagnostic confidence from pre- to post-discussion was evaluated in a similar manner, fitting the model to a binary variable for whether the panelist was “very confident.” To determine the extent changes in panelists’ diagnoses were beneficial, we estimated logistic regression models for all panelists who changed their confidence or diagnosis from pre- to post-discussion. 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 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 pre/post 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.[18] We then used the adjusted variance estimate to generate corrected *p*-values. Also, where relevant, *p*-values were adjusted for multiple tests with the Hochberg correction.[19] McNemar’s chi-squared tests were used to assess whether consensus diagnoses were more accurate than alternative methods of group diagnosis (e.g., simple majority rule).

Results

Reliability, accuracy, and confidence of diagnosis

The accuracy of consensus diagnoses of both trainee and expert panels were superior to the individual diagnoses of their own members when considering clinical scenarios (figure 1A). Panel consensus diagnoses also were superior to those of expert raters making individual judgments (figure 1B). Consensus diagnoses were more accurate than diagnoses of 9 of the 11 individual panelists and 5 of the 6 individual expert raters, and these differences often reached significance. On average, the 12 experts individually performed better than the 5 trainee panelists, although after deliberation, both trainee and expert panels had the same diagnostic accuracy. Indeed, the trainee panel was significantly more accurate than the individual opinions of 3 of the 6 expert panelists (eFigure 1).

Individual diagnostic accuracy and confidence were high with review of FDG-PET images either 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 than 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. AD sensitivity and FTD specificity (89%-94%) were higher than AD specificity and FTD sensitivity (71% - 86%). (etable 1). As expected from previous experience, diagnostic accuracy of individuals and panels was less when considering FTD than with AD. Despite the concerns of others[3], the consensus judgments were highly reproducible across panels (two-way Kappa=0.68-0.90) in spite of differences in panel membership and diagnostic information reviewed (figure 3).

Panelists' judgments tended to converge after discussion in all situations, as indicated by the increase in average Kappa agreement scores within panels, and diagnostic confidence also increased (table 1). This increase in agreement after deliberation was not uniformly associated with beneficial changes in diagnosis or confidence, however (eTable 2). Like 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. But, compared to reviewing scenarios alone, these changes were far fewer in number and were typically not beneficial (eTable 3).

Effect of Panel Consensus Rules on Diagnostic Accuracy

Following discussion and the second vote the panel was asked to determine a single final consensus diagnosis. When 5/6 or 6/6 panelists agreed on a pre-discussion 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 4/6 to unanimity, accuracy generally improves, though gains are small and at the expense of many patients going undiagnosed (eTable 4). Re-voting after discussion allowed more patients to be diagnosed and by a larger majority. None of the alternative rules exhibited a statistically significant higher accuracy than the forced consensus rule (Etable 4).

In general, discussion caused panelists to converge around the pre-discussion majority diagnosis, regardless of whether that diagnosis was correct or incorrect. The only exceptions were three 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 pre-discussion into a majority incorrect diagnosis post-discussion. As a result, the forced consensus rule and the post-discussion 4/6-majority rule for final diagnosis differed only in that the forced consensus rule yielded a diagnosis for the 6 cases across all panels with a 3-3 split post-discussion. In these six cases, the panel was correct three times.

Panel Deliberations

The duration of panel discussions varied considerably from case to case. Trainee panel discussions of scenarios (mean 5 minutes, range 0-15) were remarkably similar in length to expert panel discussions of the same information (mean 4 minutes, range 1-15). The time expended on discussions involving images was substantially less (expert panel discussions of images alone, mean 2 minutes, range 0-9, and images with scenario, mean 2 minutes, range 0-7).

Discussion

The modified-Delphi protocol resulted in reliable consensus diagnoses across panels of varying expertise and diagnostic information. Expertise of individuals does not negate the benefit of consensus; consensus improved the accuracy of both non-expert 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 histopathological information is unavailable.

In contrast, consensus diagnoses when reviewing FDG-PET images, either 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 inter-rater agreement when reviewing images than when reviewing the clinical scenario alone (table 1). This lower diversity led to lower panel performance both in terms of relative accuracy of consensus diagnoses compared to 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 the application of our 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, or experience with particular socio-economic, 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 consensus procedures currently in use in dementia research. Other consensus procedures may not provide similar positive results as 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 pre-test probability of a single diagnosis. In this situation, diagnostic errors will change autopsy confirmation rates only slightly. On the other hand, in this study, pre-test probability of FTD was unknown to the raters, but considerably higher than in many AD research studies and thus provided an informative setting to assess consensus.

Properly constituted consensus panels are time consuming, expensive, and require considerable effort to organize. In situations where resources are limited, our 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 both the legitimacy and accuracy of the judgment is important 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 problematicbecause 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 wide variety of consensus panels, our 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 number of examinations, the detail and length of the medical record, and quality of the medical history. While 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 our study. In contrast, prospectively collected comprehensive longitudinal data would probably produce less error since diagnostic accuracy improves with longitudinal information.[21,22] 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 providing 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.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank David E. Kuhl, Sid Gilman, Henry Buchtel, David Knesper, R. Scott Turner and Kirk Frey for making images from their research available for this study; Dr. Charles DeCarli for his contributions as a site investigator for NIH grant AG22394; and Peijun Chen, Charles Davies, Shelley Hershner, and Joseph O. Nnodim for serving on the pilot trainee panel. Jeff Gill, Ryan Moore, and Diana O’Brien provided valuable statistical advice.

Funding/Support: This work was supported by NIH grant AG22394, an anonymous private donation to the Center for Alzheimer’s Care, Imaging and Research, a pilot cooperative project grant from the National Alzheimer’s

Coordinating Center (AG16976) and by the following NIH Alzheimer's Disease Research Centers: Michigan (AG08671), University of California at Davis (AG10129), University of Pennsylvania (AG10124), University of California at Irvine (AG16573), Duke University (AG01328), Indiana University (AG10133), University of Pittsburgh (AG05133), and University of Texas Southwestern (AG12300).

References

1. Ott A, Stolk RP, van Harskamp F, Pols HA, Hofman A, Breteler MM. Diabetes mellitus and the risk of dementia: The Rotterdam Study. Neurology. 1999; 53(9):1937–1942. [PubMed: 10599761]
2. Lopez OL, Becker JT, Klunk W, et al. Research evaluation and diagnosis of probable Alzheimer's disease over the last two decades: I. Neurology. 2000; 55(12):1854–1862. [PubMed: 11134385]
3. Shekelle PG, Kahan JP, Bernstein SJ, Leape LL, Kamberg CJ, Park RE. The reproducibility of a method to identify the overuse and underuse of medical procedures. N Engl J Med. 1998; 338(26): 1888–1895. [PubMed: 9637810]
4. Huttin C. The use of clinical guidelines to improve medical practice: main issues in the United States. Int J Qual Health Care. 1997; 9(3):207–214. [PubMed: 9209918]
5. Gabel MJ, Shipan CR. A social choice approach to expert consensus panels. J Health Econ. 2004; 23(3):543–564. [PubMed: 15120470]
6. Jones J, Hunter D. Consensus methods for medical and health services research. BMJ. 1995; 311(7001):376–380. [PubMed: 7640549]
7. Robert G, Milne R. A Delphi study to establish national cost-effectiveness research priorities for positron emission tomography. Eur J Radiol. 1999; 30(1):54–60. [PubMed: 10389013]
8. Fick DM, Cooper JW, Wade WE, Waller JL, Maclean JR, Beers MH. Updating the Beers criteria for potentially inappropriate medication use in older adults: results of a US consensus panel of experts. Arch Intern Med. 2003; 163(22):2716–2724. [PubMed: 14662625]
9. Drasković I, Vernooij-Dassen M, Verhey F, Scheltens P, Rikkert MO. Development of quality indicators for memory clinics. Int J Geriatr Psychiatry. 2008; 23(2):119–128. [PubMed: 17582827]
10. Rikkert, MG Olde; van der Vorm, A.; Burns, A., et al. Consensus statement on genetic research in dementia. Am J Alzheimers Dis Other Demen. 2008; 23(3):262–266. [PubMed: 18509105]
11. Foster NL, Heidebrink JL, Clark CM, et al. FDG–PET improves accuracy in distinguishing frontotemporal dementia and Alzheimer's disease. Brain. 2007; 130(Pt 10):2616–2635. [PubMed: 17704526]
12. Brook RH, Chassin MR, Fink A, Solomon DH, Kosecoff J, Park RE. A method for the detailed assessment of the appropriateness of medical technologies. Int J Technol Assess Health Care. 1986; 2(1):53–63. [PubMed: 10300718]
13. McKhann G, Drachman D, Folstein M, Katzman R, Price D, Stadlan EM. Clinical diagnosis of Alzheimer's disease: report of the NINCDS-ADRDA Work Group under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease. Neurology. 1984; 34(7):939–944. [PubMed: 6610841]
14. Neary D, Snowden JS, Gustafson L, et al. Frontotemporal lobar degeneration: a consensus on clinical diagnostic criteria. Neurology. 1998; 51(6):1546–1554. [PubMed: 9855500]
15. The Lund and Manchester Groups. Clinical and neuropathological criteria for frontotemporal dementia. J Neurol Neurosurg Psychiatry. 1994; 57(4):416–418. [PubMed: 8163988]
16. McKhann GM, Albert MS, Grossman M, et al. Clinical and pathological diagnosis of frontotemporal dementia: report of the Work Group on Frontotemporal Dementia and Pick's Disease. Arch Neurol. 2001; 58(11):1803–1809. [PubMed: 11708987]
17. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977; 33(1):159–74. [PubMed: 843571]
18. Andrews DWK. Heteroskedasticity and autocorrelation consistent covariance matrix estimation. Econometrica. 1991; 59(3):817–858.
19. Hochberg Y, Benjamini Y. More Powerful Procedures for Significance Testing. Stat Med. 1990; 9(7):811–818. [PubMed: 2218183]
20. Page, SE. The Difference. Princeton University Press; Princeton, NJ: 2007.

21. Becker JT, Boller F, Lopez OL, Saxton J, McGonigle KL. The natural history of Alzheimer's disease. Description of study cohort and accuracy of diagnosis. Arch Neurol. 1994; 51(6):585–594. [PubMed: 8198470]

22. Litvan I, Agid Y, Sastry N, et al. What are the obstacles for an accurate clinical diagnosis of Pick's disease? A cliniopathologic study. Neurology. 1997; 49(1):62–69. [PubMed: 9222171]

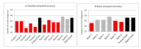

Figure 1.
Comparison of diagnostic accuracy of expert raters, and panelists with expert panels reviewing clinical scenarios
Consensus panel diagnoses (in black) based on scenarios were more accurate than diagnoses arrived at by 10 of 11 individual panelists and 5 of 6 raters. This panel superiority was statistically significant (in red) for all members of the trainee panel, 4 of members of the expert panel, and 3 of the raters ($p<.05$, Hochberg corrected). Note that one member of the trainee panel did not review 17 cases and was thus omitted from this analysis.

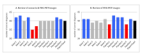

Figure 2.

Comparison of diagnostic accuracy of expert raters and panelists with panels reviewing FDG-PET images

Consensus panel diagnoses (in black) based on images alone or with clinical scenarios were more accurate than diagnoses of 0 of 6 panelists and 2 of 6 raters when reviewing images and scenarios (A) and 2 of 6 panelists and 2 of 6 raters when reviewing images alone (B). Two panelists and two raters (in red) had a statistically significant lower accuracy than each panel ($p<.05$, Hochberg corrected). In contrast, across the two panels, 6 of 12 panelists and 5 of 12 raters (in blue) had a statistically significant greater accuracy than the panel diagnoses ($p<.,05$, Hochberg corrected).

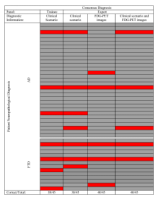

Figure 3.

Panel diagnostic accuracy by patient

Each horizontal line represents a single patient. Panel diagnoses in agreement with neuropathological diagnoses are shown in gray. Panel diagnostic errors are in red. Panels often were in error in the same patients. The pairwise Kappa agreement between diagnoses of trainee and expert panel for scenarios was 0.79 (±0.15 SE); trainee (scenario) and expert (images) panels was 0.68 (±.15 SE); trainee (scenario) and expert (scenario+images) was .79 (±.15 SE); expert (scenario) and expert (images) panels was .69 (±.15 SE); expert (images) and expert (scenario+images) panels was .79 (±.15 SE).

Table 1

Individual panelist accuracy, confidence, and agreement before and after panel deliberation

Panel	Diagnostic information	Mean panelist accuracy (range)		Mean Kappa (standard error)		% Very confident (range)	
		Pre-discussion	Post-discussion	Pre-discussion	Post-discussion	Pre-discussion	Post-discussion
Trainee*	Scenario	71% (64-78)	79% (73-82)	.45 (.08)	.73 (.07)	43% (4-76)	60% (33-73)
Expert	Scenario	79% (71-87)	82% (78-84)	.61 (.07)	.83 (.06)	46% (11-80)	58% (22-78)
Expert	Images	90% (84-96)	90% (87-93)	.81 (.03)	.91 (.05)	64% (20-87)	72% (44-87)
Expert	Scenario & Images	90% (89-93)	89% (87-91)	.84 (.07)	.97 (.02)	56% (13-93)	67% (20-87)

*Adjusted for 17 cases not reviewed by one panelist.

Panelists (trainee and expert) reviewing scenarios alone were more likely to increase than decrease their diagnostic accuracy (odds ratio=1.56, p<.01 for trainee; odds ratio=1.26, p=.03) for expert)). They were also more likely to increase their diagnostic confidence (odds ratio=1.85, p<.01 for trainee; odds ratio=1.61, p<.01 for expert) after deliberation. In contrast, panelists reviewing scenario and images or images alone were no more likely to increase than decrease their diagnostic accuracy (odds ratio=.87, p=.16 for scenario and images; odds ratio=1.0, p=1.0 for images) after deliberation. However, these panelists were more likely to increase their diagnostic confidence (odds ratio=1.25, p=.04 for scenario and images; odds ratio=1.45, p<.01 for images only). The mean two-way inter-panelist Kappa increased with deliberation on all panels (p<.01, difference in means test). Standard errors for Kappa were calculated using the bias-corrected bootstrap method.

Arch Neurol. Author manuscript; available in PMC 2011 December 1.