PEER REVIEW HISTORY

BMJ Open publishes all reviews undertaken for accepted manuscripts. Reviewers are asked to complete a checklist review form (see an example) and are provided with free text boxes to elaborate on their assessment. These free text comments are reproduced below. Some articles will have been accepted based in part or entirely on reviews undertaken for other BMJ Group journals. These will be reproduced where possible.

ARTICLE DETAILS

TITLE (PROVISIONAL)	Specificity and sensitivity of transcranial sonography of the substantia nigra in the diagnosis of Parkinson's disease: prospective cohort study in 196 patients
AUTHORS	Weber, Wim; Bouwmans, Angela; Vlaar, Annemarie; Mess, Werner; Kessels, Alfons

VERSION 1 - REVIEW

REVIEWER	Ron Postuma Associate Professor of Neurology McGill University Montreal, Quebec, Canada I have nothing to disclose in relation to this review
REVIEW RETURNED	24-Jan-2013

THE STUDY	Only the evaluation of SPECT could be reconsidered according to the 'Are the patients representative
GENERAL COMMENTS	This study deals with an important area (diagnosis of PD), using a promising technique (transcranial ultrasound, and to a lesser extent, SPECT). The results are, unfortunately negative. Negative results of this nature on a critical question are extremely important, and need to be communicated with the broader community. Whenever there is a negative study, one can easily generate a long list of potential biases, errors, etc. that can explain the absence of differences. An important strength of this study is that the authors clearly designed the study carefully to minimize sources of bias (blinding, diagnostic decisions, etc.). Nonetheless, some of these are very critical issues (e.g. any diagnostic error would almost inevitably bias towards an appearance of poor sensitivity/specificity), and it is still possible that they could explain some of the negative findings. The 40% dropout between visits, the absence of a diagnostic gold standard, and potential technician-dependence, although impossible to eliminate in any similar study design, are the most critical concerns. Suggestions for improvement: SPECT scanning in MSA, DLB and PSP is usually abnormal and can be indistinguishable from PD. However, it is usually normal in non-parkinsonian conditions. It would be important to analyze SPECT scanning separately for those diagnoses that are relevant to how it would eventually be applied. Few would suggest using CIT-SPECT diagnostically to distinguish PD from MSA, DLB, CBD or PSP; the absence of diagnostic utility at distinguishing APS from PD should be documented. However, the SPECT performed much better at separating PD from non-parkinsonian conditions (ET, DIP,

	etc.) - this therefore warrants a separate analysis (perhaps even the primary analysis). It appears that the 0.20 cm cutoff for abnormal TCS was predetermined. Ultimately for a test to be practically useful, I agree that one needs a prespecified cutoff that is constant across different environments, technicians, etc. However, one concern might be that differences in machines/techniques, etc. might change the ideal cutoff. (note that the cutoff chosen represents the 75th percentile in healthy controls - this essentially guarantees a specificity lower than 75%). This is why the ROC curve to find the best-performing cutoff was important. Can the authors provide the best sensitivity/specificity estimates that they were able to obtain at the best-performing cutoff? The statement in the introduction that the disorders 'demand vastly differing therapies' seems overstated. Initial treatment of most APS is in fact rather quite similar to PD (motor treatment with levodopa and non-motor treatment as symptoms emerge). I would suggest qualifying or softening this statement. Several points that do not require revision, but might be useful to emphasize in the discussion: This study considered diagnosis 'not established' in 199/241 patients - therefore, it really includes the majority of patients with early disease. More selective inclusion of those for whom the diagnosis would really be a 'toss-up' may provide different results (perhaps even lower sensitivity/specificity). The utility of TCS is obviously technician-dependent, a point noted by the authors. It does appear that the investigators were well-trained. As discussed, TCS error can still be an important source of bias that would reduce sensitivity and specificity estimates. On the other hand, the fact that the investigators are so well-trained may imply that real-world utility would be even lower than found. Just to emphasize a point that was appropriately dealt with by the authors, it is also likely that many DLB patients will have abnormal TCS and dopaminergic functional imaging. There is considerable debate about the utility of the distinction between DLB and PD - both are synucleinopathies, both have extremely similar prodromal and non-motor features, and both clinically overlap. Therefore, it may be inappropriate to consider the diagnosis of DLB instead of PD to be a diagnostic error - the sensitivity analysis excluding these patients is worthy of emphasis.

REVIEWER	Kristina Laučkaitė, MD, PhD, Clinic of Neurology, Academy of Medicine, Lithuanian University of Health Sciences, Lithuania. I declare no competing interests.
REVIEW RETURNED	04-Feb-2013

THE STUDY	A few English grammar mistakes should be corrected: • “Twenty-one (10,7%) patients had a diagnosis of vascular

	parkinsonism (VP), 20 (10,2%) essential tremor (ET), 7 (3,6%) drug-induced parkinsonism (DIP) and 22 (11,2%) patients had no parkinsonism but a alternative diagnosis"- Should be corrected as <…..> an alternative diagnosis. • „One investigator per hospital and blinded to clinical information, did the ultrasounds using a SONOS 5500 (Philips, Eindhoven, The Netherlands)"- ultrasound imaging (sonography) (page 9). • „M. Alzheimer" (page 11)- Alzheimer's disease. • Everywhere the commas should be replaced by the dots, e.g. 10,7%- should be corrected as 10.7%, etc. • (Confidence Interval (CI) 0,30-0,50)- should be inserted 95% Confidence Interval everywhere in the text. As well it should be advisable to add both descriptive statistical methods, also comparative and correlative criteria used into Statistics section, not only ROC analysis. • "In this study FP-CIT (123I-ioflupane, Nycomed, Amersham, U.K.) is used as presynaptic radiotracer"(page 9)- suggested to use past tense everywhere in the text, i.e. to replace was used. As well as e.g. "The mean area of the ROC curve is 0,541"- was 0.541. • „176 patients also underwent a FPCIT-SPECT at initial work-up, around the same time when they underwent a TCD, see table 3"(page 14)- abbreviations should be checked, as authors decided to use TCS. • "We also found no significant difference for the maximum size or the sum of the area of the SN+ in the different diagnoses (see table 3, figure 1 and 2)"- maybe Table 2? • „TCS findings were concordant with SPECT findings in 89 of 176 patients (p 0,36)"- p=0.36? • „At inclusion we excluded the patients with already a clear diagnosis which has not been done in the study of Gaenslen" (page 16)- to insert reference, also everywhere further in the text. I am not a native English speaking person either, but the general written English style seems easy readable and acceptable to me.
RESULTS & CONCLUSIONS	I would have only two notes about the data presented in the tables: 1) Table 1 (page 12)- a column of a p value could be inserted or given below the table in exact numbers and statistical criteria used. In case of not normally distributed variables (non-parametric), would be better descriptive statistics to give as a median and interquartile range, or mean and SE, instead of mean and SD, as e.g. mean duration complaints in months 29.8+/-41.7 the value with -1SD becomes negative. 2) Table 2 (page 13) and Table 3 (page 14)- advisable to give percentages too, not only absolute numbers. The tables could be formatted in the same style as Table 1, i.e. all patients column given firstly and then all other groups, or vice versa. In Table 2, in order to simplify it, it could be used instead "No presence of hyperechogenic SN 0,20 cm2 or more"- to replace simple as <0.20 cm2 or the SN-, and „Presence of hyperechogenic SN 0,20 cm2 or more"- to replace

	>=0.20 cm2 or the SN+, and above, in an empty cell of the table to insert- the SN plot. In Table 3 an empty cell should be described and filled in too. Other question about descriptions in all the tables, as ET is also “No parkinsonism” category. Maybe to name the latter group “Other diseases” or “Alternative diagnoses”?
GENERAL COMMENTS	This study brings TCS from the idealistic to more realistic approach. Another very important aspect of the study, that a large number of patients underwent expensive functional radionuclide imaging, which is the most sensitive and specific diagnostic method until now for parkinsonian disorders. There are only a few other notes: 1) In Abstract section (page 2), “After two years, patients were re-examined by two movement disorder specialist neurologists for a final clinical diagnosis, that served as a surrogate gold standard for our study“- should be inserted the number of prospectively re-examined patients, i.e. (n=114). Also „Within two weeks of inclusion all patients also underwent a TCS and a 123I-ioflupane Single Photon Emission Computer Tomography (FP-CIT SPECT) scan of the brain“- should be included the number of patients to whom SPECT was performed, as not all underwent, i.e. (n=176). 2) Insufficiency of bilateral temporal acoustic windows accounted for 18.67% (45/241), which should included into abstract (if still possible due to the limits of the signs), also in Discussion section. The majority of authors usually described a much lower percentage, except in the Eastern (Asia) populations. This adds more additional valuable information about TCS limitations. 3) Introduction section (page 6) “We have now assessed the diagnostic accuracy of TCS of the SN in 196 patients referred by their general practitioner (GP) for analysis of a parkinsonian syndrome of recent onset. We used a clinical diagnosis after two years as a surrogate gold standard and also compared TCS with FP-CIT-SPECT scans“, then in Methods Patients section (page 7) “This was a prospective study testing the diagnostic accuracy of TCS of the SN in patients who are referred by their GP for a first consultation by a neurologist because of recent-onset parkinsonism of unclear origin [52]”- advisable to shift the sentence from Introduction section to Methods avoiding repeat.

REVIEWER	Nicola Pavese, MD PhD Clinical Senior Lecturer and Consultant in Neurology Imperial College London Associate Professor Aarhus University I have no competing interest to disclose
REVIEW RETURNED	09-Feb-2013

THE STUDY	Results of TS in a group of established PD patients and age-matched controls are missing in the study. A repeat TS examination at follow-up would have improved the data. The English is poor and need major revision
GENERAL COMMENTS	In this large prospective study, the authors assessed specificity and sensitivity of transcranial sonography (TS) of the substantia nigra (SN) and FP-CIT SPECT for the diagnosis of Parkinson's disease in 196 patients with clinically unclear parkinsonism. The surrogate gold standard for the study was the final diagnosis at two-year follow-up performed by two neurologists expert in movement disorders blind to the results of TS and SPECT. The presence of SN hyperechogenicity at baseline showed a 0.40 sensitivity and 0.61 specificity for the diagnosis of PD, whereas sensitivity and specificity for FP-CIT SPECT were 0.88 and 0.68, respectively. The results of the study are interesting and they would suggest that sensitivity and specificity of TS for the diagnosis of PD are much lower than previously reported. The study was well designed but they should have repeated TS at follow-up. This would have provided more information, particularly in the group of definite PD at follow-up. Results of TS in a group of established PD patients and age-matched controls are also missing in the study. They state that TS findings were concordant with SPECT findings in 89 of 176 patients. It would be interesting to expand this information, possibly in a table. It would also be interesting to know sensitivity and specificity of combined positive TS and SPECT scan for the final diagnosis. In the discussion, the sentence ' we found no significant correlation …. " is not appropriate as no formal correlation were performed. The English needs a major revision. Many sentences do not read well. What do they mean with the sentence " Apart from the real gold standard of the post-mortem examination, which seems lees feasible in modern times, we think that this gold standard diagnosis of PD is methodologically the highest achievable one" ?

VERSION 1 – AUTHOR RESPONSE

Reviewer: Ron Postuma
Associate Professor of Neurology
McGill University
Montreal, Quebec, Canada

I have nothing to disclose in relation to this review

Only the evaluation of SPECT could be reconsidered according to the 'Are the patients representative. . . ' question - see my specific comments

This study deals with an important area (diagnosis of PD), using a promising technique (transcranial ultrasound, and to a lesser extent, SPECT). The results are, unfortunately negative. Negative results of this nature on a critical question are extremely important, and need to be communicated with the broader community.

We agree fully with this statement.

Whenever there is a negative study, one can easily generate a long list of potential biases, errors, etc. that can explain the absence of differences. An important strength of this study is that the authors clearly designed the study carefully to minimize sources of bias (blinding, diagnostic decisions, etc.). Nonetheless, some of these are very critical issues (e.g. any diagnostic error would almost inevitably bias towards an appearance of poor sensitivity/specificity), and it is still possible that they could explain some of the negative findings. The 40% dropout between visits, the absence of a diagnostic gold standard, and potential technician-dependence, although impossible to eliminate in any similar study design, are the most critical concerns.

These are good points, but biases may also play a role in positive studies.

Suggestions for improvement:

SPECT scanning in MSA, DLB and PSP is usually abnormal and can be indistinguishable from PD. However, it is usually normal in non-parkinsonian conditions. It would be important to analyze SPECT scanning separately for those diagnoses that are relevant to how it would eventually be applied. Few would suggest using CIT-SPECT diagnostically to distinguish PD from MSA, DLB, CBD or PSP; the absence of diagnostic utility at distinguishing APS from PD should be documented. However, the SPECT performed much better at separating PD from non-parkinsonian conditions (ET, DIP, etc.) - this therefore warrants a separate analysis (perhaps even the primary analysis).

The reviewer asks for an additional analysis, looking at TCS's discriminatory power to discern PD from non-parkinsonian (APS) syndromes. This is a good point, and we have done that. We have thus grouped all IPD diagnoses together with all APS diagnoses vs. the rest, i.e. ET, DIP, VP, etc. we then found similar specificity and sensisitivity of TCS, respectively 0.67 and 0.43. In this analysis sensitivity of SPECT remains 0.84 and specificity increases to 0.84. We have added these data to the Results section on p.15 , line 8.

It appears that the 0.20 cm cutoff for abnormal TCS was predetermined. Ultimately for a test to be practically useful, I agree that one needs a prespecified cutoff that is constant across different environments, technicians, etc. However, one concern might be that differences in machines/techniques, etc. might change the ideal cutoff. (note that the cutoff chosen represents the 75th percentile in healthy controls - this essentially guarantees a specificity lower than 75%). This is why the ROC curve to find the best-performing cutoff was important. Can the authors provide the best sensitivity/specificity estimates that they were able to obtain at the best-performing cutoff?

This is also a good point, and we had already looked into that. We were not able to obtain better diagnostic discrimination with other TCS cut-offs. E.g., when we lowered the sensitivity threshold to an absolute minimum of 0.7, we obtained a cut-off of 0,3 cm2, but then specificity was 0.29. We have added this information to the Results section on p.13, line 36

The statement in the introduction that the disorders 'demand vastly differing therapies' seems overstated. Initial treatment of most APS is in fact rather quite similar to PD (motor treatment with levodopa and non-motor treatment as symptoms emerge). I would suggest qualifying or softening this statement.

We have corrected this: p.5, line 8.

Several points that do not require revision, but might be useful to emphasize in the discussion:

This study considered diagnosis 'not established' in 199/241 patients - therefore, it really includes the majority of patients with early disease. More selective inclusion of those for whom the diagnosis would really be a 'toss-up' may provide different results (perhaps even lower sensitivity/specificity).

We have added this consideration to the Discussion, p.17 , line 15.

The utility of TCS is obviously technician-dependent, a point noted by the authors. It does appear that the investigators were well-trained. As discussed, TCS error can still be an important source of bias that would reduce sensitivity and specificity estimates. On the other hand, the fact that the investigators are so well-trained may imply that real-world utility would be even lower than found.

We have added this consideration to the Discussion, p.18 , line 11.

Just to emphasize a point that was appropriately dealt with by the authors, it is also likely that many DLB patients will have abnormal TCS and dopaminergic functional imaging. There is considerable debate about the utility of the distinction between DLB and PD - both are synucleinopathies, both have extremely similar prodromal and non-motor features, and both clinically overlap. Therefore, it may be inappropriate to consider the diagnosis of DLB instead of PD to be a diagnostic error - the sensitivity analysis excluding these patients is worthy of emphasis.

We have added this consideration to the Discussion, p.14 , line 28.

Reviewer: Kristina Laučkaitė, MD, PhD, Clinic of Neurology, Academy of Medicine, Lithuanian University of Health Sciences, Lithuania.

I declare no competing interests.

A few English grammar mistakes should be corrected:

We have corrected all the following grammar errors:

• “Twenty-one (10,7%) patients had a diagnosis of vascular parkinsonism (VP), 20 (10,2%) essential tremor (ET), 7 (3,6%) drug-induced parkinsonism (DIP) and 22 (11,2%) patients had no parkinsonism but a alternative diagnosis”- Should be corrected as <…..> an alternative diagnosis.

• „One investigator per hospital and blinded to clinical information, did the ultrasounds using a SONOS 5500 (Philips, Eindhoven, The Netherlands)“- ultrasound imaging (sonography) (page 9).

• „M. Alzheimer“ (page 11)- Alzheimer‘s disease.

• Everywhere the commas should be replaced by the dots, e.g. 10,7%- should be corrected as 10.7%, etc.

• (Confidence Interval (CI) 0,30-0,50)- should be inserted 95% Confidence Interval everywhere in the text. As well it should be advisable to add both descriptive statistical methods, also comparative and correlative criteria used into Statistics section, not only ROC analysis.

• "In this study FP-CIT (123I-ioflupane, Nycomed, Amersham, U.K.) is used as presynaptic radiotracer"(page 9)- suggested to use past tense everywhere in the text, i.e. to replace was used. As well as e.g. "The mean area of the ROC curve is 0,541"- was 0.541.

• „176 patients also underwent a FPCIT-SPECT at initial work-up, around the same time when they underwent a TCD, see table 3"(page 14)- abbreviations should be checked, as authors decided to use TCS.

• "We also found no significant difference for the maximum size or the sum of the area of the SN+ in the different diagnoses (see table 3, figure 1 and 2)"- maybe Table 2?

• „TCS findings were concordant with SPECT findings in 89 of 176 patients (p 0,36)"- p=0.36?

• „At inclusion we excluded the patients with already a clear diagnosis which has not been done in the study of Gaenslen" (page 16)- to insert reference, also everywhere further in the text.

I am not a native English speaking person either, but the general written English style seems easy readable and acceptable to me.
Thank you.

I would have only two notes about the data presented in the tables:

1) Table 1 (page 12)- a column of a p value could be inserted or given below the table in exact numbers and statistical criteria used. In case of not normally distributed variables (non-parametric), would be better descriptive statistics to give as a median and interquartile range, or mean and SE, instead of mean and SD, as e.g. mean duration complaints in months 29.8+/-41.7 the value with -1SD becomes negative.

We understand the point, as these are not normally distributed, so we have added p-values in the text under the Table, where significant. We would prefer not to add p-values to every cell in the table, as that will affect readability. If you would want that, wew ill do so, of course.

2) Table 2 (page 13) and Table 3 (page 14)- advisable to give percentages too, not only absolute numbers. The tables could be formatted in the same style as Table 1, i.e. all patients column given firstly and then all other groups, or vice versa. In Table 2, in order to simplify it, it could be used instead "No presence of hyperechogenic SN 0,20 cm2 or more"- to replace simple as <0.20 cm2 or the SN-, and „Presence of hyperechogenic SN 0,20 cm2 or more"- to replace >=0.20 cm2 or the SN+, and above, in an empty cell of the table to insert- the SN plot. In Table 3 an empty cell should be described and filled in too. Other question about descriptions in all the tables, as ET is also "No parkinsonism" category. Maybe to name the latter group "Other diseases" or "Alternative diagnoses"?

We have changed both Tables as requested.

This study brings TCS from the idealistic to more realistic approach. Another very important aspect of the study, that a large number of patients underwent expensive functional radionuclide imaging, which is the most sensitive and specific diagnostic method until now for parkinsonian disorders.

Thank you for the compliment.

There are only a few other notes:

1) In Abstract section (page 2), “After two years, patients were re-examined by two movement disorder specialist neurologists for a final clinical diagnosis, that served as a surrogate gold standard for our study“- should be inserted the number of prospectively re-examined patients, i.e. (n=114). Also „Within two weeks of inclusion all patients also underwent a TCS and a 123I-ioflupane Single Photon Emission Computer Tomography (FP-CIT SPECT) scan of the brain“- should be included the number of patients to whom SPECT was performed, as not all underwent, i.e. (n=176).
We added this information in the Abstract

2) Insufficiency of bilateral temporal acoustic windows accounted for 18.67% (45/241), which should included into abstract (if still possible due to the limits of the signs), also in Discussion section. The majority of authors usually described a much lower percentage, except in the Eastern (Asia) populations. This adds more additional valuable information about TCS limitations.

We added this information in the Abstract, and p.18, line 13.

3) Introduction section (page 6) “We have now assessed the diagnostic accuracy of TCS of the SN in 196 patients referred by their general practitioner (GP) for analysis of a parkinsonian syndrome of recent onset. We used a clinical diagnosis after two years as a surrogate gold standard and also compared TCS with FP-CIT-SPECT scans“, then in Methods Patients section (page 7) “This was a prospective study testing the diagnostic accuracy of TCS of the SN in patients who are referred by their GP for a first consultation by a neurologist because of recent-onset parkinsonism of unclear origin [52]”- advisable to shift the sentence from Introduction section to Methods avoiding repeat.
We have deleted the sentence form the Introduction

Reviewer: Nicola Pavese, MD PhD
Clinical Senior Lecturer and Consultant in Neurology
Imperial College London

Associate Professor
Aarhus University

I have no competing interest to disclose

Results of TS in a group of established PD patients and age-matched controls are missing in the study.
A repeat TS examination at follow-up would have improved the data.

The reviewer asks for TCS data in a group of established PD patients and age-matched controls. These data would have been interesting, but obtaining those was not the purpose of this study. The point is that we already know that TCS is different from PD patients and controls. The problem is that

we do not know if this difference is big/ reliable enough, so that it may be used as a diagnostic tool. This study is a diagnostic accuracy study, in which we compared the TCS with a gold standard of follow-up clinical examination. We also do not see how a repeat TCS would have improved the data: we wanted to assess how accurate a TCS is in a patient with an unknown parkinsonian disorder of recent onset. (If you already have a definite clinical diagnosis after 2 years, what would a second TCS add?)

The English is poor and need major revision

We have been through the manuscript once again and corrected the English.

In this large prospective study, the authors assessed specificity and sensitivity of transcranial sonography (TS) of the substantia nigra (SN) and FP-CIT SPECT for the diagnosis of Parkinson's disease in 196 patients with clinically unclear parkinsonism. The surrogate gold standard for the study was the final diagnosis at two-year follow-up performed by two neurologists expert in movement disorders blind to the results of TS and SPECT.

The presence of SN hyperechogenicity at baseline showed a 0.40 sensitivity and 0.61 specificity for the diagnosis of PD, whereas sensitivity and specificity for FP-CIT SPECT were 0.88 and 0.68, respectively.

The results of the study are interesting and they would suggest that sensitivity and specificity of TS for the diagnosis of PD are much lower than previously reported.

This is exactly the main message of the paper.

The study was well designed but they should have repeated TS at follow-up. This would have provided more information, particularly in the group of definite PD at follow-up.
Results of TS in a group of established PD patients and age-matched controls are also missing in the study.

The reviewer has made this point already (See her first query)

They state that TS findings were concordant with SPECT findings in 89 of 176 patients. It would be interesting to expand this information, possibly in a table.

We are a bit puzzled as to what point the reviewer is trying to make. We did look at whether the concordant data were more present in the IPD +APS group vs. non-parkinsonian, but that was not the case. We have now added this information to the Results section on p 15 , line 9.

It would also be interesting to know sensitivity and specificity of combined positive TS and SPECT scan for the final diagnosis.

We do not see the point of this in a diagnostic accuracy study of one test. We found sensitivity of 0.31 and specificity of 0.82. We would prefer to leave this out, but can include it if the editor would like us to.

In the discussion, the sentence ' we found no significant correlation …. " is not appropriate as no formal correlation were performed.

This is correct, and we have deleted this sentence.

The English needs a major revision. Many sentences do not read well.
The reviewer has made this point already (second query): we have been through the manuscript once again and corrected the English.

What do they mean with the sentence " Apart from the real gold standard of the post-mortem examination, which seems lees feasible in modern times, we think that this gold standard diagnosis of PD is methodologically the highest achievable one" ?

We have explained this sentence on p.16 , line 7-17: when studying diagnostic accuracy in parkinsonian syndromes, the problem is what one should use as gold standard. The accepted gold standard is postmortem neuropathological examination, but this is hardly feasible anymore in modern times, as relatives are reluctant to give permission for this. So, methodologically highest achievable gold standard is clinical examination after several years. The diagnostic criteria contain several items that can only be assessed after a certain amount of time (levodopa response, progression, other diagnoses)