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)	**A multi-centre RCT on community occupational therapy in Alzheimer's disease: Ten sessions are not superior to one consultation.**
AUTHORS	Voigt-Radloff, Sebastian; Graff, Maud; Leonhart, Rainer; Schornstein, Katrin; Jessen, Frank; Bohlken, Jens; Metz, Brigitte; Fellgiebel, Andreas; Dodel, Richard; Eschweiler, Gerhard; Vernooij-Dassen, Myrra; Olde Rikkert, Marcel; Hüll, Michael

VERSION 1 - REVIEW

REVIEWER	***Peter Watson*** Statistician MRC Cognition and Brain Sciences Unit Cambridge England I am not involved in any competing interests connected to the research in this paper.
REVIEW RETURNED	16-Feb-2011

GENERAL COMMENTS	This study compares change from baseline across follow-up times in the percentage of error-free activity steps and the total of activity carer ratings in patients allocated randomly to either a therapy or control group. There is also interest in mood, quality of life and other secondary measurements. Multivariate and univariate analysis of variance is used to compare the change over time in total activity ratings between the two groups. I think the methods of analysis are clear and appropriate but I would suggest using the arcsine transform prior to analysis of variance on the percentage PRPP scores and highlight some points below which will improve understanding of this paper. Page 8. The data from seven centres has been pooled for the present study. How homogeneous are the outcome measures from these centres? Do the outcome measures behave similarly over time across the centres and did the authors consider adding centre as an additional factor in the analysis of variance to check for this? Page 12 & 15. Additional 'external' raters were used to assess the carer ratings of activity of daily functioning. I am not clear how many external raters were used and if they rated independently or as a group. If they rated independently a measure of inter-rater agreement should be given between the external assessor ratings e.g. a kappa statistic or an intra-class correlation coefficient with a comment on its magnitude. The qualifications of the 'external assessers' should also be mentioned. It is pleasing to see an

	agreement measure between the ratings of the external assessors and the carers. In particular a percentage agreement of 61% (with a quoted 50% chance agreement) is mentioned on page 15 but I am not clear how 'agreement' is defined. This seems to me also to be quite a low level of agreement (and it is described as being just above chance) and may, therefore, cast doubt on the validity of the carer ratings which were used for one of the primary outcome measures. Page 14. I would suggest referencing GPOWER free software which can carry out the power calculation and mention that it is available for free download from http://www.psycho.uni-duesseldorf.de/aap/projects/gpower/. I also wondered why 'f' is quoted as an effect size when the usual two-group effect size, Cohen's d, is quoted in the original Dutch study? Page 16. The MANOVA analysis ignores the 52 week scores for the primary outcomes but uses these to look at changes in the secondary outcomes. This appears inconsistent and I wonder why the 52 week scores were not also used to look at primary outcomes. From Table 3 it appears some primary outcomes were not recorded at week 52. Page 16. I wondered why a univariate repeated measures analysis of variance was not used to analyse primary outcomes but was used to analyse secondary outcomes. A couple of sentences should be added here motivating the use of MANOVA over univariate ANOVA. Huberty and Morris (1989) compare these two approaches and could be referenced for those interested in comparing these methods. Given the large amount of missing data did the authors consider performing random effects analysis which does not rely on the filling-in of incomplete responses over time (Hedeker D and Gibbons RD (1997))? If the PRPP percentage is used as an outcome measure the arcsine transform should be used prior to the analysis of variance to equate the group variances (Howell, 1997). Page 16. Was there a pattern of missingness - for example were there different proportions of missing responses on the 11 activity items? It would be helpful to see the range of the proportions of missing activity scores (over the 11 items) and also ranged over the secondary items. It would also be helpful to state and reference which software was used to perform the imputations and MANOVAs. From Figure 1 it appears 16 out of 71 from the intervention arm and 20 of the 70 from the control group were lost during the first 26 weeks of the study which suggests (36/141) 25% of individuals had missing data which may be worth quoting in the paper. Page 18 Table 2. Do the authors have any data on why people dropped out to complement Table 2? This may further inform the reader about the nature of the sample and the limitations of generalisations from the results. In our longitudinal studies we usually categorise why someone has dropped out based upon talking with the participant and/or their carer and issues such as 'transport', 'moved away', 'not interested' often crop up as reasons for withdrawal.

	How were the 'low', 'middle' and 'high' groups under Education in Table 2 defined? Similarly the limitation groups under finance need to be defined. Page 19. Given all the outcomes are presented in Tables 3 and 4 could a column giving the multivariate F statistic for the group by time interaction mentioned on page 19 be added as an extra column? I am not sure quoting the differences in means in Tables 3 and 4 at each time point adds any useful information when all the means are very close to each other and this merely confirms that there is no interaction between group and time on any outcome suggesting the group differences are the same (around zero) at each time point for each outcome. Page 20. I find the mention of ten data imputations ambiguous. Are the authors saying that they performed a multiple imputation filling-in ten times and averaged the estimated missing responses to obtain a predicted response and then did the MANOVA on these or are they saying they performed a MANOVA ten times on each separate set of imputed data or did they perform a MANOVA ten times, obtain ten sets of estimates and obtain an overall average estimate or did they mean something else? Are the figures which further suggest a lack of group by time interaction on page 20 and Tables 2 (page 18) 3 and 4 (page 22) based upon the imputed data or the original data? Page 20. Figures 2 and 3 show the (presumably mean?) total percentage of error-free activity steps presumably with either 95% confidence intervals or mean standard error bars? Page 21. What are categorised as adverse events? Do these relate to the described deaths and hospitalisations described later in the paragraph? Were the number of adverse events formally analysed using e.g. a chi-square test? Page 23. The authors correctly mention the bias that can result from patterns of missingness which involve choosing which responses to 'fill-in'. This is an important point because the assumption behind the imputation method is that the data is missing at random. Did the authors check to see what variables the missingness might be related to by, for example, performing a logistic regression using a 0/1 coding for a non-missing/missing variable value and seeing if the other observed variables were related to whether a particular variable was missing (as recommended by Aitkin et al, 1989)? The authors say that they do not use data which is 'completely missing' (presumably on all primary or secondary outcomes?) at a particular time point. It would be helpful if the authors could explain which variables were used for the imputations stated on page 20. For example, if an activity score was missing at one of the four time points was available data from that time point ALONE used to estimate the missing response or where variables from other time points used to estimate it as well? Similarly was a missing secondary response estimated using only

	some other non-missing secondary responses (and only non-missing primary responses used to estimate missing primary responses?) Page 20 Figure 2 has a missing 'r' in '100=erroless' A plain English improvement is 100= No errors to 0=all errors. Similarly Figure 3 would read better as 0=never needed assistance to 44=always needed assistance. References ======== Aitkin, M, Anderson D, Francis B and Hinder J (1989). Statistical modelling in GLIM. Oxford University Press. Erdfelder, E., Faul, F., & Buchner, A. (1996). GPOWER: A general power analysis program. Behavior Research Methods, Instruments, & Computers, 28, 1-11. Hedeker, D and Gibbons, RD (1997). Application of random-effects pattern-mixture models for missing data in longitudinal studies. Psychological Methods 2(1) 64-78. Howell, DC (1997). Statistical methods for psychology. Fourth Edition. Wadsworth:Belmont,CA. Huberty CJ, Morris, JD (1989). Multivariate analysis versus multiple univariate analyses. Psychological Bulletin 105(2) 302-308.

REVIEWER	***Sandrine Andrieu, MD, PhD*** UMR1027 Inserm-Toulouse University
REVIEW RETURNED	26-Feb-2011

THE STUDY	Description of participants: the exclusion criteria concerning to functional status need to be clarified as it is not clear what constitutes “patients with a major need of physical nursing care”, especially since function is the main outcome of the trial. Also, the diagnosis of AD was based on ICD-10 and not on the classical criteria (DSMV/NINCDS-ADRDA) Abstract/summary/key messages/limitations: (i) the objective given in the abstract (cross-cultural validity of an intervention) is not the same as that given in the introduction to the article and is not concordant with the article title (i.e. to compare the 10 sessions of occupational therapy with 1 session); (ii) the study is described as being single blind in the abstract, but evaluators were only blinded for one of the outcome measures which was not the primary endpoint and so overall the trial cannot be considered as single blind. However it could be stated in the abstract that this was a parallel group trial; (iii) the duration of the intervention could be given in the abstract; (iv) the number of subjects analysed needs to be stated in the abstract since it is not the same as the number randomized; (v) the results need to be more clearly explained in the abstract : the meaning of the “group time interaction effect” will not be clear to all readers; (vi) there are other more important limitations than the one given as the main limitation in the article summary (p4) (for example, that only 104 of the 141 randomised patients were

	actually included in the analysis. Statistical methods: the major problem with the statistical methods is the exclusion of 37 of the 141 patients from the primary analysis (which therefore is not a true ITT analysis). It is unclear why these 37 patients were excluded – there needs to be a definition of "valid data"? Were patients excluded because of missing data? The statement that "imputation of data completely missing at a particular measurement time point would have introduced more bias" is rather strong and unjustified. It might have been preferable to use a mixed effects model for the primary analysis rather than a repeated measures MANOVA since all of the patients could have been included in the analysis (see Gueorguieva & Krystal, Arch Gen Psychiatry 2004). The results of the secondary ITT analysis could have been presented – it would also be useful to know how the multiple imputation was performed (i.e. which variables were included in the imputation model?). Although the authors state that there were no imbalances in baseline characteristics other than financial situation, there are some differences between the analysed patients in the two groups that while even if not statistically significant may be important. Also, table 3 suggests that there may have been baseline differences between the two groups for some of the outcome measures. It would be useful to conduct sensitivity analyses adjusted for such variables. Also, it should be stated in the statistical methods what software was used to perform the analyses. Standard of written English: while most of the paper is well written, there are a few sentences that need to be re-phrased. For example, page 8 line 36 ("would significantly better improve or stablise"); page 9 sentence beginning on line 41 ("Stop criteria"); page 15 sentence beginning on line 22 ("Patients and carers were asked to avoid any talks"). The paper would benefit from being re-read by a native English speaker before publication
RESULTS & CONCLUSIONS	The conclusion that "a comprehensive one-session consultation may be recommended as standard occupational therapy intervention in the German health care system" cannot be derived from the present study. This study found no difference between a 10 session programme and a single session of occupational therapy but this does not mean that a single session can be recommended as standard therapy. The key message of this article is not clear: In certain parts of the paper, the authors refer to the difficulties of implementing an intervention tested in a different country, and in other parts they focus on the non-superioirity of a 10-session intervention compared to a single session. There are some major differences between this German study and the original Dutch study that could explain the difference in results but that are not sufficiently insisted upon: for example, the use of an active control group in the German study compared to a waiting list control group in the Dutch study; the timing of primary endpoint measurements. Given the numerous methodological differences between this study and the Dutch study, it would appear more pertinent to underline the second message (10 sessions vs. 1) as the key message of the

	paper.
REPORTING & ETHICS	Ethical approval needs to be reported in the methods section rather than the acknowledgements.
GENERAL COMMENTS	In addition to the above comments: Overall, this study is very clearly reported, closely following the CONSORT guidelines. Intervention studies of this type are not easy to implement but very useful. There are some positive methodological qualities, for example the standardisation of the intervention through the thorough training sessions prior to the start of the trial. Also, the use of blinded external assessors for one of the outcomes is a strength. However, there are a number of methodological concerns: Methods: 1) Have the IDDD and PRPP undergone validation for use in German? If so, the references need to be given. If not, details of any specific translations performed for this study should be given. 2) It is not clear how the "harms" assessed (deaths, hospitalisations) are expected to be related to the intervention Results: 1) The authors must clearly present the baseline characteristics of all subjects in the intervention and control groups. While the comparison between completers and dropouts is interesting, we need to see the characteristics of the whole group to know if the randomization procedure worked. 2) It would have been useful to know how many patients were included in each centre and how long the patients had been diagnosed with AD for prior to inclusion in the study 3) In the section on "intervention delivery", it would be helpful to have a definition as to what constitutes "hindering" or "facilitating" the intervention. Also, rather than presenting patient and caregiver adherence separately, it would be of more use to present it per dyad. It is probably unnecessary to present the section on patient/caregiver satisfaction with the intervention since these results are not discussed further in this paper. Discussion 1) there are some major elements missing from the limitations section: (i) the risk of contamination between groups since the same occupational therapists gave both the experimental and the control intervention; and (ii) it would have been useful to have included an instrument measuring behavioural problems, although perhaps this was not possible due to the fact that evaluations were not carried out by clinicians.

REVIEWER	***Dr Elizabeth England*** Clinical Lecturer Primary Care clinical Sciences University of Birmingham Edgbaston B15 2TT No competing interests to declare
REVIEW RETURNED	01-Mar-2011

GENERAL COMMENTS	Really interesting and topical paper. Highly relevant to current changes in UK GP led commissioning process and increases in

	numbers of people being diagnosed with dementia

VERSION 1 – AUTHOR RESPONSE

Dr. Richard Sands

Please elaborate a little more on the limitations of the study in the strengths and limitations section at the start. Please include a data sharing statement; if no further data available please state 'No further data available'.

Response
Please find the elaboration on limitations under article focus => strength and limitations => page 4 and the data sharing statement at the end of the paper => page 27

======
Dr. Peter Watson

Page 8. The data from seven centres has been pooled for the present study. How homogeneous are the outcome measures from these centres? Do the outcome measures behave similarly over time across the centres and did the authors consider adding centre as an additional factor in the analysis of variance to check for this?

Response
Now the results of our analysis on possible study site effects are reported under results => outcomes => page 20

Page 12 & 15. Additional 'external' raters were used to assess the carer ratings of activity of daily functioning. I am not clear how many external raters were used and if they rated independently or as a group. If they rated independently a measure of inter-rater agreement should be given between the external assessor ratings e.g. a kappa statistic or an intra-class correlation coefficient with a comment on its magnitude.

Response
The Dutch external video raters were two different persons and we now report the interrater reliability between their ratings under method => randomisation and masking => page 16.

The qualifications of the 'external assessers' should also be mentioned. It is pleasing to see an agreement measure between the ratings of the external assessors and the carers. In particular a percentage agreement of 61% (with a quoted 50% chance agreement) is mentioned on page 15 but I am not clear how 'agreement' is defined. This seems to me also to be quite a low level of agreement (and it is described as being just above chance) and may, therefore, cast doubt on the validity of the carer ratings which were used for one of the primary outcome measures.

Rsponse
Our text was not clear enough. We measured the agreement between the actual group assignment and the group assignment as guessed by the blinded assessor (the person who interviews the patient and the carer at home; this person did not perform the PRPP rating). This analysis was done to estimate whether an “unblinding” of the assessor occurred during the interview. This is now better described under method => randomisation and masking => page 15

Page 14. I would suggest referencing GPOWER free software which can carry out the power calculation and mention that it is available for free download from http://www.psycho.uni-duesseldorf.de/aap/projects/gpower/. I also wondered why 'f' is quoted as an effect size when the usual two-group effect size, Cohen's d, is quoted in the original Dutch study?

Response
We used the program of Erdfelder and cite this now. The Cohen's d effect size of the Dutch study was the background for our calculation. We used this to estimate an r-value and with this an f-value, which is - according to Faul & Erdfelder - the correct effect size for a MANOVA.

Page 16. The MANOVA analysis ignores the 52 week scores for the primary outcomes but uses these to look at changes in the secondary outcomes. This appears inconsistent and I wonder why the 52 week scores were not also used to look at primary outcomes. From Table 3 it appears some primary outcomes were not recorded at week 52.

Response
Full assessment of primary outcome was until week 26 (PRPP+ IDDD). Follow up in week 52 was only a postal assessment with carer questionnaires (IDDD).

This is now clearer described under method => outcome measures => page 13 and in tables 3 and 4 => page 22

Attrition, MANOVA versus ANOVA, multiple imputation

Page 16. I wondered why a univariate repeated measures analysis of variance was not used to analyse primary outcomes but was used to analyse secondary outcomes. A couple of sentences should be added here motivating the use of MANOVA over univariate ANOVA. Huberty and Morris (1989) compare these two approaches and could be refer-enced for those interested in comparing these methods. Given the large amount of missing data did the authors consider performing random effects analysis which does not rely on the filling-in of incom-plete responses over time (Hedeker D and Gibbons RD (1997))?
If the PRPP percentage is used as an outcome measure the arcsine transform should be used prior to the analysis of variance to equate the group variances (Howell, 1997).

Page 20. I find the mention of ten data imputations ambiguous. Are the authors saying that they performed a multiple imputation filling-in ten times and averaged the estimated missing responses to obtain a predicted response and then did the MANOVA on these or are they saying they performed a MANOVA ten times on each separate set of imputed data or did they perform a MANOVA ten times, obtain ten sets of estimates and obtain an overall average estimate or did they mean something else? Are the figures which further suggest a lack of group by time interaction on page 20 and Tables 2 (page 18) 3 and 4 (page 22) based upon the imputed data or the original data?

Page 23. The authors correctly mention the bias that can result from patterns of missingness which involve choosing which responses to 'fill-in'. This is an important point because the assumption behind the imputation method is that the data is missing at random. Did the au-thors check to see what variables the missingness might be related to by, for example, performing a logistic regression using a 0/1 coding for a non-missing/missing variable value and seeing if the other observed variables were related to whether a particular variable was missing (as recommended by Aitkin et al, 1989)?

The authors say that they do not use data which is 'completely missing' (presumably on all primary or

secondary outcomes?) at a particular time point. It would be helpful if the authors could explain which vari-ables were used for the imputations stated on page 20. For example, if an activity score was missing at one of the four time points was avail-able data from that time point ALONE used to estimate the missing response or where variables from other time points used to estimate it as well? Similarly was a missing secondary response estimated using only some other non-missing secondary responses (and only non-missing primary responses used to estimate missing primary responses?)

Page 16. Was there a pattern of missingness - for example were there different proportions of missing responses on the 11 activity items? It would be helpful to see the range of the proportions of missing activity scores (over the 11 items) and also ranged over the secondary items.
It would also be helpful to state and reference which software was used to perform the imputations and MANOVAs. From Figure 1 it appears 16 out of 71 from the intervention arm and 20 of the 70 from the control group were lost during the first 26 weeks of the study which suggests (36/141) 25% of individuals had missing data which may be worth quoting in the paper.
Page 18 Table 2. Do the authors have any data on why people dropped out to complement Table 2? This may further inform the reader about the nature of the sample and the limitations of generalisa-tions from the results. In our longitudinal studies we usually categorise why someone has dropped out based upon talking with the participant and/or their carer and issues such as 'transport', 'moved away', 'not interested' often crop up as reasons for withdrawal.

Response
We thank the reviewers for their worthwhile hints regarding the attrition and the statistical analyses. In general, we think the recommendations for im-provement would be very appropriate, if we had found significant results in our analysis within the reduced sample of patients with only valid data. Then it could have been argued, that an inclusion of patients with no valid data could reduce the effects and that this reduced effects would better represent the true effects. However, because we did not find significant effects even in the analysis of the reduced sample, we would like to state that our results are calculated quite conservatively. And thus - in our judgement - our conclusion that we cannot reject the 0-hypotheses of no group differ-ences seems to be appropriate and well based in our data and analysis.

We added info on imputation under methods => statistical methods => page 16. We also re-phrased a text part under discussion => limitation => page 23

Some hopefully clarifying explanations follow:

Attrition (number, reasons, possible bias)
We asked for reasons of withdrawal. Numbers and reasons for dropouts or withdrawal are listed in Figure 1 => page 17. The term “lost for follow up” stands for participants with assessment data but not collected within the defined time range (14 days around the planned measurement day). As soon as one assessment was not within the defined time range, the participant was categorised as “no valid data” and was excluded from the analysis (com-pletely with all assessments). We lost 14 of 141 participants (10 %) for this reason. In our judge-ment, this is acceptable in a pragmatic 1-year-trial in dementia under routine care conditions, where you may lose timely assessment data just because participants are on holiday, forget the appointment or carers are busy with their job or other important dates.
The attrition shows no systematic bias in the sense that more participants are lost only in the COTiD or only in the control group. Also numbers and rea-sons of withdrawal are nearly equal in both groups.

MANOVA versus ANOVA
Primary outcome: It is known that univariate ANOVA may lead to false significant results, when

outcomes correlate with each other. Therefore, we applied not ANOVA but MANOVA for the primary outcomes, because there was a correlation be-tween the IDDD and the PRPP data.
Secondary outcome: As we had 12 secondary outcomes and 5 secondary measurement time points, the sample with complete data in all 60 data sets (measurements x time points) was too small for an appropriate MANOVA. We used two ap-proaches to solve this problem.
(1) Separate ANOVA for each secondary out-come with slightly reduced samples size de-pending on the missings in the particular out-come (as shown in tables 3 and 4). In this uni-variate analysis, which did not consider possi-ble correlations between the outcomes, we did not find significant differences. So we hy-pothesised that it is very unlikely that a multi-variate analysis would find significant effects.
(2) However to test this assumption we applied the MANOVA after the imputation of all miss-ing data for the 104 completers (all time points, all secondary outcomes, method ac-cording to Rubin please see below). But we did not find any significant group differences.

Multiple imputation
• We followed established methods of data imputation according to Rubin DB. Inference and missing data. Biometrika, 1976(63):581-92 and Rubin D B (1996). Multiple imputation after 18+ years. Journal of the American Sta-tistical Association, 1996; 91 (434):473-89.
• We did multiple imputations with the full-information-maximum-likelihood-estimator. Because there is a random component within this procedure, we did it ten times (even more than the 5 times as recommended by Rubin)
• We did a Missing Value Analysis with SPSS (MVA) and we found no substantial patterns or differences between the groups for the base-line values.
• For the imputations, we used all available data over all measurement points of all primary and secondary variables. For this we used the im-putation possibility of SPSS.

How were the 'low', 'middle' and 'high' groups under Education in Table 2 defined? Similarly the limitation groups under finance need to be defined.

Response
Now better defined in table 2 => page 18

Page 19. Given all the outcomes are presented in Tables 3 and 4 could a column giving the multivariate F statistic for the group by time interaction mentioned on page 19 be added as an extra column? I am not sure quoting the differences in means in Tables 3 and 4 at each time point adds any useful information when all the means are very close to each other and this merely confirms that there is no interaction between group and time on any outcome suggesting the group differences are the same (around zero) at each time point for each outcome.

Response
We had already discussed it within the author group. As we are also concerned with systematic reviews, we concluded that detailed report of data would provide appropriate information for future meta-analyses.

Page 20. Figures 2 and 3 show the (presumably mean?) total percentage of error-free activity steps presumably with either 95% confidence intervals or mean standard error bars?
Page 20 Figure 2 has a missing 'r' in '100=erroless' A plain English improvement is 100= No errors to 0=all errors.
Similarly Figure 3 would read better as 0=never needed assistance to 44=always needed assistance.

Response
Now clearer described in the captions of figures 2 and 3 => page 20/21

Page 21. What are categorised as adverse events? Do these relate to the described deaths and hospitalisations described later in the paragraph? Were the number of adverse events formally analysed using e.g. a chi-square test?

Response
Yes, it relates to the description later in the paragraph. Test results are now reported under results => harms => page 21

Dr. Sandrine Andrieu

Description of participants: the exclusion criteria concerning to functional status need to be clarified as it is not clear what constitutes “patients with a major need of physical nursing care”, especially since function is the main outcome of the trial.

Response
We added “level 2 or higher according to the German Long-Term Care Insurance Act, defined by daily need of physical nursing care >= 120 min” under Methods => participants and setting => page 9

Also, the diagnosis of AD was based on ICD-10 and not on the classical criteria (DSMV/NINCDS-ADRDA)

Response
In Germany, the more precise NINCDS-ADRDA criteria are primarily used in phase-II clinical trials, but not in routine care. As we conducted a pragmatic trial, we preferred the international ICD-10 criteria actually used in German routine care. These criteria are largely congruent to the US-American DSMV criteria.

Abstract/summary/key messages/limitations: (I) the objective given in the abstract (cross-cultural validity of an intervention) is not the same as that given in the introduction to the article and is not concordant with the article title (i.e. to compare the 10 sessions of occupational therapy with 1 session); (II) the study is described as being single blind in the abstract, but evaluators were only blinded for one of the outcome measures which was not the primary endpoint and so overall the trial cannot be considered as single blind. However it could be stated in the abstract that this was a parallel group trial; (III) the duration of the intervention could be given in the abstract; (IV) the number of subjects analysed needs to be stated in the abstract since it is not the same as the number randomized; (V) the results need to be more clearly explained in the abstract: the meaning of the “group time interaction effect” will not be clear to all readers; (VI) there are other more important limitations than the one given as the main limitation in the article summary (p4) (for example, that only 104 of the 141 randomised patients were actually included in the analysis)

Response
ad (I and III to VI)
We re-phrased the abstract according to the reviewer’s comments => page 5/6

ad (II)
Here we cannot fully follow the reviewer. The assessors were blinded for group assignment, their “contamination” by patients and carers was controlled. Agreement between actual group assignment

and assessors' guess was 61 %. The two Dutch external PRPP raters were fully blinded for group assignment without any contamination. The PRPP is one indicator for daily functioning, the primary outcome. However, we re-phrased the text in the abstract => page 5 and described the masking clearer under methods => randomisation and masking => page 15

Statistical methods: the major problem with the statistical methods is the exclusion of 37 of the 141 patients from the primary analysis (which therefore is not a true ITT analysis). It is unclear why these 37 patients were excluded – there needs to be a definition of "valid data"? Were patients excluded because of missing data? The statement that "imputation of data completely missing at a particular measurement time point would have introduced more bias" is rather strong and unjustified. It might have been preferable to use a mixed effects model for the primary analysis rather than a repeated measures MANOVA since all of the patients could have been included in the analysis (see Gueorguieva & Krystal, Arch Gen Psychiatry 2004). The results of the secondary ITT analysis could have been presented – it would also be useful to know how the multiple imputation was performed (i.e. which variables were included in the imputation model?). Although the authors state that there were no imbalances in baseline characteristics other than financial situation, there are some differences between the analysed patients in the two groups that while even if not statistically significant may be important. Also, table 3 suggests that there may have been baseline differences between the two groups for some of the outcome measures. It would be useful to conduct sensitivity analyses adjusted for such variables. Also, it should be stated in the statistical methods what software was used to perform the analyses.

Response
Please see reaction on reviewer 2

Standard of written English: while most of the paper is well written, there are a few sentences that need to be re-phrased. For example, page 8 line 36 ("would significantly better improve or stablise"); page 9 sentence beginning on line 41 ("Stop criteria"); page 15 sentence beginning on line 22 ("Patients and carers were asked to avoid any talks"). The paper would benefit from being re-read by a native English speaker before publication

Response
The listed sentences are re-phrased now and a native English speaker has again corrected the paper.

The conclusion that "a comprehensive one-session consultation may be recommended as standard occupational therapy intervention in the German health care system" cannot be derived from the present study. This study found no difference between a 10 session programme and a single session of occupational therapy but this does not mean that a single session can be recommended as standard therapy.

Response
This is now re-phrased under discussion => Clinical and research implications => page 25

The key message of this article is not clear: In certain parts of the paper, the authors refer to the difficulties of implementing an intervention tested in a different country, and in other parts they focus on the non-superioirity of a 10-session intervention compared to a single session. Given the numerous methodological differences between this study and the Dutch study, it would appear more pertinent to underline the second message (10 sessions vs. 1) as the key message of the paper.

Response
This is now re-phrased under article focus => key message => page 4 and under abstract =>

objective => page 5

There are some major differences between this German study and the original Dutch study that could explain the difference in results but that are not sufficiently insisted upon: for example, the use of an active control group in the German study compared to a waiting list control group in the Dutch study; the timing of primary endpoint measurements.

Response
This is now re-phrased under discussion => Clinical and research implications => page 26

Ethical approval needs to be reported in the methods section rather than the acknowledgements.

Response
Ethical approval is now reported under methods => Design => page 9

Methods: 1) Have the IDDD and PRPP undergone validation for use in German? If so, the references need to be given. If not, details of any specific translations performed for this study should be given.

Response
Details are reported now under methods => outcome measures => page 13

Methods: 2) It is not clear how the “harms” assessed (deaths, hospitalisations) are expected to be related to the intervention

Response
The assumed relation is now described under methods => outcome measures => page 14

Results: 1) The authors must clearly present the baseline characteristics of all subjects in the intervention and control groups. While the comparison between completers and dropouts is interesting, we need to see the characteristics of the whole group to know if the randomization procedure worked.

Response
This is done now => table 2 => page 18

Results: 2) It would have been useful to know how many patients were included in each centre and how long the patients had been diagnosed with AD for prior to inclusion in the study

Response
We have no data about the period from AD diagnose to study inclusion. The numbers of included patients per centre are now reported under results => Recruitment and participant flow => page 18

Results: 3) In the section on “intervention delivery”, it would be helpful to have a definition as to what constitutes “hindering” or “facilitating” the intervention.

Response
This is one topic in our process evaluation paper. We submitted both with the same letter to the editors and stated that these papers are closely connected. Unfortunately the online submission system has led to two separate review processes. However, we added rating criteria under results => intervention delivery => page 19

Results: 3) Also, rather than presenting patient and caregiver adherence separately, it would be of more use to present it per dyad.

Response
Here we do not agree with the reviewer as our interventionists reported cases of different adherence by the patient and the carer, which is also topic of the process evaluation paper.

Results: 3) It is probably unnecessary to present the section on patient/caregiver satisfaction with the intervention since these results are not discussed further in this paper.

Response
We cancelled these data

Discussion: 1) there are some major elements missing from the limitations section: (I) the risk of contamination between groups since the same occupational therapists gave both the experimental and the control intervention; and (II) it would have been useful to have included an instrument measuring behavioural problems, although perhaps this was not possible due to the fact that evaluations were not carried out by clinicians.

Response

Ad (I)
The risk of contamination is now discussed under discussion => limitations => page 24

Ad (II)
During recruitment study physicians did exclude major behavioural problems. They also had contact in week 11 and 21 to control for medical and behavioural problems and – if indicated – to discontinue the trial participation. Actually, the study physicians did exclude no participant due to major behavioural problems.

VERSION 2 - REVIEW

REVIEWER	***Peter Watson***
REVIEW RETURNED	31-Mar-11

THE STUDY	I only have a few minor points which should be easily handled by the authors. In particular there are a few inconsistencies in the descriptions and analyses which I comment on below which I am sure can be ironed out.
GENERAL COMMENTS	This study assesses baseline change over follow-up times in Alzheimer's patients randomly assigned to either a therapy or a control group. No time change by group interactions are found suggesting there is no benefit in daily functioning using the therapeutic procedure. There is now additional information on the randomisation procedure, assessors' qualifications, figure error bars and the use of multiple imputation in SPSS for filling-in missing values. Multiple imputation is seen as an unbiased approach to missing values estimation and suitable for data missing at random. I also notice (page 16) an inter-rater agreement correlation is now given and fuller details of the locations of the study centres are now on pages 9 and 10. I think the similarity of the centre locations (namely all in urban areas) and also as checked on page 20 is sufficient to justify pooling the data.

	The results (from using all ten imputed data sets) and figures do look convincing suggesting there is no time by group interaction. There are, however, some small points of clarification which I mention below which I feel could help with the understanding of this paper. Pages 12 & 16. I would mention that for the PRPP percentage which is actually a proportion (multiplied by 100) Howell (1997) suggests using the arcsine transform. He suggests using this on proportions prior to the multivariate and univariate analyses of variance to equate the group variances (Howell, 1997). This is easy to do in SPSS using the compute statement 2*arsin(sqrt(p)) where p is the proportion. Page 14. I would again suggest referencing G*POWER free software (Erdfelder et al, 1996) which can carry out the power calculation and mention that it is available for free download from http://www.psycho.uni-duesseldorf.de/aap/projects/gpower/ and state that the power calculation relates to the between subjects factor by the within subjects factor interaction term. This is correct as we are interested in such an interaction in this study. An advantage of referencing G*POWER is that this software also has a help manual which explains how the effect sizes are computed for an interaction involving one between subjects factor (group) and one within subjects factor (time) which I believe is what is calculated on page 14 and would help other readers with their own power calculations. I would mention that the effect size 'f' is based on an (group by time) interaction with another effect size called 'd-value' which I assume is Cohen's d that is used to compare two group (main effects) but not interaction terms. I would agree with the authors in saying a value of f equal to 0.1 is conservative and a d of 2.4 large but additionally justify this by referencing Cohen (1988) who states that a f of 0.1 corresponds to a small effect size and any d over 0.8 is large. Pages 16 and 20. Some inconsistency here in whether baseline was included as a covariate upon which I would appreciate clarification: On page 16 "In the primary analysis, we did not adjust for baseline values or any other covariate" but on page 20 for the imputed data sets "baseline values for all outcome measurements" are included in MANOVAs comparing study sites. I suspect baseline wasn't adjusted for when interest is in relative group changes over time because (page 18) there were no baseline differences in variables, associated with functional decline, between the two groups being compared so that any group difference would not be due to differences in baseline. I think it would be helpful if the baseline characteristics section giving the reason for not adjusting for baseline was, therefore, moved from page 18 to the start of the statistical methods section on page 16 so we then know in advance why no baseline covariates were included in the primary (and presumably secondary) outcome MANOVAs. Page 16 & 20. Slight confusion here. On page 16 a univariate repeated measures analysis of variance was used to analyse secondary outcomes (but was not used to analyse primary outcomes) whereas on page 20 it says a MANOVA was used on the imputed data for ALL primary and secondary outcomes. Could you please clarify that the MANOVA was used for all analyses involving secondary outcomes comparing change over time (page 16). There is also some confusion about which outcome variables were imputed. On page 16 we have that imputation was performed for all

	secondary outcomes but there is no mention that it was used for primary outcomes. On page 20 it implies to me that primary as well as secondary outcomes were imputed which seems sensible and I assume is what happened in the analyses on page 16. Is this the case? Page 16. I wonder why week 52 was not analysed for primary outcomes but was for secondary ones which does not seem to tie in with the implication in the 'outcome measures' section on page 6 that postal assessment was carried out on all (primary and secondary) outcome measures but if this is the case why is there no PRPP data in table 3 (page 22) for week 52. I notice there IS data in table 3 (page 22) at 52 weeks for the IDDD (primary outcome) in both groups yet week 52 is not plotted in Figure 3 of IDDD on page 21 or analysed on page 16 for the IDDD primary outcome. Was it because of large amounts of missing data at week 52 on primary measures? I wonder if another reason for not including the obtained week 52 IDDD data in either figure 3 (page 21) or the MANOVA analysis is to make more meaningful comparisons between outcome variables because not all outcomes were obtained at week 52 via the post but if this was true then why would one collect week 52 data in the first place? Pages 20 & 21. Just to be clear I assume the graphs in Figures 2 and 3 (pages 20 & 21) are of the non-imputed data ie for the observed data only. One could in principle have graphs based on the imputed data sets analysed in the study although I don't think this is necessary here as the graphs support the conclusions using imputed data that there are time by group interactions. I did also struggle to make out the lighter line in each of Figures 2 and 3. Page 21. "The group difference was not significant" – we could tweak to say more informatively 'There was no difference between the two groups in average number of nights admitted to hospital'. Page 23 & Page 4. I would rewrite the second mitigating circumstance for the first limitation in the limitations paragraph on page 23 and the last paragraph on page 4 as something such as '(2) we used multiple data imputation which is less biased than last observation carried forward in dementia research'. One could also give a supporting reference to this on page 23. Hamer and Simpson (2009), for example, suggest limitations of the last observation carried forward (although they were thinking of psychiatric applications) because it answers a rather uninteresting question of are there differences between baseline and the subject's last recorded score as opposed to difference between baseline and subject's projected final score which may not be the same as the last one observed. References Cohen J (1988). Statistical Power Analysis for the Behavioral Sciences (second ed.). Lawrence Erlbaum Associates. Erdfelder, E., Faul, F., & Buchner, A. (1996). GPOWER: A general power analysis program. Behavior Research Methods, Instruments, & Computers, 28, 1-11. Hamer RM and Simpson PM (2009). Last Observation Carried Forward Versus Mixed Models in the Analysis of Psychiatric Clinical Trials Am J Psychiatry 166:639-641.

REVIEWER	***Sandrine Andrieu***
REVIEW RETURNED	20-Apr-2011

THE STUDY	The study limitations have been given in the study summary, but some important limitations are still missing, and the discussion of LOCF is not appropriate for this section. The authors did not clearly respond to the comment regarding the ITT analysis : all randomized patients, were not analyzed . The response concerning the management of missing data is not at all convincing: it is not because LOCF is not adapted to dementia that the missing data should not be taken into account (other methods for handling missing data are available). I proposed a mixed model analysis and remain convinced by the utility of this type of analysis.
RESULTS & CONCLUSIONS	Regarding the "blinded assessment", our concern is that caregivers were not blinded and it was them that gave the information for the patient's evaluation, particularly for the primary outcome (IDDD). The authors still refer to the comparison of the intervention in two different socio-cultural contexts but the Dutch and German studies were very different: the proposed comparison would have been possible if the two studies were strictly identical.
GENERAL COMMENTS	The authors have responded correctly to many of the comments. However, the major comments that were raised would have required a more thorough revision of the paper (e.g. new statistical analysis) rather than the fairly minor revisions that have been performed.

VERSION 2 – AUTHOR RESPONSE

Reviewer 1, Dr. Watson

Pages 12 & 16. I would mention that for the PRPP percentage which is actually a proportion (multiplied by 100) Howell (1997) suggests using the arcsine transform. He suggests using this on proportions prior to the multivariate and univariate analyses of variance to equate the group variances (Howell, 1997). This is easy to do in SPSS using the compute statement 2*arsin(sqrt(p)) where p is the proportion.

Response

Using the arcsine transform did not change results: (1) Original: F(df1=3, df2=306) = 1.40; p = .243; partial eta square = 0.0135

(2) The new variable: F(df1=3, df2=306) = 1.49; p = .216; partial eta square = 0.0148. We added under results => outcomes: “Using a special transformation for the PRPP percentage did not change results (original: p=0.243; arcsine-transform: p=0.216).”

Page 14. I would again suggest referencing G*POWER free software (Erdfelder et al, 1996) which can carry out the power calculation and mention that it is available for free download from http://www.psycho.uni-duesseldorf.de/aap/projects/gpower/ and state that the power calculation relates to the between subjects factor by the within subjects factor interaction term. This is correct as we are interested in such an interaction in this study. An advantage of referencing G*POWER is that this software also has a help manual which explains how the effect sizes are computed for an interaction involving one between subjects factor (group) and one within subjects factor (time) which I believe is what is calculated on page 14 and would help other readers with their own power calculations.

Response

We used G-power 3.1 and added the website at reference no. 31

Page 14.

I would mention that the effect size ‘f’ is based on an (group by time) interaction with another effect size called ‘d-value’ which I assume is Cohen’s d that is used to compare two group (main effects) but not interaction terms. I would agree with the authors in saying a value of f equal to 0.1 is conservative and a d of 2.4 large but additionally justify this by referencing Cohen (1988) who states that a f of 0.1 corresponds to a small effect size and any d over 0.8 is large.

Response

We added under => methods => sample size calculation: “Our assumed effect size of f = 0.10 is based on a group by time interaction and compatible to Cohen’s d = 0.20, which corresponds to a small effect size and any d over 0.8 is large.”

Pages 16 and 20. Some inconsistency here in whether baseline was included as a covariate upon which I would appreciate clarification: On page 16 “In the primary analysis, we did not adjust for baseline values or any other covariate” but on page 20 for the imputed data sets “baseline values for all outcome measurements” are included in MANOVAs comparing study sites. I suspect baseline wasn’t adjusted for when interest is in relative group changes over time because (page 18) there were no baseline differences in variables, associated with functional decline, between the two groups being compared so that any group difference would not be due to differences in baseline. I think it would be helpful if the baseline characteristics section giving the reason for not adjusting for baseline was, therefore, moved from page 18 to the start of the statistical methods section on page 16 so we then know in advance why no baseline covariates were included in the primary (and presumably secondary) outcome MANOVAs.

Response

We rephrased the sentence on page 16 under => Methods => Statistical methods to “We did not adjust for baseline values, because we found no marked group differences.”

Page 16 & 20. Slight confusion here. On page 16 a univariate repeated measures analysis of variance was used to analyse secondary outcomes (but was not used to analyse primary outcomes) whereas on page 20 it says a MANOVA was used on the imputed data for ALL primary and secondary outcomes. Could you please clarify that the MANOVA was used for all analyses involving secondary outcomes comparing change over time (page 16).

There is also some confusion about which outcome variables were imputed. On page 16 we have that imputation was performed for all secondary outcomes but there is no mention that it was used for primary outcomes. On page 20 it implies to me that primary as well as secondary outcomes were imputed which seems sensible and I assume is what happened in the analyses on page 16. Is this the case?

Response

In our first response we clarified that we did an ANOVA on the NOT-imputed data of the secondary outcomes and a MANOVA on the imputed data of the secondary outcomes and data of the primary outcomes (which had no missings).

Analyses actually done:

Primary analysis in the sample of completers

MANOVA with baseline and weeks 6, 16, 26 for the primary outcome (IDDD & PRPP, Figure 2 & 3):

ANOVA (without imputation) with baseline and weeks 6, 16, 26 AND week 52 for all secondary outcomes and the primary outcome IDDD (PRPP data are not available for week 52) => Tables 3 & 4.

Secondary analysis in the sample of completers

Data imputation only in secondary but not in primary outcomes, because there were no missing data in primary outcomes in the sample of completers => MANOVA over primary and secondary outcomes with baseline and weeks 6, 16, and 26.

We rephrased the text under methods => statistical methods

Page 16. I wonder why week 52 was not analysed for primary outcomes but was for secondary ones which does not seem to tie in with the implication in the 'outcome measures' section on page 6 that postal assessment was carried out on all (primary and secondary) outcome measures but if this is the case why is there no PRPP data in table 3 (page 22) for week 52. I notice there IS data in table 3 (page 22) at 52 weeks for the IDDD (primary outcome) in both groups yet week 52 is not plotted in Figure 3 of IDDD on page 21 or analysed on page 16 for the IDDD primary outcome. Was it because of large amounts of missing data at week 52 on primary measures?

I wonder if another reason for not including the obtained week 52 IDDD data in either figure 3 (page 21) or the MANOVA analysis is to make more meaningful comparisons between outcome variables because not all outcomes were obtained at week 52 via the post but if this was true then why would one collect week 52 data in the first place?

Response

PRPP data could only be provided by videotaping an ADL task at patient's home and thus were not collected at all at week 52. In our pragmatic trial we tried to cover a follow up period as long as possible with limited resources. This was especially in order to record nursing home placements and long-term effects one year after baseline. But a further home visit at week 52 by the assessor was not within the scope of funding.

Pages 20 & 21. Just to be clear I assume the graphs in Figures 2 and 3 (pages 20 & 21) are of the non-imputed data ie for the observed data only. One could in principle have graphs based on the

imputed data sets analysed in the study although I don't think this is necessary here as the graphs support the conclusions using imputed data that there are time by group interactions. I did also struggle to make out the lighter line in each of Figures 2 and 3.

Response

We added to the capitations at figure 1 and 2 "N=104 completers" and made the lines stronger

Page 21. "The group difference was not significant" – we could tweak to say more informatively 'There was no difference between the two groups in average number of nights admitted to hospital'.

Response

We rephrased it according to the reviewer's suggestion

Page 23 & Page 4. I would rewrite the second mitigating circumstance for the first limitation in the limitations paragraph on page 23 and the last paragraph on page 4 as something such as '(2) we used multiple data imputation which is less biased than last observation carried forward in dementia research'. One could also give a supporting reference to this on page 23. Hamer and Simpson (2009), for example, suggest limitations of the last observation carried forward (although they were thinking of psychiatric applications) because it answers a rather uninteresting question of are there differences between baseline and the subject's last recorded score as opposed to difference between baseline and subject's projected final score which may not be the same as the last one observed.

Response

We replaced the statement on LOCF by "…an additional mixed model analysis of all randomised patients did not reveal significant differences…"

Reviewer 2, Dr. Andrieu

The study limitations have been given in the study summary, but some important limitations are still missing, and the discussion of LOCF is not appropriate for this section.

The authors did not clearly respond to the comment regarding the ITT analysis: all randomized patients, were not analyzed.

The response concerning the management of missing data is not at all convincing: it is not because LOCF is not adapted to dementia that the missing data should not be taken into account (other methods for handling missing data are available). I proposed a mixed model analysis and remain convinced by the utility of this type of analysis.

Response

Since we found no significant interactions with conventional statistic and multiple imputation, we now additionally applied a mixed model analysis as you suggested. The pattern of missing values in the PRPP was: complete data for N=107 at baseline and week 6, 16 and 26; missings for N=34 right from baseline, because these patients were not willing to be videotaped. The patterns of missing values in the IDDD were: complete data for N=141 at baseline, for N=131 at week 6, for N=119 at week 16 and for N=114 at week 26. The additional analysis with SPSS mixed models showed again non-significant interactions (PRPP: F = 0.074; p=0.785. IDDD: F=0.911; p=0.340).

We added under => results => outcomes: "An additional mixed models analysis of all randomised patients (N=141) as recommended by Coley and colleagues (ref) did reveal no significant interactions for the IDDD (p=0.340) and the PRPP (p=0.785)."

Furthermore we replaced our statement on LOFC in the study summary by: “and an additional mixed model analysis of all randomised patients did not reveal significant differences.”

Regarding the "blinded assessment", our concern is that caregivers were not blinded and it was them that gave the information for the patient's evaluation, particularly for the primary outcome (IDDD).

Response

We added at the strength and limitation section: “However, patients and carer could not be fully masked.”

The authors still refer to the comparison of the intervention in two different socio-cultural contexts but the Dutch and German studies were very different: the proposed comparison would have been possible if the two studies were strictly identical.

Response

Since the COTiD programme demonstrated such highly positive effects, we judged it as appropriate to conduct not an identical replication just in another country, but rather more a twofold transfer: (1) from NL to GER and (2) from a specific RCT design in one centre with high expertise to a pragmatic multi-centre RCT design in routine care. (Please see. Zwarenstein et al. Practihc group. Improving the reporting of pragmatic trials: an extension of the CONSORT statement. BMJ. 2008 Nov 11;337:a2390.)

Furthermore, in order to reduce the focus on cross-cultural transfer, we eleminated the following sentence at the discussion part: “However, it is difficult to judge whether these measures could compensate for the potential influence of different educational backgrounds of Dutch and German occupational therapists. In the Netherlands, occupational therapy education takes four years and is more psychosocial oriented than the three years curriculum in Germany.”

VERSION 3 - REVIEW

REVIEWER	***Sandrine Andrieu***
REVIEW RETURNED	08-Jun-2011

THE STUDY	For abstract: I definitely suggest replacing the sentence : "assessors were blind for treatment allocation" by "assessors were blind for treatment allocation for one of two primary outcome" in abstract (design section) for statistical : details of the additional mixed model should be reported in the method section
RESULTS & CONCLUSIONS	The authors need to present the full results of the mixed model preferably in a table the authors still refer to the comparison of the intervention in two different socio-cultural contexts : last sentence of the abstract conclusion page 7 should be deleted the authors need to add in the discussion their response to our las comment : "Since the COTID program....Germany"

VERSION 3 – AUTHOR RESPONSE

Managing Editor, Dr. Sands

The abstract is missing a few elements to be fully in line with CONSORT. Please see http://www.consort-statement.org/consort-statement/title-and-abstract/item1b_abstract/. For example, eligibility criteria, randomisation

Response
We added the missing elements according to the latest CONSORT requirements

========

Reviewer, Dr. Andrieu

For abstract: I definitely suggest replacing the sentence : "assessors were blind for treatment allocation" by "assessors were blind for treatment allocation for one of two primary outcome" in abstract (design section)

Response
We rephrased as follows: “Patients and carers were not masked. Assessors were fully blind for treatment allocation for one of two primary outcome measurements.”

———

For statistical: details of the additional mixed model should be reported in the method section. The authors need to present the full results of the mixed model preferably in a table.

Response
We added under => methods => statistical methods: “In an additional analysis we used the linear mixed-effects models (MIXED) procedure in SPSS, which allows an unequal number of repetitions and a better handling of missing values.”
We added under => results => outcomes: “details are provided as supplementary online material.”
We added tables of results as supplementary online material.

———

The authors still refer to the comparison of the intervention in two different socio-cultural contexts: last sentence of the abstract conclusion page 7 should be deleted.

Response
We replaced the sentence by: “Further research on the transfer of complex psychosocial interventions is needed.”

———

The authors need to add in the discussion their response to our last comment: "Since the COTID program....Germany"

Response
We add under => discussion => comparison: “Since the Dutch COTiD programme demonstrated such highly positive effects, we judged it as appropriate to conduct not an identical replication, but a twofold

transfer from the source to the target country and from a mono-centre RCT design with high expertise of interventionists to a pragmatic multi-centre RCT design in routine care [ref: Zwarenstein et al.].”