PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	Quality of Life of Patients with Dementia in Acute Hospitals: A non-randomised case-control-study comparing a regular ward to a special care ward with dementia care concept.
AUTHORS	Lüdecke, Daniel; Poppele, Georg; Klein, Jens; Kofahl, Christopher

VERSION 1 - REVIEW

REVIEWER	Dr Anthony Scerri Uni of Malta
REVIEW RETURNED	22-Apr-2019

GENERAL COMMENTS	This non-randomized case-control study sought to identify the factors that predict the quality of life of patients with dementia admitted in acute hospitals wards and to evaluate whether a special care concept improves the patient's quality of life. The study is well conceived and the write-up clear. As acknowledged in the supplementary material by the authors, the lack of familiarity with the statistical analysis chosen makes it somewhat hard to follow the results especially for readers who are not accustomed to this type of analysis. Nevertheless, the supplementary material offers a helpful source of information for any reader to follow how the analysis was carried out. The study achieves interesting findings about the association between the special care concept and the quality of life of patients with dementia during hospital stay. As also recognized by the authors, the main limitation of this study is the selection of the participants coming from different hospitals. It would be appropriate to explain to the readers why was the intervention group and control group not taken from the same hospital. Could another ward in the same hospital as the intervention ward be a better option? Considering that the organizational culture can influence the quality of care and QoL of the patients, taking the intervention and control ward from the same hospital could have reduced this potential bias. These are some additional questions and comments for further clarification and to improve the study further: Page 7 Line 14-15 - Did the study nurses who collected the data, work in the same hospitals where the study was conducted? Was some independent monitoring been carried out to make sure that the study nurses are accurately filling in the questionnaire?

	Page 7 Line 44-45 – Who carried out the dementia screening in the control ward? Page 8 Line 3 – Who recorded the independent variables? Were these recorded by the study nurses or by the researchers? Page 6: Can you be more specific in terms of the staffing levels? What is the difference in staffing levels and staff: patient ratio between the intervention and control wards? Page 7: Give more detail of the inclusion/exclusion criteria used to identify the participants in the control group. Were they screened for delirium first? Give more detail about how the participants were selected in the intervention group. Was the data collected retrospectively following patient's discharge from the medical notes? Page 10: Dummy variable? Can you explain this more clearly? Page 17 Line 10: Whilst you have attributed the difference in staff to patient ratio as one of the main of the main difference between the intervention and control ward, another major difference could be the dementia training provided between the two hospitals. You could refer to international studies on the positive effects of dementia training programs for staff working in acute hospitals. Page 18: Another limitation is that the study nurses who measured the QOL could not be blinded for the intervention and the control (i.e. the two hospital wards). Whilst it is difficult to carry out blinding in these studies, this limitation should be acknowledged.

REVIEWER	Dr Inderpal Singh Aneurin Bevan University Health Board/Honorary Senior Lecturer, School of Medicine, Cardiff University
REVIEW RETURNED	30-Apr-2019

GENERAL COMMENTS	Dear Authors It is a good study design and you have chosen an excellent research question which needs to be addressed in more detail. Acute dementia patients do have poorer clinical outcomes (https://www.mdpi.com/2308-3417/4/1/7) and we do need to consider better ways to improve QoL. My suggestions are as below Patients in the intervention group were significantly younger, significantly better function and lower CCI. It will be prudent to revise following statement in the limitation section. However, since we accounted for many different patient characteristics including functional status, comorbidities and behavioural problems, we assume that a bias due to patient selection mechanisms is rather low. I feel there is a selection bias due to natural background of two hospitals by having higher staff-patient ratio as well. Your study describe need dependent model for acute dementia patients. Best wishes

REVIEWER	Desmond O'Neill Trinity College Dublin, Ireland
REVIEW RETURNED	07-May-2019

GENERAL COMMENTS	It is good to see initiatives which focus on the needs of patients with dementia in hospital. However, the study design, comparing in a non-blinded fashion in wards in two completely different hospitals (one with Emergency Department), does not allow for a valid comparison. The specifics of the training of the staff in the intervention arm are not adequately described.

REVIEWER	Le Bihan, Etienne University of Luxembourg Luxembourg
REVIEW RETURNED	28-May-2019

GENERAL COMMENTS	General comments: The article is well written, clear and the study and analyzes answer the research questions. The weak point is obviuosly the non-random sampling and the fact that patients come from two particular hospitals, which limits the generalizability of the results and does not exclude the presence of bias. Nevertheless, the authors noted and discussed these points. If, as the authors say, this research is original, it is a good starting point for further research that may involve more hospitals. p. 8 line 55: "The QUALIDEM total score applies to all severities of dementia, so all patients' scores are comparable", this sentence seems to contradict the fact that 3 dimensions have not been measured for patients with very severe dementia. I don't understand how the total score was calculated for all patients. p. 10 line 36: Simple imputation may overestimate the accuracy of the estimated parameters. I would suggest to use data augmentation which does not have this drawback and is (normally) easy to implement with the MCMC algorithm. p 11 line 24: Could the authors give more information on the 'main diagnosis' variable: how many different diagnoses are there? do they differ between the two hospitals? what are the frequencies of these diagnoses among the sampled patients? p. 14 line 12: is there a particular reason for choosing a 89% probability? table 1: please indicate the unit of the variable "length of stay" and delete the underscores in line labels.

VERSION 1 – AUTHOR RESPONSE

Reviewer 1

1) As also recognized by the authors, the main limitation of this study is the selection of the participants coming from different hospitals. It would be appropriate to explain to the readers why was the intervention group and control group not taken from the same hospital. Could another ward in the same hospital as the intervention ward be a better option? Considering that the organizational culture can influence the quality of care and QoL of the patients, taking the intervention and control ward from the same hospital could have reduced this potential bias.

Answer #1:

The hospital from the intervention group had two wards for internal medicine, the one with the special care concept, and a second one (regular care). One aim of the hospital was the direct admission from patients with dementia into the special care ward in order to provide the “most appropriate treatment” for this particular patient group. Thus, nearly no patients with dementia were treated in the second internal medicine ward from that hospital. This is why we needed to recruit patients for the control group from another hospital. We added a paragraph to the section “Intervention Group” to explain this to the readers on page 8:

“The Protestant Hospital Alsterdorf has a second ward for internal medicine, however, patients with dementia were usually immediately transferred to the special care ward after admission to hospital. Thus, as almost no patients with dementia were treated in the second internal medicine ward, the control group was taken from another hospital.”

2) Page 7 Line 14-15 - Did the study nurses who collected the data, work in the same hospitals where the study was conducted? Was some independent monitoring been carried out to make sure that the study nurses are accurately filling in the questionnaire?

Answer #2:

Yes, study nurses were working at those hospitals where data was collected. We had no (financial and logistic) opportunities to either perform interrater reliability checks with additional study nurses or intensive monitoring of the study nurses when filling out questionnaires. However, we did plausibility checks on the data, and we checked scales like QUALIDEM for internal consistency and scalability. We added a sentence to the limitation-section on page 22 to point this out:

“However, due to financial and logistic limitations, it was not possible to monitor the complete data collection and accurate completion of questionnaires. Hence, we cannot give evidence on the interrater reliability apart from the intense training of the study nurses.”

3) Page 7 Line 44-45 – Who carried out the dementia screening in the control ward?

Answer #3:

The study nurse carried out the dementia screening in the control ward. She reported that in most cases, a dementia diagnosis was available upon hospital admission of the patients. The screening was required just for a few patients, and ambiguous cases regarding the assessment of the dementia severity were excluded, as there were enough patients with dementia to be included in the study. We added this information to the “Data collection and participants” section on page 9:

“A short dementia screening was used carried out by the study nurse to assess the severity of dementia for patients who had no clarified dementia diagnosis, in order to identify further patients who qualify for the study.”

4) Page 8 Line 3 – Who recorded the independent variables? Were these recorded by the study nurses or by the researchers?

Answer #4:

All items/variables were filled out by study nurses. The researchers had no direct contact to patients due to privacy protection. Many of the items and the required data and information respectively had been copied from the patient files (where possible and appropriate).

5) Page 6: Can you be more specific in terms of the staffing levels? What is the difference in staffing levels and staff: patient ratio between the intervention and control wards?

Answer #5:

Thanks for this comment, which refers to an important characteristic of the hospital wards. We added a sentence including this information to the section “Intervention Group” on page 9:

“With respect to the total number of full-time equivalents [FTE] nurses, the staff-patient-ratio is one FTE nurse per 39 patients.”

Furthermore, we added a paragraph including this information to the section “Control Group” on page 8:

“The staff-patient-ratio in the regular care ward is approximately one FTE nurse per 130 patients. However, since the internal medicine ward in this hospital also treats patients from the emergency ambulance, the staff-patient-ratio related to the number of patients who stayed longer in hospital (three days and more) is lower. Unfortunately, the hospital management was not willing to provide more detailed information beside the publicly available quality reports, so we cannot quantify the staff-patient-ratio exactly.”

6) Page 7: Give more detail of the inclusion/exclusion criteria used to identify the participants in the control group. Were they screened for delirium first? Give more detail about how the participants were selected in the intervention group. Was the data collected retrospectively following patient’s discharge from the medical notes?

Answer #6:

For the control group, patients who already had a dementia diagnosis were included in the study. Patients with no clarified diagnosis, who seemed to have cognitive impairments, were screened using a short three-item-test (Kaiser et al. 2014). There was no initial delirium screening. The CAM (Confusion Assessment Method) was used to assess a possible development of a delirium during the hospital stay. A possible delirium was found in about three percent of all patients during the hospital stay. We revised the paragraph in the section “Data collection and participants” on page 9 to provide more information and make this clearer to the readers:

“For the regular care ward (control group), patients who already had a diagnosed dementia or cognitive impairments were included in the study. A short dementia screening was carried out by the study nurse to assess the severity of dementia of patients who had no clarified dementia diagnosis, and to identify further patients who qualify for the study.”

For the intervention group, almost all patients were included in the study, as “dementia” was a requirement for the admission to the special care ward. We have added a sentence including this information in the “Data collection and participants” section on page 9:

“In the special care ward (intervention group) all patients were assessed because a diagnosed dementia was a requirement for admission to that hospital. Hence, the participation rate for the special care ward was about 94% and excluded only a few patients that were not responsive.”

Furthermore, complete immobility was an exclusion criterion, i.e. if patients were only lying in bed for the whole hospital stay. We added a short paragraph with the additional information on the inclusion and exclusion criteria (page 9):

“For both the intervention and control group, patients were excluded when they were completely confined to bed due to severe health-related dependency. As both care wards had no particular selection criteria for patients like age, mobility, or the main diagnosis that lead to hospital admission, no further exclusion criteria for the study were defined.”

Most of the data was collected during the hospital stay, e.g. in the “Measures” section on page 9, we write for the QUALIDEM: “After observing patients for about one week (depending on the length of stay), the study nurses rated their QoL.” Similarly, assessments like MMSE, Barthel-Index or CCI were carried out during the hospital stay of the patients. In most cases, information on data like age and gender were also recorded by the study nurses during the hospital stay, sometime this was recorded (i.e. copied from patient files) shortly after discharge.

7) Page 10: Dummy variable? Can you explain this more clearly?

Answer #7:

Thank you for pointing this out. We have revised the paragraph in the “Measures” section on page 12, which now should be clearer regarding the variable we used in our model:

“While these variables already cover many different aspects that have an effect on the QoL, we decided to add a further predictor as proxy for the intervention to the model. Therefore, we included a binary variable with two categories (“control” as reference and “intervention”) representing the two hospitals, to estimate the impact of the special care concept.”

8) Page 17 Line 10: Whilst you have attributed the difference in staff to patient ratio as one of the main of the main difference between the intervention and control ward, another major difference could be the dementia training provided between the two hospitals. You could refer to international studies on the positive effects of dementia training programs for staff working in acute hospitals.

Answer #8:

We agree that dementia training programmes are another key element to improve quality of care and quality of life for patients with dementia. Though we have mentioned the training programmes in the description of the intervention group, we indeed could have discussed this issue more in detail and therefore added a section to the discussion on page 19:

“Furthermore, dementia-specific educational programmes, as implemented in the special care ward, have positive effects on nurses regarding their interaction with patients with dementia. Trained nurses can improve their coping skills in handling challenging behaviour of these patients, and better attend to the patients’ unmet physical and psychological needs. Studies suggest that the use of both physical and chemical restraints is reduced for nurses who completed a dementia-specific training as opposed to nurses who did not complete such an educational programme. Trained nurses had better skills in providing patient-centred care and thus improving the QoL for patients with dementia.”

9) Page 18: Another limitation is that the study nurses who measured the QOL could not be blinded for the intervention and the control (i.e. the two hospital wards). Whilst it is difficult to carry out blinding in these studies, this limitation should be acknowledged.

Answer #9:

We have acknowledged the aspect of lack of blinding for the study nurses and added a paragraph to the limitations on page 22:

“Finally, due to the nature of the study design, it was not possible that study nurses in the intervention and control group were blinded. This might affect the results insofar as study nurses may have generated more generous responses for the assessment scales.”

Reviewer 2

10) Patients in the intervention group were significantly younger, significantly better function and lower CCI. It will be prudent to revise following statement in the limitation section. However, since we accounted for many different patient characteristics including functional status, comorbidities and behavioural problems, we assume that a bias due to patient selection mechanisms is rather low. I feel there is a selection bias due to natural background of two hospitals by having higher staff-patient ratio as well.

Answer #10:

Thank you for this comment. We revised the according section in the limitations on pages 20-21 and described in more detail the potential bias of the different structures between the control and intervention group, and what we tried to reduce this bias as much as possible, given our data. The section now reads:

“Our study has several limitations. One concerns the structural differences between the two hospitals. The hospital with the special care ward is much smaller than the hospital that hosted the control group. A second control group or an intervention group in a hospital of a similar size as the hospital with the regular care ward may have permitted a more distinct comparison. We tried to keep the impact of the structural differences as minimal as possible, for instance by accounting for many different patient characteristics including functional status, comorbidities and behavioural problems. Furthermore, the main diagnoses of patients were also considered in the analysis. We assume that we could at least partly adjust our analysis for a bias due to patient selection mechanisms. To validate our assumptions, we investigated to which extent the association between patient characteristics and QoL is affected by differences between the control and intervention group (details shown in Supplementary File 3). Results suggest that our data provides no strong evidence for noticeably differences between the intervention and control group regarding the association between complexity of patients’ needs and QoL. However, although we adjusted our analysis for many patient characteristics, we cannot eliminate a potential bias due to different hospital structures. In particular, the higher mean age and stronger functional limitations in the control group may indicate a selection bias in our sample. We suggest that further studies should take a second control group or a more comparable intervention group into account to gain more insight into potential biases due to structural differences of the control and intervention group.”

Furthermore, we revised another section of the limitations to emphasize that the different staff-patient-ratio is not only a “structural” difference, but an intended element of the special care concept. We think that such an intervention automatically leads to structural differences, however, modifying

hospital structures to improve QoL (and quality of care) is often a side effect of an intervention. Thus, we consider the higher staff-patient-ratio at least not completely as a selection bias. The section now reads (page 21):

“Another structural difference between the intervention and control group that certainly affects the results are the different staff-patient-ratios. In the special care ward, nurses have to care for fewer patients than in the regular care ward. Although we assume that this aspect probably has the highest impact on the outcomes in QoL, this is not a “selection bias” per se rather than a core component of the intervention. A higher staff-patient-ratio, dementia-specific training programmes, or a specific architectonical design are key elements of the special care concept, which, in their entirety, are reflected in the resulting differences between hospitals.”

Reviewer 3

11) It is good to see initiatives which focus on the needs of patients with dementia in hospital. However, the study design, comparing in a non-blinded fashion in wards in two completely different hospitals (one with Emergency Department), does not allow for a valid comparison.

Answer #11:

We agree that the different hospital structures are a limitation of our study and may introduce selection bias. Therefore, we tried to address this issue and reducing this potential bias as much as possible:

- We adjusted our regression analysis for many different patient characteristics to rule out differences between hospitals just because of very different patient populations. We substantially revised the limitations section to discuss this issue in more details, acknowledging that the different hospitals are not perfectly comparable (see also Answer #10 for further details on our revisions in this regard).

- Furthermore, we conducted additional analyses in order to check if the associations between patients’ characteristics and QoL differ by hospital (see Supplementary File 3). The results suggest that there are differences in QoL depending on the outcome of patients’ characteristics (e.g. lower QoL for lower mobility or more frequent psychotropic drug use), however, these differences do not differ significantly by hospital ward. Although this does not allow to infer that we have no selection bias in our data, we at least found no strong evidence for a potential bias. We have added a paragraph discussing this in the “limitations” section.

- We now also discuss the aspect of non-blinding in the limitations section. Please see Answer #9 for further details.

- We have also described more detailed the main diagnoses of patients, and how these differ by hospitals (see Results section on page 15 and Supplementary File 1).

- One of the main differences between the hospital wards is the staff-patient-ratio. We have added more information about this aspect to the “Methods” section on pages 7-8, and discussed this issue also in the “limitations” section.

12) The specifics of the training of the staff in the intervention arm are not adequately described.

Answer #12:

We added more details to the training programmes for employees in the special care ward and revised the section “Intervention Group” in the “Methods” section on page 7, which now reads:

“b) doctors, nurses and service staff are trained in coping with challenging behaviour and other dementia related issues, like basal stimulation or validation therapy, but also included case conferences to discuss issues with current patients; duration of training courses and case conferences was about one hour and were provided on a monthly basis by external instructors; additionally, twice per year, an internal training course was offered for employees, lasting for half a day;”

Reviewer 4

13) p. 8 line 55: "The QUALIDEM total score applies to all severities of dementia, so all patients’ scores are comparable", this sentence seems to contradict the fact that 3 dimensions have not been measured for patients with very severe dementia. I don't understand how the total score was calculated for all patients.

Answer #13:

The recommendations from the authors of the QUALIDEM is to build a sum score of all subscales (i.e. the sum score of the six scales for patients with very severe dementia, and a sum score from the nine subscales for patients with mild to severe dementia). The sum score is then “normalized” to a consistent range, from 0 to 100. In case of the nine subscales, each of the 37 items can get a score from 0 to 3, so the total score ranges from 0 to 111. The sum score was then divided by 111 and multiplied by 100 in order to get a score ranging from 0 to 100. The same was performed for the sum score of the six subscales, which ranges from 0 to 54, so this score was multiplied by 100 and divided by 54. We revised the paragraph to describe the normalization of the QoL-index more clearly on pages 10-11:

“For regression analyses, a QoL index was calculated by summing up and normalizing the QUALIDEM subscales (six subscales for patients with very severe dementia, nine subscales for the remaining patients) to a range from 0 to 100 points. A higher score indicates better QoL. Due to normalization of the QUALIDEM total score for all severities of dementia, all patients’ scores are consistent and comparable”.

14) p. 10 line 36: Simple imputation may overestimate the accuracy of the estimated parameters. I would suggest to use data augmentation which does not have this drawback and is (normally) easy to implement with the MCMC algorithm.

Answer #14:

The brms-package we used to fit the model allows for multiple imputation by a flexible posterior sample draw stacking method, or by jointly model the missing values, so missing value could be imputed during the model-fitting process. However, we decided to apply a multiple imputation algorithm (which used several iterations per imputation step) because we did not impute a complete sum score per patient, but rather single items (item-level data) of a scale. The sum score was computed after imputing missing values for single items. This means, we did not impute missing values during model fitting, because for single item imputation, we would have had to include the single items instead of the sum score itself in the model. Imputing missing values during model fitting

would only make sense when applied to the sum score. We explain the reason for this approach using an example with the MMSE: There were patients who had three missing items in the MMSE. If their sum score without these three missing items was already 20, the final sum score has a possible range from 20 to 23. Imputing the missing values of the single items (item-level data) and computing the sum score afterwards gives a feasible result within that range. However, setting the MMSE sum score for these patients to missing and imputing the complete MMSE sum score may result in sum scores being lower than 20 or larger than 23, and thus being less precise. We imputed missing data both on item-level data and on the complete MMSE sum score and compared the results from both approaches (see Figure 1 below in this response letter). While there are some patients where the imputed MMSE sum score does not differ between these two approaches, there are also many patients where we see large deviations, especially implausible deviations (based on what could be expected from non-missing data, i.e. available information we had for single items). Burns et al. recommend imputing missing item-level data as an effective method, so we decided to take this route.

15) p 11 line 24: Could the authors give more information on the 'main diagnosis' variable: how many different diagnoses are there? do they differ between the two hospitals? what are the frequencies of these diagnoses among the sampled patients?

Answer #15:

Thanks for this comment. We added more information about the main diagnosis to the section “Measures” on page 11:

“Details about the distribution of the main diagnoses among patients and by hospitals are shown in the Supplementary File 1. If a main diagnosis was mentioned no more than one time in both hospital wards, it was recoded into the category “other”. The final variable “main diagnosis” comprised 20 different diagnoses.”

Furthermore, we added a description of the distribution to the “Results” section on page 15:

“In most cases, the distribution of main diagnoses of patients were comparable between the two hospital wards (see Supplemental File 1). Most frequent were pneumonia (13.5% in the intervention group and 11.9% in the control group), a worsening medical condition of patients (8.7% and 7.2%) or exsiccosis (4.8% and 6.7%). Noticeable differences between the two wards were found in urinary tract infections (UTI) (9.9% in the intervention group and 3.1% in the control group) or dyspnoea (1.2% and 7.8%).”

Finally, since the presentation of the distribution of main diagnoses is rather complex, we decided to add a figure as Supplementary File 1 (see also Figure 2 in this response letter, below). We refrained from integrating this figure into the main manuscript, because this would have put too much emphasis on this single variable.

16) p. 14 line 12: is there a particular reason for choosing a 89% probability?

Answer #16:

The 89% intervals are more stable than, for instance, 95% intervals (Kruschke 2015). Moreover, we follow a suggestion of McElreath (2015), who suggests using non-standard percents “to constantly remind the reader that conventions like 95% and 5% are arbitrary. Furthermore, boundaries [for Bayesian posterior distributions, DL] are meaningless. There is a continuous change in probability as we move away from the expected value. So one side of the boundary is almost equally probable as

the other side. Also, 89 is a prime number, [which makes it easy to remember]. That doesn't mean it is a better choice than any other number here, but it's no less silly than using a multiple of 5".

17) table 1: please indicate the unit of the variable "length of stay" and delete the underscores in line labels.

Answer #17:

We have removed the underscores and added the unit of the length of stay (days) to the table.

References:

Burns RA, Butterworth P, Kiely KM, et al. Multiple imputation was an efficient method for harmonizing the Mini-Mental State Examination with missing item-level data. Journal of Clinical Epidemiology 2011;64:787–93. doi:10.1016/j.jclinepi.2010.10.011

Kaiser AK, Hitzl W, Iglseder B. Three-question dementia screening: Development of the Salzburg Dementia Test Prediction (SDTP). Zeitschrift für Gerontologie und Geriatrie 2014;47:577–82. doi:10.1007/s00391-013-0568-7

Kruschke JK. Doing Bayesian data analysis: a tutorial with R, JAGS, and Stan. 2. ed. Amsterdam: Elsevier, Academic Press 2015.

McElreath R. Statistical rethinking: a Bayesian course with examples in R and Stan. Boca Raton London New York: CRC Press, Taylor & Francis Group 2016.

Figure 1: Differences between missing-value-imputed MMSE-Scores, based on imputation for total sum score and imputation on item-level data

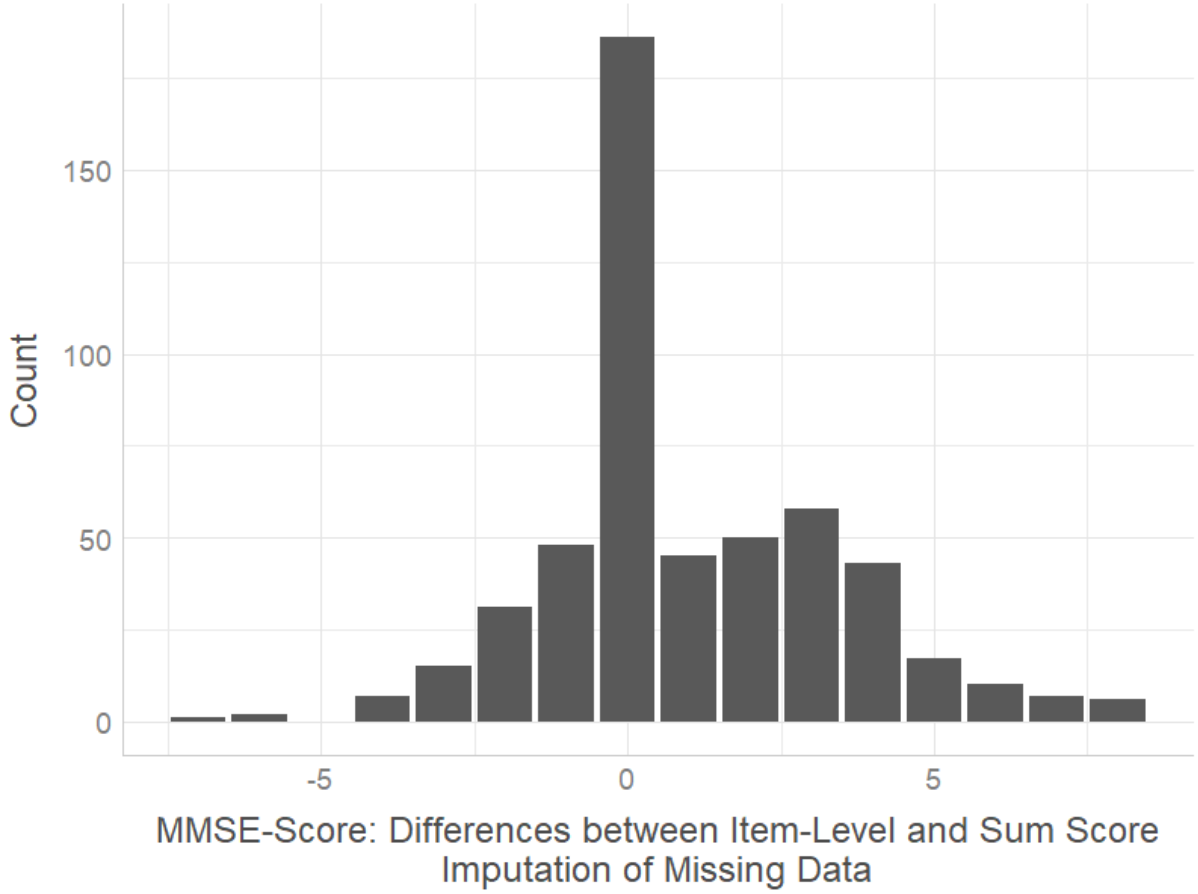
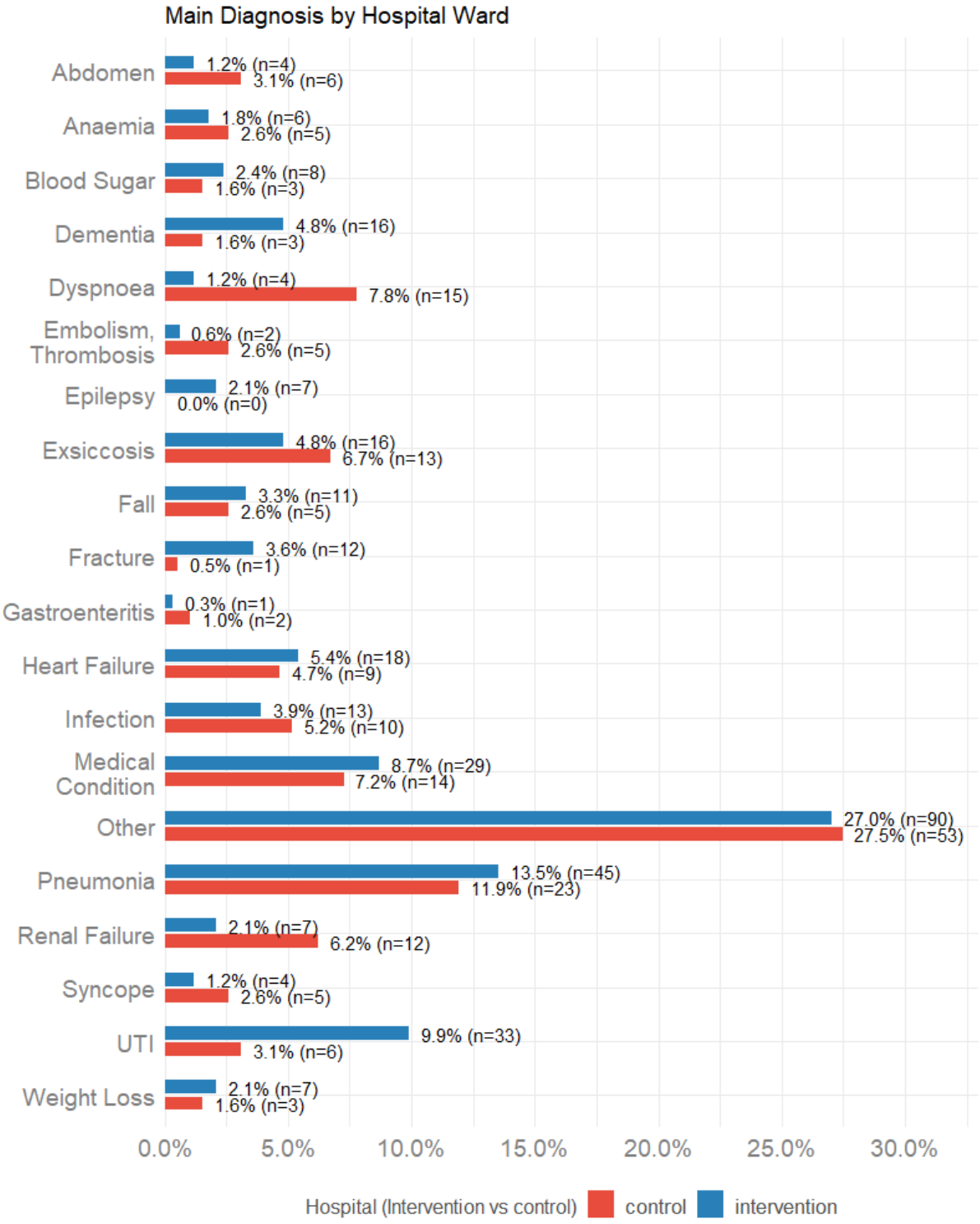


Figure 2: Distribution of Main Diagnoses from Patients by Hospital Wards*



* Figure shows the main diagnoses of patients; for a small proportion of patients, the “main diagnosis” according to the patient files was “Dementia”, but for most patients, dementia was a secondary diagnosis

VERSION 2 – REVIEW

REVIEWER	Dr Anthony Scerri Uni of Malta
REVIEW RETURNED	27-Jun-2019

GENERAL COMMENTS	Thank you for considering my suggestions and comments

REVIEWER	Le Bihan, Etienne University of Luxembourg
REVIEW RETURNED	16-Jul-2019

GENERAL COMMENTS	The calculation of the total qualidem score is now clear. Nevertheless, I have some reservations about this calculation since it is the same as to assigning to the missing scores of patients with very severe dementia the average score obtained on the other 6 sub-scales. Although this corresponds to the recommendations of the authors of the scale, this can be seen as a missing value issue and a simple imputation procedure, which can be biased and underestimate the measurement error variance for the observations of patients with very severe dementia. If this is feasible, I would suggest calculating the total score from the subscales in the MCMC algorithm and leaving missing the 3 subscales that are not measured in patients with very severe dementia (obviously for this group only), so they can be treated with data augmentation or any more adequate method for handling missing data. The answers to my other remarks are satisfactory and I thank the authors for the details of the diagnoses.

VERSION 2 – AUTHOR RESPONSE

Reviewer 4

1) The calculation of the total QUALIDEM score is now clear. Nevertheless, I have some reservations about this calculation since it is the same as to assigning to the missing scores of patients with very severe dementia the average score obtained on the other 6 sub-scales. Although this corresponds to the recommendations of the authors of the scale, this can be seen as a missing value issue and a simple imputation procedure, which can be biased and underestimate the measurement error variance for the observations of patients with very severe dementia. If this is feasible, I would suggest calculating the total score from the subscales in the MCMC algorithm and leaving missing the 3 subscales that are not measured in patients with very severe dementia (obviously for this group only), so they can be treated with data augmentation or any more adequate method for handling missing data.

Answer #1:

We agree that omitting three of nine subscales is equivalent to mean value imputation for these three subscales and understand the concerns related to such an approach. In order to see to which extent results would change when the missing data for the three subscales for patients with very severe dementia will be imputed, we calculated another regression model, using a new QUALIDEM total score as outcome. This new QUALIDEM total score is based on imputation of missing values for all

nine subscales for patients with very severe dementia. The results for the new regression model is shown in Table 1 below. Table 1 also shows the comparison to the main model used in our manuscript. Results do not differ substantially. This suggests that using just the six subscales for patients with very severe dementia had no significant bias for our analyses and results.

We furthermore compared QUALIDEM total scores for the subgroup of patients with very severe dementia only (n=126). The mean value for the regular QUALIDEM total score for the subgroup of patients with very severe dementia was 38.5 (SD = 17.7), while the mean value for the new QUALIDEM total score (where missing data for the three omitted subscales have been imputed) was 40.8 (SD = 13.0) (see also Figure 1 in this response). The difference between both QUALIDEM total scores for this subgroup is not significant (t-test, p = 0.5745).

Since the results of the regression models do not remarkably differ, we decided to keep our first model in the main manuscript and to follow the recommendations of the QUALIDEM authors. Another reasoning for this decision is also to provide comparability of results with other studies that may use the QUALIDEM as outcome in statistical analyses. Nevertheless, we think that your remark is an important statistical issue that should be addressed or at least be considered in future research using the QUALIDEM, and thus we decided to publish these results in the Supplemental File 5.

We also added a paragraph to the limitation-section on page 22 to point this out:

“Another debatable issue regarding the QUALIDEM concerns the computation concept of the total score for patients with very severe dementia. We followed the QUALIDEM authors’ instruction to use only six of the nine subscales to calculate the total score for this group. Technically, this is similar to mean value imputation for the missing scores of the three omitted subscales. This, however, may result in biased and/or underestimated measurement error variance for this group. Therefore, we also calculated a regression model with a QUALIDEM total score based on imputation for missing values for all nine subscales for patients with very severe dementia (see Supplementary File 5). In the results section, we have provided the analyses as suggested by the QUALIDEM authors for comparability reasons. In order to meet different views on the computation concept, we also provide the results of the alternative analysis in the Supplementary File 5. These are very similar to the first analysis and do not differ significantly.”

Table 1: Comparison of Models for two different QoL-Scores, N=526

	New Model with full imputed QoL-Score for very severe dementia			Main Model		
Term	Estimate	SE	89% HDI	Estimate	SE	89% HDI
Length of Stay	-0.0	0.1	-0.2 – 0.1	-0.1	0.1	-0.2 – 0.1
Age	0.8	0.5	-0.0 – 1.6	1.2	0.5	0.4 – 2.1
Moderate Dementia	1.1	1.9	-1.8 – 4.1	1.2	2.0	-1.8 – 4.6
Severe Dementia	-0.3	1.9	-3.2 – 2.7	-0.4	2.0	-3.6 – 2.7
Female	0.6	1.0	-1.0 – 2.2	0.2	1.1	-1.6 – 1.9
Barthel Score	2.0	0.4	1.3 – 2.7	2.0	0.5	1.3 – 2.8
Physical Restraints (yes)	-4.1	1.3	-6.1 – -2.1	-4.9	1.2	-7.0 – -2.8
Special Care Ward (Intervention)	5.2	1.1	3.4 – 6.9	5.7	1.2	3.8 – 7.6
PAS-Score	-2.7	0.2	-3.0 – -2.4	-2.9	0.2	-3.2 – -2.7
Charlson’s Comorbidity Index	0.0	0.3	-0.5 – 0.5	-0.1	0.3	-0.6 – 0.5

Psychotropic Drug Use (yes, as-needed)	-4.3	1.3	-6.3 – -2.3	-4.4	1.4	-6.5 – -2.1

Figure 1: Distribution of QUALDEM total scores for patients with very severe dementia (n=126), regular QUALIDEM total score and new QUALIDEM total score with imputed data for all nine subscales for patients with very severe dementia