## PEER REVIEW HISTORY

BMJ Open publishes all reviews undertaken for accepted manuscripts. Reviewers are asked to complete a checklist review form (**http://bmjopen.bmj.com/site/about/resources/checklist.pdf**) and are provided with free text boxes to elaborate on their assessment. These free text comments are reproduced below.

## ARTICLE DETAILS

| | |
|---|---|
| **TITLE (PROVISIONAL)** | Assessing the Readability and Patient Comprehension of Rheumatology Medicine Information Sheets: A Cross-Sectional Health Literacy Study |
| **AUTHORS** | Oliffe, Michael; Thompson, Emma; Johnston, Jenny; Freeman, Dianne; Bagga, Hanish; Wong, Peter |

## VERSION 1 – REVIEW

| | |
|---|---|
| **REVIEWER** | Reviewer name: Dr William Hunt<br>Institution and Country: South Warwickshire NHS Foundation Trust<br>Competing interests: None Declared. |
| **REVIEW RETURNED** | 13-Jun-2018 |

| | |
|---|---|
| **GENERAL COMMENTS** | Title:<br>As per journal guidance it advises that all articles should include the research question and study design.<br>In this case it would be reasonable to add that it is cross-sectional in a nature.<br>Abstract:<br>Structured abstract conforms to journal specification.<br>Page 3: line 32. It would be worthwhile being clear in the objectives that patient comprehension was assessed in Australia only using the Australian MIS and not assessed in the other counties.<br>Page 3: line 23. In regards to the randomisation it was not quite clear about the methodology. Please see comments in the methodology section and clarify.<br>Page 3: line 35. In the results section it is suggested that the mean grade level for the MIS across the countries was as reported. However, from which instrument is this taken from Gunning Fog or SMOG? Please clarify.<br>Page 3. Line 42. “Overall, 10-79% of patients failed to correctly answer all five simple multiple choice questions assessing MIS comprehension.” Please make this clearer if possible.<br>Page 3: line 48. In the conclusion it is stated that the mean readability of the MIS from Australia and the UK was greater than 8. However, is this not also true of Canada from the results section? Please clarify. |

Introduction:

Page 5: line 42. The Australian reference (12) to the readability of healthcare related information suggests that information should be grade 8 level. However, is there any national guidance since this is from SA?

Furthermore, the other reference (11) regarding the level of grade 8 for health related information appears to be a secondary reference for the American Medical Association (AMA), National Institutes of Health (NIH) and Centers for Disease Control and Prevention (CDC). On going to the primary sources, the AMA suggest no greater than a sixth-grade reading level, the NIH between 7th or 8th grade level, and the reference scrutinised for the CDC did not clearly give a specific grade.

I appreciate that it difficult given that there are multiple conflicting guidelines in the literature and most these specific target levels for readability for health information originate from the US; however, it is important to be clear why a particular value was chosen. Furthermore, if it is to be used for looking at percentage compliance in your particular level, what constitutes a level needs to be defined (e.g. less then 6.9 based on the AMA guidance).

Page 5: line 49. Are you able to define what “low literacy skills” constitutes (i.e. reading age)?

Page 5: line 53. Reference 20 appears to be a secondary reference to a 1994 government study. Is there anything more up to date? If not please reference primary reference and articulate the age of the information may not represent the current situation.

Page 5: line 53. “48% of Canadians fall into the lowest two literacy categories and 26% lack skills to” should be clear that the 26% is part of the 48% and not an additional 26%.

Page 5: line 55. The reference 21 appears to be a secondary reference to a 2003 study. There is a more up to date data here:

https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/623399/11-1367-2011-skills-for-life-survey-findings.pdf

This demonstrates improved readability. Suggest saying “just under one in six adults have the literacy level of an 11-year-old” or stating direct percentage.

Page 6: line 18. Considering the article for reference 24 is over 15 years old, suggest making this clear.

Methods:

Page 7. Line 4. From the appendices it appears that multiple readability tests were used in the analysis. Why in the main paper are only 3 described in full. Please provide rationale.

Page 7: line 9. For Flesch Reading Ease, it is possible to also gain minus scores and scores over 100. The standard scale is between 0-100.

Page 7: line 16. Gunning Fog from my understanding is based on mean sentence length in addition to the number of complex words (more than 3 syllables), not the number of sentences only. Please clarify.

In general: as per STROBE. How did you arrive at the number of patients to enrol in the study? This does not appear to be mentioned.

Page 7: Line 39. Patients were randomly selected. However, in line 48 it appears consecutive sampling was employed. Please clarify. Was it that they were randomly selected and then if they consented their comprehension was assessed for the MIS on their clinic day? If so would you be able to make this clearer. Furthermore, if random selection obtained, how was this undertaken?

Page 7: Line 47. To clarify was only one MIS was assessed per patient? No patients were assessed for more than one MIS?

Page 8. Line 34. Given that you likely tested for normality using a statistical test, it is worth including the test used.

Page 8. Line 34. Furthermore, for the data that is non-normally distributed the medians should be presented also. It is unclear which data was not normally distributed in the paper.

Results:

In general: for figures do you think it would be of benefit to provide some of the sample graphs generated in readability studio for the article?

Page 9. Line 9. “The mean grade level for the ARA MIS calculated..” was this with SMOG or Gunning Fog? Please clarify.

Page 9. Line 10. Please highlight the number of MIS to develop the mean.

Page 9: Line 30: It would be worthwhile giving the baseline characteristics of the patients included at the in prose as well as in the table.

Page 9. Line 35. Please include the number of patients assessed for comprehension per MIS.

Page 9. Line 38. Was there any correlation between correct answers regarding the MIS and education level of the participants?

Page 10. Line 15. How were these differences assessed? Was it the mean of all the respective MIS? If so using Gunning Fog or SMOG or Flesch? Also was comparing means acceptable, since some of the data was not normal – would medians have been more representative of the data? This area needs to be clarified.

Page 10. Line 30. Please define “complex words”.

Discussion

Page 10: line 44. Please review the reference re eight-grade level.

| | |
|---|---|
| | Page 11. Line 37. Very valid.<br>Page 12. Line 7. Agree.<br>Page 12. Line 42. Also do you think it is reasonable to highlight that the ARA MIS were compared only to a sample of UK and Canadian MIS. Was this sample you obtained of the other countries MIS representative of their entire MIS available? Please comment. Does this affect the generalizability of the findings?<br>Conclusion<br>Page 13: line 18. The mean level for Canada (SMOG or Gunning Fog?) also exceeded the value of 8 and therefore also this needs to be mentioned. All mean score for MIS across the countries exceeded the stated guidance grade-level. Please clarify.<br>References:<br>The link supplied for reference 12 appears to be not accessible. Please provide an alternative / up to date link. |

| | |
|---|---|
| **REVIEWER** | Reviewer name: Morten Pilegaard<br>Institution and Country: Dept. Communication & Culture, Aarhus University, Denmark<br>Competing interests: None declared |
| **REVIEW RETURNED** | 03-Aug-2018 |

| | |
|---|---|
| **GENERAL COMMENTS** | Methods: Page 6, lines 3-18: The authors acknowledge part of the criticism raised against readability formulas (quantitative measurements of lexical and syntactical complexity), but, surprisingly, still use these 50+-year old formulas, even if they are imperfect predictors of text readability and understandability or comprehension. First, such formulas largely ignore situational or contextual parameters, which lie at the heart of any act of communication. Any account of readability failing to recognize the social and cultural construction of MIS as an act of communication falls short of adequacy. Although some account is given of contextual factors, the results would have been more useful had we had more detailed socio- and demographic information on e.g. age (as elderly are known to have lower health literacy levels than younger patients), disease stage, years with disease, etc., as these factors profoundly affect patients' comprehension. Second, the process of providing information and obtaining consent takes place within the context of a conversation between a healthcare professional and a potential participant in that research. This process is essentially dialogical and must take into account the linguistic complexity of the act of communication, the fact that the act of asking the patient to read the MIS usually follows after a series of communicative acts which together form a polyphonic, heteroglossial discourse presenting many, potentially conflicting, perspectives. Third, because readability formulas are composed of the variables of words and sentence length, they characterize the surface structure of a text, not its deeper syntactic and semantic structures. These shortcomings should to be addressed in more detail (re. Clerehan, Buchbinder & Moodie (ref. 36); not least since they also affect the simpler issue of assessing patients' literal comprehension. |

## VERSION 1 – AUTHOR RESPONSE

Reviewer: 1

Reviewer Name: Dr William Hunt

As per journal guidance it advises that all articles should include the research question and study design.

In this case it would be reasonable to add that it is cross-sectional in a nature.

We had already done this in the Abstract under “Design” but have added this to the title as follows: “Assessing the Readability and Patient Comprehension of Rheumatology Medicine Information Sheets: A Cross-Sectional Health Literacy Study”

Abstract: Structured abstract conforms to journal specification.

Page 3: line 32. It would be worthwhile being clear in the objectives that patient comprehension was assessed in Australia only using the Australian MIS and not assessed in the other counties.

We have added “Australian” to this line to clarify the geographic origin of the study population. Page 4, line 23 and line 4 of “Outcome Measures” in the Abstract have also been amended to reflect this.

Page 3: line 23. In regards to the randomization it was not quite clear about the methodology. Please see comments in the methodology section and clarify.

Please see our responses under the “Methods” section below.

Page 3: line 35. In the results section it is suggested that the mean grade level for the MIS across the countries was as reported. However, from which instrument is this taken from Gunning Fog or SMOG? Please clarify.

Mean grade level was the mean of FORCAST, Gunning Fog and SMOG grade level. Tables 1, 4, 5 have been amended accordingly. Line 2-3 of para. 1 of “Results” has been added to clarify this.

Page 3. Line 42. “Overall, 10-79% of patients failed to correctly answer all five simple multiple choice questions assessing MIS comprehension.” Please make this clearer if possible.

We have clarified this.

Page 3: line 48. In the conclusion it is stated that the mean readability of the MIS from Australia and the UK was greater than 8. However, is this not also true of Canada from the results section? Please clarify.

We agree and have amended this line accordingly.

Introduction:

Page 5: line 42. The Australian reference (12) to the readability of healthcare related information suggests that information should be grade 8 level. However, is there any national guidance since this is from SA?

As per the Australian Commission on Safety and Quality in Health Care, there is no national recommendation about this.

Furthermore, the other reference (11) regarding the level of grade 8 for health related information appears to be a secondary reference for the American Medical Association (AMA), National Institutes of Health (NIH) and Centers for Disease Control and Prevention (CDC). On going to the primary sources, the AMA suggest no greater than a sixth-grade reading level, the NIH between 7th or 8th grade level, and the reference scrutinised for the CDC did not clearly give a specific grade.

I appreciate that it difficult given that there are multiple conflicting guidelines in the literature and most these specific target levels for readability for health information originate from the US; however, it is important to be clear why a particular value was chosen. Furthermore, if it is to be used for looking at percentage compliance in your particular level, what constitutes a level needs to be defined (e.g. less then 6.9 based on the AMA guidance).

We have clarified this to better reflect the variation in recommendations (para. 3 of the “Introduction”). Eighth grade level was chosen as it is the upper limit of the recommended level of difficulty.

Page 5: line 49. Are you able to define what “low literacy skills” constitutes (i.e. reading age)?

We have clarified this as “minimum required for individuals to meet the complex demands of everyday life”.

Page 5: line 53. Reference 20 appears to be a secondary reference to a 1994 government study. Is there anything more up to date? If not please reference primary reference and articulate the age of the information may not represent the current situation.

We have updated the reference to a more recent one.

Page 5: line 53. “48% of Canadians fall into the lowest two literacy categories and 26% lack skills to” should be clear that the 26% is part of the 48% and not an additional 26%.

We have removed this section as we have used more recent data.

Page 5: line 55. The reference 21 appears to be a secondary reference to a 2003 study. There is a more up to date data here:

https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/623399/11-1367-2011-skills-for-life-survey-findings.pdf

This demonstrates improved readability. Suggest saying “just under one in six adults have the literacy level of an 11-year-old” or stating direct percentage.

We appreciate the reference and have updated accordingly (reference 26).

Page 6: line 18. Considering the article for reference 24 is over 15 years old, suggest making this clear.

We have addressed this in the text as “ In 2002,”

Methods:

Page 7. Line 4. From the appendices it appears that multiple readability tests were used in the analysis. Why in the main paper are only 3 described in full. Please provide rationale.

We have expanded para. 2 of “Methods” and added para. 3 to “Methods” to address this.

Page 7: line 9. For Flesch Reading Ease, it is possible to also gain minus scores and scores over 100. The standard scale is between 0-100.

This has been added to para. 2 of “Methods”.

Page 7: line 16. Gunning Fog from my understanding is based on mean sentence length in addition to the number of complex words (more than 3 syllables), not the number of sentences only. Please clarify.

This is correct. We have amended the following sentence in para. 2 of “Methods” accordingly: “The Gunning Fog formula calculates grade level and reader age based on number of sentences, their mean length and number of complex words (containing three or more syllables).”

In general: as per STROBE. How did you arrive at the number of patients to enrol in the study? This does not appear to be mentioned.

We aimed for a study sample of 100. A power calculation was not necessary as we were not comparing groups.

Page 7: Line 39. Patients were randomly selected. However, in line 48 it appears consecutive sampling was employed. Please clarify. Was it that they were randomly selected and then if they consented their comprehension was assessed for the MIS on their clinic day? If so would you be able to make this clearer. Furthermore, if random selection obtained, how was this undertaken?

We have clarified this as follows: “All consecutive patients scheduled for a randomly selected consulting day were contacted via telephone…” Suitable consulting days were selected depending on availability of the investigators.

Page 7: Line 47. To clarify was only one MIS was assessed per patient? No patients were assessed for more than one MIS?

We have clarified this as follows “ about the content of the one ARA MIS….” No patients were assessed on more than one MIS to avoid patient fatigue and inconvenience to them.

Page 8. Line 34. Given that you likely tested for normality using a statistical test, it is worth including the test used.

This was visually assessed using GraphPad Prism 6.

Page 8. Line 34. Furthermore, for the data that is non-normally distributed the medians should be presented also. It is unclear which data was not normally distributed in the paper.

The only parameter that was not normally distributed was median total score when literal comprehension was assessed using five multiple choice questions (median total score 4/5) as contained in Table 3.

Results:

In general: for figures do you think it would be of benefit to provide some of the sample graphs generated in readability studio for the article?

While this is easy to do, we don’t think it is worthwhile as they would not add much to the manuscript. We are also concerned about manuscript length - already six Tables, two Figures and 10 attachments (five comprehension question sheets and five MIS).

Page 9. Line 9. “The mean grade level for the ARA MIS calculated…” was this with SMOG or Gunning Fog? Please clarify.

We have added the following line to clarify this: “These were obtained by calculating the mean of the FORCAST, Gunning Fog and SMOG mean grade level and reading age.”

Page 9. Line 10. Please highlight the number of MIS to develop the mean.

See reply to the previous query.

Page 9: Line 30: It would be worthwhile giving the baseline characteristics of the patients included at the in prose as well as in the table.

As requested, we have added the following line:

“Mean age of participants was 60 ± 13.2 (mean ± SD) years, with 71/95 (75%) females and 24/95 (25%) males (Table 3). Only 19/95 (20%) had a university degree (Table 3).”

Page 9. Line 35. Please include the number of patients assessed for comprehension per MIS.

While it is generally best to avoid stating data in prose when it is contained in a Table, we have lengthened this section to include patient numbers.

Page 9. Line 38. Was there any correlation between correct answers regarding the MIS and education level of the participants?

This is an excellent point which we had not considered. In fact, there was, and following further analysis, we have added the following line “Highest level of education achieved (r=0.33, p =0.001) and age (r= -0.3, p=0.0002) correlated moderately strongly with a higher comprehension score.” An additional line has been added to “Statistical analyses” and to the “Discussion” (para.1, line 8).

Page 10. Line 15. How were these differences assessed? Was it the mean of all the respective MIS? If so using Gunning Fog or SMOG or Flesch? Also was comparing means acceptable, since some of the data was not normal – would medians have been more representative of the data? This area needs to be clarified.

Grade levels were normally distributed and so means were compared using Student’s t-test. We have added “mean” to this line for clarification.

Page 10. Line 30. Please define “complex words”.

We have added “containing three or more syllables” for clarification.

Discussion

Page 10: line 44. Please review the reference re eight-grade level.

We have updated the references and removed secondary references.

Page 11. Line 37. Very valid.

Page 12. Line 7. Agree.

Page 12. Line 42. Also do you think it is reasonable to highlight that the ARA MIS were compared only to a sample of UK and Canadian MIS. Was this sample you obtained of the other countries MIS representative of their entire MIS available? Please comment. Does this affect the generalizability of the findings?

We have added “of a sample of commonly prescribed Rheumatology medications” and “These 10 MIS were representative of the MIS available on both these websites.” to the last para. of “Assessment of readability” under “Methods”.

Conclusion

Page 13: line 18. The mean level for Canada (SMOG or Gunning Fog?) also exceeded the value of 8 and therefore also this needs to be mentioned. All mean score for MIS across the countries exceeded the stated guidance grade-level. Please clarify.

We have added “and Canada” to this line for clarification.

References:

The link supplied for reference 12 appears to be not accessible. Please provide an alternative / up to date link.

We have updated the link (now reference 15).

Reviewer: 2

Reviewer Name: Morten Pilegaard

Methods: Page 6, lines 3-18: The authors acknowledge part of the criticism raised against readability formulas (quantitative measurements of lexical and syntactical complexity), but, surprisingly, still use these 50+-year old formulas, even if they are imperfect predictors of text readability and understandability or comprehension.

First, such formulas largely ignore situational or contextual parameters, which lie at the heart of any act of communication. Any account of readability failing to recognize the social and cultural construction of MIS as an act of communication falls short of adequacy. Although some account is given of contextual factors, the results would have been more useful had we had more detailed socio- and demographic information on e.g. age (as elderly are known to have lower health literacy levels than younger patients), disease stage, years with disease, etc., as these factors profoundly affect patients’ comprehension.

While we agree these are old readability formulae with recognized limitations, they are still in widespread use - see Refs 18, 32, 34, 35. As we have pointed out in the “Results” section, “As the validity of the above readability assessment measures has been questioned”…..” we proceeded to assess patient literal comprehension of the ARA MIS”. It is precisely because of their limitations that we proceeded to assess the most important outcome – direct patient comprehension. This is something which had not usually been done - see Refs 18, 32, 34, 35.).

As outlined in Table 3, the age and highest level of education achieved were recorded and as discussed above in our response to Reviewer 1 - “Highest level of education achieved (r=0.33, p =0.001) and age (r= -0.3, p=0.0002) correlated moderately strongly with a higher comprehension score.”

Second, the process of providing information and obtaining consent takes place within the context of a conversation between a healthcare professional and a potential participant in that research.

This process is essentially dialogical and must take into account the linguistic complexity of the act of communication, the fact that the act of asking the patient to read the MIS usually follows after a series of communicative acts which together form a polyphonic, heteroglossial discourse presenting many, potentially conflicting, perspectives.

This is all correct. However, this is the usual way that patient recruitment occurs for clinical trials/research. We were careful to ensure the investigators (MO, ET) who contacted potential study participants were not involved in clinical care of the patients.

Third, because readability formulas are composed of the variables of words and sentence length, they characterize the surface structure of a text, not its deeper syntactic and semantic structures. These shortcomings should to be addressed in more detail (re. Clerehan, Buchbinder & Moodie (ref. 36); not least since they also affect the simpler issue of assessing patients' literal comprehension.

We agree and have expanded our critique of the formulae in the "Discussion". Despite their limitations, we believe they do have some use in the development of written patient information.

Authors must include a statement in the methods section of the manuscript under the sub-heading 'Patient and Public Involvement'.

This has been added. We have thanked all patient participants under "Acknowledgements".

We trust these changes meet with your approval and look forward to hearing from you.

## VERSION 2 – REVIEW

| | |
|---|---|
| **REVIEWER** | Reviewer name: Dr William Hunt<br>Institution and Country: Dermatology Registrar, Bristol Royal Infirmary, United Kingdom<br>Competing interests: None declared |
| **REVIEW RETURNED** | 01-Oct-2018 |

| | |
|---|---|
| **GENERAL COMMENTS** | Many thanks for your comments. I have reviewed the paper and your comments.<br><br>I have a few queries still. I have listed them in order:<br><br>1. In regards to the random selection of day for consecutive sampling, did you use a particular approach for this?<br>2. I would be slightly cautious of using a mean of several instruments (SMOG, Gunning Fog and Forcast) for the data (Pg 3, line 39). I can see why you have done a mean of means, however, it is rather an imprecise value and might not always be representative of the data. As you mention yourself on Pg 7, line 48 "Consequently, SMOG may produce grade level scores one to two grades higher than other formulae". The instruments are somewhat heterogenous. I think it would be worthwhile just signposting in the results section that the means of the different instruments are available in the tables (perhaps at Pg 11, line 13) and perhaps highlighting the pros / cons of the approach briefly in the discussion.<br>3. Regarding normality testing to ensure the data is parametric, is there a particular reason you assessed it visually as opposed to testing for normality statistically? For example the Shapiro–Wilk test. This would be more robust methodologically.<br>4. The link for reference 15 does link to the source in question. |

| | |
|---|---|
| | 5. For the student t test (Pg 10, line 25) was this paired or unpaired?<br>Many thanks. You have already answered most of my queries and adjusted the manuscript accordingly. |

| | |
|---|---|
| **REVIEWER** | Reviewer name: Morten Pilegaard<br>Institution and Country: Aarhus University, Denmark<br>Competing interests: None declared |
| **REVIEW RETURNED** | 09-Oct-2018 |

| | |
|---|---|
| **GENERAL COMMENTS** | Thanks for reviewiing the paper and taking my suggested comments into account. I find the paper publishable. Found a few typos (please see track change markings in enclosed manuscript)<br>The reviewer also provided a marked copy with additional comments. Please contact the publisher for full details. |

## VERSION 2 – AUTHOR RESPONSE

Reviewer: 1

Reviewer Name: Dr William Hunt

Institution and Country: Dermatology Registrar Bristol Royal Infirmary United Kingdom

1. In regards to the random selection of day for consecutive sampling, did you use a particular approach for this?

No.

2. I would be slightly cautious of using a mean of several instruments (SMOG, Gunning Fog and Forcast) for the data (Pg 3, line 39). I can see why you have done a mean of means, however, it is rather an imprecise value and might not always be representative of the data. As you mention yourself on Pg 7, line 48 "Consequently, SMOG may produce grade level scores one to two grades higher than other formulae". The instruments are somewhat heterogenous. I think it would be worthwhile just signposting in the results section that the means of the different instruments are available in the tables (perhaps at Pg 11, line 13) and perhaps highlighting the pros / cons of the approach briefly in the discussion.

As suggested, in para. 1 of “Results” we have added “Due to the heterogeneity of these instruments, the means of each of these measures are available in the relevant Table”.

3. Regarding normality testing to ensure the data is parametric, is there a particular reason you assessed it visually as opposed to testing for normality statistically? For example the Shapiro–Wilk test. This would be more robust methodologically.

It’s usually visually obvious if a parameter is normally distributed when plotted in a statatistics programme such as Stata. If there’s any doubt about this, I tend to err on the side of conservatism and use a non- parametric test, eg Kruskal-Wallis or Wilcoxon Rank Sum.

4. The link for reference 15 does link to the source in question.

I assume this means the link does not link to the source in question.

Depending on the security settings of the computer used, the link may not work if clicking directly on the hyperlink. However, if the hyperlink is pasted directly into the browser URL space, the link definitely works. (I've just done it).

5. For the student t-test (Pg 10, line 25) was this paired or unpaired?

Unpaired. We have added this to the text.

Many thanks. You have already answered most of my queries and adjusted the manuscript accordingly.

We thank the reviewer for his helpful comments.

Reviewer 2:

Reviewer Name: Morten Pilegaard

Institution and Country: Aarhus University, Denmark

Please state any competing interests or state 'None declared': None declared

Thanks for reviewing the paper and taking my suggested comments into account. I find the paper publishable. Found a few typos (please see track change markings in enclosed manuscript)

We have amended these accordingly as follows:

"Patient and Public Involvement" under Methods, line 3

Discussion, para. 5, line 11.

We trust these changes meet with your approval and look forward to this paper being published.