PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	COVID-19 prevention and treatment information on the Internet: a systematic analysis and quality assessment
AUTHORS	Fan, Ka Siu; Ghani, Shahi; Machairas, Nikolaos; Lenti, Lorenzo; Fan, Ka Hay; Richardson, Daniel; Scott, Aneya; Raptis, Dimitri

VERSION 1 – REVIEW

REVIEWER	Ray Jones University of Plymouth UK
REVIEW RETURNED	01-Jun-2020

GENERAL COMMENTS	This is an interesting and timely paper. I have some concerns that could be addressed fairly rapidly. 1. I wonder how representative the identification of the 100 websites is for different countries and different people. This was done by a website scraping tool from a server in Texas. The authors say that no preferences were made to limit searches by geographical region, but we know that the results of individual Google searches made by humans will depend on their previous search histories, geographical location (and perhaps other things that Google does not tell us about!). So we know that a person who has previously searched for (say) BMJ, NICE, SAGE etc based in London will get a very different set of results to someone who has previously searched for Chinese virus, Chinese conspiracy, malaria treatment, ultra violet light to treat coronavirus, and Make America Great Again, living in New York (:>). I wonder if the authors are able to respond to the implication of this on their findings and conclusions? 2. It seems likely that UK, Australian and Canadian users are more likely to choose and read a website from their own country. (Indeed in Discussion the authors state ".. as access and usage of online health information is known to vary between different demographic populations, it is paramount to create and provide targeted and effective educational material for public use.[61,62]". It seems a shame therefore that their analysis focussed so heavily on US (Table 4) rather than compared the four countries. This seems important given that the top scoring websites (Table 5) do not (eg) include the UK Dept of Health website – which is the most promoted website in the UK. 3. But this statement is difficult to interpret as we are not given a table or data to know what I was surprised (Figure 1) at the lack of duplication from the 1275 website identified using the 12 search terms. These only reduced to 1013 and most of the reduction to 321 came about from applying the inclusion/exclusion criteria. I

	would have expected much more duplication from the search terms. Can the authors comment on this? 4. Many people will access websites not by Google but by following a link embedded in social media such as Twitter, Facebook, or more personally WhatsApp rather than search on Google. So although this paper can discuss those websites that may have been found by people on Google this will represent only a proportion of all the website views on the topic. 5. In their discussion they claim that the overall scores for COVID were low….. but compared to what? Probably the scores on these three assessment tools are low whatever topic is being investigated. It would be useful to know (i) if the best websites were MUCH better than the rest, ie are there some very good evidenced based sites but perhaps the mean scores are dragged down by the poor websites? So seeing the distribution of scores would be useful. (ii) how this compared to (say) some well-known and respected cancer, diabetes (or similar) websites. Is COVID any worse using thee assessment criteria, or is it just that these criteria represent a high standard that very few websites meet. 6. I have not made any detailed notes on presentation, but there were quite a number of grammatical/typo errors that I overlooked as I presume would be picked up at copy edit stage.

REVIEWER	Alla Keselman National Library of Medicine, USA
REVIEW RETURNED	19-Jun-2020

GENERAL COMMENTS	The study evaluates the quality of online Covid-19 treatment and prevention information for the general public, which is a very important topic. It uses three different tools to evaluate websites and analyzes 321 websites generated from 12 different search terms – good breadth of analysis. I commend the authors for taking this on. My major concern with the study is its uncritical assumption that the three quality evaluation tools, developed when the digital ecosystem was very different, are proper instruments for evaluating information quality of Covid-19 information sources for the public. An article with the focus on the appropriateness of these sources, with an attempt to supplement / modify them, would be very useful under the circumstances. I am particularly concerned about application of these tools to evaluation information in news sources. I am also concerned about excluding videos, which now constitute a very significant source for health information for the public. I recommend addressing these concerns in the paper. Additionally, at this time, I think it is very important to understand the specific qualitative nature of the problem with information quality. I'd like to see quantitative results supplemented with quality narrative analysis. I thank the authors for the opportunity to review their work and wish them the best in their future endeavors. The rest of my comments in by section.

	Introduction: - Authors criticize sources of information for the public as not peer-reviewed, but primary research sources are not a good source of public health information for the general public. I would suggest not focusing on primary sources and peer-review. Methods: - The authors state that the keywords for the searches used to generate the websites for their analysis came from Google Adwords Keyword Planner and Google Trends. I'd like to see a more detailed description of how each tool was used and which keywords came from which tool. I'd also suggest including a brief description of each tool. - I'd like to see a statement about where the taxonomy of types of sources come from. Also, what was the level of agreement in classifying sources – e.g., was it always easy to distinguish between a non-profit organization and a patient group? - What did News Service as a tool include? Is this primarily news articles? - I would've liked to see more details about the EQIP tool - Very importantly, I think the article should include a discussion of relative appropriateness of the three tools employed in the study for evaluating the type of websites used in the study. For example, DISCERN was developed for evaluating patient education materials - such as pamphlets – about treatment options. It may not be appropriate for evaluating quality of news articles, because news articles follow a different convention: their coverage may include a specific narrow focus and not aim to be comprehensive (e.g., aim to discuss treatment, but not prevention). It is also not customary to include citations in news articles. For this reason, a rather strong news website may come out looking poorly when evaluated via DISCERN. - I'd like to know more about how inter-rater agreement was assessed - My concern about excluding videos, as this format is widely used – what is the justification? Also, what percentage were videos – how many were excluded? Results - "Website demographic and search trends" section probably belongs in Methods, consider moving it there - How much overlap was there among high-scoring websites by different tools? - It is very difficult to interpret the results without any qualitative picture. Where were score points lost? When it came to the content proper, were there more omissions or inaccuracies? Were low scores truly indicators of poor quality, or poor fit of available tools? Discussion - As mentioned earlier, I consider appropriateness of evaluation tools for this task a major question, so it should be addressed in depth in the discussion. Current online data sources are much more multi-faceted than they were when the tools used in the study were developed. - The discussion talks about free access to original Covid-19 research and the public's access to preprints. Such discussion is incomplete without the discussion of health literacy and scientific literacy prerequisites for obtaining information from such primary sources.

	Minor comments: - Page 12 mentions “additional hits on the last page” – could you please clarify the last page of what.

VERSION 1 – AUTHOR RESPONSE

Reviewer comments

Reviewer 1
Comment: I wonder how representative the identification of the 100 websites is for different countries and different people. This was done by a website scraping tool from a server in Texas. The authors say that no preferences were made to limit searches by geographical region, but we know that the results of individual Google searches made by humans will depend on their previous search histories, geographical location (and perhaps other things that Google does not tell us about!). So we know that a person who has previously searched for (say) BMJ, NICE, SAGE etc based in London will get a very different set of results to someone who has previously searched for Chinese virus, Chinese conspiracy, malaria treatment, ultra violet light to treat coronavirus, and Make America Great Again, living in New York (:>). I wonder if the authors are able to respond to the implication of this on their findings and conclusions?

Answer: The effects of geographical variation of search pattern/behaviour is a valid concern and will impact our results. However, it is an unavoidable limitation as completing the queries using a server in each country would be technically challenging and would still be subject to the same effects. We have now addressed these concerns in the limitation and a new search to compare the top results, which reveals very small differences between querying on different servers.

Comment: It seems likely that UK, Australian and Canadian users are more likely to choose and read a website from their own country. (Indeed in Discussion the authors state “.. as access and usage of online health information is known to vary between different demographic populations, it is paramount to create and provide targeted and effective educational material for public use.[61,62]”. It seems a shame therefore that their analysis focussed so heavily on US (Table 4) rather than compared the four countries. This seems important given that the top scoring websites (Table 5) do not (eg) include the UK Dept of Health website – which is the most promoted website in the UK.

Answer: We agree with the lack of subgroup analysis on the other countries, however, it was not feasible to conduct a full subgroup analysis of governmental websites of each country as the study and search terms were not tailored to consider government websites specifically. Our approach was to include any governmental websites into the analysis and hold them to equal standards with all other online information by using the same set of tools. Table 4 was set up to compare global vs USA governmental websites simply due to nearly half of them belonging to the USA. The comparison was performed to understand the effects of having a high number of government websites from numerous local state governments. As opposed to the countries with most websites (United Kingdom, Canada and Australia), the scarcity of government websites did not allow for a fair statistical comparison. It is important to note that we did not assume government websites to be the standard quality level of online health information and that much of the government websites simply did not score high enough to be placed within the top 5 websites of the tools.

Comment: But this statement is difficult to interpret as we are not given a table or data to know what I was surprised (Figure 1) at the lack of duplication from the 1275 website identified using the 12 search terms. These only reduced to 1013 and most of the reduction to 321 came about from

applying the inclusion/exclusion criteria. I would have expected much more duplication from the search terms. Can the authors comment on this?

Answer: We have now addressed this point in methods, results and discussion. The webscraper tool identifies all unique links within the first 10 pages of each search term and automatically excludes duplicates for the data output. The final output of 1275 unique websites already has its intra-term duplicates removed and is reduced to 1013 after removing inter-term duplications. The final reduction to 321 websites is achieved by removing all websites with no health information and resources for professional. The exclusion of websites at the different stages is evaluated in the discussion, and likely attributable to the nature of websites rather than their overlapping between search terms. This was evidenced by the majority of exclusions being due to the lack of health information.

Comment: Many people will access websites not by Google but by following a link embedded in social media such as Twitter, Facebook, or more personally WhatsApp rather than search on Google. So although this paper can discuss those websites that may have been found by people on Google this will represent only a proportion of all the website views on the topic

Answer: The access of health information through social media is indeed an increasingly important topic, however, it is not within the scope of our study. As evidenced by current literature, the public does not find health information on social media to be trustworthy and is more likely to utilise popular search engines in their health-seeking behaviour, our study aimed to evaluate the information quality specific to those seeking health information. The health information accessed through social media, though important, will likely affect the public different than health information found through search engines.

Comment: I n their discussion they claim that the overall scores for COVID were low….. but compared to what? Probably the scores on these three assessment tools are low whatever topic is being investgated. It would be useful to know (i) if the best websites were MUCH better than the rest, ie are there some very good evidenced based sites but perhaps the mean scores are dragged down by the poor websites? So seeing the distribution of scores would be useful. (ii) how this compared to (say) some well-known and respected cancer, diabetes (or similar) websites. Is COVID any worse using thee assessment criteria, or is it just that these criteria represent a high standard that very few websites meet.

Answer: We have evaluated the difference in performance between high and low scoring websites as it is unfair to set our scoring standards directly against other diseases/conditions/treatments which are much more established. With regards to score comparison against diabetes or cancer, we cannot comment on the research that is not planned by us or performed elsewhere. However, based on the available literature using any combination of EQIP, JAMA and DISCERN, it is likely that poor content will still score low. In our discussion, we also included discussion on poorly performing websites and comparison of subsections. The distribution of scores is now included, along with further qualitative analysis of how and why websites performed better than others. The distribution of scores is also included. We also further analysed the correlation between the different tools, as shown by the high inter-class correlation, corroborating that the validated tools also apply here. Therefore, our conclusion remains that while the overall quality of information was poor, the websites with good content still scored highly.

Comment: I have not made any detailed notes on presentation, but there were quite a number of grammatical/typo errors that I overlooked as I presume would be picked up at copy edit stage.

Answer: Have addressed grammatical and typo errors in revision.

Reviewer 2
Comment: My major concern with the study is its uncritical assumption that the three quality evaluation tools, developed when the digital ecosystem was very different, are proper instruments for evaluating information quality of Covid-19 information sources for the public. An article with the focus on the appropriateness of these sources, with an attempt to supplement / modify them, would be very useful under the circumstances. I am particularly concerned about application of these tools to evaluation information in news sources.

Answer: There is indeed no specific tool to assess the rapidly developing online health information on pandemics, however, these tools used here have all been validated to assess the quality of online health information. They were previously used in a variety of studies, ranging for surgical emergencies, to cancer and elective cosmetic surgeries. An important consideration for us is that we should hold all health information to the same standards, regardless of its source, as the material will be read by the same population nonetheless. The concern with evaluating news sources is indeed important as it is arguable to consider them as resources to provide health information. To address this concern, our methodology is to include all websites, news article or not, only if they provide health information. Specifically, we have excluded news articles that provide a dashboard of COVID-19 news and any other news articles that do not discuss the prevention or treatment. A large portion of websites was of excluded for irrelevant content and news update/dashboards concerning infection rates/mortality.

Comment: I am also concerned about excluding videos, which now constitute a very significant source for health information for the public. I recommend addressing these concerns in the paper.

Answer: With video content being a popular source of information, it is indeed relevant to explore their use in this day and age. However, as now addressed in our limitations, no tool has been validated in assessing video-based information, especially for pandemic information. Our toolset is all validated in assessing the quality of written information, rather than in video form, hence, its results are unlikely to be reflective of the true informative value of videos. The standard of information quality would also likely differ between video and written information. Further, the inclusion of video content will likely yield very low scores using the current toolset as videos typically do not include as much information as written information.

Comment: Additionally, at this time, I think it is very important to understand the specific qualitative nature of the problem with information quality. I'd like to see quantitative results supplemented with quality narrative analysis.

Answer: We have now supplemented discussion with both qualitative and quantitative analysis for each tool.

Comment: Authors criticize sources of information for the public as not peer-reviewed, but primary research sources are not a good source of public health information for the general public. I would suggest not focusing on primary sources and peer-review.

Answer: We agree that the source is primary or peer-reviewed is not of priority here, however, we aim to address how well information is conveyed to the layperson. An issue here is that there is no formal assessment of quality in the review process, for peer-reviewed or not, the content is produced by one person and reviewed by another, which is not a standardised process. The aim here is to evaluate the information produced and whether it still holds up a good standard of quality when assessed using standardised tools. We have also added to these concerns in our discussion.

Comment: The authors state that the keywords for the searches used to generate the websites for their analysis came from Google Adwords Keyword Planner and Google Trends. I'd like to see a more detailed description of how each tool was used and which keywords came from which tool. I'd also suggest including a brief description of each tool.

Answer: We have addressed these in our methodology: Google Adwords Keyword Planner was used to identify all our search terms and to visualise its popularity trend, Google Trends was used.

Comment: I'd like to see a statement about where the taxonomy of types of sources come from. Also, what was the level of agreement in classifying sources – e.g., was it always easy to distinguish between a non-profit organization and a patient group?

Answer: Taxonomy of types of sources was based on similar available literature. We have included a more detailed methodology of how to classify websites, with specific comparisons drawn between similar groups: patient group vs charity/non-governmental organisations, practitioner vs industry, academic centre vs professional societies. Using these criteria, there was no disagreement on source categorisation.

Comment: What did News Service as a tool include? Is this primarily news articles?

Answer: News service includes both primary and secondary news articles that are not written for professionals and contain health information about COVID-19 prevention or treatment. All news service websites that only consist of a news dashboard, mortality/infection rates or health policies were excluded as it does not provide health information.

Comment: I would've liked to see more details about the EQIP tool

Answer: More details are included under EQIP Tool of methodology as well as evaluation of its performance in the discussion.

Comment: Very importantly, I think the article should include a discussion of relative appropriateness of the three tools employed in the study for evaluating the type of websites used in the study. For example, DISCERN was developed for evaluating patient education materials - such as pamphlets – about treatment options. It may not be appropriate for evaluating quality of news articles, because news articles follow a different convention: their coverage may include a specific narrow focus and not aim to be comprehensive (e.g., aim to discuss treatment, but not prevention). It is also not customary to include citations in news articles. For this reason, a rather strong news website may come out looking poorly when evaluated via DISCERN.

Answer: We fully agree, and because this is a novel topic, we wanted to avoid loss of assessment and hence, used three assessment tools. We feel that patients need to be educated to the same degree regardless of the source of information, and so expectations were also set to be the same between all sources. News articles that focussed only on news, and not patient education/health information, then it is not a good source of patient information and was excluded as such.

Comment: I'd like to know more about how inter-rater agreement was assessed

Answer: Inter-rater agreement was performed to find not significant bias of all assessors (all within 95%CI). Details are included in supplementary data.

Comment: My concern about excluding videos, as this format is widely used – what is the justification? Also, what percentage were videos – how many were excluded?

Answer: With video content being a popular source of information, it is indeed relevant to explore their use in this day and age. However, as now addressed in our limitations, no tool has been validated in assessing video-based information, especially for pandemic information. Our toolset is all validated in assessing the quality of written information, rather than in video form, hence, its results are unlikely to be reflective of the true informative value of videos. The standard of information quality would also likely differ between video and written information. Further, the inclusion of video content will likely yield very low scores using the current toolset as videos typically do not include as much information as written information. A total of 35 (2.7%) videos were excluded from our search, none were duplicated.

Comment: Website demographic and search trends" section probably belongs in Methods, consider moving it there

Answer: We have the relevant parts of "Website demographic and search trends" section to Methods as suggested.

Comment: How much overlap was there among high-scoring websites by different tools?

Answer: High-scoring websites are websites that scored highly, within the 75th percentile, within all three tools, returning a total of 74 websites were considered high-scoring. We have performed an intra-class analysis between the different assessment tools to show a good score correlation between all the tools.

Comment: It is very difficult to interpret the results without any qualitative picture. Where were score points lost? When it came to the content proper, were there more omissions or inaccuracies? Were low scores truly indicators of poor quality, or poor fit of available tools?

Answer: We have evaluated the difference in performance between high and low scoring websites as it is unfair to set our scoring standards directly against other diseases/conditions/treatments which are much more established. With regards to score comparison against diabetes or cancer, we cannot comment on the research that is not planned by us or performed elsewhere. However, based on the available literature using any combination of EQIP, JAMA and DISCERN, it is likely that poor content will still score low. In our discussion, we also included discussion on poorly performing websites and comparison of subsections. The distribution of scores is now included, along with further qualitative analysis of how and why websites performed better than others. The distribution of scores is also included. We also further analysed the correlation between the different tools, as shown by the high inter-class correlation, corroborating that the validated tools also apply here. Therefore, our conclusion remains that while the overall quality of information was poor, the websites with good content still scored highly.

Comment: As mentioned earlier, I consider appropriateness of evaluation tools for this task a major question, so it should be addressed in depth in the discussion. Current online data sources are much more multi-faceted than they were when the tools used in the study were developed

Answer: We agree that the online data sources are now multi-faceted and need further consideration in using assessment tools. Our study aimed to evaluate the information quality specific to those seeking health information, regardless of its source. We believe that patients, and the public, should all have access to the same standard of information and that would require the use of standardised tools. While EQIP, JAMA and DISCERN are not specifically designed for health information of pandemics, it has been validated in numerous settings and we believe it can be justified that these are among the best tools available for this task. Additionally, the changing digital ecosystem does

mean other sources should be assessed too. Video content is increasingly popular and undoubtedly should be assessed too, however, to hold it to the same standards as written content may be unfair due to their vastly different nature as well as the lack of a validated tool to assess video-based health information.

Comment: The discussion talks about free access to original Covid-19 research and the public's access to preprints. Such discussion is incomplete without the discussion of health literacy and scientific literacy prerequisites for obtaining information from such primary sources.

Answer: We have taken the advice to explore the role of health literacy, of not just the public, but also its role in health journalism.

Comment: Page 12 mentions "additional hits on the last page" – could you please clarify the last page of what.

Answer: We have rephrased and clarified this in methodology. The webscraper tool identifies all unique links within the first 10 pages of each search term and automatically excludes duplicates for the data output. The final output of 1275 unique websites already has its intra-term duplicates removed and is reduced to 1013 after removing inter-term duplications. The final reduction to 321 websites is achieved by removing all websites with no health information and resources for professional.