**Peer Review File**

Article Information: https://dx.doi.org/10.21037/jss-22-57

**Reviewer A**

**Comment 1:** The authors present an interesting paper appraising Evicore's guidelines for imaging in the setting of lower extremity pain with neurological symptoms. Overall, the topic is certainly worth investigating; however, there were only 3 reviewers used. This seems to be an underpowered study. The authors would have more success if this were conducted as a survey to more physicians.

**Reply 1:** Thank you for this recommendation. We have added 2 additional appraisers per this reviewer's recommendation.

**Changes in the text:** Methods section has been updated to reflect the addition of the two reviewers. Results section has been updated with the additional scores. Discussion section has also been updated to encompass the feedback from all five reviewers.

**Reviewer B**

**Comment 1:** Thank you for the opportunity to review this manuscript. This manuscript analyzes whether the guidelines put in place by Evicore for prior authorization of lumbar spine MRI with or without back pain are transparent, methodologically sound, and based on high quality evidence. This is an important consideration for spine surgeons who may deal with frequent prior authorization denials due to the arduous burden placed on them to demonstrate that the patient requires a spine procedure. Three authors analyzed multiple different factors that may improve a guidelines efficacy and transparency and came to the conclusion that the current guidelines require at minimum, some modification, but possibly a complete overhaul. I believe the following minor edits may improve the quality of the manuscript:

INTRO

Line 44 of introduction: there is a small typo. It should read "The guidelines produced by Evicore address areas of over-utilization and unnecessary spend(ing) as well as pinpoint areas to improve care and increase savings (1).

**Reply 1:** Thank you so much for this very helpful feedback. We agree with all of your suggestions. The typo in the intro has been corrected from "spend" to "spending."

**Changes in the text:** See referenced quote above, found in the third sentence of the second paragraph in the introduction.

**Comment 2:** This is a suggestion and the authors may take it or leave it as they wish. However, I believe some additional background in the introduction or discussion about the rate of prior authorization denials is important in the context of this manuscript. This may strengthen why

this manuscript is an important topic to address in the spine literature. Currently, the rate of denials is increasing, partly due to issues with prior authorization.

**Reply 2:** We have added information on the rate of prior authorization denials.

**Changes in the text:** Please see the first paragraph of our introduction.

**Comment 3:**

RESULTS

Please include the results of the intraclass correlation coefficient in the results section and link it to table 4.

**Reply 3:** The results of the intraclass correlation coefficient were added to the results section and linked table 4.

**Changes in the text:** See last sentence of the results section.

**Comment 4:** For table 3, please provide the domain name instead of numbers in the first column so the readers can easily identify what domain is being evaluated without having to reference the text. It also would be easier to have the AGREE II parameter table listed before the results table (currently table 2).

**Reply 4:** We have updated what was previously known as table 3 (now table 1) to include the domain name so that readers won't have to reference the text. We have also reordered our tables so that the AGREE II parameters are listed in table 1, overall assessment scores are listed in table 2, and comprehensive scores are listed in table 3. We agree that this order is much more intuitive.

**Changes in the text:** See attached tables.

**Comment 5:** Lines 122-123 of the results section is redundant with the methodology section. Thus, please remove this from the results section "A clinical guideline was considered satisfactory if it scored at least 50% on all six domains."

**Reply 5:** We have removed the phrase "A clinical guideline was considered satisfactory if it scored at least 50% on all six domains" due to redundancy.

**Changes in the text:** This was removed from the results section.

**Comment 6:** For the first line of the discussion line 131-132, I'm not sure "escalate" is the correct word. Perhaps "becomes paramount" is more apt.

**Reply 6:** We have changed the word "escalate" to "becomes paramount" per your suggestion.

**Changes in the text:** Please see the first sentence of the discussion section.

**Comment 7:**

DISCUSSION

The following sentence in lines 157-160 is confusing "As the success of a guideline depends in part on the suggested strategies used to improve compliance, absence of this information additionally lends itself to inconsistency in the rate of application among providers". Please rephrase it.

**Reply 7:** We have rephrased what was formerly lines 157-160 for improved clarity. This sentence has been changed to "Guideline characteristics that increase the chance of utilization by providers include intelligibility, convenience, and accessibility of required resources").

**Changes in the text:** Please see the "applicability" paragraph in the discussion section.

**Comment 8:** Also please rephrase "The editorial independence domain similarly has been acknowledged as a universal concern as a consequence of its high variability." Can you clarify what you mean by high variability and why it is a universal concern. Is the universal concern that it is rarely implemented/followed by people making guidelines? Is the high variability referring to whether or not people follow this domain?

**Reply 8:** We have deleted the above quoted sentence and clarified our intended sentiment - that questionable editorial independence in formulating a healthcare guideline is a problem that is widespread, (we have cited sources in the paper to justify this claim as well). The sentence now reads "These problems [i.e. issues pertaining to editorial independence] are not unique to Evicore's guidelines, as questionable editorial independence and failure to disclose conflicts of interest are widespread problems throughout the healthcare industry that can have negative consequences (27, 28)."

**Changes in the text:** Please see editorial independence paragraph in the discussion section.

**Comment 9:** Please rephrase the following sentence "Many argue that inclusion of patient preference in guidelines is unnecessary due to individual variation and patients may prefer to not be involved in their care, while others contend patient input should be incorporated due to its association with improved medical outcomes (15, 16)." Something along the lines of "Inclusion of a patients' preference when creating guidelines may be unnecessary due to preference variations and/or patients may prefer to not be involved in their care; however, patient input may ultimately improve patient care and clinical outcomes." I believe this will be a less controversial statement then gauging how many people argue one side versus the other. If these points are included, I believe this topic will be of interest to readers and can be an important addition to the literature.

**Reply 9:** Thank you again for your thoughtful feedback. We have rephrased "Many argue that inclusion of patient preference in guidelines is unnecessary due to individual variation and patients may prefer to not be involved in their care, while others contend patient input should be incorporated due to its association with improved medical outcomes" to the suggested wording that this reviewer recommended.

**Changes in the text:** Please see fourth sentence in the "stakeholder involvement" paragraph in the discussion section.

**Reviewer C**

Points For Improvement:

**Comment 1:** "Since its 2009 release, the AGREE II has been used to evaluate hundreds of guidelines and has been extensively validated (3,4,5)." This is a strong comment saying the word 'hundreds' then backed up by 3 references. As a reader, I'd be interested in the similar studies or relevant guidelines it has been used to review, and results around the similar studies.

**Reply 1:** Thank you for your feedback. Though we understand that this comment is strongly worded, we have kept the wording as is for the following reasons. We did not add additional sources to substantiate our claim that the AGREE II tool has been used to evaluate hundreds of guidelines and has been extensively validated, as it is not practical to cite the hundreds of high-quality papers that have used the AGREE II tool. A query search through PubMed substantiates our claim – 1,090 papers include the words "AGREE II" in the title or abstract and searching this term in all fields populates 6,772 results. In line with your suggestion, we have added mention of relevant studies (i.e. ones pertaining to back pain) that have used the AGREE II tool – please see sentence added to intro "More specifically, this tool has been utilized to evaluate guidelines about surgical and nonsurgical management of back pain and spinal cord injury. These appraisals have demonstrated room for improvement in knowledge translation and health system changes while also highlighting strengths in guideline development" (see Stochkendahl et al., 2018, Layne et al., 2018, Martin et al., 2018, Bussieres et al., 2018 for utilization of the AGREE II tool to manage back pain guidelines).

**Changes in the text:** The above changes can be found in the second to last paragraph of the introduction.

**Comment 2:** "MRI is essential to treating many conditions of the lumbar spine." MRI is a diagnostic tool with margin for error, it in itself, has never treated any spinal conditions.

**Reply 2:** We have removed this sentence.

**Changes in the text:** See introduction.

**Comment 3:** Is there any reason there is only 3 reviewers? For example, if they are all from the same institution, the bias created is extremely large. I feel like this paper would be taken with a lot more weight, if there was a larger sample size of reviewers. Also, some history about who they are - eg a medical student vs a 10 year Spinal Surgeon creates very different results.

**Reply 3:** We have added 2 additional appraisers to our study to reduce the bias that this reviewer mentioned. Per the AGREE II manual and instructions found on the AGREE II website (https://www.agreetrust.org/resource-centre/agree-ii/agree-ii-instructions/), the AGREE II tool is designed to be used by "guideline developers, policy makers, health administrators, program managers, professional organizations, stakeholders (including patients/consumers, health professionals, researchers, educators and other stakeholders interested in the development and update of clinical guidelines." The tool is designed such that years of clinical experience should not affect how the tool is employed; this is substantiated by

the fact that this tool is recommended for such a broad audience. Practically, it should not matter whether a medical student or a spinal surgeon with years of experience used this tool. It should also not matter what institution an appraiser is from, as all appraisers should get similar AGREE II results regardless of experience level and training given how the tool is designed.

**Changes in the text:** Methods section has been updated to reflect the addition of the two reviewers. Results section has been updated with the additional scores. Discussion section has also been updated to encompass the feedback from all five reviewers.

**Comment 4:** Where did the figure of 50% come from? Is this from the Authors? Is this from the producers? This needs to be explicitly stated.

**Reply 4:** The origin of the figure of 50% to assess whether a guideline domain was/was not satisfactory is the AGREE II manual… we have added this information to our paper for clarification.

**Changes in the text:** Please see the last sentence of the third paragraph in the methods section.

**Comment 5:** Extremely long passage in the discussions session is very hard to follow, this needs to be separated into domains and discussed.

**Reply 5:** We have separated our discussion sentence into domains and added additional information to our analyses for each domain.

**Changes in the text:** Please see discussion section.

**Comment 6:** It is not clear who recommended 4 reviewers for the AGREE tool. Is this the producers of the tool? Does this have evidence backing. This is also unclear.

**Reply 6:** The AGREE II manual recommends 4 reviewers.

**Changes in the text:** We have added this to our discussion section in the second sentence of the "limitations" paragraph.

**Comment 7:** The conclusion about the tool is only discussed very briefly in the end of the conclusion. Unfortunately, these discussion points are extremely generic, and I feel they lack and specificity.

**Reply 7:** We have added to our discussion section to incorporate a more robust discussion of the Evicore guidelines and our suggestions for their improvement. We have incorporated specific feedback such as "this guideline would benefit from a clear acknowledgement section that explicitly states individuals from relevant professional groups," for example.

**Changes in the text:** Please see the entirety of the discussion section, as many changes have been made in response to this comment.

**Comment 8:** Overall, the flow jumps from talking about the AGREE tool, then the guidelines, then the AGREE tool. It would benefit from discussion about the tool, then more in-depth

discussion about what the tool is actually evaluating. I feel the majority of this paper is discussing the tool itself rather than the guidelines.

**Reply 8:** We have separated our discussion sentence into domains and made extensive edits to improve readability and clarity. In addition, we have added to our discussion section to incorporate a more robust discussion of the Evicore guidelines and our suggestions for their improvement.

**Changes in the text:** Please see discussion section.

**Comment 9:** Lots of the bold points they make aren't strongly backed by evidence or data, or at least are not explicitly made obvious.

**Reply 9:** Thank you for this feedback. In considering this author's feedback we have changed our wording in several places to soften the claims we make or have included additional references if we felt that our original strong wording was substantiated in evidence.

**Changes in the text:** Many changes were made throughout the text that are in line with this reviewer's helpful feedback.

**Reviewer D**

**Comment 1:** The authors provide an insightful study evaluating the quality of the Evicore spinal imaging guidelines. Their study is well written, succinct, and clear; they should be commended for their work. My only suggestion is that if 4 reviewers are recommended for the Agree II tool, the authors might be able to incorporate one additional reviewer in order to ensure optimum validity.

**Reply 1:** Thank you for your comment – we appreciate it. We have added two additional reviewers.

**Changes in the text:** Methods section was updated to reflect additional reviewers and their information. Results section is updated with the additional scores. Discussion section was also updated to encompass the feedback from all five reviewers.

**Reviewer E**

**Comment 1:** The concept of this paper is good, but the discussion is rough and lacking. The discussion has old literature and is not well argued. Also, there are many critical opinions, but few useful ones.

**Reply 1:** Thank you for your feedback – we agree that our discussion section needed work. We have updated our discussion session with more recent literature. We have also supplemented our critiques with suggestions for guideline improvement per your recommendation.

**Changes in the text:** Please see entirety of discussion section.

**Comment 2:** Moreover, the authors should examine intraobserver intraclass correlation coefficients.

**Reply 2:** Similar to previous studies that have employed the AGREE II tool to evaluate guidelines (see references 2-8 for examples), we chose to examine only the *inter-observer* intraclass correlation coefficient. This intraclass correlation coefficient measures reproducibility of numerical measurements made by different people measuring the same thing. Given that the AGREE II tool is an internationally recognized and validated assessment tool that recommends each rater evaluate the guideline only once, we did not go against this recommendation and have individual raters repeat their assessments in order to calculate an *intra-observer* correlation score. We believe that our existing statistical calculations are appropriate as is.

**Changes in the text:** Not applicable.