The first steps in the evaluation of a "black-box" decision support tool: a protocol and feasibility study for the evaluation of Watson for Oncology

Lotte Keikes, Stephanie Medlock, Daniël J. van de Berg, Shuxin Zhang, Onno R. Guicherit, Cornelis J.A. Punt, Martijn G.H. van Oijen

Corresponding author
Lotte Keikes, M.D., Department of medical oncology, Cancer Center Amsterdam, Academic Medical Center, University of Amsterdam, Meibergdreef

Handling editor:
Michal Heger
Department of Experimental Surgery, Academic Medical Center, University of Amsterdam, Amsterdam, the Netherlands

Review timeline:

Received: 21 February, 2018
Editorial decision: 1 March, 2018
Revision received: 7 June, 2018
Editorial decision: 28 June, 2018
Published online: 27 July, 2018

1st editorial decision:

Date: 1-Mar-2018

Ref.: Ms. No. JCTRes-D-18-00009
The first steps in the validation of a cognitive decision support tool: a protocol and feasibility study for the evaluation of Watson for Oncology
Journal of Clinical and Translational Research

Dear authors,

All reviewers have now commented on your manuscript. Two reviewers are advising a major revision, while one reviewer has recommended a reject. The editorial board would like to see the authors' standpoint on the issues raised by the reviewers, and is willing to reconsider the decision granted that all comments are properly addressed in a point-by-point manner. Please pay extra attention to the following items:

1) Please address in the manuscript which other "rigorous evaluation methods" exist for content validity and provide the reasons for why none of these methods were used to evaluate Watson for Oncology. This is really a critical point as multiple reviewers had trouble with your approach. This does not mean your approach is faulty per se, but premises should be given for choosing to deviate from conventions;
2) Point 4 raised by reviewer 1;
3) Please provide a sound, scientific basis for your evaluation method and scoring scheme that is non-overlapping with the reason for why none of the existing evaluation methods were used

(my first point);
4) Please clearly discern between the methodological hypothesis and testing part versus the actual validation part.
5)) Please include a comparison of Watson vs NCCN guidelines.

For your guidance, reviewers' comments are appended below.

If you decide to revise the work, please submit a list of changes or a rebuttal against each point which is being raised when you resubmit your work.

Your revision is due by May 31, 2018.

To submit a revision, go to https://jctres.editorialmanager.com/ and log in as an Author. You will see a menu item called Submission Needing Revision. You will find your submission record there.

Yours sincerely,

Michal Heger
Editor-in-Chief
Journal of Clinical and Translational Research

Reviewers' comments:

Reviewer #1: Manuscript: JCTRes-D-18-00009

Assessment:
1- This is not, as the authors contend, a study of content validity. There are rigorous evaluation methods for examining content validity and the design of this study is incompatible with any of these methods. The examination of concordance is not a suitable alternative for assessing validity.

2- Concordance is a useful concept, but there needs to be a more substantial discussion of the notion of "correctness" relative to what computer programs suggest. There is no discussion, for example, of the lack of a gold standard nor what is known about the accuracy and/or quality of the Dutch guidelines offered through Oncoguide.

3- The authors claim to have developed a new approach to examining concordance but fail to provide a discussion of: existing methods for examining concordance, the shortcomings of these approaches, or the added value of their approach.

4- The manuscript demonstrates a significant lack of biomedical informatics expertise, as evidenced by:
i. the absence of a discussion of the science and corpus of knowledge on computer-based clinical decision support systems and their evaluation. There is a broad and rich literature on this subject.
ii. the perfunctory approach to addressing the complex subject of usability including the use of cognitive walkthrough without offering any important citations in this field and with results based on descriptive comments rather than formal qualitative data

5- The authors provide no logic or principles to support their scoring scheme.

6- There is no indication that the descriptions of the Watson for Oncology treatment advisor are accurate or current since confirmation by the technology provider is not mentioned and reference to the version of the technology is not offered.

Reviewer #2: With pleasure I reviewed your paper on evaluation of cognitive decision support tools and feasibility to test the content and usability of Watson using this approach.

General:
1. Overall my most major issue is with the overall aim of the paper; At every stage of the paper it is challenging for the reader to discern the methodological hypothesis and testing, and separating it from the more 'practical' validation of the content of IBM's Watson for Oncology tool.

Introduction
2. In line with Ewout Steyerberg's work (although more reflecting on predictive and less on diagnostic decision support tools), I would suggest to use internal and external validation terminology; This validation takes place in another country, according to more local standards standards. I miss a clear connection to available knowledge and guidelines on hot to evaluate cognitive decision support tools.

Methods en Results
3. I do not see the ethical challenges of using real (retrospective) cases compared to synthetic cases. Can you elaborate?
4. In order to validate the use of synthetic cases as a tool compared to real cases some type of validation is needed; Can we replace real cases like this since we will extrapolate results of validation processes like this to the real world.
5. The walk through approach really helps, thank you. However, why was it not chosen to do traditional accuracy testing like (AU)ROC characteristics?

Discussion and conclusions
6. What about the issue that Watson was trained at Sloan Kettering? We see that results from the Watson can be off in different local settings; different decision making can take place where guidelines are absent/different or ethical choices lead away from the setting the algo was trained? Is the general and dogmatic approach of one algo, one training, and no retraining to local standards desirable? I think your conclusions can be more stronger in this respect but I agree with the line of reasoning.
7. Do you have suggestions for the 'black box' approach of Watson?

Reviewer #3: The authors describe a validation of a stand-alone implementation of IBM Watson for Oncology on a synthetic set of colon cancer patients after surgery. In the experiment, the recommendations of Watson were compared with the Dutch guidelines. Besides, the output of the IBM algorithm, also an interface "created by MRDM" was evaluated using three users.
I believe the evaluation of Watson is an important scientific contribution and I appreciate the

design of the experiment. However, the presentation of the results in the article is open to improvement.

1. Please provide a to-the-point, verifiable description of Watson. Although the reasoning of Watson is "opaque to the user", the system is described in wording that in my opinion would better fit the IBM website. I suggest to drop descriptions as "cognitive computer system", case-based reasoning, neural networks and the like. Also the assumed approach of Watson integrating natural language process with guidelines and studies do not contribute to the article. The extraction of data from the EMR is not studied here and therefore out of the scope of this article. The authors suggest that Watson uses the American guidelines in its reasoning. This also seems not to be verifiable.
So, I believe that at best, we can describe Watson as a black box algorithm that assumes a predefined set of parameters and produces recommendations for chemo therapy.
Also, I assume Watson to be a system that evolves over time. Can you indicate which version of Watson you have evaluated?
2. Rationale of use case. You have selected post-surgery recommendations for colon cancer patients. I understand that for this use case a well-defined ground truth (oncoguide) is available, but I do wonder if there are other reasons for this use case? Does Watson only cover recommendations in Medical Oncology? Would a use case be supported where multiple types of therapies can be an option (e.g. in prostate cancer or breast cancer)? In colon cancer, why did you restrict to the post-surgery use case? And, why did you exclude Stage IV patients?
3. Role of NCCN Guidelines in the paper. You hypothesize that Watson is based on the NCCN guidelines, I think we are all willing to assume this. But, while you present some differences between Dutch and American recommendations, it remains unclear whether Watson indeed presents recommendations according to these NCCN guidelines. The usage of guidelines in Watson however, is central in your argumentation as "localization of Watson with adjustments based on national or local guidelines" is needed. Also Watson seems to perform better in the Asian studies, one of the arguments being that those local guidelines may "more closely parallel the US guidelines". Therefore, I believe the impact and reasoning of our work would substantially improve if you would indeed not only compare Watson with the Oncoguide but also the NCCN guidelines.
4. Usability Test. I have doubts or have not fully understood the impact of the usability experiment conducted. You provide the initials of the users, I do hope they no co-authors of your article. The set-up does not seem to mimic the usage of Watson in the clinic, as no integration with the EMR is provided. I believe you provide the users with a list of items to fill in on the UI, including "contraindication oxilaplatin", which we will most likely not encounter in any patient record. Also, the interface is not offered by IBM but by MRDM. The consequence of a well- or ill-usable interface for me is unclear, which makes it hard for me to appreciate this section. In any way, it would be interesting to include a screen shot, so we better understand the interface you are piloting.
5. Experiment design and sharing of use cases. A big contribution is that you describe a test set for evaluation clinical decision support systems in oncology as well as a well-designed mechanism to score the performance. I think it would be a big contribution you could make the test set as well as the scoring tool available to the community. It will help to compare against other systems and also help to measure performance of future versions of Watson.

So, in summary, I appreciate the work of the research team and experimental work. The work however has more potential as the important results can be better presented.

Authors' rebuttal

Dear Editor,
We like to thank you for the opportunity to revise our original article *'The first steps in the evaluation of a "black box" support tool: a protocol and feasibility study for the evaluation of Watson for Oncology'* according to the reviewers' comments.
We thank the reviewers for the critical reading of the manuscript and insightful comments. We feel that the incorporation of the suggestions of the reviewers has substantially improved the value of our manuscript.
Please find below our point-by-point reply. We used track changes in our revised manuscript and added a new adjusted version with processed track changes ("clean copy").
Thank you for considering our revised original article for publication in The *Journal of Clinical and Translational Research.*

Manuscript: JCTRes-D-18-00009
The editorial board would like to see the authors' standpoint on the issues raised by the reviewers, and is willing to reconsider the decision granted that all comments are properly addressed in a point-by-point manner. Please pay extra attention to the following items:

A. Please address in the manuscript which other "rigorous evaluation methods" exist for content validity and provide the reasons for why none of these methods were used to evaluate Watson for Oncology. This is really a critical point as multiple reviewers had trouble with your approach. This does not mean your approach is faulty per se, but premises should be given for choosing to deviate from conventions;
Reply: We agree that this is a critical point, and we have substantially changed the manuscript accordingly. First, we realize that the term "content validity" was confusing, as it means different things in different fields. Therefore we have changed the manuscript to use the term "evaluation", and explain in the text what we are evaluating. We also give additional background on the usual approach to evaluation of decision support systems, and why a different approach is needed for systems such as Watson:
We added the following paragraphs to the Introduction:
'...This brings unique challenges in evaluating software like Watson prior to clinical use. Evaluation of software is often divided into two steps: verification and validation. (1) Verification is checking whether the system was built according to specification. In a typical rule-based decision support system, this would involve testing the individual rules to ensure that the system provides the expected output for a given set of inputs. The second step is validation, or checking that the system meets user expectations. In a typical decision support system, this would involve giving set of cases (selected based on the range of expected inputs and outputs) to clinicians and asking them to compare the output of the system to their own assessment. This testing should be done before conducting a clinical trial, which is aimed to assess the impact of the system on actual clinical decision-making. (2) However, in a system like Watson, this approach is not possible.
The exact inputs (the data that the system uses in its reasoning) and expected outputs are opaque to the user. Previous efforts in evaluating Watson have thus far only been reported as conference abstracts (3-5) and these short reports indicate that the evaluations consisted of comparing the output of Watson in actual clinical cases against the evaluation of clinical experts. However, although this approach approximates the validation step of a typical evaluation, a selection of consecutive clinical cases probably represents only a small sample of possible cases. Common cases are likely to be overrepresented. Unusual cases may not

appear at all. Since machine learning systems tend to perform better when they have been trained with more data, Watson may also perform better in common cases than in unusual ones – and since it is precisely these unusual cases where clinicians may seek advice, a systematic approach to testing is needed before performing an impact study. Smith et al. suggested a general approach to evaluation of such systems which involves comparing the performance of such systems to a validated gold standard. (6) However, as is often the case in medicine, no gold standard exists in oncology. Furthermore, Watson's use of free text data complicates the analysis by introducing uncertainty about the spectrum of cases which should be tested.'

B. Point 4 raised by reviewer 1;

Reply: We agree that our previous version did not adequately place this study in the field of evaluation research in decision support systems. We now explain this more clearly in the introduction (point A above). We extend this explanation with the following text in the discussion, including addressing our choice of cognitive walkthrough as our usability evaluation method:

'Clinical decision support systems are typically evaluated in clinical trials, and evaluate whether the system changes the process of care in ways which could affect clinical outcomes. (7) However, before determining whether advice is followed, it is first necessary to ensure that the system is providing the right advice. As outlined in the introduction, this is typically done by first comparing the system to the clinical knowledge on which it was based, then comparing the output of the system to the judgment of clinical experts in a defined set of test cases. In Watson and other neural network systems, there is no specification to perform this first step. Thus, following the recommendations of Smith et al. (6) we have chosen to compare to another system. OncoGuide is intended to represent the standard of evidence-based care in the Netherlands (as it is a representation of the Dutch colorectal cancer guidelines in decision trees), thus it is a logical choice for evaluating the system for use in the Netherlands. Comparison against an objective standard also adds value over only comparing with clinician judgment. (8) Given the large number of test cases we expect to generate, this also allows us to perform the evaluation more efficiently, as we can reasonably assume that if Watson and OncoGuide agree then the clinician will also agree. Thus clinicians can focus on cases where Watson and OncoGuide disagree, and determine which advice is better in their judgment.

We elected to use a relatively simple, expert-based usability evaluation, the cognitive walkthrough approach. We considered this approach to be appropriate to the circumstances of this evaluation: the tasks to be accomplished in the system were well-defined, and the main goal was to identify usability issues that could be a barrier to use of the system in a trial setting with naive users. (9) Furthermore, the interface itself is relatively simple, and more qualitative methods such as thinkaloud were, in the authors' view, unlikely to yield additional insights.'

C. Please provide a sound, scientific basis for your evaluation method and scoring scheme that is non-overlapping with the reason for why none of the existing evaluation methods were used (my first point)

Reply: The parts that are new to this protocol are the definition of synthetic cases based on the guideline and expert opinion, and the scoring system. We have added further explanation of this to the discussion:

'This protocol introduces two new methods for evaluating artificial intelligence-based decision support systems: a method for generating synthetic cases, and the scoring system for assessing agreement. Synthetic cases are generated based on the known inputs and outputs of the comparison system (in our example, OncoGuide), and input from clinical experts on variables which might indicate a justifiable departure from the guideline. This approach should capture both cases where Watson is likely to agree with the guideline, and cases where Watson may be able to offer better advice than the guideline. As with other decision support systems, simple "agreement" is not sufficient to describe the performance of this system. (1) Friedman and Wyatt proposed the use of contingency tables when evaluating decision support systems, to make explicit the difference between false positive and false negative classifications. Their reasoning is that a false-positive error, such as erroneously suggesting a diagnosis for a healthy patient, may be less serious than a falsenegative error, which may in turn be less serious than proposing the wrong diagnosis entirely. (1) Likewise, a suggestion from the system that the clinician "consider" a treatment which is in fact contraindicated is a less serious error than "recommending" use of that treatment. Thus, we have extended the notion of a contingency table to express the idea that some disagreements have greater consequences than others.'

D. **Please clearly discern between the methodological hypothesis and testing part versus the actual validation part.**

Reply: We agree that this is an important distinction for our readers. We have re-labeled the headings and added some additional explanatory text to the methods. We have also re-structured the results to more clearly reflect which results derive from following the protocol, and which result from additional methods that were only followed for the feasibility study.

E. **Please include a comparison of Watson vs NCCN guidelines.**

Reply: We agree with the reviewers that the incorporation of a comparison of Watson with the NCCN guidelines improves the value of our paper. We therefore conducted additional analyses and incorporated the results in our paper. We added the following sentences to the Methods, Results and Discussion section and added Figure 3b:

'2.3 Additional analyses

In addition to following the protocol in part 1, we performed two additional analyses for our feasibility study: a comparison of the results to the US guidelines (to gain a sense of the degree to which non-concordance between Watson and Oncoguide is attributable to non-concordance between Dutch and US guidelines), and a usability assessment.'

…

'Next, we compared Watson's advice directly with the NCCN guideline recommendations using the same methods as earlier prescribed in paragraph 1.3. (Comparison with the Dutch guidelines).

…

Comparison to US guidelines

Overall concordance scores ranged between a minimum score of -4 (9 cases) to a maximum concordance score of 12 (24 cases) and concordance scores ranged per cancer stage (Figure 3). The median concordance score was 5. No orange or red flags were reported for the comparison between
Watson and the NCCN guidelines.'

…

'We performed an additional comparison of Watson versus the NCCN guidelines to gain a sense of the degree to which non-concordance between Watson and Oncoguide was attributable to nonconcordance between Dutch and US guidelines. We identified variety in the concordance scores in both situations, but no orange or red flags were reported for the comparison between Watson and the NCCN guidelines. This supports that disagreements between Watson and Oncoguide are (partially) attributable to guideline differences.'

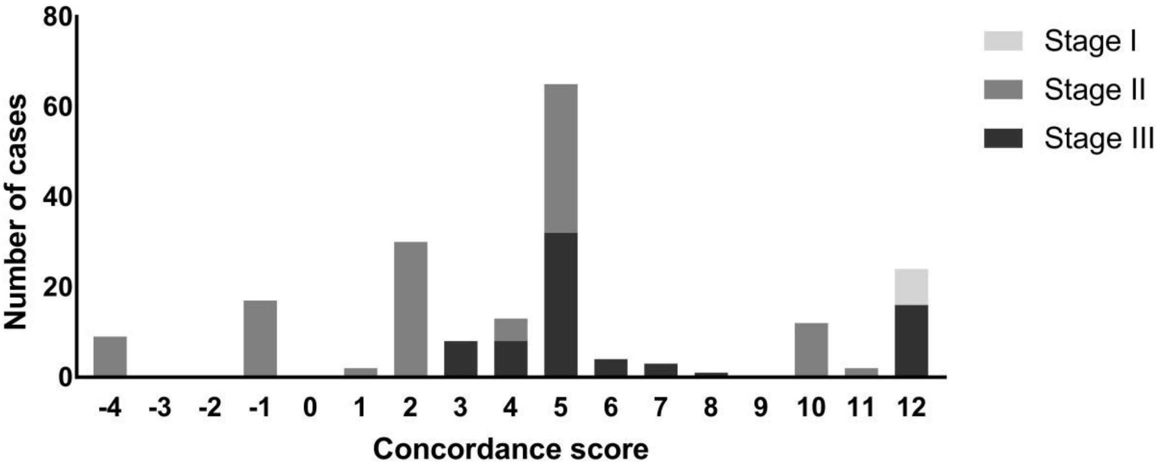

Figure 3b. Concordance scores (Watson versus NCCN) differentiated by tumor stage

Comments – Reviewer #1:

Assessment

1. **This is not, as the authors contend, a study of content validity. There are rigorous evaluation methods for examining content validity and the design of this study is incompatible with any of these methods. The examination of concordance is not a suitable alternative for assessing validity. Reply:** We agree that this is a critical point, and we have substantially changed the manuscript accordingly. First, we realize that the term "content validity" was confusing, as it means different things in different fields. Therefore we have changed the manuscript to use the term "evaluation", and explain in the text what we are evaluating. We also give additional background on the usual approach to evaluation of decision support systems, and why a different approach is needed for systems such as Watson. We have added new paragraphs to the introduction as prescribed in response to comment A of the editorial board.

2. **Concordance is a useful concept, but there needs to be a more substantial discussion of the notion of "correctness" relative to what computer programs suggest. There is no discussion, for example, of the lack of a gold standard nor what is known about the accuracy and/or quality of the Dutch guidelines offered through Oncoguide.**

Reply: We agree that this is an important point. We have mentioned the lack of a gold standard in the Introduction and Discussion:

‘Smith et al. suggested a general approach to evaluation of such systems which involves comparing the performance of such systems to a validated gold standard. (6) However, as is often the case in medicine, no gold standard exists in oncology.

...

Thus, following the recommendations of Smith et al. (6), we have chosen to compare to another system. Oncoguide is intended to represent the standard of evidence-based care in the Netherlands (as it is a representation of the Dutch colorectal cancer guidelines in decision trees), thus it is a logical choice for evaluating the system for use in the Netherlands. (10) Comparison against an objective standard also adds value over only comparing with clinician judgment. (8)’

3. The authors claim to have developed a new approach to examining concordance but fail to provide a discussion of: existing methods for examining concordance, the shortcomings of these approaches, or the added value of their approach.

Reply: We agree that this discussion is necessary for a paper of this type. We have clarified the starting points for our method and the reasoning behind our extensions to existing methods (Discussion):

'Friedman and Wyatt proposed the use of contingency tables when evaluating decision support systems, to make explicit the difference between false positive and false negative classifications. Their reasoning is that a false-positive error, such as erroneously suggesting a diagnosis for a healthy patient, may be less serious than a false-negative error, which may in turn be less serious than proposing the wrong diagnosis entirely. (1) Likewise, a suggestion from the system that the clinician "consider" a treatment, which is in fact contraindicated, is a less serious error than "recommending" use of that treatment. Thus, we have extended the notion of a contingency table to express the idea that some disagreements have greater consequences than others.'

4. The manuscript demonstrates a significant lack of biomedical informatics expertise, as evidenced by:

i. the absence of a discussion of the science and corpus of knowledge on computer-based clinical decision support systems and their evaluation. There is a broad and rich literature on this subject. ii. the perfunctory approach to addressing the complex subject of usability including the use of cognitive walkthrough without offering any important citations in this field and with results based on descriptive comments rather than formal qualitative data

Reply: Please find our reply below comment B of the editorial board.

5. The authors provide no logic or principles to support their scoring scheme.

Reply: The editors also recommended providing justification of our scoring system; please see our reply below comment C of the editorial board.

6. There is no indication that the descriptions of the Watson for Oncology treatment advisor are accurate or current since confirmation by the technology provider is not mentioned and reference to the version of the technology is not offered.

Reply: The authors are not affiliated with IBM or MRDM, and to assure scientific integrity, we do not plan to seek approval of the text from the technology provider. However, we have added references to our description of Watson, and elided text that was based on non-refereed sources. The revised description now reads:

In the last decade, several "artificial intelligence" or "cognitive computing" medical decision support initiatives have gained attention. One of these tools is IBM Watson for Oncology (abbreviated as Watson).(11-13) Watson uses natural language processing to extract data from free text in medical records and select treatments from consensus guidelines.(14) Its selection of treatments is refined using machine learning, trained by specialists from New York's Memorial Sloan Kettering Cancer Center.(14)

Additionally, we have added information about the versions:
(to the protocol): 'The newest versions of both systems will be used.'

(to the feasibility study): ‘The latest versions of the software were used (Oncoguide 1.1.0 and Watson 17.3).’

Comments – Reviewer #2:

General:

1. **Overall my most major issue is with the overall aim of the paper; At every stage of the paper it is challenging for the reader to discern the methodological hypothesis and testing, and separating it from the more 'practical' validation of the content of IBM's Watson for Oncology tool.**

Reply: We agree that this distinction is important, and should be clear for our readers. The editorial board also suggested this change, and the response can be found under comment D above.

Introduction

2. **In line with Ewout Steyerberg's work (although more reflecting on predictive and less on diagnostic decision support tools), I would suggest to use internal and external validation terminology; This validation takes place in another country, according to more local standards standards. I miss a clear connection to available knowledge and guidelines on hot to evaluate cognitive decision support tools.**

Reply: We agree that the use of the term "content validity" was confusing, and that we did not adequately place this work in the context of existing methods for evaluation of decision support systems. This was also pointed out by the editors, and the response can be found under comment A and B above. We have chosen to refer to our goal as "evaluation," as this is the term used most often in the field of decision support. We agree that there are many parallels to external validation of prediction models, and have added this to the Discussion. ‘In many respects, a machine learning system can be viewed as a prediction model: the system "predicts" which treatment experts would recommend for this patient. Thus this evaluation can also be viewed as an "external validation" of this model: US-based experts trained the system, and it is not known if its recommendations will be valid in another setting.’

Methods en Results

3. **I do not see the ethical challenges of using real (retrospective) cases compared to synthetic cases. Can you elaborate?**

Reply: Ethical challenges of using real (retrospective) cases may concern patients with extraordinary features that somehow could be identified based on their clinical features. However, the main reason to choose for synthetic cases was to exhaustively test for differences with the guideline recommendations. Rather than add more exposition to the Methods, we therefore changed the following sentence in our manuscript: ‘We decided to generate synthetic patient cases instead of real patient data to exhaustively test for differences with the guideline recommendations.’

4. **In order to validate the use of synthetic cases as a tool compared to real cases some type of validation is needed; Can we replace real cases like this since we will extrapolate results of validation processes like this to the real world.**

Reply: We agree that this choice needs additional explanation. As described above, the main reason to choose for synthetic cases was to exhaustively test for differences with guideline recommendations. We have added the following text to explain the need for this kind of testing (Introduction):

‘Previous efforts in evaluating Watson have thus far only been reported as conference abstracts (3-5), and these short reports indicate that the evaluations consisted of comparing the output of Watson in actual clinical cases against the evaluation of clinical experts. However, although this approach approximates the validation step of a typical evaluation, a selection of consecutive clinical cases probably represents only a small sample of possible cases. Common cases are likely to be overrepresented. Unusual cases may not appear at all. Since machine learning systems tend to perform better when they have been trained with more data, Watson may also perform better in common cases than in unusual ones – and since it is precisely these unusual cases where clinicians may seek advice, a systematic approach to testing is needed before performing an impact study.’

We agree with the reviewer that an important next step is to test the system with real patient cases. However, all real patient cases will somehow fit in the decision tree and will thus be comparable with a synthetic patient case as used in our study.

5. **The walk through approach really helps, thank you. However, why was it not chosen to do traditional accuracy testing like (AU)ROC characteristics?**

Reply: We agree that this requires further explanation. We have added the following text to explain our choice:

‘As with other decision support systems, simple "agreement" is not sufficient to describe the performance of this system. (1) Friedman and Wyatt proposed the use of contingency tables when evaluating decision support systems, to make explicit the difference between false positive and false negative classifications. Their reasoning is that a false-positive error, such as erroneously suggesting a diagnosis for a healthy patient, may be less serious than a false-negative error, which may in turn be less serious than proposing the wrong diagnosis entirely. (1) Likewise, a suggestion from the system that the clinician "consider" a treatment, which is in fact contraindicated, is a less serious error than "recommending" use of that treatment. Thus, we have extended the notion of a contingency table to express the idea that some disagreements have greater consequences than others.’

Discussion and conclusions

6. **What about the issue that Watson was trained at Sloan Kettering? We see that results from the Watson can be off in different local settings; different decision making can take place where guidelines are absent/different or ethical choices lead away from the setting the algo was trained? Is the general and dogmatic approach of one algo, one training, and no retraining to local standards desirable? I think your conclusions can be more stronger in this respect but I agree with the line of reasoning.**

Reply: Thank you for this suggestion. We prefer to be a bit cautious with our conclusions at this time, as the evaluation we performed was only for a small, specific part of the guideline. However, we agree that re-training should be mentioned, and have added a sentence to our conclusion:

‘This may imply that Watson needs to be re-trained by local experts to reflect differences in the local care setting.’

7. Do you have suggestions for the 'black box' approach of Watson?

Reply: We agree that this is a potentially interesting topic for the Discussion, and have added a few sentences reflecting on this problem:

'"Black box" systems such as Watson impose a risk that other decision support systems do not, in that we cannot know exactly how the system arrives at its conclusions. For example, Oncoguide does not consider the patient's age in its recommendations; its users are aware of this and compensate accordingly. Watson *may or may not* be considering age, and its end users have no way to know when it does, or whether this may change with a new version update. A partial solution could be to maintain a suite of test patients as we've proposed in our protocol, and to run these tests regularly.

Then clinicians could be made aware if its recommendations change for some groups of patients.'

Comments – Reviewer #3:

1. Please provide a to-the-point, verifiable description of Watson. Although the reasoning of Watson is "opaque to the user", the system is described in wording that in my opinion would better fit the IBM website. I suggest to drop descriptions as "cognitive computer system", casebased reasoning, neural networks and the like. Also the assumed approach of Watson integrating natural language process with guidelines and studies do not contribute to the article. The extraction of data from the EMR is not studied here and therefore out of the scope of this article. The authors suggest that Watson uses the American guidelines in its reasoning. This also seems not to be verifiable.

So, I believe that at best, we can describe Watson as a black box algorithm that assumes a predefined set of parameters and produces recommendations for chemotherapy.

Also, I assume Watson to be a system that evolves over time. Can you indicate which version of Watson you have evaluated?

Reply: We have improved our description of Watson by adding citations, and eliding text that was based on non-refereed sources. We agree that terms such as "cognitive computing," although widely used, are poorly defined. We also agree that Watson is best described as a "black box" system, and have used that wording in our description. The new description reads:

'In the last decade, several "artificial intelligence" or "cognitive computing" medical decision support initiatives have gained attention. One of these tools is IBM Watson for Oncology (abbreviated as Watson).(11-13) Watson uses natural language processing to extract data from free text in medical records and select treatments from consensus guidelines.(14) Its selection of treatments is refined using machine learning, trained by specialists from New York's Memorial Sloan Kettering Cancer Center.(14)'

2. Rationale of use case. You have selected post-surgery recommendations for colon cancer patients. I understand that for this use case a well-defined ground truth (oncoguide) is available, but I do wonder if there are other reasons for this use case? Does Watson only cover recommendations in Medical Oncology? Would a use case be supported where multiple types of therapies can be an option (e.g. in prostate cancer or breast cancer)? In colon cancer, why did you restrict to the post-surgery use case? And, why did you exclude Stage IV patients?

Reply: Oncoguide, decision-tree based software representing the Dutch colorectal cancer guideline is available for all colorectal cancer stages and is not limited to post-surgery cases. We used synthetic patient cases in the adjuvant setting as a first example to evaluate concordance between the Dutch guidelines and Watson. It provided us with a first impression

of concordance levels and recurrent deviations. We intend to repeat our study with more complicated patient cases (e.g. stage IV patients). Besides, there are clear guideline recommendations available for the adjuvant setting (colorectal cancer guideline 2014), which facilitated comparison with Watson. Watson was available for the following tumor types during the conduction of our study: breast, lung, colon, rectal, gastric, cervical and ovarian cancer. For some of these tumor type are multidisciplinary recommendations available, but not for colon cancer yet. We revised the following sentences of our manuscript to clarify the choice for patients in the adjuvant setting (Methods – part 2. Feasibility study):
'To assess the practicality of and illustrate the application of our proposed protocol, we performed a feasibility study using the Dutch colorectal cancer guidelines for the adjuvant setting. Following the protocol outlined in Part 1, we generated cases simulating patients that underwent resection of stage I-III colon cancer with curative intent and who might be eligible for adjuvant treatment with chemotherapy. We used this patient category as a first example to evaluate concordance between the Dutch guidelines and Watson. We also chose for this patient category as clear and straightforward guideline recommendations are available in the most recent Dutch guideline from 2014 (15) which facilitates comparison with Watson's treatment advice. '

Additionally, we have clarified in the Discussion:
'We intend to repeat our study with patients with more complicated features.'

3. Role of NCCN Guidelines in the paper. You hypothesize that Watson is based on the NCCN guidelines; I think we are all willing to assume this. But, while you present some differences between Dutch and American recommendations, it remains unclear whether Watson indeed presents recommendations according to these NCCN guidelines. The usage of guidelines in Watson however, is central in your argumentation, as "localization of Watson with adjustments based on national or local guidelines" is needed. Also Watson seems to perform better in the Asian studies, one of the arguments being that those local guidelines may more closely parallel the US guidelines". Therefore, I believe the impact and reasoning of our work would substantially improve if you would indeed not only compare Watson with Oncoguide but also the NCCN guidelines.

Reply: We agree with the reviewer that the incorporation of a comparison of Watson with the NCCN guidelines improves the value of our paper. We therefore conducted additional analyses and incorporated the results in our paper as described below comment E. of the editorial board.

4. Usability Test. I have doubts or have not fully understood the impact of the usability experiment conducted. You provide the initials of the users, I do hope they no co-authors of your article. The set-up does not seem to mimic the usage of Watson in the clinic, as no integration with the EMR is provided. I believe you provide the users with a list of items to fill in on the UI, including "contraindication oxaliplatin", which we will most likely not encounter in any patient record. Also, the interface is not offered by IBM but by MRDM. The consequence of a well- or ill-usable interface for me is unclear, which makes it hard for me to appreciate this section. In any way, it would be interesting to include a screen shot, so we better understand the interface you are piloting. **Reply:** We agree that the usability assessment was poorly explained. We have added some additional explanation of the method, as well as

some explanation for why we chose this method. Cognitive walkthrough is an assessment performed by experts, who in this case are authors of the article. None of the authors are affiliated with IBM or MRDM, thus this does not introduce a conflict of interest. To our knowledge, no direct interface is available between Watson and our hospital information system at this time. MRDM aims to be the platform that would facilitate the connection between EMRs and Watson in the future. We have also added text explaining why usability testing was important as a part of our feasibility study (Methods and Discussion):
'Another important aspect of evaluating a decision support system is evaluation of the system's usability. Serious usability issues could lead to inability to use the system, or misinterpretation of the results. The interface offered for use in the Netherlands is a relatively simple form-based interface provided by MRDM. Patient data must be copied and entered into the form. As the primary goal of this evaluation was to determine whether this interface would be usable in subsequent testing, a cognitive walkthrough method was chosen. (9, 16) Cognitive walkthrough is an evaluation performed by experts, in which a set of goals is specified along with the actions required to complete the goals.'
...
'We elected to use a relatively simple, expert-based usability evaluation, the cognitive walkthrough approach. We considered this approach to be appropriate to the circumstances of this evaluation: the tasks to be accomplished in the system were well-defined, and the main goal was to identify usability issues that could be a barrier to use of the system in a trial setting with naive users. (9) Furthermore, the interface itself is relatively simple, and more qualitative methods such as thinkaloud were, in the authors' view, unlikely to yield additional insights. Although no usability problems were identified and the system is usable for our proposed evaluation, the workflow of hand-entering data is cumbersome. Direct interoperability with a patient record database would be preferable, but it is not yet available for the electronic health record in use at our hospital nor for synthetic cases.'
5. Experiment design and sharing of use cases. A big contribution is that you describe a test set for evaluation clinical decision support systems in oncology as well as a well-designed mechanism to score the performance. I think it would be a big contribution you could make the test set as well as the scoring tool available to the community. It will help to compare against other systems and also help to measure performance of future versions of Watson.
Reply: Thank you for this suggestion. We agree that making the data set with synthetic patient cases including the analysis of all cases would be useful to our readers. We have therefore uploaded it as supplementary files. Our scoring tool (without data) is represented in Table 1 and examples of analyzed cases are presented in Table 5. Further explanation of our scoring tool is given in Table 2 and Figure 1 (the concordance matrix).

References used in 'response to reviewers'
(please note that the numbers of these references do not correspond to the reference numbers in our manuscript as not all references of our manuscript are used in the 'response to reviewers' section)

1. Friedman CP WJ. Evaluation Methods in Biomedical Informatics, 2nd edition. 2006
2. Wyatt J. Quantitative evaluation of clinical software, exemplified by decision support systems. Int J Med Inform. 1997;47(3):165-73.
3. Suthida Suwanvecho HS, Montinee Sangtian, Andrew D Norden, Alexandra Urman, Annette Hicks, Irene Dankwa-Mullan, Kyu Rhee, Narongsak Kiatikajornthada. Concordance assessment of a cognitive computing system in Thailand.: J Clin Oncol; 2017.

4. Catherine Sarre-Lazcano AAA, Fidel David Huitzil Melendez, Oscar Arrieta, Andrew D Norden, Alexandra Urman, Mariana Perroni, Alice Landis-Mcgrath, Heriberto Medina-Franco. Cognitive computing in oncology: A qualitative assessment of IBM Watson for Oncology in Mexico.: J Clin Oncol; 2017.

5. S.P. Somashekhar M-JS, Andrew D Norden, Amit Rauthan, Kumar Arun, Poonam Patil, Ramya Y Ethadka, Rohit C Kumar. Early experience with IBM Watson for Oncology (WFO) cognitive computing system for lung and colorectal cancer treatment.: J Clin Oncol 2017.

6. Smith AE, Nugent CD, McClean SI. Evaluation of inherent performance of intelligent medical decision support systems: utilising neural networks as an example. Artif Intell Med. 2003;27(1):1-27. 7. Kaplan B. Evaluating informatics applications--clinical decision support systems literature review. Int J Med Inform. 2001;64(1):15-37.

8. McNair JB. Handbook of Evaluation Methods for Health Informatics. 2005.

9. Jaspers MW. A comparison of usability methods for testing interactive health technologies: methodological aspects and empirical evidence. Int J Med Inform. 2009;78(5):340-53.

10. Van Oijen MG VX, Van Vegchel T, Nagtegaal ID, Lahaye M, Méndez Romero, A, Rütten H, De Bruijn S, Verheul HM, Tanis PJ, Punt CJA, Keikes L. Improving visualization and adherence by converting the Dutch colorectal cancer guidelines into decision trees: The Oncoguide project. Annals of Oncology. 2017;28(10):1093.

11. https://www.ibm.com/watson/health/oncology-and-genomics/oncology/.

12. Allain JS. From Jeopardy to Jaundice: The Medical Liability Implications of Dr. Watson and
Other Artificial Intelligence Systems. La L Rev. 2012;73:1049.

13. Khan OF BG, Alimohamed NA. Artificial intelligence in medicine What oncologists need to know about its potential - and its limitations. Oncology exchange 2017;16(4):8-13.

14. Bach P ZM, Gucalp A, Epstein AS, Norton L, Seidman AD, Caroline A, Grigorenko A, Bartashnik A, Wagner I, Keesing J, Kohn M, Hsiao F, Megerian M, Stevens RJ, Malin J, Whitney J, Kris MG. Beyond Jeopardy!: Harnessing IBM's Watson to Improve Oncology Decision Making. J Clin Oncol.
2013;31:(suppl; abstr 6508).

15. Dutch colorectal cancer guideline 2014 [Available from:
http://www.oncoline.nl/colorectaalcarcinoom.

16. Wharton C RJ, Lewis C, Polson P "The cognitive walkthrough method: a practitioner's guide" in J. Nielsen & R. Mack "Usability Inspection Methods" 1994:105-40.

Please find below an overview of the references we've added to our revised manuscript: (the numbers of these references correspond to the reference numbers in our manuscript)

4. Khan OF BG, Alimohamed NA. Artificial intelligence in medicine What oncologists need to know about its potential - and its limitations. Oncology exchange 2017;16(4):8-13.

5. Bach P ZM, Gucalp A, Epstein AS, Norton L, Seidman AD, Caroline A, Grigorenko A, Bartashnik A, Wagner I, Keesing J, Kohn M, Hsiao F, Megerian M, Stevens RJ, Malin J, Whitney J, Kris MG. Beyond Jeopardy!: Harnessing IBM's Watson to Improve Oncology Decision Making. J Clin Oncol.
2013;31:(suppl; abstr 6508).
8. Friedman CP WJ. Evaluation Methods in Biomedical Informatics, 2nd edition. 2006
9. Wyatt J. Quantitative evaluation of clinical software, exemplified by decision support systems. Int J Med Inform. 1997;47(3):165-73.
13. Smith AE, Nugent CD, McClean SI. Evaluation of inherent performance of intelligent medical decision support systems: utilising neural networks as an example. Artif Intell Med. 2003;27(1):1-27. 19. Wharton C RJ, Lewis C, Polson P "The cognitive walkthrough method: a practitioner's guide" in J. Nielsen & R. Mack "Usability Inspection Methods" 1994:105-40.
20. Jaspers MW. A comparison of usability methods for testing interactive health technologies: methodological aspects and empirical evidence. Int J Med Inform. 2009;78(5):340-53.
21. Kaplan B. Evaluating informatics applications--clinical decision support systems literature review. Int J Med Inform. 2001;64(1):15-37.
23. McNair JB. Handbook of Evaluation Methods for Health Informatics. 2005.

2nd editorial decision

Date: 28-Jun-2018

Ref.: Ms. No. JCTRes-D-18-00009R1
The first steps in the evaluation of a "black-box" decision support tool: a protocol and feasibility study for the evaluation of Watson for Oncology
Journal of Clinical and Translational Research

Dear authors,

I am pleased to inform you that your manuscript has been accepted for publication in the Journal of Clinical and Translational Research.

This was a collective decision by multiple editorial board members because reviewer #1 kept insisting on a rejection, despite the fact that the majority of his objections regarding your initial submission were properly addressed in our view and the view of reviewers #2 and 3. What ultimately made us decide in favor of accepting this manuscript for publication was the explicitly mentioned preliminary character of your method (starting with the title), which automatically warns the reader that the support tool was developed in the context of a feasibility framework without all the ultimately desirable clinical validation steps (as had been advocated by reviewer #1).

You will receive the proofs of your article shortly, which we kindly ask you to thoroughly review for any errors.

Thank you for submitting your work to JCTR.

Kindest regards,

Michal Heger
Editor-in-Chief
Journal of Clinical and Translational Research

Comments from the editors and reviewers:

Reviewer #1: Authors needs to do proper prospective study as suggested previously in our comments and suggestions

Reviewer #2: The comments have been taken on across the board and for me the manuscript is now acceptable for publication.
The comments about the importance of identifying the differences and similarities between the different test populations (build up and constitution of cases) remains essential to the results oif the validation. It remains a difficult discussion and hard to reflect in results.
I thank the authors and editor for the opportunity to read and review about this interesting issue.

Reviewer #3: Dear authors, I would like to congratulate you on the tremendous improvement of your manuscript. It was a pleasure to read the responses on the comments by the reviewers and to see your willingness to conduct additional research.