PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	KEY DRIVERS OF INNOVATIVENESS APPRAISAL FOR MEDICINES: THE ITALIAN EXPERIENCE AFTER THE ADOPTION OF THE NEW RANKING SYSTEM
AUTHORS	Galeone, Carlotta; Bruzzi, Paolo; Jommi, Claudio

VERSION 1 – REVIEW

REVIEWER	Maximilian Salcher-Konrad London School of Economics and Political Science, United Kingdom
REVIEW RETURNED	17-Jul-2020

GENERAL COMMENTS	In this article on an interesting and timely topic, the authors empirically reviewed the first 3 years of the Italian Medicine Agency's innovativeness appraisal framework for new medicines and found that only two of the three parameters for assessing innovation (i.e., added therapeutic value and quality of evidence) were driving "innovative" drug status, with added therapeutic value the most important factor. This study adds to existing literature about innovativeness of medicines, typically appraised by European HTA bodies (as referenced by the authors), by providing an account of the Italian experience. By reviewing the characteristics of drugs assessed by AIFA for innovativeness, the authors aimed to identify the main drivers of an "innovative" status of drugs. Given that the Italian innovativeness appraisal framework is relatively new, such a study has its merit to better understand how the framework is implemented, but the aims of the study should be clearly stated, and conclusions should be made within the remit of the study. Instead, the authors described their study as a critical review, and provided an overall assessment of the performance of the framework as enhancing transparency, accountability and predictability. In my view, the analysis fell short of a critical review. Instead of critically scrutinising the framework, its assumptions, and resulting assessments of new drugs, the analysis was conducted within the parameters of the framework. It feels like a missed opportunity not to dig deeper into whether this framework is fit for purpose. If this is outside the scope of the study, then the way the authors present their review and its implications should be revised. A fundamental methodological question about this study is whether there is selection bias into the drugs being assessed for innovativeness. The authors state this process is typically started by industry (although no data on who initiated the assessment was available). This is important: if industry already know whether a drug addresses an unmet therapeutic need (there is reason to believe

	they would, since there would likely already be an external assessment of how FDA and EMA assess the role of the drug in the therapeutic landscape) and only submit drugs where such unmet need is relatively certain, then this becomes the most important criterion of whether a drug is considered innovative or not. Drugs that do not address an unmet need would then be effectively excluded from the sample of drugs being assessed for innovativeness. Indeed, only 3 of 53 drugs in the sample had a "poor" rating for unmet therapeutic need, and all others had at least a moderate rating. Unmet therapeutic need therefore appears to be a bottleneck for drugs to be eligible for an innovativeness assessment. This is briefly mentioned in the discussion section, but the "decision tree" (Figure 3) and the overall conclusion do not reflect the importance of unmet need. An important finding, in my view, is that there were some drugs with moderate added therapeutic value (i.e., they are not better than existing alternatives with respect to patient-relevant endpoints) and low quality of evidence that were assessed as conditionally innovative. A critical review of these drugs and how they can be considered "innovative" when the evidence base for their therapeutic effect is highly uncertain would be an important contribution to the literature. Is it the case that, for these drugs, the unmet need is considered to be so high that a new drug would effectively have to be proven to worsen a patient's life to not be considered innovative? If so, what is the definition of a sufficiently high unmet need to warrant such a low bar for innovation (note that the added therapeutic value for these was "moderate", meaning that alternative treatments do exist – although these may not be working very well)? Some more specific comments and questions: - P11, para 3: the description of the recursive algorithm is too short to be replicated. Given that this plays a big role in the results section, the methods should be expanded. - Table 1: There were only 5 appraisals in 2019, compared to more than 20 in the two previous years. Is there an explanation for this? Were reports for some of the appraisals conducted in 2019 not available at the time the database was searched? Was there a change in policy in terms of eligibility, or a change in incentives for companies to apply for innovative status? - Some results from Table 1 are described in the text despite there not being a statistically significant difference between innovative and non-innovative drugs ("More recently assessed medicines, orphan drugs, pediatric/mixed indications, and medicines approved with at least one RCT…"). This omits other results from Table 1 that, while not statistically significant either, would warrant commenting on. Most importantly, the absence of RCT evidence was associated with a higher chance of a drug being assessed as innovative (although the sample for this subgroup was relatively small). - P16: "In oncological setting, innovative drugs provided on average more RCT evidence in support of the application when compared to non-oncological ones." This statement is not accurate: the proportion of oncology drugs relying on non-RCT data was approximately one third, while there was only 1 non-oncology drug that relied on non-RCT data. Further, of those drugs that were deemed innovative, one third of oncology drugs only had non-RCT data, while this was the case for only 1 of 11 innovative non-oncology drugs. - Table 2: What was the reasoning behind comparing the mean (and median) values of unmet therapeutic need, added therapeutic value and evidence quality in addition to the proportions of drugs in each

	category? These are not continuous measures, so I was not sure what the meaning of a mean value of, e.g., 2.5 or 3.6 for added therapeutic value is. I found the proportions much more interesting and intuitive to understand. For example, none of the assessed drugs was found to have maximum added therapeutic value, and the largest category for innovative drugs was moderate therapeutic value. Similarly, the quality of evidence was overall quite low, with more drugs being assessed to have low quality evidence vs. high quality. Given that this is the first empirical assessment of this framework, these results are worth highlighting. - The study would benefit from reflecting more on the existing critical literature on therapeutic value of new drugs (not just in terms of direct comparison of results with other European countries). A non-exhaustive list of references assessing the value of drugs, often intended for life -threatening or seriously debilitating diseases (similar to the entry criteria for AIFA's innovativeness framework), that the authors may want to consider: o Aggarwal, A., Fojo, T., Chamberlain, C., Davis, C., & Sullivan, R. (2017). Do patient access schemes for high-cost cancer drugs deliver value to society?-lessons from the NHS Cancer Drugs Fund. Annals of Oncology, 28(8), 1738–1750. https://doi.org/10.1093/annonc/mdx110 o Grössmann, N., Del Paggio, J. C., Wolf, S., Sullivan, R., Booth, C. M., Rosian, K., Emprechtinger, R., & Wild, C. (2017). Five years of EMA-approved systemic cancer therapies for solid tumours—a comparison of two thresholds for meaningful clinical benefit. European Journal of Cancer, 82, 66–71. https://doi.org/10.1016/j.ejca.2017.05.029 o Gyawali, B., Hey, S. P., & Kesselheim, A. S. (2019). Assessment of the Clinical Benefit of Cancer Drugs Receiving Accelerated Approval. JAMA Internal Medicine, 179(7), 906–913. https://doi.org/10.1001/jamainternmed.2019.0462 o Trotta, F., Mayer, F., Barone-Adesi, F., Esposito, I., Punreddy, R., Da Cas, R., Traversa, G., Perrone, F., Martini, N., Gyawali, B., & Addis, A. (2019). Anticancer drug prices and clinical outcomes: A cross-sectional study in Italy. BMJ Open, 9(12). https://doi.org/10.1136/bmjopen-2019-033728 o Wieseler, B., McGauran, N., & Kaiser, T. (2019). New drugs: where did we go wrong and what can we do better

REVIEWER	Joerg Ruof r-connect ltd; Basel
REVIEW RETURNED	17-Jul-2020

GENERAL COMMENTS	Thanks very much for the opportunity to review this interesting paper analysing the new AIFA innovation framework that was introduced in 2017. The authors analysed 54 (53, respectively) AIFA appraisal reports on innovativeness that were published between 2017 - 2019. The impact of the three innovation criteria i) unmet medical need; ii) added therapeutic value and iii) quality of evidence on AIFA's final innovation categorization was analysed. The added therapeutic value turned out to be the most important driver of AIFA's final decision. My comments on the paper are: 1) What I'm missing most when reading the paper is a list/ table with the 54 (53 respectively) procedures. The list should contain a minimum of information such as e.g. name of the medicine/ innovation status/ outcome for each of the three criteria (unmet

	need/ added value/ quality of evidence)/ and e.g. disease characteristics (onc/ non-onc/ Orphan). 2) The three criteria are i) innovative; ii) conditionally innovative; iii) not innovative. Table 2 of the paper does not include the 'conditionally innovative' category and it remains unclear to me how this category factors into the whole message of the paper. Is the key conclusion of the paper (added value is the most important driver of innovativeness decision) equally valid for both categories ('fully innovative'; 'conditionally innovative')? 3) The authors claim that since the introduction of the new process in Italy in 2017 this is the first investigation of the drivers of innovativeness. I think this statement is not correct. I'm aware of at least a couple of other publications/ posters of this particular subject and authors should clearly describe if and how their approach differs from previous assessments (i) Rova A, Cioni L, Urbinati D. The Impact of the new innovative Algorithm in Italy; Poster at ISPOR Europe 2018 Nov 10 - 14 in Barcelona/ Spain; ii) Slight S. Italy's New Pharmaceutical Innovation Ranking System: Key Criteria for Successfully Achieving Innvation Status. Poster PHP274, Europe 2018 Nov 10 - 14 in Barcelona/ Spain; iii) Fortinguerra F, Tafuri G, Trotta F, Addis A. Using GRADE methodology to assess innovation of new medicinal products in Italy. PJCP 2019 DOI:10.1111/bcp.14138; Meremetidis A, Ferrario M, Giuliani G. The new AIFA innovation framework to recognize innovative drugs: a preliminary analysis of key drivers of evaluation. Poster PHP61, ISPOR Europe 2018 Nov 10 - 14 in Barcelona/ Spain) 4) Within the methods section the authors describe that they analyzed categorical data Chi-square of Fisher's eact test as appropriate. Continuous data instead were analyzed using by Student's T test (if normally distributed) or by Wilcoxon rank-sum test otherwise. The authors should be explicit which analyses were conducted and the respective test method per each of the relevant analysis. The whole testing prcedure remains totally unclear to me. Also, as scoring for all three innovation criteria is categorical it remains unclear when Student's T - test was applied. 5) The authors developed a recursive algorithm for innovativeness. Details of this algorithm remain totally unclear. Methods are not clearly described. What are the details of the algorithm; was a regression analysis conducted? Do we have any insights into the predictive capacity of any of the contributing variables etc. etc? What was the R square? The authors also state that their decision tree accounted for 43 out of 53 cases (81%). For the other 10 appraisals it was not possible to discern factors determining the final appraisal. It remains unclear why this was not possible. Is this a value judgement by the authors? As this relates to almost 20% of the cases there seems to be a considerable source of bias. Also: authors claim a high level of internal consistency. This statement should be substantiated by facts/ numbers. 6) While I tend to agree with most of the authors' conclusions I think the conclusion section is too long and some of the raised points should be moved to the discussion.

VERSION 1 – AUTHOR RESPONSE

Reviewer: 1

Reviewer Name: Maximilian Salcher-Konrad Institution and Country: London School of Economics and Political Science, United Kingdom Please state any competing interests or state 'None declared': None declared.

In this article on an interesting and timely topic, the authors empirically reviewed the first 3 years of the Italian Medicine Agency's innovativeness appraisal framework for new medicines and found that only two of the three parameters for assessing innovation (i.e., added therapeutic value and quality of evidence) were driving "innovative" drug status, with added therapeutic value the most important factor.

Reply:
We thank Professor Salcher-Konrad for his interesting comments that, together with Professor Ruof's comments, helped us to improve our paper. Before replying to specific comments, we underline that we have revised the analyses and the whole manuscript by including updated information released by AIFA during 2020 on 23 additional appraisal documents. The total number of documents is therefore now 77. This improved the knowledge of this new topic and the key-messages remained substantially unchanged after this update.

#1. This study adds to existing literature about innovativeness of medicines, typically appraised by European HTA bodies (as referenced by the authors), by providing an account of the Italian experience. By reviewing the characteristics of drugs assessed by AIFA for innovativeness, the authors aimed to identify the main drivers of an "innovative" status of drugs. Given that the Italian innovativeness appraisal framework is relatively new, such a study has its merit to better understand how the framework is implemented, but the aims of the study should be clearly stated, and conclusions should be made within the remit of the study. Instead, the authors described their study as a critical review, and provided an overall assessment of the performance of the framework as enhancing transparency, accountability and predictability. In my view, the analysis fell short of a critical review. Instead of critically scrutinising the framework, its assumptions, and resulting assessments of new drugs, the analysis was conducted within the parameters of the framework. It feels like a missed opportunity not to dig deeper into whether this framework is fit for purpose. If this is outside the scope of the study, then the way the authors present their review and its implications should be revised.

Reply to #1:
We thank for this crucial comment and to help us to better clarify the aims of our manuscript. We agree our project is not a "critical review". We deleted it from the title and through the manuscript. We revised section "Introduction" and "Discussion/Conclusion" (also as suggested by Editor, see #2) making more explicit that the aim of our study was understanding how the new Italian innovativeness appraisal framework was implemented and covering three information gaps: the role played by the three criteria on the final decision, if these criteria have been consistently used over time and if other variables influence the innovativeness status.

#2. A fundamental methodological question about this study is whether there is selection bias into the drugs being assessed for innovativeness. The authors state this process is typically started by industry (although no data on who initiated the assessment was available). This is important: if industry already know whether a drug addresses an unmet therapeutic need (there is reason to believe they would, since there would likely already be an external assessment of how FDA and EMA assess the role of the drug in the therapeutic landscape) and only submit drugs where such unmet need is relatively certain, then this becomes the most important criterion of whether a drug is considered innovative or not. Drugs that do not address an unmet need would then be effectively excluded from the sample of drugs being assessed for innovativeness. Indeed, only 3 of 53 drugs in the sample had a "poor" rating for unmet therapeutic need, and all others had at least a moderate rating. Unmet therapeutic need therefore appears to be a bottleneck for drugs to be eligible for an

innovativeness assessment. This is briefly mentioned in the discussion section, but the “decision tree” (Figure 3) and the overall conclusion do not reflect the importance of unmet need.

Reply to #2:
This is comment is very interesting and was deeply discussed by the working group during all phases of our project.
From a statistical point of view, we have no methods to consider the selection bias, since we have no available data on which considerations/decisions were performed by industries before applying for innovativeness. We do believe, like the reviewer, that the unmet need is an important criterium for innovativeness, but probably it is a-priori condition as we stated in section “Discussion”. The decision tree reported in the manuscript was merely data-driven (i.e., the analyses were ex-post) and, also in the updated analyses here reported on 77 appraisals, emerged that added therapeutical value and quality of evidence are crucial for the decision on innovativeness *status.* This result is not interpreted as ‘the unmet need does not count’ but as ‘companies are likely to apply for innovativeness if they consider important the unmet need”. Poor rating for five cases could derive from the circumstance that the ‘unmet need’ assessment is still quite controversial. This has been quoted in the conclusions.

#3. An important finding, in my view, is that there were some drugs with moderate added therapeutic value (i.e., they are not better than existing alternatives with respect to patient-relevant endpoints) and low quality of evidence that were assessed as conditionally innovative. A critical review of these drugs and how they can be considered “innovative” when the evidence base for their therapeutic effect is highly uncertain would be an important contribution to the literature. Is it the case that, for these drugs, the unmet need is considered to be so high that a new drug would effectively have to be proven to worsen a patient’s life to not be considered innovative? If so, what is the definition of a sufficiently high unmet need to warrant such a low bar for innovation (note that the added therapeutic value for these was “moderate”, meaning that alternative treatments do exist – although these may not be working very well)?

Reply to #3:
We thank the reviewer for this interesting comment. As also replied to Reviewer #2 (Reply#2), we would have been very pleased to have a further insight into “conditionally innovative” vs “fully innovative”, instead of aggregating them into “innovative” vs "not innovative". However, the available appraisals are scanty to allow comparisons between three groups. We will address this interesting research question as soon as the available appraisals increase. We added this issue in the limitation of the study in section “Discussion”.

Some more specific comments and questions:
#4. P11, para 3: the description of the recursive algorithm is too short to be replicated. Given that this plays a big role in the results section, the methods should be expanded.

Reply to #4:
We thank for this comment that helped to improve the manuscript. The recursive algorithm is based on a deterministic approach. This approach was merely data-driven, and the univariate analyses on the role played by the three domains on innovative status were the starting point to create the decision tree. In these univariate analyses we found that added therapeutic value ($p<0.01$) and quality of evidence ($p=0.03$) were associated to innovativeness status and for these reasons we developed this data-driven decision tree using a deterministic approach. The probabilistic approach, using multivariate logistic regression models was considered, but too few data were available to apply these models. We reported these details in section “Method” and “Results”.

#5 Table 1: There were only 5 appraisals in 2019, compared to more than 20 in the two previous years. Is there an explanation for this? Were reports for some of the appraisals conducted in 2019 not available at the time the database was searched? Was there a change in policy in terms of eligibility, or a change in incentives for companies to apply for innovative status?

Reply to #5:

The explanation is that there is a lag-time between CTS meeting date (i.e., the date when AIFA decides the innovation status) and the publication of the appraisal document on the website, which takes place at the same time as that the price and reimbursement status is published on the Gazzetta Ufficiale of the Italian Government.
However, since we updated the analysis there are now 24 appraisals released during 2019. No changes in policy in terms of eligibility or in incentives for companies were applied during the last year.

#6 Some results from Table 1 are described in the text despite there not being a statistically significant difference between innovative and non-innovative drugs ("More recently assessed medicines, orphan drugs, pediatric/mixed indications, and medicines approved with at least one RCT…"). This omits other results from Table 1 that, while not statistically significant either, would warrant commenting on. Most importantly, the absence of RCT evidence was associated with a higher chance of a drug being assessed as innovative (although the sample for this subgroup was relatively small).

#7 P16: "In oncological setting, innovative drugs provided on average more RCT evidence in support of the application when compared to non-oncological ones." This statement is not accurate: the proportion of oncology drugs relying on non-RCT data was approximately one third, while there was only 1 non-oncology drug that relied on non-RCT data. Further, of those drugs that were deemed innovative, one third of oncology drugs only had non-RCT data, while this was the case for only 1 of 11 innovative non-oncology drugs.

Reply to #6 and #7:
We revised all the section "Results", also in consideration of new data following the 2020 update of appraisals. We added a sentence on all variables reported in Table 1. Moreover, we agreed to comment #7 and we deleted the sentence on RCT.

#8 Table 2: What was the reasoning behind comparing the mean (and median) values of unmet therapeutic need, added therapeutic value and evidence quality in addition to the proportions of drugs in each category? These are not continuous measures, so I was not sure what the meaning of a mean value of, e.g., 2.5 or 3.6 for added therapeutic value is. I found the proportions much more interesting and intuitive to understand. For example, none of the assessed drugs was found to have maximum added therapeutic value, and the largest category for innovative drugs was moderate therapeutic value. Similarly, the quality of evidence was overall quite low, with more drugs being assessed to have low quality evidence vs. high quality. Given that this is the first empirical assessment of this framework, these results are worth highlighting.

Reply to #8:
We presented both measures (i.e., proportions and continuous data) to provide more detailed information. We agree with the Referee that the proportions are more interesting, but the distributions are often based on small frequencies (e.g., there are only 5 evaluations of "poor" unmet therapeutic need, only 1 "very poor" added therapeutic value, etc.) and therefore the corresponding tests for comparison between groups has low statistical power.

#9 The study would benefit from reflecting more on the existing critical literature on therapeutic value of new drugs (not just in terms of direct comparison of results with other European countries). A non-exhaustive list of references assessing the value of drugs, often intended for life -threatening or seriously debilitating diseases (similar to the entry criteria for AIFA's innovativeness framework), that the authors may want to consider:
Aggarwal, A., Fojo, T., Chamberlain, C., Davis, C., & Sullivan, R. (2017). Do patient access schemes for high-cost cancer drugs deliver value to society?-lessons from the NHS Cancer Drugs Fund. Annals of Oncology, 28(8), 1738–1750. https://doi.org/10.1093/annonc/mdx110
Grössmann, N., Del Paggio, J. C., Wolf, S., Sullivan, R., Booth, C. M., Rosian, K., Emprechtinger, R., & Wild, C. (2017). Five years of EMA-approved systemic cancer therapies for solid tumours—a comparison of two thresholds for meaningful clinical benefit. European Journal of Cancer, 82, 66–71. https://doi.org/10.1016/j.ejca.2017.05.029

Gyawali, B., Hey, S. P., & Kesselheim, A. S. (2019). Assessment of the Clinical Benefit of Cancer Drugs Receiving Accelerated Approval. JAMA Internal Medicine, 179(7), 906–913. https://doi.org/10.1001/jamainternmed.2019.0462
Trotta, F., Mayer, F., Barone-Adesi, F., Esposito, I., Punreddy, R., Da Cas, R., Traversa, G., Perrone, F., Martini, N., Gyawali, B., & Addis, A. (2019). Anticancer drug prices and clinical outcomes: A cross-sectional study in Italy. BMJ Open, 9(12). https://doi.org/10.1136/bmjopen-2019-033728
Wieseler, B., McGauran, N., & Kaiser, T. (2019). New drugs: where did we go wrong and what can we do better

Reply to #9:
We do thank the reviewer for this comment. We have included in the discussion a short summary of evidence on value frameworks. We still consider appropriate confronting our findings with the application of value frameworks in other countries.

Reviewer: 2

Reviewer Name: Joerg Ruof
Institution: r-connect ltd; Basel
Please state any competing interests or state 'None declared': No Competing Interest regarding this Paper

Thanks very much for the opportunity to review this interesting paper analysing the new AIFA innovation framework that was introduced in 2017. The authors analysed 54 (53, respectively) AIFA appraisal reports on innovativeness that were published between 2017 - 2019. The impact of the three innovation criteria i) unmet medical need; ii) added therapeutic value and iii) quality of evidence on AIFA's final innovation categorization was analysed. The added therapeutic value turned out to be the most important driver of AIFA's final decision.

Reply:
We thank Professor Rouf for his useful revision to our manuscript. His comments, together with those of Professor Salcher-Konrad, helped us to improve our paper, especially on methodological issue of our data analysis. Before replying to specific comments, we underline that we have revised the analyses and the whole manuscript by including updated information released by AIFA during 2020 on 23 additional appraisal documents. The total number of documents is therefore now 77. This improved the knowledge of this new topic; the key-messages remained substantially unchanged after this update.

My comments on the paper are:

#1. What I'm missing most when reading the paper is a list/ table with the 54 (53 respectively) procedures. The list should contain a minimum of information such as e.g. name of the medicine/ innovation status/ outcome for each of the three criteria (unmet need/ added value/ quality of evidence)/ and e.g. disease characteristics (onc/ non-onc/ Orphan).

Reply #1:
We thank for this comment that gives us the opportunity to provide the readers with more detailed information. As suggested, we prepared a Table with detailed information on medicines, innovation status, outcome for each of the three criteria and other useful information for readers, i.e., disease group - onco vs non onco, rare disease, target population - adult vs paediatric, CTS appraisal date). We included this Table as Supplementary file.

#2. The three criteria are i) innovative; ii) conditionally innovative; iii) not innovative. Table 2 of the paper does not include the 'conditionally innovative' category and it remains unclear to me how this

category factors into the whole message of the paper. Is the key conclusion of the paper (added value is the most important driver of innovativeness decision) equally valid for both categories ('fully innovative'; 'conditionally innovative')?

Reply #2:
"Fully innovative" and "conditionally innovative" appraisals were grouped together in a single "innovative" group. This definition was better clarified in the revised version of the manuscript and we added a note in Table 1 and Table 2. Univariate analyses reported in Table 1 and Table 2 include comparisons between innovative and non-innovative since the available innovativeness appraisals are scanty to allow comparisons between three groups. We will perform these analyses as soon as the available appraisals increase. We added this issue in the limitation of the study in section "Discussion".

#3. The authors claim that since the introduction of the new process in Italy in 2017 this is the first investigation of the drivers of innovativeness. I think this statement is not correct. I'm aware of at least a couple of other publications/ posters of this particular subject and authors should clearly describe if and how their approach differs from previous assessments (i) Rova A, Cioni L, Urbinati D. The Impact of the new innovative Algorithm in Italy; Poster at ISPOR Europe 2018 Nov 10 - 14 in Barcelona/ Spain; ii) Slight S. Italy's New Pharmaceutical Innovation Ranking System: Key Criteria for Successfully Achieving Innvation Status. Poster PHP274, Europe 2018 Nov 10 - 14 in Barcelona/ Spain; iii) Fortinguerra F, Tafuri G, Trotta F, Addis A. Using GRADE methodology to assess innovation of new medicinal products in Italy. PJCP 2019 DOI:10.1111/bcp.14138; Meremetidis A, Ferrario M, Giuliani G. The new AIFA innovation framework to recognize innovative drugs: a preliminary analysis of key drivers of evaluation. Poster PHP61, ISPOR Europe 2018 Nov 10 - 14 in Barcelona/ Spain)

Reply #3:
We thank for this comment. We added these references in section "Introduction" and we reported that previous studies were based on a few innovativeness appraisals (max 20, updated to 2018) and mainly reported descriptive analyses. We also modified the statement on the strength of our study.

#4. Within the methods section the authors describe that they analyzed categorical data Chi-square of Fisher's eact test as appropriate. Continuous data instead were analyzed using by Student's T test (if normally distributed) or by Wilcoxon rank-sum test otherwise. The authors should be explicit which analyses were conducted and the respective test method per each of the relevant analysis. The whole testing prcedure remains totally unclear to me. Also, as scoring for all three innovation criteria is categorical it remains unclear when Student's T - test was applied.

Reply #4:
As suggested, we better clarified in section "Methods" the statistical tests used for analyses. We also added a note in Table 1 and Table 2 to better clarify this important issue.
We performed Chi-square or Fisher's exact test to compare groups (i.e., innovative vs. non-innovative outcome) in case of categorical data (e.g., CTS appraisal year 2017, 2018, 2019). In case of comparison between groups according to continuous variables (e.g., Number of SoF or mean score of all three innovation criteria) we used the appropriate statistical tests, i.e. Student's T test or Wilcoxon rank-sum test.

#5. The authors developed a recursive algorithm for innovativeness. Details of this algorithm remain totally unclear. Methods are not clearly described. What are the details of the algorithm; was a regression analysis conducted? Do we have any insights into the predictive capacity of any of the contributing variables etc. etc? What was the R square?

Reply #5:
We thank for this comment that helped to improve the methodological details of our manuscript (see also our reply to comment #4 of Reviewer 1). The recursive algorithm for innovativeness was based on a deterministic approach. It was not based on a regression analysis (no R square can be calculated) but was data-driven. In the construction of this algorithm, the univariate analyses on the

role played by the three domains on innovative status (reported in Table 2) were crucial as a starting point to create the decision tree. In the univariate analyses we found that added therapeutic value (p<0.01) and quality of evidence (p=0.03) were associated to innovativeness status and for these reasons we developed this data-driven decision tree using a deterministic approach. The probabilistic approach, using multivariate logistic regression models was considered, but too few data were available to apply these models. We reported these details in section "Method" and "Results".

#6. The authors also state that their decision tree accounted for 43 out of 53 cases (81%). For the other 10 appraisals it was not possible to discern factors determining the final appraisal. It remains unclear why this was not possible. Is this a value judgement by the authors? As this relates to almost 20% of the cases there seems to be a considerable source of bias.
Also: authors claim a high level of internal consistency. This statement should be substantiated by facts/ numbers.

Reply #6:
We updated the decision tree after the inclusion of new appraisals released by AIFA and its good descriptive performances were confirmed, as 82% of appraisals fitted with the decision tree based only on two criteria. Please consider that this is only a description of the evaluations released to date, and there is thus no judgement by our working group. With further reference to the statement of internal consistency, we modified the Abstract and Discussion sections to comply with the Referee's comment.

#7. While I tend to agree with most of the authors' conclusions I think the conclusion section is too long and some of the raised points should be moved to the discussion.

Reply #7:
As suggested, we moved most of points from section "Conclusion" to "Discussion"

VERSION 2 – REVIEW

REVIEWER	Maximilian Salcher-Konrad London School of Economics and Political Science, UK
REVIEW RETURNED	29-Sep-2020

GENERAL COMMENTS	I want to thank the authors for revising the manuscript thoroughly and addressing my and the second reviewer's comments. Substantial additional work has gone into this revision through the extraction and analysis of 23 additional appraisals. My comments regarding the framing of the manuscript have been addressed. The aims of this work are now much clearer. However, it is still not clear to me how the question about consistent use of the criteria over time has been answered. I would expect some more detailed analysis of what happens over time to answer this question (going beyond the univariate analysis of year of assessment). Maybe this is just an issue of how this particular aim is worded (i.e., were the authors really interested in changes over time, or just consistency within the total sample of appraisals to date?). While the framing of the paper is now clear, some issues remain with respect to the analyses conducted. Univariate analyses (Table 2): in the analysis for the updated, larger sample, there is now a discrepancy for the results on quality of evidence (in terms of statistical significance) depending on whether the analysis is based on proportions (p=0.11) or the continuous

	measure of quality of evidence (p=0.03). In the previous version of the manuscript, the results were consistent across these analyses, so it did not matter too much which one the interpretation was based on. Given the discrepancy, a clear rationale for relying on one vs. the other is needed. This is missing from the methods section and should ideally have been specified a priori (note that the proportions results would assign the same importance to unmet therapeutic need and quality of evidence, so the choice of outcome measure is key to the conclusions of the study). Also note that this again raises the issue of why the two types of measures (proportions and continuous measures) are presented in the Table in the first place. The objective behind developing the algorithm for innovativeness assessment remain a little unclear to me, as do the methods for doing this (even after the revisions to address reviewer 2's and my comments). This is described as a deterministic exercise, yet the algorithm fails to account for almost 20% of cases. The problems with this approach and the interpretation of its findings become apparent on Page 21, para 3: "When the added therapeutic value was rated as "poor" or "absent", or when the GRADE evaluation was "very low", the indication is never considered innovative. Innovativeness resulted from an at least a "moderate" added therapeutic value combined with an at least a "moderate" GRADE evaluation". These statements are not accurate, as there was one (ultra-rare disease) case where GRADE was very low and the drug was still considered conditionally innovative, and another four where GRADE was low and the drugs were conditionally innovative. I understand that this is what the algorithm produces, but its weakness becomes obvious when looking at the flow chart and seeing these cases that directly contradict the statements quoted above.

REVIEWER	Joerg Ruof r-connect ltd
REVIEW RETURNED	18-Sep-2020

GENERAL COMMENTS	Thanks very much for all the work that went into the revision of the manuscript, the development of the supplementary table, the clarificaiton on methodology etc.. I think the manuscript improved a lot and timely publication would add to the scientific discussion. I've just two minor comment to the authors. 1) Post-revision the intro has 4 pages and the methods section 2 pages. To improve proportions authors might consider to move all the description of the new value framework into the methods section 2) Just as an aside: what is desparately missing with the Italian HTA Assessment is the availability of an english version. France does provide for most of the HAS decisions an english version. G-BA recently started to provide english versions. Perhaps authors might consider the inclusions of a respective comment somewhere in their discussion section - considering the size and relevance of the Italian market and the recent introduction of the innovation scheme @ AIFA the development of short english summaries of the AIFA appraisals

	would be helpful and timely.

VERSION 2 – AUTHOR RESPONSE

Reviewer: 2

Reviewer: 2
Reviewer Name: Joerg Ruof

Institution and Country: r-connect ltd
Please state any competing interests or state 'None declared': no

Please leave your comments for the authors below
Thanks very much for all the work that went into the revision of the manuscript, the development of the supplementary table, the clarificaiton on methodology etc..
I think the manuscript improved a lot and timely publication would add to the scientific discussion.

Reply:
We thank Professor Rouf for helping us to improve the manuscript.

I've just two minor comment to the authors.

1) Post-revision the intro has 4 pages and the methods section 2 pages. To improve proportions authors might consider to move all the description of the new value framework into the methods section

Reply:
As suggested, we moved the description of the new value framework into the beginning of methods section.

2) Just as an aside: what is desparately missing with the Italian HTA Assessment is the availability of an english version. France does provide for most of the HAS decisions an english version. G-BA recently started to provide english versions. Perhaps authors might consider the inclusions of a respective comment somewhere in their discussion section - considering the size and relevance of the Italian market and the recent introduction of the innovation scheme @ AIFA the development of short english summaries of the AIFA appraisals would be helpful and timely.

Reply:
We agree with this comment. We added the following sentence in the manuscript "These appraisals are written in Italian only. An English version should be desirable to allow greater dissemination of information outside Italy."

Reviewer: 1

Reviewer Name: Maximilian Salcher-Konrad
Institution and Country: London School of Economics and Political Science, UK
Please state any competing interests or state 'None declared': None declared

Please leave your comments for the authors below
I want to thank the authors for revising the manuscript thoroughly and addressing my and the second reviewer's comments. Substantial additional work has gone into this revision through the extraction and analysis of 23 additional appraisals.

#1: My comments regarding the framing of the manuscript have been addressed. The aims of this work are now much clearer. However, it is still not clear to me how the question about consistent use of the criteria over time has been answered. I would expect some more detailed analysis of what happens over time to answer this question (going beyond the univariate analysis of year of assessment). Maybe this is just an issue of how this particular aim is worded (i.e., were the authors really interested in changes over time, or just consistency within the total sample of appraisals to date?).

Reply #1:
We thank Professor Salcher-Konrad for helping us to improve the clarity and the key-messages of our manuscript. We are potentially interested in changes over time in AIFA decisions, but the current system adopted by AIFA does not allow it now. In fact, AIFA releases an appraisal after the end of the negotiation iter, but sometimes it takes more than one year for publication of the decision after the innovativeness appraisal.In-depth analyses and consideration of the changes over time will be performed in the next future, when all the innovativeness appraisals that have been done during these years by the CTS will be released on the website. In this first set of analyses, we can only preliminarly evaluate the consistency within the total sample of available appraisals.

#2: While the framing of the paper is now clear, some issues remain with respect to the analyses conducted.
Univariate analyses (Table 2): in the analysis for the updated, larger sample, there is now a discrepancy for the results on quality of evidence (in terms of statistical significance) depending on whether the analysis is based on proportions (p=0.11) or the continuous measure of quality of evidence (p=0.03). In the previous version of the manuscript, the results were consistent across these analyses, so it did not matter too much which one the interpretation was based on. Given the discrepancy, a clear rationale for relying on one vs. the other is needed. This is missing from the methods section and should ideally have been specified a priori (note that the proportions results would assign the same importance to unmet therapeutic need and quality of evidence, so the choice of outcome measure is key to the conclusions of the study).
Also note that this again raises the issue of why the two types of measures (proportions and continuous measures) are presented in the Table in the first place.

Reply #2:
Our study is a first analysis on this topic (preliminary analysis) since it is still based on a limited number of innovativeness appraisals. This a major limitation, and we reported it in the "Discussion" section. We hope to apply multivariate analyses as soon as more appraisal is released.
However, we believe that this article, based on a limited number of appraisals, is useful to disseminate how AIFA is working on innovativeness.
As known, the power of Fisher's exact test to detect an association is limited, i.e., the probability of obtaining false-negative conclusions (type II error) is high. For this reason, we had decided a-priori that the more powerful test for continuous variables had to be used as the first choice to analyze the association between the judgments of the three domains and the innovativeness status.
We specified it in section "Material and Methods", and we deleted in Table 2 the p-value from Fisher's exact tests.

#3: The objective behind developing the algorithm for innovativeness assessment remain a little unclear to me, as do the methods for doing this (even after the revisions to address reviewer 2's and my comments). This is described as a deterministic exercise, yet the algorithm fails to account for almost 20% of cases. The problems with this approach and the interpretation of its findings become apparent on Page 21, para 3: "When the added therapeutic value was rated as "poor" or "absent", or when the GRADE evaluation was "very low", the indication is never considered innovative. Innovativeness resulted from an at least a "moderate" added therapeutic value combined with an at least a "moderate" GRADE evaluation". These statements are not accurate, as there was one (ultra-rare disease) case where GRADE was very low and the drug was still considered conditionally innovative, and aother four where GRADE was low and the drugs were conditionally innovative. I

understand that this is what the algorithm produces, but its weakness becomes obvious when looking at the flow chart and seeing these cases that directly contradict the statements quoted above.

Reply #3:
This deterministic, data-driven approach attempts to go beyond the univariate analysis approach. In fact, the most appropriate multivariate data analysis approach was not applicable due to the limited number of appraisals (and consequently, the probable problem of convergence in the estimations).
We are aware of the limitation of this approach (e.g, the decision tree did not explicate all the appraisals final decisions; there is a lack of a formal measure of consistency, …). For this reason, we decided to tone down the conclusion on this topic throughout the manuscript and to report the decision tree as a Supplementary File.
*Finally, we deleted the wrong sentence "*When the added therapeutic value was rated as "poor" or "absent", or when the GRADE evaluation was "very low", the indication is never considered innovative. Innovativeness resulted from an at least a "moderate" added therapeutic value combined with an at least a "moderate" GRADE evaluation". We appreciate the careful reading of the Reviewer since this sentence was not updated after the upgrading to the 77 appraisals (instead of 54 in the first submitted version of our paper).

VERSION 3 – REVIEW

REVIEWER	Maximilian Salcher-Konrad London School of Economics and Political Science, UK
REVIEW RETURNED	09-Dec-2020

GENERAL COMMENTS	Thank you for addressing my comments. I appreciate that the revised manuscript now better reflects the early nature of this work, given that the new appraisal mechanism has only been in place for a relatively short period. With respect to the choice of primary outcome (continuous vs. categorical measures of the three assessment domains), reference to a documentation of the a priori decision to focus on continuous measures (e.g., in a protocol) would increase confidence in the results, if available. I appreciate that, through reference to statistical power, there is now a clear rationale for selecting the continuous outcome. Some possible typos and grammatical errors that should be looked at: Abstract, conclusions section: "the accurancy of the appraisal process" (accuracy) p17, para 1: "Rare disease and pediatric/mixed indications were appraised innovative by a larger proportion, although not statistical significant." (not statistically significant?) p21, para 2: "for examples" (lose the s). p21, para 2: "Despite there is a general consensus ..." (Despite there being a general consensus?) p22, para 2: "our results, besides being the first one published on the Italian-case, cannot be fully compared with that of our countries." (compared with other countries?)

REVIEWER	Joerg Ruof r-connect ltd; Switzerland
REVIEW RETURNED	17-Dec-2020

GENERAL COMMENTS	Thanks for the second revision of this paper - which is from my point of view now ready for publication.