PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	The CCEDRRN COVID-19 Infection Score (CCIS): development and validation in a Canadian cohort of a clinical risk score to predict SARS-CoV-2 infection in patients presenting to the emergency department with suspected COVID-19
AUTHORS	McRae, Andrew; Hohl, Corinne M.; Rosychuk, Rhonda; Vatanpour, Shabnam; Ghaderi, Gelareh; Archambault, Patrick M.; Brooks, Steven C.; Cheng, Ivy; Davis, Philip; Hayward, Jake; Lang, Eddy; Ohle, Robert; Rowe, Brian; Welsford, Michelle; Yadav, Krishan; Morrison, Laurie J.; Perry, Jeffrey

VERSION 1 – REVIEW

REVIEWER	Jeffrey Kline Indiana University Bloomington
REVIEW RETURNED	16-Aug-2021

GENERAL COMMENTS	Thank you for the chance to read and comment on this important work. This paper adds to current literature by developing a set of criteria to predict SARS-CoV-2 at the bedside without laboratory or radiological testing. The data are from a multicenter sample from Canada and are compared with the CORC rule. I believe the work has merit and should be published, but I have some comments I would like to see addressed. 1. The Wynant review (reference 2), while relevant, is now obsolete as it was published in April, 2020, and many new rules have been developed since then, so it is not really a great basis for the introduction any longer 2. Page 9 line 16. Semantics/syntax: The word consecutive is often misused (or at least overused) in diagnostic literature. Strictly speaking, consecutive means “one after the next without interruption.” I doubt that is what you mean…that you collected consecutive patients at all 32 institutions, all starting at the same time, and none interrupted. For example, reference 15 refers to period 1 and period 2, and exclusion of asymptomatic patients, both of which would suggest lack of consecutive sampling of all patients tested. This point is relevant later when you compare CCEDRRN rule against the CORC rule, and assert that a strength of the CCEDRRN is that the data for CORC were not consecutive. The data from each hospital participating in the RECOVER registry (the data used for the CORC rule) were collected consecutively, but at different start/end time points from different hospitals. The RECOVER registry also excluded patients tested for reasons defined as “administrative”, for example,a patient admitted for, say, a cholecystectomy and no suspicion of COVID.(PMID: 33392542) Accordingly, I do not believe it is accurate to claim on page 16, line that the CCEDRRN is consecutive, and the CORC rule sample was “non-consecutive” as a basis of superiority.

	3. Page 12 line 8. Minor issue. It might be helpful to include a Hosmer Lemeshow P value to allow more direct comparison of calibration with other models in the literature. 4. Page 14 lines 18-22 and 34-36. Syntax: Specificity is the proportion of disease negative patients with a negative test. The text of these lines, as written seems to describe predictive value positive, rather than specificity. I would either change the numbers to PPV, or if you want to keep the specificity numbers, rewrite to say something like “indicating a low frequency of false positives” 5. METHODS/RESULTS: The biggest criticism I have is about rigor/reproducibility of this rule, mainly because of the epidemiological requirement. a. Do you have interrater reliability for rule? In particular, the “publicly available epidemiology data” in reference 19? If not, this must be stated as a limitation. I tried to use the website for about 15 minutes and could not find a way to exactly find the data by a 7 day average from a postal code. It is a cool resource though. b. How would this be used day-to-day in the ED—is the expectation that an administrator will post this somewhere or does the doctor have to look it up on every shift? c. Some of these regions are huge; I would suspect there was within-region variability of positivity. What if a patient lived in one zip code and works in another? d. This may limit the use of the rule to Canada, which is fine, but on page 16 lines 18-24, you imply that your rule is more useful than the CORC rule in other countries based on “international diversity”. The CORC rule was derived in a more heterogenous sample, and included rural samples. 6. RESULTS first paragraph. This is important: What was the timeframe of the sample? Can you show diagnostic performance from the first half of 2020 compared with a later sample, say in the fall of 2020? If not, this is a limitation. 7. RESULTS, page 12 last paragraph. Comparison with CORC: Why did you not include race/ethnicity? Trying to understand your position, I have been told that some ethics committees (e.g. in Germany) will not even allow recording of patient race in research. Maybe that is the case in Canada, and that would be a reason not to include race in CORC. But as written, the wholesale removal of race/ethnicity is not scientifically justified, removes important variables from CORC, and introduces imprecision and inaccuracy to the comparison. These race/ethnicity variables were justified to include in CORC, at least in the US, based upon the epidemiological finding of a significantly higher infection rate among persons of color, versus White patients. At least in the US, this is a more available method of assessing prevalence than using a website to assess prevalence in a zip code. Of course, the US has more Black/ethnic folks than Canada, but still the numbers are not trivial. I just did a quick Google search, and this revealed that the Canadian census suggests that 3.5% of Canadians are Black, and about the same percentage identify as Latino/Hispanic. 8. Reference 26 addresses inequities and injustice in various algorithms caused by the erroneous use of Black Race. This may unfairly categorize the CORC rule. The authors of the CORC rule strongly considered this issue and ultimately decided that their inclusion could increase testing in the most vulnerable patients, and therefore was the “least worst” option. Excluding them would be lower in terms of social justice than including them.

REVIEWER	Michele Bartoletti University of Bologna

REVIEW RETURNED	09-Sep-2021

GENERAL COMMENTS	In this paper McRae et al aimed to develop a 10-item score that may identify the risk of SARS-CoV-2 infection among patients presenting at Emergency department. Overall the paper Is interesting. I have the following comments: - The major limitation in my opinion is that the score include prevalence of cases of SARS-CoV-2 in that area as predictor of SARS-CoV-2 diagnosis. This may introduce less applicability as not all physicians at ED may know the actual current incidence rate at the moment they are visiting a patient. Additionally, as several parameters to assess accuracy of scores (e.g. PPV, NPV) are correlated with prevalence this may introduce a statistical bias - Authors stated that missing values were very low (appendix table 2). In the methods section, page 9 lines 31-32 they state that "Patients with missing data for categorical variables were assumed to have the reference value for that categorical variable". This should clarified. Similarly, according to the tables roughly 5-6% of patients had altered consciousness or confusion, 3% had dementia and 6% other neurological disorder. However, all were able to report symptoms (appendix table 2), including anosmia/dysgeusia. This seems poorly reliable. - The main endpoint was positivity for SARS-CoV-2 at NAAT. As you know NAAT using pharyngeal and nasal swab has sensitivity of 80-85%. Can you provide the number of patients that were NAAT positive using nasal or pharyngeal swab or NAAT performed on deeper respiratory samples (BAL, etc.) and the number of patients that were positive at retest (within 14 days as stated). Additionally, can you provide the accuracy of your score in this subgroup of patients?

REVIEWER	He Yang Weill Cornell Medicine, Pathology and Laboratory Medicine
REVIEW RETURNED	10-Sep-2021

GENERAL COMMENTS	Thank you for the opportunity to review the manuscript entitled "Development and validation of a clinical risk score to predict SARS-CoV-2 infection in emergency department patients: the CCEDRRN COVID-19 Infection Score (CCIS)". This study developed a clinical risk score based on patient symptoms, institutional living, working as a healthcare provider, close personal contact with infected individuals. The multiple regression model was developed using data from 21,743 patients from multiple hospitals during March 1 to Oct 30, 2021. Please see my specific questions and comments below: While the authors criticized other machine learning models were not ready for widespread use, this study does not show clear clinical utility either. The COVID-19 has evolved significantly in the past year and now the variants are widely spread. In the meanwhile, many people, especially healthcare workers, have been vaccinated. Therefore, a model based on data from 2020 may not be suitable for the current situation. It is necessary to validate the model performance using new data of 2021. It would be nice to show the longitudinal change of model performance from 2020 to 2021. I wonder if some predictors, e.g., living conditions and household contacts, are empirically correlated to the SARS-Cov-2 infection at the initial outbreak of COVID-19, without inherent causal relation to

	the disease, which limits the generalization ability of the proposed model to later pandemic scenarios. For instance, the spread of Delta variant in 2021 could lead to dramatic changing in these non-clinical features/predictors. In addition, in 2020 when vaccine was not developed, being a healthcare work posed a significant risk of infection. I wonder if this is still true now. Last but not least, since the features are not specific to COVID-19, the model may not be able to differentiate SARS-CoV-2 and other respiratory viruses, such as influenza, RSV, etc. As we know, there was remarkably low rate of influenza infection in 2020 winter, however, we are not sure about the situation in 2021 winter. Average daily incidences of SARS-Cov-2 infections appears a strong prior to the prediction, as well as the top 4 predictors in Table 2. These statistics may not be available or reliable during the early stage of the outburst. Instead of examining the co-linearity of candidate predictors, how about performing a PCA analysis on all predictors before training the logistic regression model? I am a little surprised that patient demographic information is not included in the model. Would that improve model performance? This paper missed citing some closely related work, such as: 1) Bayat V et al., A SARS-CoV-2 prediction model from standard laboratory tests. Clinical Infectious Diseases 2) Yang HS et al., Routine laboratory blood tests predict SARS-CoV-2 infection using machine learning. Clinical Chemistry Overall, this study lacks novelty since there has been many similar publications on this topic. It would be necessary to show model performance in recent data in 2021 as well as the longitudinal change in the past 1.5 years. I don't think the features selected for the model are specific to COVID-19, so the model may not differentiate COVID-19 and other respiratory infection diseases.

REVIEWER	Vera Clérigo Hospital de São Bernardo, Pulmonology
REVIEW RETURNED	13-Sep-2021

GENERAL COMMENTS	Dear authors, thank you for the opportunity to read your manuscript. The authors derived and validated a user-friendly 10-item risk prediction tool that uses clinical variables available at the time of a patient's initial presentation, that accurately excludes COVID-19 infection in one-third of patients and accurately rules in COVID-19 infection in high-risk patients. Diagnosis models are, now, more useful for under/low resource settings as RT-PCR is widely available. However, common publicly available measure of community COVID-19 incidence may be underestimated in these settings, posing a challenge to this score full usefulness in places where resources are scarce. This limitation could also be acknowledged in the manuscript. I hope that these data could one day be incorporated into data from other countries so that we can form a robust database, with internal and external validity from which we could draw comprehensive conclusions.

VERSION 1 – AUTHOR RESPONSE

Reviewer 1	
1. The Wynant review (reference 2), while relevant, is now obsolete as it was published in April, 2020, and many new rules have been developed since then, so it is not really a great basis for the introduction any longer	We agree. We now reference the February 2021 update to the systematic review, with appropriately edited text.
2. The word consecutive is often misused (or at least overused) in diagnostic literature. Strictly speaking, consecutive means "one after the next without interruption." I doubt that is what you mean…that you collected consecutive patients at all 32 institutions, all starting at the same time, and none interrupted. For example, reference 15 refers to period 1 and period 2, and exclusion of asymptomatic patients, both of which would suggest lack of consecutive sampling of all patients tested. This point is relevant later when you compare CCEDRRN rule against the CORC rule, and assert that a strength of the CCEDRRN is that the data for CORC were not consecutive. The data from each hospital participating in the RECOVER registry (the data used for the CORC rule) were collected consecutively, but at different start/end time points from different hospitals. The RECOVER registry also excluded patients tested for reasons defined as "administrative", for example,a patient admitted for, say, a cholecystectomy and no suspicion of COVID.(PMID: 33392542) Accordingly, I do not believe it is accurate to claim on page 16, line that the CCEDRRN is consecutive, and the CORC rule sample was "non-consecutive" as a basis of superiority.	We have edited the methods section for further clarity. CCEDRRN enrolled consecutive patients, independent of symptoms, at each site starting at the date each site began data collection (specified in supplementary appendix). We did not exclude "administrative" testing from enrolment as part of the funding mandate of CCEDRRN was to quantify yield of testing in asymptomatic patients. We have removed consecutive enrolment from the discussion of the comparison of the CCIS to the CORC score.
3. Page 12 line 8. Minor issue. It might be helpful to include a Hosmer Lemeshow P value to allow more direct comparison of calibration with other models in the literature.	Thank you for this suggestion. Our view is that the preferred approach to comparing calibration is using visual plots, and that the Hosmer-Lemeshow p-value is of limited utility in comparing calibration between models given the low power of the test (Harrell, F. Regression modelling strategies. 2021. https://hbiostat.org/doc/rms.pdf. Accessed September 28, 2021), and the statistic is sensitive to the grouping strategy meaning different software packages produce different results based on their choice of grouping algorithm (Hosmer, D.W., Hosmer, T., Le Cessie, S., Lemeshow, S.: A comparison of goodness-of- fit

	tests for the logistic regression model. Statistics in Medicine 16, 965–980 (1997)).
4. Page 14 lines 18-22 and 34-36. Syntax: Specificity is the proportion of disease negative patients with a negative test. The text of these lines, as written seems to describe predictive value positive, rather than specificity. I would either change the numbers to PPV, or if you want to keep the specificity numbers, rewrite to say something like "indicating a low frequency of false positives"	Thank you. We have adopted this wording suggestion.
5a Do you have interrater reliability for rule? In particular, the "publicly available epidemiology data" in reference 19? If not, this must be stated as a limitation. I tried to use the website for about 15 minutes and could not find a way to exactly find the data by a 7 day average from a postal code. It is a cool resource though.	As we have not evaluated the score in an implementation study, we do not have inter-rater reliability for the score itself. We have validated a number of the clinical variables (see ref 15). Ref 19 is the source of our epidemiologic data, but in practice this would be ascertained from a regional public health unit (in Canadian regions, this data is published daily on health unit websites). We have discussed this further in the limitations section.
b. How would this be used day-to-day in the ED—is the expectation that an administrator will post this somewhere or does the doctor have to look it up on every shift?	We have added to the discussion how this could be implemented in practice.
5c. Some of these regions are huge; I would suspect there was within-region variability of positivity. What if a patient lived in one zip code and works in another?	The reviewer raises a concern that, for Canadians who may live in one postal code and work in another, the rule may not work if the case rate is different between the home region and the region an individual may work. We have addressed this in the discussion and suggest modifying the score to use the higher of the two regional case rates for the small number of affected individuals.
5d. This may limit the use of the rule to Canada, which is fine, but on page 16 lines 18-24, you imply that your rule is more useful than the	This sample also included patients enrolled in smaller rural regions with very low population

CORC rule in other countries based on "international diversity". The CORC rule was derived in a more heterogenous sample, and included rural samples.	densities and their associated regional hospitals. We have revised the discussion and comparison to the CORC score with respect to race and regional disease burden
6. RESULTS first paragraph. This is important: What was the timeframe of the sample? Can you show diagnostic performance from the first half of 2020 compared with a later sample, say in the fall of 2020? If not, this is a limitation.	The recruitment period is shown in the supplementary appendix. We have added this information to the results section.
7. RESULTS, page 12 last paragraph. Comparison with CORC: Why did you not include race/ethnicity? Trying to understand your position, I have been told that some ethics committees (e.g. in Germany) will not even allow recording of patient race in research. Maybe that is the case in Canada, and that would be a reason not to include race in CORC. But as written, the wholesale removal of race/ethnicity is not scientifically justified, removes important variables from CORC, and introduces imprecision and inaccuracy to the comparison. These race/ethnicity variables were justified to include in CORC, at least in the US, based upon the epidemiological finding of a significantly higher infection rate among persons of color, versus White patients. At least in the US, this is a more available method of assessing prevalence than using a website to assess prevalence in a zip code. Of course, the US has more Black/ethnic folks than Canada, but still the numbers are not trivial. I just did a quick Google search, and this revealed that the Canadian census suggests that 3.5% of Canadians are Black, and about the same percentage identify as Latino/Hispanic.	The reasons for not including race/ethnicity are several-fold: 1) Race and ethnicity are not routinely collected in Canadian hospital clinical records or administrative data 2) Race and ethnicity are frequently conflated, and—depending on method of ascertainment--not reliable 3) We agree that race/ethnicity is an important predictor in the CORC model. But, given that it is a surrogate for other socioeconomic predictors of risk of infection, is a suboptimal indicator of risk of SARS-CoV-2 infection. In our revision we address the reasons for not including race/ethnicity in our evaluation of the CORC score (reason 1 above), and the reasons for not including race/ethnicity in the CCIS.
8. Reference 26 addresses inequities and injustice in various algorithms caused by the erroneous use of Black Race. This may unfairly categorize the CORC rule. The authors of the CORC rule strongly considered this issue and ultimately decided that their inclusion could increase testing in the most vulnerable patients, and therefore was the "least worst" option. Excluding them would be lower in terms of social justice than including them.	Thank you. We did not mean to imply that the CORC score developers had not carefully considered the use of race and/or ethnicity in the score. We also respect their decision to include the variable in heir model. Our point is that, given that race is poorly collected as a variable and not biologically associated with the risk of SARS-CoV-2 infection

	but rather an association confounded by sociodemographic factors related to infection risk, access to testing, and other factors, using local incidence data is preferable.
Reviewer 2	
1. The major limitation in my opinion is that the score include prevalence of cases of SARS-CoV-2 in that area as predictor of SARS-CoV-2 diagnosis. This may introduce less applicability as not all physicians at ED may know the actual current incidence rate at the moment they are visiting a patient. Additionally, as several parameters to assess accuracy of scores (e.g. PPV, NPV) are correlated with prevalence this may introduce a statistical bias	Thank you. We have expanded the text addressing the clinical applicability of incidence data in the discussion. We agree that predictive values are not appropriate for assessing accuracy of scores, and have removed these columns from the tables.
2. Authors stated that missing values were very low (appendix table 2). In the methods section, page 9 lines 31-32 they state that “Patients with missing data for categorical variables were assumed to have the reference value for that categorical variable”. This should clarified. Similarly, according to the tables roughly 5-6% of patients had altered consciousness or confusion, 3% had dementia and 6% other neurological disorder. However, all were able to report symptoms (appendix table 2), including anosmia/dysgeusia. This seems poorly reliable.	We have clarified in the text that patients with categorical variables having a value of “not documented” (eg if the use of supplemental oxygen was not documented in the clinical record) were assumed to have the reference value (ie no). We note that neurologic disorders refer to chronic conditions such as epilepsy or a prior cerebrovascular accident that do not have impaired cognition, and also that most patients with dementia are able to reliably report symptoms. We conducted a validation of the reliability of symptom data collection compared to prospective data collection as part of our pilot implementation (see ref 15) and found these variables to be reliable.
3. The main endpoint was positivity for SARS-CoV-2 at NAAT. As you know NAAT using pharyngeal and nasal swab has sensitivity of 80-85%. Can you provide the number of patients that were NAAT positive using nasal or pharyngeal swab or NAAT performed on deeper respiratory samples (BAL, etc.) and the number of patients that were positive at retest (within 14 days as stated). Additionally, can you provide the accuracy of your score in this subgroup of patients?	Thank you Of the 1167 patients who tested positive, only 4 initial samples were from deep sites (BAL, etc.) Nasopharyngeal swabs were standard practice at the participating sites during the study period. The laboratories at each site used established quality control approaches to ensure optimal test performance for nasopharyngeal swabs.

	Of the 1167 patients who tested positive, 1133 were positive on the initial test, and 34 tested positive on subsequent testing after initial negative (27) or indeterminate (7) samples. We have added this information to the first paragraph of the results section. Given the small number of SARS-CoV-2 patients who had an initial negative test, a subgroup or sensitivity analysis in this group would not be informative.
Reviewer 3	
While the authors criticized other machine learning models were not ready for widespread use, this study does not show clear clinical utility either. The COVID-19 has evolved significantly in the past year and now the variants are widely spread. In the meanwhile, many people, especially healthcare workers, have been vaccinated. Therefore, a model based on data from 2020 may not be suitable for the current situation. It is necessary to validate the model performance using new data of 2021. It would be nice to show the longitudinal change of model performance from 2020 to 2021.	These are excellent points and we appreciate the opportunity to discuss them. We believe that publication of this rule is an essential starting point, particularly given that the discrimination is superior to other published models. Many countries that could potentially use this rule have very low vaccination rates and are reporting incidence rates accessible on the internet. Even countries with high vaccination rates have inordinately high vaccine hesitancy rates that in some cases is geographical for a variety of reasons and the rule could help with efficiencies and processes of care in emergency departments serving these regions. Finally, as variants of concern emerge and vaccinations increase, the rule will need to be validated on additional cohorts.
1. I wonder if some predictors, e.g., living conditions and household contacts, are empirically correlated to the SARS-Cov-2 infection at the initial outbreak of COVID-19, without inherent causal relation to the disease, which limits the generalization ability of the proposed model to later pandemic scenarios. For instance, the spread of Delta variant in 2021 could lead to dramatic changing in these non-clinical features/predictors. In addition, in 2020 when vaccine was not developed, being a healthcare work posed a significant risk of infection. I wonder if this is still true now. Last but not least, since the features are not specific to COVID-19, the model may not be able to differentiate SARS-CoV-2 and other respiratory viruses, such as influenza, RSV, etc. As we know, there was remarkably low rate of influenza infection	We agree that there likely are temporal variations in risk factors for SARS-CoV-2 infection, particularly with the advent of vaccines. Unfortunately, due to the rapidly evolving epidemiology of SARS-CoV-2, any published evidence involving the epidemiology of infection will be out of step with contemporary viral variants, immunization rates of the population and the relative effectiveness of current vaccines against dominant variants at the time it is published. However, we believe that fundamental aspects of the disease biology will remain the same and the derivation and validation of our rule provides

in 2020 winter, however, we are not sure about the situation in 2021 winter.	important insights into clinical factors that predict test positivity. Our study will be of vital importance for those refining screening criteria for testing and those developing clinical policy around efficient and effective testing strategies
2. Average daily incidences of SARS-Cov-2 infections appears a strong prior to the prediction, as well as the top 4 predictors in Table 2. These statistics may not be available or reliable during the early stage of the outburst.	Thank you for this comment. We agree average daily incidence (effectively a measure of pre-test probability) is important. While it may not have been widely reported early in the pandemic, we believe that case tracking such as this will be commonly done in future outbreaks or pandemics. The other predictors are derived largely from the empirical experience in the first wave, and from the ISARIC case report form, so we believe these measures will be reliably captured in future studies.
3. Instead of examining the co-linearity of candidate predictors, how about performing a PCA analysis on all predictors before training the logistic regression model?	Thank you for this comment. In the end, only one variable was excluded because of collinearity. We considered using PCA analysis but decided on an alternative strategy. A large number of candidate predictor variables were available and not necessarily all were important. Creating principal components would mean a coefficient would be assigned to each predictor for each principal component and this would not be easily transparent to users. It would have required a larger number of predictors for a user to have available and enter, and if not all predictors were available, may have limited usability. Further, we wanted the key predictors in the score to be those immediately available at the bedside. Finally, PCA may have lead to challenges in interpretability of the coefficients in the final model and would likely require an online calculator, which we hoped to avoid.
4. I am a little surprised that patient	Thank you. Demographic information including

demographic information is not included in the model. Would that improve model performance?	age and sex were included as potential predictors. But, they did not have significant associations with the outcome after adjusting for other predictors, and were excluded by the model using an *a priori* selection process.
5. This paper missed citing some closely related work, such as: 1) Bayat V et al., A SARS-CoV-2 prediction model from standard laboratory tests. Clinical Infectious Diseases 2) Yang HS et al., Routine laboratory blood tests predict SARS-CoV-2 infection using machine learning. Clinical Chemistry	Thank you for these suggested references. However, we do not believe they add substantially to our introduction or discussion. We chose *a priori* not to incorporate laboratory testing into our model to ensure that the score can be applied at hospital arrival, prior to the performance of any laboratory or imaging investigations.
6. Overall, this study lacks novelty since there has been many similar publications on this topic. It would be necessary to show model performance in recent data in 2021 as well as the longitudinal change in the past 1.5 years. I don't think the features selected for the model are specific to COVID-19, so the model may not differentiate COVID-19 and other respiratory infection diseases.	While we agree that several other risk scores to predict a positive SARS-CoV-2 test have been published, we would point out that this study addresses several of the limitations of previously-published scores. It includes the derivation and external validation of the risk score in a large sample of consecutively-enrolled patients from a geographically distributed clinical network with rigorous data quality protocols. It uses only variables easily available at the time of triage, does not require imaging or additional laboratory testing, and does not require use of an electronic calculator or electronic medical record for implementation.
Reviewer 4	
Diagnosis models are, now, more useful for under/low resource settings as RT-PCR is widely available. However, common publicly available measure of community COVID-19 incidence may be underestimated in these settings, posing a challenge to this score full usefulness in places where resources are scarce. This limitation could also be acknowledged in the manuscript.	Thank you for this comment. We have expanded the discussion of the clinical utility of incidence measures in the strengths and limitations.
1. I hope that these data could one day be incorporated into data from other countries so that we can form a robust database, with internal and external validity from which we could draw comprehensive conclusions.	Thank you. We agree that validation in other countries would be a useful effort. See previous comments.

VERSION 2 – REVIEW

REVIEWER	Jeffrey Kline Indiana University Bloomington
REVIEW RETURNED	07-Nov-2021

GENERAL COMMENTS	Thank you for addressing my comments.

REVIEWER	Michele Bartoletti University of Bologna
REVIEW RETURNED	29-Oct-2021

GENERAL COMMENTS	Previous Comments were adequately adressed. I have no further suggestions to make