PEER REVIEW HISTORY

This paper was submitted to a another journal from BMJ but declined for publication following peer review. The authors addressed the reviewers' comments and submitted the revised paper to BMJ Open. The paper was subsequently accepted for publication at BMJ Open.

(This paper received three reviews from its previous journal and all three reviewers agreed to published their review.)

ARTICLE DETAILS

TITLE (PROVISIONAL)	Development and validation of an early warning score to identify COVID-19 in the emergency department based on routine laboratory tests: a multicenter case-control study
AUTHORS	Boer, Arjen-Kars; Deneer, Ruben; Maas, Maaike; Ammerlaan, Heidi; van Balkom, Roland; Thijssen, Wendy; Bennenbroek, Sophie; Leers, Mathie; Martens, Remy; Buijs, Madelon; Kerremans, Jos; Messchaert, Muriël; van Suijlen, Jeroen; van Riel, Natal; Scharnhorst, Volkher

VERSION 1 – REVIEW

REVIEWER	Thomas, Ben The University of Sheffield, ScHARR
REVIEW RETURNED	07-Sep-2021

GENERAL COMMENTS	This paper presents the CoLab score its derivation, internal validation, external validation, and temporal validation. The CoLab score uses a panel of ten routinely collected variables, plus age, from ED labs to dentine the likelihood of a person having COVID-19. The score aims to be balance accuracy with ease of use and is more optimally used to rule out COVID-19 before the use of a PCR test. This is in general an interesting and compelling article that is appropriate for publication from my perspective, with the appropriate edits. The key message is clear and it appears that the CoLab score has potential use in EDs for quicker diagnosis/ruling out of COVID-19. My major concerns are two fold: 1) to confirm that under Dutch ethical processes their is not any further approval required to use patient data without consent, i.e. such the approval from CAG under UK regulations. This may not be a) required or b) is part of the approval as described in the paper. Suggest this is clarified to the editor. 2) From reading the paper I am unsure as to why their is use of data stretching back to July 2019. I appreciate that only data from 24/02/2020 was used in the modelling but even the justification for this particular date is unclear. This should be justified and/or made clearer to the reader why these date choices

	were made. Unless there is a compelling reason I would advice removing the description of the data from July 2019 for clarity. Issues to be considered for amendment: The strength of this study using lab based tests could be further highlighted, in paragraph beginning line 104, in comparison to non lab based risk tools such as the PRIEST EWS. On line 141 can it be made clear whether or not cases were excluded for having any one of the 10 routine panel missing, e.g. Presentations with any missing value in the routine panel were excluded. This is made clear later but would be useful at this point. I suggest making it clear what known variant or variants of COVID was circulating in the Netherlands during the data collection period. It is stated what has 'recently' been circulating, but it is not clear this was during the data collection. Further, were there known difference between the derivation and external validation variants? It is stated that different variants should be detectable due to it being based on the host response. Any evidence to this affect could be added. If so this would negate the previous point of different variants between the derivation and validation cohorts. There is some inconsistencies in who large numbers are recorded.

REVIEWER	Soltan, Andrew A. S. University of Oxford, Division of Cardiovascular Medicine
REVIEW RETURNED	08-Sep-2021

GENERAL COMMENTS	In the development and validation of an early-warning score for COVID-19, the authors identify an important clinical problem. There is no doubt that delayed result times are placing significant operational pressures on hospitals, and I commend the authors for an innovative approach to the problem. However, there are some unconsidered high-level limitations that limit utility of this score, in brief: a) Lateral flow testing: This is now widely used in emergency departments and provides rapid COVID-19 screening results. LFTs now form the direct competitor of such a score, rather than PCR, however the value add over lateral flow testing is not considered or shown in the validation. b) Vaccinations: The overwhelming majority of the Dutch, and much of the Western European population, are now vaccinated and this has been shown to reduce disease severity. Reduced disease severity will impact on degree of derangement of blood test results, and therefore likely have a significant impact on the sensitivity of this score. No mention or consideration is given to the impacts of vaccination - and what effects this may have on generalisability. c) What is the value add of this work over existing works? The authors justify this over existing works (machine learning based) by saying that their model is easier to implement in hospital IT systems. What evidence can they offer to show that this is the case?

	d) Some concerns about the requirement for some non-routine blood tests in inclusion criteria. e) Exclusion of patients with previous COVID-19 is problematic - as this population is probably closest in performance to previously vaccinated, and does not help a physician decide whether a patient is infectious. Please see some more specific comments below: Background: - The authors cite Wynants et al. but I feel this fast-moving space mandates a more up to date search to give fair comment to today's literature. The limitation of Wynants' review is that it was last updated in early April 2020 (only 3 months after SARS-CoV-2 was first described!), and was therefore written at a timepoint where sufficient time for development & validation of models had not passed. The age of this review is particularly problematic as the field has been very fast moving, and as Wynants' living review has not been updated, it now offers a poor view on todays' landscape of COVID-19 diagnostic models. Please elaborate in the background on more recent uses of blood-based COVID-19 screening. - Lateral flow testing which provides results in 30 minutes is now widely adopted in emergency departments and is readily accessible. How does this work add value over lateral flow testing which now forms the existing standard of care, and addresses this problem? Methods: - The exclusion of previous confirmed COVID-19 patients from analysis is problematic, as a physician wishes to know whether a patient is (a) infected with COVID-19, and if (b) is infectious to others. Patients re-infected with COVID-19 are likely to still be infectious (albeit less so, and at lower risk of severe disease) - and therefore to justify a use for such a tool in infection-control, there needs to be consideration of reinfected patients. - Why were patients with 'extreme' results (>10x SD from mean) excluded? Many patients presenting to an emergency department will attend because they are unwell - and therefore are at highest probability of an extreme abnormal result (e.g. a CRP >100 where a reference range is <4 would likely be excluded - but is an important group). Was the exclusion criterion >10 times the study population mean, or a laboratory reference range? Please elaborate. - No specific consideration is given to patients who have been vaccinated. This raises two significant problems, as vaccination reduces severity of disease. First, it is reasonable therefore to presume that the extent of abnormality in blood-tests considered by the score will be reduced - and therefore that sensitivity of the score in vaccinated patients will be lower. Second, this greatly compromises generalisability as >80% of the Dutch population are now vaccinated (similar figures seen also across Western Europe). - Selection of lab panel constituents: Some of the lab tests which make up inclusion criterion - for example, LDH, CK, ɣ GT - are not widely considered to be routine tests performed for patients attending the emergency department prior to the pandemic. Use of these particular tests only became more widespread after emergence of the pandemic and early reports that they were deranged. However, prior to the pandemic, they were only performed in specific subgroups of patients - for example,

	lymphoma patients or patients who had had falls with long lies. During the pandemic, they are selectively performed where there is potential concern regarding COVID-19. There is therefore a problem in using pre-pandemic patients for model development and excluding patients missing these results - this potentially introduced a very significant bias in the composition of the development dataset; the very presence of a 'CK' result will bias towards a potential case. This may in part also explain why the rate of missing data was higher in the pre-pandemic population (figure 1a). Can the authors confirm that LDH, CK, ɣ GT measurements were considered part of the routine blood panel for allpatients attending ED? Similarly for validation, if the score was only applied to patients who had had a CK/LDH measured - this may bias towards patients where a clinician suspected COVID-19 and therefore arranged for this to be measured. Results: - The importance of high NPV is clear, however when quoting NPV it would be helpful to simultaneous mention PPV. For example, when stating a CoLab score of 0 having an NPV of 0.996, it is relevant that PPV is 0.06 (meaning that only around 1 in 16 patients judged positive by this score would be true positives). - On the individual patient level, how does this perform for patients with certain comorbidities, e.g. haematological malignancy? Discussion: - The authors identify existing works that have performed the same function as the work presented here (references 19 and 25). The authors say that the value add of their work is that their proposed score is easier to implement in current hospital IT systems than machine learning models - and that this warrants a trade-off in performance. What evidence can they offer that this is the case? Machine learning models can be mathematically reduced to weights attached to variables, and so theoretically could be similar in difficulty to implement to this model. - Please see previous comments on potential inclusion bias; the primary focus here is whether CK/LDH/gamma GT were as routinely performed prior to the pandemic as during the pandemic. The authors report that the full lab panel was most frequently missing for surgical patients - in my view this is evidence of bias, as these patients are at lower probability of testing positive for COVID-19 than patients presenting with acute COVID-19 symptoms. - Reference is made to data that is not shown in the discussion - please either qualify this by presenting this data in the supplement, or adapt comments to remove this.

REVIEWER	Hoo, Zhe Hui University of Sheffield, School of Health and Related Research (ScHARR)
REVIEW RETURNED	09-Sep-2021

GENERAL COMMENTS	This is an interesting study and it is clear that the authors have put in considerable effort in data collection and analyses. My main concerns regarding the statistical methods are as follow: • Both the derivation and validation analyses used correlated (or clustered) data in that each participant can have >1 presentation, yet there is no mention how the logistic regression model and calculation of diagnostic accuracy account for the correlated data.

	The authors should consider using mixed-effect modelling for the logistic regression and logistic random-effects models or generalised estimating equations to calculate the diagnostic accuracy values (see https://doi.org/10.1148/radiol.12120509). At the very least, the authors should perform sensitivity analyses whereby only the first presentation from each participant is used to determine whether their results are robust to the effect of potential clustering effect. • The fact that many of the participants did not have PCR testing is a limitation of this retrospective analysis. Both the derivation and validation analyses assumed that untested participants do not have Covid-19. There is no justification why such an assumption is valid. Among participants with PCR testing, participants at Catharina Hospital had positivity rate of 279/1224 (22.8%), participants at 'Center 1' had positivity rate of 52/501 (10.4%), participants at 'Center 2' had positivity rate of 99/1051 (9.4%) and participants at 'Center 2' had positivity rate of 336/839 (40.0%). Whilst patients selected to have PCR testing may be more likely to have a positive test, patients not selected for PCR may still have Covid-19. The high positivity rates (especially at 'Centre 3') suggest that a substantial proportion of people with Covid-19 remain untested. At the very least, the authors should perform sensitivity analyses using data only from participants with PCR testing to determine whether their results are robust. In particular, it seems unlikely that 1/3 of participants with PCR testing will have CoLab-score of "0". Perhaps what is required is different thresholds of CoLab-score to guide PCR testing according to different pre-test probabilities. My other comments are: • For the abstract, i. The aims of developing and validating a clinical score should be explicitly stated in the 'Introduction' (page 5 lines 16-20) instead of just stating that such a score is desirable. ii. The method of validation (by calculating diagnostic accuracy values) should be explicitly stated in the 'Methods' (page 5 lines 30-35). iii. Both the sensitivity and specificity values should be provided in the 'Results' (page 5 lines 47-52). Providing sensitivity and negative predictive values meant that results of false positives are not being presented. Providing specificity and positive predictive values meant that results of false negatives are not being presented. iv. In the 'conclusions', it is important to acknowledge that the proportion of patients not requiring PCR based on their CoLab-score may vary according to the prevalence of Covid-19. Perhaps it is better to say that "Depending on the community prevalence, COVID-19 may be safely ruled-out in more than one third of ED presentations" (page 5 line 59 to page 6 line 4) instead of "With this score, COVID-19 can be safely ruled-out in more than one third of ED presentations". • In the 'Methods' (page 9 line 9), it should be stated that the study is a retrospective case-control study, instead of the study design only being stated in the 'Discussion' (page 20 line 49). • All abbreviations for the tests in Table 2 should be explained in the 'laboratory tests' subsection (page 10 line 40 to page 11 line 8). I am uncertain what "AF" is referring to. Is it alpha-fetoprotein? • Ideally, the rationale for evaluating the tests in the 'laboratory tests' subsection (page 10 line 40 to page 11 line 8) should be

	clearly stated. Is there evidence from existing literature that the tests differentiate between those with and without Covid-19? • Confidence intervals for β should be provided in Table 2. • In Table 3, the cross tabulation of the index test results by the results of the reference standard should be provided. For example, the values of TP, TN, FP and FN can be provided as they were in Table 4. • In Table 4, the confidence intervals for the diagnostic accuracy values should be provided, as they were in Table 3. • The demographics for participants in the validation dataset should be provided, at least as supplementary material if the table limit for the main text has been reached.

VERSION 1 – AUTHOR RESPONSE

Reviewer(s') Comments to Author:
Reviewer: 1

Comments to the Author

1. My major concerns are two fold: 1) to confirm that under Dutch ethical processes there is not any further approval required to use patient data without consent, i.e. such the approval from CAG under UK regulations. This may not be a) required or b) is part of the approval as described in the paper. Suggest this is clarified to the editor.

This is part of the approval obtained from the Medical research Ethics Committees United (MEC-U) and the internal hospital review board (IRB). The data collected for this study is described in the approved research protocol.

2. From reading the paper I am unsure as to why their is use of data stretching back to July 2019. I appreciate that only data from 24/02/2020 was used in the modelling but even the justification for this particular date is unclear. This should be justified and/or made clearer to the reader why these date choices were made. Unless there is a compelling reason I would advice removing the description of the data from July 2019 for clarity.

The reason for including data from July 2019 until July 2020 for the development cohort is twofold. Firstly, to limit the effect of seasonal variation in the ED patient population, and secondly, to include a large representative sample of pre-pandemic controls. We have made this more clear in the manuscript by adding this to the “Materials and Methods” section under “Development dataset”.

Issues to be considered for amendment:

3. The strength of this study using lab based tests could be further highlighted, in paragraph beginning line 104, in comparison to non lab based risk tools such as the PRIEST EWS.

We were not aware of the PRIEST EWS, thank you for this suggestion, we have added this reference to the introduction.

4. On line 141 can it be made clear whether or not cases were excluded for having any one of the 10 routine panel missing, e.g. Presentations with any missing value in the routine panel were excluded. This is made clear later but would be useful at this point.

Patients with missing values in any of the 28 laboratory test results, were excluded from the development dataset. This is made more clear in the manuscript in the “Materials and Methods”

section under “Development dataset”. After feature selection by the model, only 10 test results remain that are required to calculate the CoLab-score. Hence, hereafter the inclusion criterion for the temporal and external validation datasets are limited to the 10 test results of the CoLab-score.

5. I suggest making it clear what known variant or variants of COVID was circulating in the Netherlands during the data collection period. It is stated what has 'recently' been circulating, but it is not clear this was during the data collection. Further, were there known difference between the derivation and external validation variants?

This was one of the major shortcoming of our study and has been addressed in this revision. The temporal validation period has been extended from July 2020 until October 2021. This period now covers the emergence of new SARS-CoV-2 variants in the Netherlands, as well as widespread vaccinations. By merging our study data with data from the Dutch national institute of public health it is shown that the diagnostic performance of the CoLab-score is preserved under the rise of variant B.1.1.7 (Alpha), variant B.1.617.2 (Delta) and vaccinations. Please refer to Supplemental Material 2 for more details.

6. It is stated that different variants should be detectable due to it being based on the host response. Any evidence to this affect could be added. If so this would negate the previous point of different variants between the derivation and validation cohorts.

See response to previous comment. We would like to refer to the results in Supplemental Material 2.

7. There is some inconsistencies in who large numbers are recorded.

These have been addressed.

Reviewer: 2

Comments to the Author
Notes

1. Lateral flow testing: This is now widely used in emergency departments and provides rapid COVID-19 screening results. LFTs now form the direct competitor of such a score, rather than PCR, however the value add over lateral flow testing is not considered or shown in the validation.

Lateral flow tests (LFTs) are not part of routine care in the ED of our center, therefore we cannot directly compare the diagnostic performance with the CoLab-score. Nevertheless, we agree that LFTs are a direct competitor of the CoLab-score and that the added value has be considered in the manuscript. We argue that the advantage of the CoLab-score lies in the fact that its outcome is continuous, rather than binary. Low CoLab-scores (0 or 1) offer higher sensitivity than LFTs (as reported by other studies) and are therefore more suitable to rule-out COVID-19 when PCR-testing is not available. LFTs however, offer higher specificity. We have added a paragraph to the discussion highlighting these considerations with references to two large studies that report on the diagnostic performance of LFTs.

2. Vaccinations: The overwhelming majority of the Dutch, and much of the Western European population, are now vaccinated and this has been shown to reduce disease severity. Reduced disease severity will impact on degree of derangement of blood test results, and therefore likely have a significant impact on the sensitivity of this score. No mention or consideration is given to the impacts of vaccination - and what effects this may have on generalisability.

This is an important consideration that was also raised by reviewer 1. As stated in response to comment 5 by reviewer 1, we have extended the temporal validation period to cover the months when widespread vaccination took place. Although vaccination status was not registered for the vast majority of ED presentations, no evidence was found that the sensitivity of the CoLab-score is reduced in the months when widespread vaccination was achieved (see Table 2 in Supplemental Material 2). Moreover, 8 of 12 patients who were registered as vaccinated and were tested COVID-19 positive, had the highest CoLab-score (see Figure 2 in Supplemental Material 2). Whilst more evidence may be needed to show with statistical significance that the score is not affected by vaccinations, we hope that the results presented provide more confidence in the generalizability of the CoLab-score.

3. What is the value add of this work over existing works? The authors justify this over existing works (machine learning based) by saying that their model is easier to implement in hospital IT systems. What evidence can they offer to show that this is the case?

This depends on the health information system provider of the hospital. In the Netherlands, the vast majority of hospital systems use either software provided by ChipSoft or Epic. For ChipSoft it is only possible to perform “simple” operations such as multiplication and addition, which is the only requirement for calculating the CoLab-score. Whilst machine learning models can be reduced or approximated by simpler models, this is somewhat more involved and requires an extra step before implementation. We are not aware of any health information system providers that can readily implement machine learning models.

4. Some concerns about the requirement for some non-routine blood tests in inclusion criteria.

This is an important concern that we would like to clarify as this highlights a strength of our study. All 28 blood tests, as listed under “Laboratory tests” and Table 1, are part of the routine ED laboratory panel. ED physicians request this panel when a patient is presenting at the ED, rather than selecting tests individually. Therefore, all 28 tests are routine in the sense that they are part of the ED panel of our center, i.e. requiring CK or LDH or gGT as an inclusion criterion should not cause any bias as these are part of the routine panel. In some cases exceptions occur, e.g. for obstetric or pediatric patients, and tests are chosen individually. Nevertheless, these are exceptions and occur in about 10% of presentations as can be seen in Figure 1.

5. Exclusion of patients with previous COVID-19 is problematic - as this population is probably closest in performance to previously vaccinated, and does not help a physician decide whether a patient is infectious.

We agree that the CoLab-score should also alert physicians to patients who are re-infected with COVID-19. The main reason for excluding re-presenting patients is that during the first wave of COVID infections (i.e. the development cohort), the majority of patients who were discharged after treatment for COVID-19, but later re-presented at the ED, were re-presenting within 12 days after first presentation. Therefore, these were most likely not patients who had recovered and re-infected at a later point in time, but patients who had worsened after the initial presentation/treatment. Given that the CoLab-score follows the host-immune response, the score is time sensitive and is most accurate between 1 and 10 days after onset of symptoms (see Supplemental Material 4). Including these patients would impact the specificity of the CoLab-score as patients in a later phase of the disease show different biomarker profiles.

Please see some more specific comments below:

Background:

6. The authors cite Wynants et al. but I feel this fast-moving space mandates a more up to date search to give fair comment to today's literature. The limitation of Wynants' review is that it was last updated in early April 2020 (only 3 months after SARS-CoV-2 was first described!), and was therefore written at a timepoint where sufficient time for development & validation of models had not passed. The age of this review is particularly problematic as the field has been very fast moving, and as Wynants' living review has not been updated, it now offers a poor view on todays' landscape of COVID-19 diagnostic models. Please elaborate in the background on more recent uses of blood-based COVID-19 screening.

We are unsure if the reviewer has accessed the latest update of the paper by Wynants et al.. Since the release in April 2020, the paper has been updated twice, most recently on 3 February 2021, covering 232 prediction models for diagnosis and prognosis of COVID-19. We do however, agree that the paper does not focus on blood-based COVID-19 screening models, but is referenced to illustrate the vast amount of models developed to date and the varying quality. We are aware of two other studies with a similar goal, namely by Plante et al. and Soltan et al. these are discussed and compared to our study in the discussion.

7. Lateral flow testing which provides results in 30 minutes is now widely adopted in emergency departments and is readily accessible. How does this work add value over lateral flow testing which now forms the existing standard of care, and addresses this problem?

Please see the reply under comment 1.

8. The exclusion of previous confirmed COVID-19 patients from analysis is problematic, as a physician wishes to know whether a patient is (a) infected with COVID-19, and if (b) is infectious to others. Patients re-infected with COVID-19 are likely to still be infectious (albeit less so, and at lower risk of severe disease) - and therefore to justify a use for such a tool in infection-control, there needs to be consideration of reinfected patients.

Please see the reply under comment 5.

9. Why were patients with 'extreme' results (>10x SD from mean) excluded? Many patients presenting to an emergency department will attend because they are unwell - and therefore are at highest probability of an extreme abnormal result (e.g. a CRP >100 where a reference range is <4 would likely be excluded - but is an important group). Was the exclusion criterion >10 times the study population mean, or a laboratory reference range? Please elaborate.

The exclusion criterion is >10 times SD from the median, both the median and SD are calculated from the study population. The reason for excluding these patients is to stabilize estimation of regression coefficients. Leaving extreme results in the model dataset can cause failure of model convergence and unstable estimates of regression coefficients, negatively affecting future predictions. There is only a small group that is excluded by this criterion, the exclusion limits are shown in Table 2 to prevent users from calculating the CoLab-score for 'extreme' patients. We have added this to the Materials and Methods section under "Development dataset".

10. No specific consideration is given to patients who have been vaccinated. This raises two significant problems, as vaccination reduces severity of disease. First, it is reasonable therefore to presume that the extent of abnormality in blood-tests considered by the score will be reduced - and therefore that sensitivity of the score in vaccinated patients will be lower. Second, this greatly

compromises generalisability as >80% of the Dutch population are now vaccinated (similar figures seen also across Western Europe).

Please see the reply under comment 2.

11. Selection of lab panel constituents: Some of the lab tests which make up inclusion criterion - for example, LDH, CK, ɣ GT - are not widely considered to be routine tests performed for patients attending the emergency department prior to the pandemic. Use of these particular tests only became more widespread after emergence of the pandemic and early reports that they were deranged. However, prior to the pandemic, they were only performed in specific subgroups of patients - for example, lymphoma patients or patients who had had falls with long lies. During the pandemic, they are selectively performed where there is potential concern regarding COVID-19. There is therefore a problem in using pre-pandemic patients for model development and excluding patients missing these results - this potentially introduced a very significant bias in the composition of the development dataset; the very presence of a 'CK' result will bias towards a potential case. This may in part also explain why the rate of missing data was higher in the pre-pandemic population (figure 1a). Can the authors confirm that LDH, CK, ɣ GT measurements were considered part of the routine blood panel for allpatients attending ED? Similarly for validation, if the score was only applied to patients who had had a CK/LDH measured - this may bias towards patients where a clinician suspected COVID-19 and therefore arranged for this to be measured.

Please see the reply under comment 4.

12. The importance of high NPV is clear, however when quoting NPV it would be helpful to simultaneous mention PPV. For example, when stating a CoLab score of 0 having an NPV of 0.996, it is relevant that PPV is 0.06 (meaning that only around 1 in 16 patients judged positive by this score would be true positives).

The PPV has been added where the NPV is mentioned, and vice versa.

13. On the individual patient level, how does this perform for patients with certain comorbidities, e.g. haematological malignancy?

The CoLab-score is based on a large representative sample of ED presentations, it is therefore reasonable to assume that patients with a haematological malignancy are also included in this sample and that the model is able to deal with these patients, i.e. not generating false negatives or false positives. Nevertheless, it is true that the score can be affected by extreme comorbidities. For example erythropenia can lead to false negative results, which was the motivation for setting the exclusion limit for erythrocytes < 2.9 /pL. The CoLab-score serves as an early warning score, not a replacement for a clinical examination.

14. The authors identify existing works that have performed the same function as the work presented here (references 19 and 25). The authors say that the value add of their work is that their proposed score is easier to implement in current hospital IT systems than machine learning models - and that this warrants a trade-off in performance. What evidence can they offer that this is the case? Machine learning models can be mathematically reduced to weights attached to variables, and so theoretically could be similar in difficulty to implement to this model.

Please see the reply under comment 3.

15. Please see previous comments on potential inclusion bias; the primary focus here is whether CK/LDH/gamma GT were as routinely performed prior to the pandemic as during the pandemic. The authors report that the full lab panel was most frequently missing for surgical patients - in my view this

is evidence of bias, as these patients are at lower probability of testing positive for COVID-19 than patients presenting with acute COVID-19 symptoms.

Please see the reply under comment 4. Furthermore, there is indeed about 10% of missingness that is mainly represented by surgical, pediatric and obstetric patients. The a-priori probability of these patients of testing positive for COVID-19 could indeed be lower than other patients. This missingness is unavoidable and the result of the retrospective nature of this study. However, we argue that this should not be a concern for bias in the regression coefficients itself but could mean that probabilities predicted by the model may be over-estimated, since the actual prevalence is lower when these patients would be taken into account. I.e. this will have an effect on the model calibration rather than the discriminative ability or bias in regression coefficients.

16. Reference is made to data that is not shown in the discussion - please either qualify this by presenting this data in the supplement, or adapt comments to remove this.

We have added the data to Supplemental Material 4, Figure 2.

Reviewer: 3

Comments to the Author

1. Both the derivation and validation analyses used correlated (or clustered) data in that each participant can have >1 presentation, yet there is no mention how the logistic regression model and calculation of diagnostic accuracy account for the correlated data. The authors should consider using mixed-effect modelling for the logistic regression and logistic random-effects models or generalised estimating equations to calculate the diagnostic accuracy values (see https://doi.org/10.1148/radiol.12120509). At the very least, the authors should perform sensitivity analyses whereby only the first presentation from each participant is used to determine whether their results are robust to the effect of potential clustering effect.

This is an important observation from a statistical point of view. It is indeed possible for a single patient to appear twice or more in the dataset, depending on the number of ED presentations within the time period. However, there are two reasons that mixed-effects modeling nor generalized estimation equations were used. First, 86% of patients have only one ED presentation, 10% have two ED presentations and the remaining 4% of patients have 3 or more presentations. Therefore, the vast majority of patients have only a single visit. Secondly, we assume that ED presentations are independent observations. The median time between re-presentations is 38 days, most likely resulting in variations in laboratory values between presentations, and hence, only weak correlations. We believe mixed-effects modeling is more suitable when the majority of patients have multiple visits that are close in time or with strong autocorrelation. Nevertheless, we have performed a sensitivity-analysis only taking the first presentation (N = 8610) of each patient. The AUC did not differ (0.937, 95% CI: 0.923 - 0.957), nor did the diagnostic performance of the CoLab-scores (cf. Table 3 in the manuscript):

	Sensitivity	Specificity	PPV	NPV	% of pop.
0	0.980 (0.970 - 0.997)	0.455 (0.398 - 0.607)	0.141 (0.122 - 0.182)	0.996 (0.994 - 1.000)	42 (37 - 56)
≤ 1	0.926 (0.889 - 0.943)	0.802 (0.773 - 0.856)	0.298 (0.259 - 0.368)	0.992 (0.987 - 0.994)	74 (71 - 79)
≤ 2	0.867 (0.843 - 0.909)	0.885 (0.871 - 0.909)	0.407 (0.365 - 0.470)	0.987 (0.984 - 0.991)	82 (81 - 84)

≤ 3	0.790 (0.752 - 0.842)	0.949 (0.943 - 0.960)	0.585 (0.542 - 0.664)	0.980 (0.977 - 0.986)	89 (88 - 90)
≤ 4	0.667 (0.592 - 0.746)	0.976 (0.972 - 0.983)	0.714 (0.678 - 0.798)	0.970 (0.963 - 0.978)	92 (91 - 93)

2. The fact that many of the participants did not have PCR testing is a limitation of this retrospective analysis. Both the derivation and validation analyses assumed that untested participants do not have Covid-19. There is no justification why such an assumption is valid. Among participants with PCR testing, participants at Catharina Hospital had positivity rate of 279/1224 (22.8%), participants at 'Center 1' had positivity rate of 52/501 (10.4%), participants at 'Center 2' had positivity rate of 99/1051 (9.4%) and participants at 'Center 2' had positivity rate of 336/839 (40.0%). Whilst patients selected to have PCR testing may be more likely to have a positive test, patients not selected for PCR may still have Covid-19. The high positivity rates (especially at 'Centre 3') suggest that a substantial proportion of people with Covid-19 remain untested. At the very least, the authors should perform sensitivity analyses using data only from participants with PCR testing to determine whether their results are robust. In particular, it seems unlikely that 1/3 of participants with PCR testing will have CoLab-score of "0". Perhaps what is required is different thresholds of CoLab-score to guide PCR testing according to different pre-test probabilities.

It is indeed possible that some patients in the "Untested" population are missed by clinicians and are in fact PCR positive. However, we do not believe this will have a large impact on the model and resulting CoLab-score. Untested patients form, together with pre-pandemic and PCR negative patients, the control group. The control group is much larger than the case/PCR-positive group (N = 10138, versus N = 279), therefore a small number of false negatives in the control group is not expected to result in a large change of regression coefficients or performance of the CoLab-score since these are outnumbered by true negatives. In any case, we have also evaluated the performance of the CoLab-score on the subgroup of PCR-tested patients. Discriminative ability is slightly lower with an AUC of 0.9077 (95% CI: 0.8887-0.9268) however confidence intervals overlap. In the diagnostic performance of the CoLab-scores only minor differences are observed (PPV and NPV not taken into account due to different pre-test probabilities). The fraction of patients with score 0 is somewhat lower with 31% versus 38% in the development cohort.

	Sensitivity	Specificity	PPV	NPV	% of pop.
0	0.983 (0.969 - 0.990)	0.390 (0.291 - 0.502)	0.320 (0.279 - 0.381)	0.988 (0.974 - 0.994)	31 (23 - 38)
≤ 1	0.911 (0.898 - 0.949)	0.719 (0.672 - 0.770)	0.487 (0.430 - 0.548)	0.965 (0.958 - 0.981)	58 (54 - 62)
≤ 2	0.856 (0.821 - 0.892)	0.821 (0.798 - 0.862)	0.582 (0.535 - 0.650)	0.951 (0.939 - 0.967)	67 (65 - 71)
≤ 3	0.757 (0.704 - 0.803)	0.915 (0.896 - 0.932)	0.722 (0.668 - 0.767)	0.928 (0.914 - 0.943)	76 (74 - 78)
≤ 4	0.610 (0.533 - 0.696)	0.955 (0.943 - 0.968)	0.797 (0.755 - 0.860)	0.893 (0.870 - 0.920)	83 (80 - 85)

My other comments are:

3. For the abstract,

i. The aims of developing and validating a clinical score should be explicitly stated in the 'Introduction' (page 5 lines 16-20) instead of just stating that such a score is desirable.

ii. The method of validation (by calculating diagnostic accuracy values) should be explicitly stated in the 'Methods' (page 5 lines 30-35).
iii. Both the sensitivity and specificity values should be provided in the 'Results' (page 5 lines 47-52). Providing sensitivity and negative predictive values meant that results of false positives are not being presented. Providing specificity and positive predictive values meant that results of false negatives are not being presented.
iv. In the 'conclusions', it is important to acknowledge that the proportion of patients not requiring PCR based on their CoLab-score may vary according to the prevalence of Covid-19. Perhaps it is better to say that "Depending on the community prevalence, COVID-19 may be safely ruled-out in more than one third of ED presentations" (page 5 line 59 to page 6 line 4) instead of "With this score, COVID-19 can be safely ruled-out in more than one third of ED presentations".

Thank you for these suggestions, we have edited the abstract accordingly.

4. In the 'Methods' (page 9 line 9), it should be stated that the study is a retrospective case-control study, instead of the study design only being stated in the 'Discussion' (page 20 line 49).
• All abbreviations for the tests in Table 2 should be explained in the 'laboratory tests' subsection (page 10 line 40 to page 11 line 8). I am uncertain what "AF" is referring to. Is it alpha-fetoprotein?
• Ideally, the rationale for evaluating the tests in the 'laboratory tests' subsection (page 10 line 40 to page 11 line 8) should be clearly stated. Is there evidence from existing literature that the tests differentiate between those with and without Covid-19?

These additions have been made. AF has been changed to ALP, which refers to alkaline phosphatase, its abbreviation is now in accordance to the Materials and Methods section. The rationale for selecting the tests in the laboratory tests is that these are part of the routine ED panel and do therefore not suffer from selection bias. Findings are compared to existing literature in the discussion.

5. Confidence intervals for β should be provided in Table 2.

Unfortunately confidence intervals for (adaptive) lasso regression is an open issue in statistics. Therefore we have supplied the relative importance as a substitute.

6. In Table 3, the cross tabulation of the index test results by the results of the reference standard should be provided. For example, the values of TP, TN, FP and FN can be provided as they were in Table 4.

To keep Table 3 in line with Table 4 we have omitted TP, TN, FP and FN numbers from Table 4. These can be calculated from the sensitivity and specificity and would result in a table outside the margins when given with 95% confidence intervals.

7. In Table 4, the confidence intervals for the diagnostic accuracy values should be provided, as they were in Table 3.

The 95% confidence intervals have been added this table.

8. The demographics for participants in the validation dataset should be provided, at least as supplementary material if the table limit for the main text has been reached.

Aside from age and gender, the only other demographic information available for this study is the specialism for which the patient was admitted. We have added this information to table one. More

demographic information can unfortunately not be made available as the approved study protocol does not specify this.

VERSION 2 – REVIEW

REVIEWER	Zhe Hui Hoo The University of Sheffield, School of Health and Related Research (ScHARR)
REVIEW RETURNED	13-Nov-2021

GENERAL COMMENTS	Thank you for inviting me to re-review this manuscript. The authors have iterated the manuscript taking into account the earlier set of reviews. Some limitations remain and I recommend the following changes: • Lateral flow testing is now discussed in the manuscript but lateral flow testing is now an important aspect of Covid-19 testing and the value of the scoring system as compared to lateral flow testing should be mentioned in the 'Article summary' and ideally in the Abstract as well. • Whilst the 28 blood tests are routine for the ED department of Catharina Hospital Eindhoven (see the underlined response from authors to the comments by Reviewer #2), these tests may well not be routine in other ED departments. Also, not everyone presenting to ED will require a blood test. Even for the study subjects from Catharina Hospital Eindhoven, the lab panel was incomplete in 5231/30368 (17.2%) of all presentations (Figure 1). It is important for the authors to consider the practicality and cost of the required blood tests against lateral flow testing if the aim is to guide the need for PCR testing. • Neither vaccination status nor genomic information of SARS-CoV-2 variant were available for study subjects (page 12 lines 32-35), though the authors have the vaccination status for a small subgroup of 12 study subjects (page 23 line 32). Therefore, the temporal validation period was divided into three phases according to extent of vaccination and the dominant variant of each phase (page 12 lines 34-46). This indirect method of evaluating the test characteristics is vulnerable to confounding – it is entirely plausible that people with different vaccination status have different likelihood of being Covid-19 positive. Even with this indirect evaluation, there is already some differences in test characteristics, for example CoLab-score of ≤2 has a sensitivity of 0.864 (95% CI 0.826 to 0.902) during the second phase and sensitivity of only 0.690 (95% CI 0.569 to 0.810) during the third phase. It must also be noted that overlapping confidence intervals does not necessarily indicate a lack of difference (doi: 10.1067/mva.2002.125015). Unless there is direct evidence that the score behaves in the same way among people with different vaccination status and people infected by different variants of SARS-CoV-2, the authors must acknowledge that "No evidence was found that the performance was affected by vaccinations and new SARS-CoV-2 variants" ('Article summary') simply reflects an absence of evidence rather than evidence of absence. • In terms of the results being unaffected by study subjects with >1 presentation and study subjects without PCR testing, the authors should perform a sensitivity analysis of only the first presentation among those with PCR testing and include that sensitivity analysis as one of the supplemental material.

	• Omitting the TP, TN, FP and FN from the diagnostic accuracy tables (Tables 3 & 4) is a poor practice. Recommendation #23 on the 2015 STARD checklist is “Cross tabulation of the index test results (or their distribution) by the results of the reference standard”. If there is insufficient space in the diagnostic accuracy results, the cross tabulation results can be presented in the supplemental material.

REVIEWER	Andrew A Soltan University of Oxford
REVIEW RETURNED	28-Dec-2021

GENERAL COMMENTS	Thank you very much for inviting me to re-review the revised manuscript, and I’m grateful to the authors for their time. There is much that is done well in this study, such as the statistical aspects of the score derivation, the external validation, and additional analyses in the supplement. Although below I’ve focussed on areas where improvement/clarification is needed, my overall impression of the work is positive and I thank the authors for their additions. I have some areas for clarification/improvement on scientific aspects of the study: 1. I remain concerned about the exclusion of all patients with any of the 28 (or 10 final) laboratory tests missing, partly as there are some hints in the data that these selected tests are not truly routine for all patient presentations (evidenced by the below-expected representation of surgical specialties shown in Table 1), and partly as missingness is probable to be informative where it is present (for example - that additional lab tests are likely to be collected at external centres where there is a clinical suspicion of COVID-19). One way to resolve this could be i) reporting the fraction of missing data for each lab parameter by COVID-19 PCR status (therefore showing if there is asymmetry in the missing data) and ii) performing sensitivity analysis where imputation has been performed, but there may also be other approaches.- The Figure 1 legend reads: “For the development dataset, completeness of the lab panel was assessed for all the 29 laboratory tests (see Table 1), for the temporal validation dataset this was only necessary for 10 laboratory tests (see Table 2).”. However I am unable to find rates of missing data reported for lab parameters in Tables 1 & 2 - apologies if I have missed this elsewhere, or if it is in a supplementary file which I haven’t received in the review portal, but I would be grateful if I could please be pointed in the right direction as these data would go some way to allaying concerns on the above point. 2. The very high level of exclusions (~50% in Centre 1, ~60% in centre 2, and ~50% in Centre 3 - as per Figure 2) is concerning as it shows that the score has not generalised well when applied to other centres, and that missing lab data were the major reason for this. This goes back to my earlier concern about the selected tests not being truly routine, and missingness potentially being informative of low-clinical suspicion. Additional analyses which might address this concern could be reporting of rates of missing data for each lab parameter by PCR status, and looking at whether this reveals any asymmetry or informativeness in missing data. A further approach/sensitivity analysis could try impute the missing lab data using - for example - training population median values, and assessing whether model performance is lower on the ~50% of validation patients who were excluded. In both cases, I think

	these missing validation populations should be more closely examined, and also more fully discussed in limitations in the discussion.- As with the previous comment, apologies if the fraction of missing lab parameters was reported here as per the figure legend but I have not been able to find it in the review portal. 3. As COVID-19 can present asymptomatically, to justify the benefits set out in the background & discussion the tool would need to be applied to all patients being admitted to hospital. Therefore it is problematic if the tool is (or is shown to be) less generalisable to surgical patients owing to missing data, and I also suspect that specificity of CoLab 0 and 1 would fall further if applied to all patients (which is potentially problematic given the very high false positive rate/low PPV for CoLab 0, though I appreciate this is just one configuration). One possible solution could be to more strictly define the scope & utility of CoLab as being for 'acute medical/respiratory admissions' and excluding surgical/ObGyn/Paediatric admissions entirely - however the current implementation (where some surgical admissions are included, but a majority is asymmetrical excluded) may potentially be seen as problematic. 4. The exclusion criterion of 10 SDs from the median feels a little unusual to me (and perhaps the manuscript should be clarified to make clear this is 10SD from median rather than mean, or perhaps could consider median and 1st / 99th centiles), but notwithstanding this the cut-off thresholds in Table 2 seem clinically reasonable to me and so I am not concerned about this biasing the results. (From my understanding of Table 2, I cannot see an upper CRP cut off, which is also reassuring) 5. What is meant by relative importance in Table 2? If it means relative importance to model prediction, I would have expected total importances to sum to 100%. I would be grateful for some clarification on what is meant by this, and perhaps some information in the supplement about how these were calculated 6. A clinical/biochemical characterisation of false positive and false negative patients (perhaps in the supplement) would be very helpful. 7. Could I please ask for clarification on how the estimate of the percentage of vaccinated patients attending the ED is calculated? My initial understanding of figure 1 in Supplementary Material 2 was that the blue representing estimated fraction of patients attending the ED who are vaccinated was estimated from general population data. Can I please confirm if estimate of vaccination rate in ED was simply inferred from the national rate of vaccination for the given age distribution at the presentation date? If vaccination rates are inferred by looking at the fraction of vaccinated individuals in the general population, attention must be given to the study group being patients more likely to have severe disease (i.e. attending the ED) - and therefore that unvaccinated individuals will be dramatically over-represented in the ED population relative to the general population (by a factor that can be determined by vaccine efficacy). If the % of the general population vaccinated at the age group/date was used, this would be neglecting the protective effects of vaccine against severe disease and be a major issue with the analysis. Apologies if I have misunderstood how the estimate has been calculated, but some clarification here would be helpful. If there is a problem with the inference of likely percentage vaccinated, some other confirmation of CoLabs' performance in vaccinated patients (as in the initial round of comments) would please be needed.

	A couple of only brief notes about the background and discussion: 1. I still feel that the paper would benefit from fuller discussion of competing & prior work in the background. Although I appreciate comments regarding Wynants et al., the update of Feb 2021 covers only models to June 2020 which is still within the very first months of the pandemic, and the review's very ambitious remit in an exponentially growing literature space means it has struggled to keep pace. I feel the field has developed very significantly in the latter half of 2020 & throughout 2021 such that readers would benefit from an updated background, which mind find a much less pessimistic picture than Wynants et al. presents in early 2020. 2. I think the low PPV of using the low CoLabs 0/1 as a rule-out (and its clinical meaning should be raised as a limitation in the discussion as this has implications on logistical benefit - i.e. In the centre 2 validation a cut off of CoLabs 0 called 19 false positives for each true positive! 3. Noting and agreeing with the response on implementational advantages of a simpler model, the authors might be interested in EPIC's AppOrchard platform which allows for app-style deployment of EHR plugins Thank you very much for the opportunity to rereview this paper - and for its strengths also - and I hope to work further with the authors as they iterate towards a final version.

VERSION 2 – AUTHOR RESPONSE

Reviewer: 1

1. Lateral flow testing is now discussed in the manuscript but lateral flow testing is now an important aspect of Covid-19 testing and the value of the scoring system as compared to lateral flow testing should be mentioned in the 'Article summary' and ideally in the Abstract as well.

As commented by the editor, the "Article summary" is unfortunately only allowed to relate specifically to the methods. Since lateral flow tests (LFTs) were not part of the methods we cannot mention any comparison to LFTs in the "Article summary". We have therefore included a line the abstract reflecting that the CoLab-score can be used in addition to LTFs: "The CoLab-score is continuous, in contrast to the binary outcome of lateral flow testing, and can guide PCR testing and triage ED patients."

2. Whilst the 28 blood tests are routine for the ED department of Catharina Hospital Eindhoven (see the underlined response from authors to the comments by Reviewer #2), these tests may well not be routine in other ED departments. Also, not everyone presenting to ED will require a blood test. Even for the study subjects from Catharina Hospital Eindhoven, the lab panel was incomplete in 5231/30368 (17.2%) of all presentations (Figure 1). It is important for the authors to consider the practicality and cost of the required blood tests against lateral flow testing if the aim is to guide the need for PCR testing.

It is indeed not likely that all 28 blood tests are routine for other ED departments. Therefore, feature selection with adaptive lasso regression was performed. This results in the CoLab-score requiring only 10 blood tests (see Table 2). The full lab panel was indeed incomplete for 17.2% of presentations. However, 7.7% of missingness was due to analytical errors, leaving ≈10% of actual missingness, mainly due to surgical/ob-gyn/pediatric presentations. The CoLab-score is only valid for an ED presentation that requires a blood test, this is an inclusion criterion. Therefore, if an ED presentation does not require a blood test, it would not be practical nor valid to use the CoLab-score.

In this case lateral flow testing is a reasonable alternative. We have more strictly defined the target population of the CoLab-score, as this concern was also shared by the other reviewer,

3. Neither vaccination status nor genomic information of SARS-CoV-2 variant were available for study subjects (page 12 lines 32-35), though the authors have the vaccination status for a small subgroup of 12 study subjects (page 23 line 32). Therefore, the temporal validation period was divided into three phases according to extent of vaccination and the dominant variant of each phase (page 12 lines 34-46). This indirect method of evaluating the test characteristics is vulnerable to confounding – it is entirely plausible that people with different vaccination status have different likelihood of being Covid-19 positive. Even with this indirect evaluation, there is already some differences in test characteristics, for example CoLab-score of ≤2 has a sensitivity of 0.864 (95% CI 0.826 to 0.902) during the second phase and sensitivity of only 0.690 (95% CI 0.569 to 0.810) during the third phase. It must also be noted that overlapping confidence intervals does not necessarily indicate a lack of difference (doi: 10.1067/mva.2002.125015).
Unless there is direct evidence that the score behaves in the same way among people with different vaccination status and people infected by different variants of SARS-CoV-2, the authors must acknowledge that “No evidence was found that the performance was affected by vaccinations and new SARS-CoV-2 variants” (‘Article summary’) simply reflects an absence of evidence rather than evidence of absence.

We agree that people with different vaccination status have a different likelihood of presenting at the ED, i.e. that the estimated fraction of ED population vaccinated (as depicted in Supplemental Material 2, Figure 1) is most likely an over-estimation. To avoid confusion, we have renamed the y-axis to “Age matched fraction vaccinated”. Nevertheless, the CoLab-score is aimed at patients presenting at the ED, if vaccinated patients have a lower likelihood of presenting at the ED than this results in a change in the ED patient population. What we aim to show, by splitting the data in three phases, is not that the CoLab-score performs equally in vaccinated and unvaccinated patients, but that the discriminative ability is not affected due to changes in the underlying ED patient population as a result of vaccinations. To emphasize this we have rewritten this section of the discussion to: “Moreover, there is no evidence that the discriminative ability of the CoLab-score is reduced by a change in the ED patient population as a result of widespread vaccination.”
In the results we have added a sentence that reflects the lower sensitivity in the third phase:
“Diagnostic performance is preserved in terms of sensitivity and specificity, except a moderately reduced sensitivity of scores ≥ 3 in the third phase as compared to the first phase.”

4. In terms of the results being unaffected by study subjects with >1 presentation and study subjects without PCR testing, the authors should perform a sensitivity analysis of only the first presentation among those with PCR testing and include that sensitivity analysis as one of the supplemental material.

We have included these sensitivity analyses in Supplemental Material 4, Table 1. We have used the temporal validation cohort to assess the performance of the CoLab-score under different inclusion criteria. The reason for doing so is that this cohort is not used in model fitting so the result most closely resembles real-world performance. We have performed two analyses: i)using only the first presentation of each subject (i.e. no repeated presentations) ii) including only subject with a PCR test. We have compared the performance to the original temporal validation cohort in Supplemental Material 4, Table 1. From these analyses we can conclude that including re-presentations does not have an influence on the reported performance, but that the performance is reduced when the CoLab-score is used in a PCR-tested population only. This does not change the conclusion of the study since the CoLab-score is not meant as a replacement for PCR-testing. We have added lines in the discussion referring to these results.

5. Omitting the TP, TN, FP and FN from the diagnostic accuracy tables (Tables 3 & 4) is a poor practice. Recommendation #23 on the 2015 STARD checklist is “Cross tabulation of the index test results (or their distribution) by the results of the reference standard”. If there is insufficient space in the diagnostic accuracy results, the cross tabulation results can be presented in the supplemental material.

These have been re-added, including 95% CIs.

Reviewer: 2

1. I remain concerned about the exclusion of all patients with any of the 28 (or 10 final) laboratory tests missing, partly as there are some hints in the data that these selected tests are not truly routine for all patient presentations (evidenced by the below-expected representation of surgical specialties shown in Table 1), and partly as missingness is probable to be informative where it is present (for example - that additional lab tests are likely to be collected at external centres where there is a clinical suspicion of COVID-19). One way to resolve this could be i) reporting the fraction of missing data for each lab parameter by COVID-19 PCR status (therefore showing if there is asymmetry in the missing data) and ii) performing sensitivity analysis where imputation has been performed, but there may also be other approaches.

It is correct that the fraction missingness in the PCR-tested group is significantly lower than in the untested group ($\chi 2$-test p-value <0.001). This true for the development cohort, as well the validation cohorts (all $\chi 2$-test p-values <0.001). Rather than imputing we have asked the ED clinicians for a more strict definition of patients that fulfill the inclusion criteria for a full ED laboratory panel and, as a consequence, for which patients the full panel is not requested. This is summarized as follows and added to the “Methods”, “Development dataset” section:
“The routine ED laboratory panel is requested for (adult) patients presenting with abdominal pain, chest pain, shortness of breath, syncope, sepsis or other non-specific complaints, or for patients (including non-adult patients) presenting with specific complaints where a suspected diagnosis has to be ruled-in or ruled-out.”
In the discussion we have also added a section which addresses the missingness in the external centers, please refer to the response to question 3.

2. The Figure 1 legend reads: “For the development dataset, completeness of the lab panel was assessed for all the 29 laboratory tests (see Table 1), for the temporal validation dataset this was only necessary for 10 laboratory tests (see Table 2).”. However I am unable to find rates of missing data reported for lab parameters in Tables 1 & 2 - apologies if I have missed this elsewhere, or if it is in a supplementary file which I haven’t received in the review portal, but I would be grateful if I could please be pointed in the right direction as these data would go some way to allaying concerns on the above point.

The tables are referenced to, to list the names of the laboratory tests that were assessed for completeness. So, our apologies for the confusion as there is no table with missingness per lab parameter, moreover 29 should read 28. We have removed the references in the text and replaced 29 with 28. We have added a line to the discussion indicating that the missingness is significantly lower in the PCR-tested group versus the untested group.

3. The very high level of exclusions (~50% in Centre 1, ~60% in centre 2, and ~50% in Centre 3 - as per Figure 2) is concerning as it shows that the score has not generalised well when applied to other centres, and that missing lab data were the major reason for this. This goes back to my earlier concern about the selected tests not being truly routine, and missingness potentially being informative of low-clinical suspicion. Additional analyses which might address this concern could be reporting of

rates of missing data for each lab parameter by PCR status, and looking at whether this reveals any asymmetry or informativeness in missing data. A further approach/sensitivity analysis could try impute the missing lab data using - for example - training population median values, and assessing whether model performance is lower on the ~50% of validation patients who were excluded. In both cases, I think these missing validation populations should be more closely examined, and also more fully discussed in limitations in the discussion.
As with the previous comment, apologies if the fraction of missing lab parameters was reported here as per the figure legend but I have not been able to find it in the review portal.

We agree that the high rates of missingness in the laboratory parameters of the external centers warrants a more full discussion. The indications for requesting a panel of blood tests required for the CoLab-score, can vary between centers. This indeed has its implications for the interpretation of the external validation results. We have therefore added this limitation to the discussion:
“Due to these high levels of missingness, the results of the external centers cannot be used to show that the CoLab-score generalizes to the entire ED population. Rather, the results show that for the vast majority of COVID-19 positive patients that are presenting at the ED, a routine laboratory panel is available from which the CoLab-score can be calculated, and that the performance of the CoLab-score in this population is comparable to the development population.”

4. As COVID-19 can present asymptomatically, to justify the benefits set out in the background & discussion the tool would need to be applied to all patients being admitted to hospital. Therefore it is problematic if the tool is (or is shown to be) less generalisable to surgical patients owing to missing data, and I also suspect that specificity of CoLab 0 and 1 would fall further if applied to all patients (which is potentially problematic given the very high false positive rate/low PPV for CoLab 0, though I appreciate this is just one configuration). One possible solution could be to more strictly define the scope & utility of CoLab as being for ‘acute medical/respiratory admissions’ and excluding surgical/ObGyn/Paediatric admissions entirely - however the current implementation (where some surgical admissions are included, but a majority is asymmetrical excluded) may potentially be seen as problematic.

We agree that is too optimistic to state that the CoLab-score can be applied to all patients admitted to the hospital. We have more strictly defined the scope and utility in several places in the manuscript:
1. In the final paragraph of the introduction indicated that the score is based on “patients who undergo routine laboratory testing at presentation.”
2. In the “Methods” section under “Development dataset” given a more strict definition when routine laboratory is requested by ED clinicians (see answer to question 1).
3. In the “Discussion” the second paragraph more strictly defines the target population.
4. In the “Discussion” the final paragraph (conclusion) the target population is also more strictly defined.
5. In the “Article Summary” we have added a statement regarding the limited generalizability to other centers due to the missingness.

5. The exclusion criterion of 10 SDs from the median feels a little unusual to me (and perhaps the manuscript should be clarified to make clear this is 10SD from median rather than mean, or perhaps could consider median and 1st / 99th centiles), but notwithstanding this the cut-off thresholds in Table 2 seem clinically reasonable to me and so I am not concerned about this biasing the results. (From my understanding of Table 2, I cannot see an upper CRP cut off, which is also reassuring)

The reason for choosing a 10 SD from the median, rather than mean, is that some distributions of laboratory parameters are skewed and/or have outliers, the median is preferred as it is more resistant to outliers. This was stated in the “Methods” section under “Development Dataset” but in parentheses.

We have made this more explicit and stated the reason for choosing the median rather than the mean.

6. What is meant by relative importance in Table 2? If it means relative importance to model prediction, I would have expected total importances to sum to 100%. I would be grateful for some clarification on what is meant by this, and perhaps some information in the supplement about how these were calculated

Relative importance represents the importance of each variable, relative to the most important variable (in this case basophils, which explains why this variable has an importance of 100%). These are calculated from the unscaled coefficients, where each unscaled coefficient is divided by the maximum unscaled coefficient (basophils). This is different from cumulative importance, where the importance sums up to 100%. We have added a section to Supplemental Material 1 and a sentence in the legend of Table 2 as clarification.

7. A clinical/biochemical characterisation of false positive and false negative patients (perhaps in the supplement) would be very helpful.

We assume that PCR-positive patients with a CoLab-score of 0 are implied with “false negatives” and PCR-negative patients with a CoLab-score of 5 are implied with “false positives”. What we have observed is that chemotherapy, which causes myeloid suppression, will decrease neutrophilic, basophilic and eosinophilic counts and thereby “falsely” increasing the CoLab-score, resulting in false positives. Conversely, COVID-19 patients with severe anemia could have “falsely” lowered CoLab-scores (since erythrocyte concentration is a variable in the CoLab-core). To minimize false negatives, we have therefore advised to report CoLab-scores only when the concentration of erythrocytes is ≥ 2.9 /pL. We have added this to the discussion.

8. Could I please ask for clarification on how the estimate of the percentage of vaccinated patients attending the ED is calculated? My initial understanding of figure 1 in Supplementary Material 2 was that the blue representing estimated fraction of patients attending the ED who are vaccinated was estimated from general population data. Can I please confirm if estimate of vaccination rate in ED was simply inferred from the national rate of vaccination for the given age distribution at the presentation date? If vaccination rates are inferred by looking at the fraction of vaccinated individuals in the general population, attention must be given to the study group being patients more likely to have severe disease (i.e. attending the ED) - and therefore that unvaccinated individuals will be dramatically over-represented in the ED population relative to the general population (by a factor that can be determined by vaccine efficacy). If the % of the general population vaccinated at the age group/date was used, this would be neglecting the protective effects of vaccine against severe disease and be a major issue with the analysis. Apologies if I have misunderstood how the estimate has been calculated, but some clarification here would be helpful. If there is a problem with the inference of likely percentage vaccinated, some other confirmation of CoLabs’ performance in vaccinated patients (as in the initial round of comments) would please be needed.

This question overlaps with question 3 from the first reviewer. In short, the fraction of vaccinated patients was indeed inferred from the national rate of vaccination for the given age distribution at the presentation date. This would most likely be an over-estimation since unvaccinated patients have a higher likelihood of presenting at the ED. We have therefore renamed the y-axis to “Age matched fraction vaccinated”. We also agree that this does not prove that the CoLab-score performs equally in vaccinated and un-vaccinated patients. However, this does show that the CoLab-score still performs similarly in the ED patient population after widespread vaccination. The discriminative ability is not affected, by which we argue that the CoLab-score is still valid in patients presenting at the ED,

regardless of widespread vaccination. Please refer to the answer to question 3 from the first reviewer for the changes that were made to the manuscript. Finally, confirmation that the CoLab-score performs similarly in vaccinated and unvaccinated patients presenting at the ED is only available from a subgroup of 13 patients for whom vaccination status was registered, see Figure 2 in Supplemental Material 2.

9. I still feel that the paper would benefit from fuller discussion of competing & prior work in the background. Although I appreciate comments regarding Wynants et al., the update of Feb 2021 covers only models to June 2020 which is still within the very first months of the pandemic, and the review's very ambitious remit in an exponentially growing literature space means it has struggled to keep pace. I feel the field has developed very significantly in the latter half of 2020 & throughout 2021 such that readers would benefit from an updated background, which mind find a much less pessimistic picture than Wynants et al. presents in early 2020.

In the "exponentially growing literature space" we have decided to limit prior work, to studies that serve a similar goal to our study. Specifically, prediction models or risk scores that aim to identify COVID-19 positive patients, from all other patients presenting at the ED, based on routine (laboratory) data. We are aware of only two other studies that have been published (the study by Plante et al. and Soltan et al.) that fulfill these criteria.Aside from mentioning these in the discussion, we have also mentioned them in the introduction, illustrating that these are the only similar studies that were found. We agree that the current picture is less pessimistic than early 2020, we have therefore nuanced in the introduction: "methodological shortcomings of early models.".

10. I think the low PPV of using the low CoLabs 0/1 as a rule-out (and its clinical meaning should be raised as a limitation in the discussion as this has implications on logistical benefit - i.e. In the centre 2 validation a cut off of CoLabs 0 called 19 false positives for each true positive!

It is true that CoLabs 0/1 have a low specificity, however, we do not imply nor advise that all patients with a score > 0 should undergo PCR-testing as this would indeed lead to many false positives. Rather, the CoLab-score is a continuous score that is best used in such manner, i.e. the higher the score, the higher the clinical suspicion. Since the lowest score has a sensitivity greater than or equal to a single PCR-test, our reasoning is that using score 0 to rule-out an infection (when PCR-testing is not available) would not result in more patients being missed than PCR-testing all presenting patients. The logistical benefit arises from the fact that these patients can be safely excluded (unless clinicians decide otherwise!), not that all other patients should be PCR-tested.

11. Noting and agreeing with the response on implementational advantages of a simpler model, the authors might be interested in EPIC's AppOrchard platform which allows for app-style deployment of EHR plugins

We are glad to see that some EHR software vendors are focusing on platforms that allow for more advanced EHR plugins. In the era of artificial intelligence we welcome this development.

VERSION 3 – REVIEW

REVIEWER	Zhe Hui Hoo The University of Sheffield, School of Health and Related Research (ScHARR)
REVIEW RETURNED	13-Mar-2022

GENERAL COMMENTS	The authors have further amended their manuscript taking into account comments from reviewers. Most of my previous comments have been addressed but some issues remain:

	• For the 'Article summary', I suggest that the first bullet point should be expanded to: "A comprehensive panel of 28 laboratory tests was measured for 10.417 emergency department (ED) presentations and combined with SARS-CoV-2 PCR test results, but there was no comparison with results of lateral flow testing." • It should be made explicit in the limitations paragraphs (page 23 line 23 to page 25 line 34) that the Co-Lab score is only applicable to those admitted to ED and required routine blood tests. For example, at page 23 end of line 50, the following sentence could be included: "It should also be noted that the CoLab-score is only valid for an ED presentation that requires a blood test and may not be generalise to those presenting to ED but was otherwise well and did not require a blood test." • "A moderately reduced sensitivity of scores ≥ 3 in the third phase as compared to the first phase" (page 18 line 59) contradicts the statements that: "Diagnostic performance is sustained in periods with different dominant variants." (page 25 lines 20-22) "Moreover, there is no evidence that the discriminative ability of the CoLab-score is reduced by a change in the ED patient population as a result of widespread vaccination." (page 25 lines 22-27). Please can the authors re-interpret their findings accordingly? • It is interesting that the authors considered the reduced performance of CoLab-score among the PCR-tested population as inconsequential because "the CoLab-score is not meant as a replacement for PCR-testing." The issues here are: (a) The CoLab-score is meant to guide PCR testing (b) The CoLab-score requires validation prior to clinical use (c) A study validating the CoLab-score will be strengthened by the use of an objective endpoint. Though PCR is not a perfect test for Covid-19 infection, at least it is objective and widely accepted as the 'gold standard' for Covid-19 infection. In page 24 line 60 to page 25 line 3, the authors state that: "Note the performance of the CoLab-score in a suspected/PCR-tested cohort is not equal to the". This sentence appears to be hanging and I am unsure of the message here. The authors have demonstrated differential performance of the CoLab-score among those with PCR testing and those without PCR testing. My interpretation is that it is incorrect to assume everyone without PCR testing did not have Covid-19. It is a fact that asymptomatic Covid-19 infection is possible yet the authors did not account for this in their validation, which is a major methodological limitation. My suggestion would be to present the results only for those with PCR testing, unless the authors can account for possible asymptomatic Covid-19 infection among those without PCR testing. Otherwise the performance of the CoLab-score would be over-estimated. • I thank the authors for including the values of TP, TN, FP and FN in the diagnostic accuracy tables. However, these values need to be carefully checked. For example, in Table 3 (page 17 line 26), it is stated that TP = 133 and FN = 0 for CoLab-score ≤0. Sensitive

	= TP / (TP + FN), so sensitivity = 1.000 when FN = 0. Yet the sensitivity was listed as 0.984. Interesting, the sum of TP, TN, FP and FN varies for all five scores presented in Table 3 when the total population for the derivation dataset should be fixed. In contrast, Table 4 (page 20) presents the same sum of TP, TN, FP and FN (N = 14,080) for the corresponding five-CoLab scores.

VERSION 3 – AUTHOR RESPONSE

Reviewer: 1

1. For the 'Article summary', I suggest that the first bullet point should be expanded to:

"A comprehensive panel of 28 laboratory tests was measured for 10.417 emergency department (ED) presentations and combined with SARS-CoV-2 PCR test results, but there was no comparison with results of lateral flow testing."

We have added a separate bullet point to reflect this limitation: "The score was not directly compared to lateral flow testing.".

2. It should be made explicit in the limitations paragraphs (page 23 line 23 to page 25 line 34) that the Co-Lab score is only applicable to those admitted to ED and required routine blood tests. For example, at page 23 end of line 50, the following sentence could be included:

"It should also be noted that the CoLab-score is only valid for an ED presentation that requires a blood test and may not be generalise to those presenting to ED but was otherwise well and did not require a blood test."

We have added this line to the limitations at the bottom of page 23: "Important to note is that the CoLab-score is only valid for ED presentations where routine blood testing is requested, and as a consequence does not generalize to the ED population who is otherwise well and does not undergo routine blood testing."

3. "A moderately reduced sensitivity of scores ≥ 3 in the third phase as compared to the first phase" (page 18 line 59) contradicts the statements that:

"Diagnostic performance is sustained in periods with different dominant variants." (page 25 lines 20-22) "Moreover, there is no evidence that the discriminative ability of the CoLab-score is reduced by a change in the ED patient population as a result of widespread vaccination." (page 25 lines 22-27).

Please can the authors re-interpret their findings accordingly?

Stating that the diagnostic performance is sustained may indeed be too strong, as there are some discrepancies in individual scoring categories. We have therefore changed "diagnostic performance" to "discriminative ability", since there is no evidence that the discriminative ability in terms of area-

under-the-ROC-curve (AUC) significantly different between periods (see Supplemental Material 2 Table 2).

4. It is interesting that the authors considered the reduced performance of CoLab-score among the PCR-tested population as inconsequential because “the CoLab-score is not meant as a replacement for PCR-testing.” The issues here are:

(a) The CoLab-score is meant to guide PCR testing

(b) The CoLab-score requires validation prior to clinical use

(c) A study validating the CoLab-score will be strengthened by the use of an objective endpoint. Though PCR is not a perfect test for Covid-19 infection, at least it is objective and widely accepted as the ‘gold standard’ for Covid-19 infection.

In page 24 line 60 to page 25 line 3, the authors state that: “Note the performance of the CoLab-score in a suspected/PCR-tested cohort is not equal to the”. This sentence appears to be hanging and I am unsure of the message here.

The authors have demonstrated differential performance of the CoLab-score among those with PCR testing and those without PCR testing. My interpretation is that it is incorrect to assume everyone without PCR testing did not have Covid-19. It is a fact that asymptomatic Covid-19 infection is possible yet the authors did not account for this in their validation, which is a major methodological limitation.

My suggestion would be to present the results only for those with PCR testing, unless the authors can account for possible asymptomatic Covid-19 infection among those without PCR testing. Otherwise the performance of the CoLab-score would be over-estimated.

Thank you for pointing out the incomplete sentence in the discussion, the sentence has been re-written to read: “Using the CoLab-score in a symptomatic/PCR-tested cohort also results in different diagnostic performance characteristics, as compared to using the score on the full ED cohort (see Supplemental Material 4 Table 1).”

It is correct that there might be some asymptomatic COVID cases in the untested group. This is a limitation but in our opinion the impact on the CoLab-score remains limited and is not unique to our study. Our arguments are as follows:

1. For this reason we have included a large pre-pandemic control group. The majority of all controls are presentations prior to the first case of COVID-19 in the Netherlands (labeled as “Pre-COVID” in Table 1, N = 5890) and are therefore guaranteed to be COVID negative.

2. While it is impossible to know exactly the number of asymptomatic cases that were missed in the untested group during COVID, we have found one other study that screened all asymptomatic patients admitted to the ED during the peak of the pandemic, and found no cases in 1814 PCR tested asymptomatic ED patients [1]. Another study during low-prevalence found two cases in 1246 PCR tested asymptomatic ED patients [2]. Therefore we have reason to assume that the order of magnitude remains small in relation to the large sample size of the study (N = 12879, 279 PCR positive).

3. This limitation is also present in studies where a subgroup of PCR-tested patients is compared. Since it is possible for some patients to test positive only after the second, third or fourth PCR test,

these patients would be labeled as “negative” if only a single PCR test is taken into account. This effect is most likely larger in order of magnitude than missed asymptomatic cases, as in our study 9% of positive patients tested negative in their first PCR test. We have addressed this in our study by including all PCR tests within 1 week after presentation.

To conclude, in our experience, clinical data always contains some unavoidable ‘noise’ in the form of misregistrations, misdiagnoses or patients who were missed. We have tried to mitigate this by including a large pre-pandemic control group and including all PCR tests within 1 week after discharge. We prefer to keep the results of the full ED population in the main article as this is in line with the aim of the score, the development and all validation cohorts. We have however added a line to the discussion stating that we cannot rule-out that any asymptomatic COVID patients could be present in the untested control group.

1. Ravani, Pietro, et al. "COVID-19 screening of asymptomatic patients admitted through emergency departments in Alberta: a prospective quality-improvement study." Canadian Medical Association Open Access Journal 8.4 (2020): E887-E894.

2. Ford, James S., et al. "Testing Asymptomatic Emergency Department Patients for Coronavirus Disease 2019 (COVID-19) in a Low-prevalence Region." Academic emergency medicine: official journal of the Society for Academic Emergency Medicine 27.8 (2020): 771-774.

5. I thank the authors for including the values of TP, TN, FP and FN in the diagnostic accuracy tables. However, these values need to be carefully checked. For example, in Table 3 (page 17 line 26), it is stated that TP = 133 and FN = 0 for CoLab-score ≤0. Sensitive = TP / (TP + FN), so sensitivity = 1.000 when FN = 0. Yet the sensitivity was listed as 0.984. Interesting, the sum of TP, TN, FP and FN varies for all five scores presented in Table 3 when the total population for the derivation dataset should be fixed. In contrast, Table 4 (page 20) presents the same sum of TP, TN, FP and FN (N = 14,080) for the corresponding five-CoLab scores.

We thank the reviewer for pointing out that the TP, TN, FP and FN numbers do not add up. There are two reasons that these numbers do not add up in Table 3 (as opposed to Table 4).

First, there was an error in the script related to the formatting of the table, where only 3 significant digits were shown, this has been addressed.

Second, the diagnostic performance reported in Table 3 is obtained through internal validation via bootstrapping. In bootstrapping, the entire dataset is repeatedly resampled (with replacement) to obtain replicates from the original dataset. However, the sampling is random and each replicate (resampled dataset) will therefore contain a different number of controls and cases. The results are then aggregated in a final step. Although this is explained in the “Internal validation” section we have rounded the TP, TN, FP, FN numbers to one digit in Table 3, so as to make clear these are bootstrapped numbers and we added an explanation to the caption of Table 3.

VERSION 4 – REVIEW

REVIEWER	Zhe Hui Hoo The University of Sheffield, School of Health and Related Research (ScHARR)
REVIEW RETURNED	22-May-2022

GENERAL COMMENTS	I thank the authors for iterating their manuscript taking into account all my previous comments. It does feel that the iterated manuscript over-emphasises 'positive' findings at the expense of the 'negative' findings. For example, whilst it is true that: "… that the discriminative ability is sustained in periods with different dominant variants" (page 25 lines 37-39), the fact remains that there was "a moderately reduced sensitivity of scores ≥ 3 in the third phase as compared to the first phase" (page 19 lines 8-10). The CoLab-score is a screening tool where sensitivity is particularly important. Therefore, if sustained discriminative ability is being highlighted, the potential reduced sensitivity should also be mentioned for balance. On the same note, I am unsure of the evidence that we can expect the Co-Lab score to be sensitive to future SARS-CoV-2 variants on the basis that the Co-Lab score reflects the host response to the virus (page 25 lines 34-37). What is the evidence that different SARS-CoV-2 variants (e.g. Omicron vs the original variant) elicit the same host response when there is differing morbidity and mortality rates from the different variants? Despite the different diagnostic performance characteristics among the symptomatic/PCR-tested cohort versus the pre-pandemic or asymptomatic cohort, strong, formidable, convincing arguments were made to include the full ED population. The reasoning for including the full ED population should be made explicit in the manuscript, in particular the argument that "clinical data always contains some unavoidable noise". It should also be acknowledged in the limitations paragraph that: 1. The overwhelming majority of participants (9193/10417, 88% of the derivation dataset; 5337/7728, 69% of the validation dataset) cannot possibly be tested positive for Covid-19, either because Covid-19 did not exist at the time or no confirmatory test was performed. 2. Therefore the results of this study and the Co-Lab score are most applicable to the pre-pandemic population or the very low risk population whereby any Covid-19 test unlikely to be positive (the papers by Ravani et al and Ford et al should be cited). From a clinical perspective, probably no screening tests for Covid-19 are actually required for the population where the Co-Lab score is most applicable. By extension, the 'Article summary' should include the statement that "the Co-Lab score may be most applicable for the population where no screening tests for Covid-19 are actually required". The conclusion of the 'Abstract' should also include the sentence: "However, the Co-Lab score may be most applicable for the population where no screening tests for Covid-19 are actually required".

VERSION 4 – AUTHOR RESPONSE

Question 1

It does feel that the iterated manuscript over-emphasises 'positive' findings at the expense of the 'negative' findings. For example, whilst it is true that: "… that the discriminative ability is sustained in periods with different dominant variants" (page 25 lines 37-39), the fact remains that there was "a

moderately reduced sensitivity of scores ≥ 3 in the third phase as compared to the first phase" (page 19 lines 8-10).

The CoLab-score is a screening tool where sensitivity is particularly important. Therefore, if sustained discriminative ability is being highlighted, the potential reduced sensitivity should also be mentioned for balance.

On the same note, I am unsure of the evidence that we can expect the Co-Lab score to be sensitive to future SARS-CoV-2 variants on the basis that the Co-Lab score reflects the host response to the virus (page 25 lines 34-37). What is the evidence that different SARS-CoV-2 variants (e.g. Omicron vs the original variant) elicit the same host response when there is differing morbidity and mortality rates from the different variants?

Response 1

To keep the Discussion of the results on page 25 in line with the Results, we have added a sentence to the paragraph in the Discussion stating that "… the sensitivity of scores ≥ 3 is somewhat lower in the third phase." Also we have scrapped the sentence: "Moreover, there is no evidence that the discriminative ability of the CoLab-score is lowered by a change in the ED patient population as a result of widespread vaccination.", since this is not mentioned in the results.

With regards to the second point, at this point in time we do not have enough data from the omicron (B.1.1.529) variant to make a statement with regards to the performance. We agree that a continuous assessment of the CoLab-score is required after implementation, since new variants might emerge and the host's immune response may change. More general, this is recommended for all risk scores that are used in healthcare. We have added this sentence to the discussion: "Continuous assessment of the performance of the CoLab-score is required due to the emergence of new variants and changes in the host's immune response."

Question 2

Despite the different diagnostic performance characteristics among the symptomatic/PCR-tested cohort versus the pre-pandemic or asymptomatic cohort, strong, formidable, convincing arguments were made to include the full ED population. The reasoning for including the full ED population should be made explicit in the manuscript, in particular the argument that "clinical data always contains some unavoidable noise". It should also be acknowledged in the limitations paragraph that:

1. The overwhelming majority of participants (9193/10417, 88% of the derivation dataset; 5337/7728, 69% of the validation dataset) cannot possibly be tested positive for Covid-19, either because Covid-19 did not exist at the time or no confirmatory test was performed.

2. Therefore the results of this study and the Co-Lab score are most applicable to the pre-pandemic population or the very low risk population whereby any Covid-19 test unlikely to be positive (the papers by Ravani et al and Ford et al should be cited). From a clinical perspective, probably no screening tests for Covid-19 are actually required for the population where the Co-Lab score is most applicable.

By extension, the 'Article summary' should include the statement that "the Co-Lab score may be most applicable for the population where no screening tests for Covid-19 are actually required". The conclusion of the 'Abstract' should also include the sentence: "However, the Co-Lab score may be most applicable for the population where no screening tests for Covid-19 are actually required".

Response 2

Thank you for being open to our arguments for using the full ED patient population, we have also added the arguments to the Discussion: “Clinical data always contains some unavoidable ‘noise’ in the form of misregistrations, misdiagnoses or patients who were missed. We have tried to mitigate this by including a large pre-pandemic control group and including all PCR tests within 1 week after discharge.”

1. We have added this sentence to the Discussion: “The vast majority of controls were not tested for COVID-19, because they were either pre-pandemic or asymptomatic/untested patients (89% in the development dataset).”

2. The references to Ford et al. and Ravani et al. have been added to the discussion. As mentioned in the previous response, the CoLab-score applies to the full ED population so we prefer to not cause any confusion by stating that the CoLab-score “is applicable” only to the population where no screening tests for Covid-19 are actually required. We agree that given the current low-prevalence of COVID-19 in the ED, the CoLab-score is mostly used in a population where screening for COVID-19 is not strictly required. However, the prevalence of COVID-19 (in the ED) can vary greatly over time and, by extent, the need for screening for COVID-19 in the ED patient population. Therefore we prefer to not make any strong statements about the “applicability” since a) this might lead to confusion with readers and b) the future with respect to COVID-19 is uncertain.