PEER REVIEW HISTORY

BMJ Open publishes all reviews undertaken for accepted manuscripts. Reviewers are asked to complete a checklist review form (**http://bmjopen.bmj.com/site/about/resources/checklist.pdf**) and are provided with free text boxes to elaborate on their assessment. These free text comments are reproduced below.

ARTICLE DETAILS

TITLE (PROVISIONAL)	Chronic lung lesions in COVID-19 survivors: predictive clinical model
AUTHORS	Carvalho, Carlos; Chate, Rodrigo; Sawamura, Marcio; Garcia, Michelle; Lamas, Celina; Cardenas, Diego; Lima, Daniel Mario; Scudeller, Paula; Salge, João; Nomura, Cesar; Gutierrez, Marco

VERSION 1 – REVIEW

REVIEWER	Patricia Kipnis Kaiser Permanente
REVIEW RETURNED	26-Nov-2021

GENERAL COMMENTS	Please see my questions, edits and comments in the attached version of the file. The paper needs to be edited to ensure the English is correct. The paper describes two models (DL and ML) but it is not clear what cohorts were used for each, sample size, outcomes or predictors. It is not clear when each data point is collected relative to the outcome until the discussion section when they describe there is a limitation in the interval between CXR and CT. The sample size is small. Is there a possibility of expanding the sample size with new data? Maybe using more recent data for validation? There are many other comments in the attached document. The authors claim it is very important to assess the probability of lung problems but they don't describe how the model results would be used in practice.

REVIEWER	Deepak Nag Ayyala Augusta University, Population Health Sciences
REVIEW RETURNED	29-Nov-2021

GENERAL COMMENTS	The manuscript describes a logistic regression(LR) machine learning (ML) model for predicting lung lesions due to COVID-19 infection. The authors have used a new data set, reporting various outcome measures and a LR-based odds ratio calculator. While the overall impact of the model is significant, there are a few concerns that need to be addressed: Major concerns: 1. The demographic, anthropometric and comorbidities variables were summarized in Supplementary Tables S2-S6 which indicate significance differences when comparing between patients with/without pulmonary changes and involvement. Is there a rationale for not including these variables in the predictive model?

Severity of infection or comorbidities could have a significant effect on occurrence of lung lesions and should be considered.

2. In addition to the predictive modeling equation, the authors must report the odds ratios, 95% confidence intervals of the coefficient estimates and the p-values assessing their significance. This is standard output from R/Python and will provide valuable information about the strength of the coefficients in the model.

3. The authors add l2-regularization to the logistic regression model. Did you also perform variable selection? Or consider including a l1-penalty to do feature selection?

4. Did the authors check for potential multicollinearity amongst the variables? There are several variables which are strongly correlated, e.g. BMI and diabetes, smoking and COPD, etc.

5. In the predictive model equation, pCT represents the log-odds ratio or a 0-1 binary value? Please clarify.

6. From Page 9, Line 29: Only 257 patients had all the variables observed to fit the ML models (please correct me if I misread that paragraph). This should be properly stated in the abstract as it is misleading in the “Design, settings and participants” section (which states 749, but not all subjects’ data is used for the ML model).

Minor concerns:

7. Page 7 – Line 48: “Normally distributed continuous variables,… “ You can drop “normally distributed..” and just state “Continuous variables were expressed as means and standard deviations..” Also, the tests used to assess normality and the non-parametric tests used for comparison should be specificied.

8. Page 5 – Line 5: “.. to found Long COVID studies.” Should be “to fund long COVID studies.”

9. Page 4 – Line 54: The numbers don’t add up. 4 million deaths and 233 million recovered out of 221 million confirmed cases? Kindly check.

10. For the CXR evaluation and resolution of disagreements by consensus: are all the images rated by both the radiologists? Also, provide the agreement rate.

11. In the equation, why not express the term “2.16 x mMRC/4” as “0.54 x mMRC”?

12. Also, in the equation: The coefficient of SpO2 should be “0.679” and not “0,679”.

VERSION 1 – AUTHOR RESPONSE

Reviewer 1: The paper needs to be edited to ensure the English is correct.

Answer: The entire manuscript was revised to ensure language and grammar accuracy by “Editage” company, according to the certificate attached to the pdf (Supplemental Material for Editors).

Reviewer 1: The paper describes two models (DL and ML) but it is not clear what cohorts were used for each, sample size, outcomes or predictors. It is not clear when each data point is collected relative to the outcome until the discussion section when they describe there is a limitation in the interval between CXR and CT.

Answer: The cohort used for both models considered the 257 patients with CXR e CT images. Dl model was previously trained using SIIM-RSNA dataset to detect radiographic patterns of COVID-19 pneumonia (Supplemental Methods). The predictors of the DL and ML models were the probabilities that CXR and CT images, respectively, presented findings related to COVID-19 sequelae. The CXR was performed during the general evaluation (Figure 2 and Page 7, Line 46). Patients who meet at least one of the criteria during the general evaluation were enrolled to undergo CT (Page 8, Line17).

Reviewer 1: The sample size is small. Is there a possibility of expanding the sample size with new data? Maybe using more recent data for validation?

Answer: We appreciate the reviewer's suggestion and agree that it would be important to have a larger sample. In our predictive clinical model, were included 257 patients that had complete results for all the five exams used (oximetry, spirometry, mMRC, CRX and CT). We could verify other studies published in important journals that used much less than 257 patients on multivariable analysis[1-4], since it is common to have missing of patients during the follow-up course. However, in our study the sample size does not implicate in the reliability of the prediction. We adopted a five-fold cross-validation strategy for model training and validation, reinsuring the method accuracy. Thus, we could obtain a robust prediction equation, with sensitivity, 0.85±0.08; specificity, 0.70±0.06; F1-score, 0.79±0.06; and AUC, 0.80±0.07.

Reviewer 1: The authors claim it is very important to assess the probability of lung problems but they don't describe how the model results would be used in practice.

Answer: We appreciate the reviewer's suggestion and agree that it would be important to discusses how the model would be used in practice. Thus, we added two paragraphs on the manuscript discussion about it (Page 14; Line 56). We described an initiative that we already have to implement this protocol in Brazil as an example of its practical use and summarized in a flowchart a suggestion for lung lesion case-finding in COVID-19 survivors (Figure 4).

Other comments in the attached document:

-Abstract (Page 3): The abstract was rewritten to address the reviewer suggestions and fill the word limit. According to the reviewer suggestion we added a time frame on "Outcome Measures", and remove the word "chronic" from the "Conclusion".

-Strengths and limitations of this study (Page 4): This section was rewritten to address the Editor comments.

INTRODUCTION

-Page 5 Line 43: The word "have" was changed by "has", according to the reviewer suggestion (Page 5; Line 49).

-Page 6 Line 5: The word "found" was changed by "fund", according to the reviewer suggestion (Page 6; Line 15).

-Reviewer1: Not clear how this DL model is different from the Castiglione and Wang models.

Answer: Although these previous works presented promising results, they were focused on images of patients in acute phase of COVID-19. However, as the pandemic is still ongoing with limited knowledge on long COVID-19 consequences, a more comprehensive protocol for screening COVID-19 patients and assessing the risk of chronic pulmonary changes in recovered patients has not been validated to date.

METHODS

Study design and eligibility

- Reviewer1: Patients were admitted only once, right?

Answer: Yes, we considered only the first patient admission. A statement about it was added in Page 7 Line 7.

- Reviewer1: were the patients still in the hospital?

Answer: The patients were invited to participate in the study six months after admission. At this point, they were already discharged. A statement about it was added in Page 7 Line 15.

General Evaluation

- Reviewer1: at what point were these data collected? 6 months after hospitalization? During hospitalization? The timing of predictors and outcomes is not clear.

Answer: We rewrote the data collection method in order to clarify this issue (Study Design and Eligibility, Page 7, Line 20). The clinical data (comorbidities, cardiorespiratory symptoms, and smoking history), including the length of ICU stay and the need for IMV, were retrospectively collected from the electronic medical records of HCFMUSP.

- Reviewer1: what is the outcome for this model? What were the predictors? What are these previous classifications by radiographs? What are radiographs? Are these the ones in the paragraph above or different ones? This is confusing.

Answer: The predictors of the DL model were the probabilities that CXR images, respectively, presented findings related to COVID-19 sequelae. The radiographs (CXR images) were previously evaluated by two chest radiologists (MVYS and RCC, have 7 and 16 years of experience in thoracic radiology, respectively) working on dedicated workstations (page7, Line 49). Then, these images were used train and validate a DL algorithm to predict the probability that the CXR has findings related to COVID-19 sequelae. (Page 8, Line 3)

- Reviewer1: what is this DL model predicting and when does it predict? How many observations were used?

Answer: The cohort considered the 257 patients with CXR e CT images. Dl model was previously trained using SIIM-RSNA dataset to detect radiographic patterns of COVID-19 pneumonia and the DL model predicted the probability that a CXR image had finds related to COVID-19 sequelae (Supplemental Methods).

Chest CT

- Reviewer1: All a, b, c and d or at least one of these?

Answer: At least one of these. A statement about it was added in Page 8 Line 17.

- Reviewer1: who assigned these and based on what data?

Answer: Categorization of the CT features and score assignment were blindly and independently performed by the same two thoracic radiologists who evaluated the CXR (MVYS and RCC). A statement about it was added in Page 8 Line 42.

Machine learning (ML) model

This section was rewritten to address the reviewer suggestions and clarify this issue (Page 8 Line 55).

Statistical analysis

- Reviewer1: how was the cut off selected for these calculations?

Answer: The performance of the predictive model, expressed in terms of mean ± standard deviation and 95% Confidence Interval (CI), were considered: sensitivity, 0.85±0.08 (95% CI [0.77, 0.94]); specificity, 0.70±0.14 (95% CI [0.55, 0.85]); F1-score, 0.79±0.06 (95% CI [0.73, 0.85]); and AUC, 0.80±0.07(95% CI [0.72, 0.87]). (Supplemental Methods).

RESULTS

-Reviewer1: what is this number?

Answer: The percentage symbol (%) was added to the number (Page 9 Line 57).

-Reviewer1: At what time point?

Answer: The vital signs were collected during the hospitalization period. A statement about it was added in Page 10 Line 10.

-Reviewer1: what are these numbers vs. lower and upper limits of 5.4 and 12.9?

Answer: The “6.7-8.5” are the interquartile range; and the “5.4 and 12.9” are the lower and upper limits of the median interval between hospital admission and consultation. It was rewritten on the manuscript in order to clarify this issue (Page 10 Line 15).

-Reviewer1: how did you deal with these missing values for these patients? Were they not used for the model?

Answer: 348 patients underwent CT. However, the CT score was obtained from 328 patients since 20 patients were excluded because of low CT scan quality or had motion artifacts. It was rewritten on the manuscript in order to clarify this issue (Page 10; Line 59).

-Reviewer1: How many patients were used for this model? Seems like a very low number for a model if you had to drop 91 patients. Need to show a table with the number used in the actual model, separated by those who had the outcome and those who didn't.

Answer: Among the 328 patients with CT score, we included in the predictive clinical model 257 patients. These patients were selected to the predictive clinical model since they had complete results for all the five exams (oximetry, spirometry, mMRC, CRX and CT score). The Supplemental Table S6 compares the groups of patients included (n=257) and excluded (n=91) of the predictive clinical model. We verified few differences between these groups that indicated the patients included in the predictive clinical model (n=257) were more severe ill than the excluded, showing we did not have an important loss of chronic patients' characterization (Supplemental Table S6). We observed the tendency to include the severe ill patients since the follow-up beginning, when we could verify that severe ill patients were the ones that had criteria to undergo the CT (Supplemental Table S2). Also, we agree that would be important to have a larger sample. Nevertheless, we could verify other studies published in important journals that used much less than 257 patients on multivariable analysis[1-4], since it is common to have missing of patients during the follow-up course. However, in our study the sample size does not implicate in the reliability of the prediction. We adopted a five-fold cross-validation strategy for model training and validation, reinsuring the method accuracy. Thus, we obtained a robust prediction equation, with sensitivity, 0.85±0.08; specificity, 0.70±0.06; F1-score, 0.79±0.06; and AUC, 0.80±0.07.

-Reviewer1: This belongs to the methods section.

Answer: This phrase was removed from the results section, according to the reviewer suggestion.

-Reviewer1: what is expressed as means and SD – this sentence is not clear.

Answer: This sentence was rewritten in order to clarify this issue (Page 11; Line 42).

-Reviewer1: what is prxn?

Answer: The variables p_{CXR0} to p_{CXR4} are the probabilities that the CXR image has findings related to sequelae from COVID-19, obtained in each fold (0 to 4) during a 5-folds cross validation. (Supplemental Methods)

DISCUSSION

-Page 13 Line 5: The word "COVID-19" was added to the phrase according to the reviewer suggestion.

-Page 13 Line 7: The word "herein" was changed to "in-person" according to the reviewer suggestion.

-Page 13 Line 15: The expression “among COVID-19 hospitalized patients” were added to the phrase according to the reviewer suggestion.

-Reviewer1 (Page 11 Line 47): Maybe you can use this part for the intro – better written

Answer: The phrase “In this context….” was removed from the discussion, rewritten and added on the introduction of the manuscript (Page 6; Line 7), according to the reviewer suggestion.

-Reviewer1 (Page 12 Line 12): This is much better motivator for the study than introduction. Consider using some of these ideas to start the paper

Answer: The phrase “Thus, we demonstrated…..” was removed from the discussion, rewritten and added on the introduction of the manuscript (Page 6; Line 46), according to the reviewer suggestion.

-Reviewer1: What can be done clinically once you know a patient is at high risk of pulmonary changes? Why is this important?

Answer: We appreciate the reviewer’s suggestion and agree that it would be important to discusses about what can be clinically done and how it can be practically used. Thus, we added two paragraphs on the manuscript discussion about it (Page 14; Line 56). We described an initiative that we already have to implement this protocol in Brazil as an example of its practical use and summarized in a flowchart a suggestion for lung lesion case-finding in COVID-19 survivors (Figure 4).

Reviewer: 2 Dr. Deepak Nag Ayyala, Augusta University

Comments to the Author:

MAJOR CONCERNS:

-Reviewer 2: 1. The demographic, anthropometric and comorbidities variables were summarized in Supplementary Tables S2-S6 which indicate significance differences when comparing between patients with/without pulmonary changes and involvement. Is there a rationale for not including these variables in the predictive model? Severity of infection or comorbidities could have a significant effect on occurrence of lung lesions and should be considered.

Answer: We appreciate the reviewer’s suggestion and agree that variables such as the severity of infection or comorbidities have a significant effect on occurrence of lung lesion. Currently several previous cohort studies about COVID-19 survivors’ sequelae have already showed that old individuals, comorbidities, severity of infection, ICU admission, ICU length of stay, among other demographic and clinical variables have a significant effect on occurrence of lung lesion[2-6]. Corroborating these studies, we also verified herein that patients with pulmonary involvement were older, had more comorbidities, and had a higher ICU need and ICU length of stay than patient with no pulmonary involvement (Supplemental Table S2). Once we aimed to propose a screening clinical protocol to identify COVID-19 survivors that have lung lesions or are at risk to evolve to a chronic lung lesion, our predictive clinical model focused on simple, low cost and accessible exams, that could support the clinical decisions on practice in different countries worldwide, reducing radiation exposure and the conduct of costly imaging examinations. Our results led us to propose a flowchart for lung lesion case-finding in COVID-19 survivors that were included in the manuscript (Figure 4).

-Reviewer 2: 2. In addition to the predictive modelling equation, the authors must report the odds ratios, 95% confidence intervals of the coefficient estimate and the p-values assessing their significance. This is standard output from R/Python and will provide valuable information about the strength of the coefficients in the model.

Answer: Include the estimates of the logistic regression function (Supplemental Table 4)

Table 4. Estimates of the logistic regression function.

Variable	Estimated regression coefficient (β)	Estimated Standard Error	p-value	95% CI for regression coefficient (β)		Estimated odds ratios
FVC^*	-0.3705	0.3210	0.248	-0.9990	0.2580	0.6904
$mMRC^*$	-2.2807	0.3020	<0.001	-2.8730	-1.6890	0.1022
S_pO_2	-0.7450	0.2320	0.001	-1.2010	-0.2890	0.4747
p_{CXR0}	1.1257	0.4150	0.007	0.3120	1.9400	3.0824
p_{CXR1}	1.4960	0.4160	<0.001	0.6810	2.3110	4.4638
p_{CXR2}	1.0761	0.3390	0.002	0.4120	1.7410	2.9332
p_{CXR3}	0.7328	0.3380	0.030	0.0710	1.3950	2.0809
p_{CXR4}	-0.7613	0.4580	0.096	-1.6590	0.1360	0.4671

Forced vital capacity (FVC); modified Medical Research Council dyspnea scale (mMRC); radiographic probabilities (Pcxr0 to Pcxr4).

-Reviewer 2: 3. The authors add l2-regularization to the logistic regression model. Did you also perform variable selection? Or consider including a l1-penalty to do feature selection?

Answer: A Machine Learning (ML) model based on a Logistic Regression (LR) with L2 regularization to prevent overfitting was adopted to detect the presence of COVID-19-related chronic lung lesions. The L1 regularization was not included due to the variable selection by statistical significance that removed irrelevant and correlated attributes.

-Reviewer 2: 4. Did the authors check for potential multicollinearity amongst the variables? There are several variables which are strongly correlated, e.g. BMI and diabetes, smoking and COPD, etc.

Answer: We appreciate the reviewer's suggestion and recognize that some variables such as BMI/Diabetes and Smoking/COPD, ICU admission/ICU length of stay/IMV, potentially have collinearity. However, once the focus of our study was to propose a screening clinical protocol to identify COVID-19 survivors that have lung sequalae, using simple, low cost and accessible exams, instead of demographic and clinical descriptive variables; the multicollinearity evaluation does not seem essential to achieve our objective.

-Reviewer 2: 5. In the predictive model equation, pCT represents the log-odds ratio or a 0-1 binary value? Please clarify.

Answer: p_{CT} is the probability [0,1] of the presence of abnormalities on CT images (Supplemental Results)

-Reviewer 2: 6. From Page 9, Line 29: Only 257 patients had all the variables observed to fit the ML models (please correct me if I misread that paragraph). This should be properly stated in the abstract as it is misleading in the "Design, settings and participants" section (which states 749, but not all subjects' data is used for the ML model).

Answer: We added this information to the abstract, according to the reviewer suggestion.

MINOR CONCERNS:

-Reviewer 2: 7. Page 7 – Line 48: “Normally distributed continuous variables,... “ You can drop “normally distributed..” and just state “Continuous variables were expressed as means and standard deviations..” Also, the tests used to assess normality and the non-parametric tests used for comparison should be specified.

Answer: The alterations suggested were added to the “Statistical analysis” section. We also specified the tests used to assess normality and the non-parametric tests used, according to the reviewer suggestion.

-Reviewer 2: 8. Page 5 – Line 5: “.. to found Long COVID studies.” Should be “to fund long COVID studies.”

Answer: The word “found” was substituted by “fund”, according to the reviewer suggestion.

-Reviewer 2: 9. Page 4 – Line 54: The numbers don’t add up. 4 million deaths and 233 million recovered out of 221 million confirmed cases? Kindly check.

Answer: The numbers were checked and actualized with the December’ data.

-Reviewer 2: 10. For the CXR evaluation and resolution of disagreements by consensus: are all the images rated by both the radiologists? Also, provide the agreement rate.

Answer: The agreement rate was 75% and 2/3 of the CRX images were analysed by both the radiologists. The imagens evaluated by only one radiologist were predominantly normal. This information was added on “General Evaluation” section (Page 7; Line 60).

-Reviewer 2: 11. In the equation, why not express the term “2.16 x mMRC/4” as “0.54 x mMRC”?

Answer: Clinical variables were normalized by dividing the mMRC values by 4 (resulting in values between 0 and 1) and the FVCResting by twice the FVCmin (resulting in a minimum value of 0.257 and a maximum value of 0.847). (Supplemental Dataset and normalization of clinical data).

-Reviewer 2: 12. Also, in the equation: The coefficient of SpO2 should be “0.679” and not “0,679”.

Answer: The coefficient was changed to “-0.7450”, after revision of the results.

VERSION 2 – REVIEW

REVIEWER	Patricia Kipnis Kaiser Permanente
REVIEW RETURNED	04-Jan-2022

GENERAL COMMENTS	The reviewer provided a marked copy with additional comments. Please contact the publisher for full details.

REVIEWER	Deepak Nag Ayyala Augusta University, Population Health Sciences
REVIEW RETURNED	19-Jan-2022

GENERAL COMMENTS	This is a review of the revision submitted by the authors for the manuscript titled “Chronic lung lesions in COVID-19 survivors: predictive clinical model”. The authors have provided detailed responses to my concerns. However, there are a few responses which need further justification. Major concern: 1. Regarding the inclusion of demographic variables along with the variables included in the model, the authors agree that several studies have found demographic and anthropometric measures to be significantly associated with occurrence of lung lesion. If so, there are two concerns a. If these variables are significantly associated, then you will observe a improvement in model accuracy by including them. And since there is no additional cost to observe these variables, the authors should provide a stronger rationale (reduction in prediction accuracy) to justify dropping these variables. b. One of the main standing points of the manuscript is the low-cost examinations used in building the predictive model. Consider a model using ONLY the anthropometric and demographic variables (which can be done with no additional cost). The proposed model can be only be considered as significant if it achieves better predictive power than the model using only anthropometric and demographic variables. The authors need to justify how the examinations-based exam compares to the demographics based model. 2. In response to the definition of pCT, the authors note that it is the probability which takes values between 0 and 1. The equation presented in the manuscript and supplementary materials both take a linear form, which is not guaranteed to lie between 0 and 1. Example, consider all CXR probabilities equal to zero. The pCT will be negative. A logistic regression (LR) model studies linear association between the predictor variables and the log-odds, not the actual probability.

VERSION 2 – AUTHOR RESPONSE

Reviewer: 1 Dr. Patricia Kipnis, Kaiser Permanente

Author: We would like to thank Dr Patricia Kipnis for her comments that certainly contributed to improve the study scientific quality. All your suggestions were included in the manuscript and we addressed all your questions below.

Introduction

- **Page 4 Line 37:** This phrase was changed according to the reviewer suggestion: **“**One study performed chest computed tomography (CT) in 171 patients 4 months after hospital discharge and showed abnormalities in 75.5% of the patients who required invasive mechanical ventilation (IMV)."

- **Page 4 Line 47:** The word "The" was added to the phrase according to the reviewer suggestion.
- **Page 4 Line 38:** The word "works" was substituted by "studies" according to the reviewer suggestion.
- **Page 5 Line 39:** The word "were" was removed according to the reviewer suggestion.
- **Page 5 Line 46:** This phrase was rewritten in order to clarify what the data included in the predictive clinical model were the results of mMRC dyspnoea scale, oximetry, spirometry, and CXR examinations. "Thus, this study aimed to propose a predictive clinical model to detect the presence of radiologic chronic lung lesions due to SARS-CoV-2 infections based on the results of simple and accessible examinations, such as the mMRC dyspnoea scale, oximetry, spirometry, and CXR."

Methods

Reviewer 1: What was the time lag between positive result and admission?

Answer: The RT-PCR-confirmed SARS-CoV-2 infection was obtained at hospital admission day. A statement about it was added to Page 7 Line 7.

- Page 7 Line 46: The expression "It was" was changed by the word "We" according to the reviewer suggestion.

- Page 7 Line 19: The expression "at the face-to-face consultations" was included in the sentence according to the reviewer suggestion.

- Page 7 Line 33: This phrase was rewritten in order to clarify what patients underwent the face-to-face consultation. "Patients who agreed to participate in the study signed an informed consent form and underwent a face-to-face consultation during the collection of anthropometric data and a pulmonary assessment, with an emphasis on respiratory symptoms."

- Page 7 Line 46: This phrase was rewritten in order to clarify that the CXR were performed in the same patients that underwent the face-to-face consultation. "At the same face-to-face consultation described above, the same patients underwent a posteroanterior and lateral CXR according to standard guidelines."

- Page 8 Line 3: This phrase was rewritten in order to clarify that the previous classifications of radiographs by radiologists were described in the paragraph above (Page 8, Paragraph 3). "After the consensus classification performed by the radiologists (described above), the dataset with classified CXR were used to train and validate a DL algorithm developed to predict the probability that the CXR had findings related to sequelae of COVID-19."

- Reviewer 1: What are the predictors of the CXR model? I don't see them described here nor in the supplement.

Answer: The CXR images were previously classified by the radiologists as normal (n=145) or with findings related to COVID-19 (n=112) and randomly distributed in training and validation sets (214 patients) and a test set (n=43). This dataset is described in the Supplement as the InRad dataset. The InRad dataset was used to train and validate a DL algorithm to predict the probability that the CXR has findings related to COVID-19 sequelae, based on the classification by the radiologists.

-Reviewer 1: Machine Learning Model - Yes, L1 removes irrelevant and correlated variables – why is this a problem here? Was there no variable selection then?

Answer: Previously, we did a selection of variables by statistical significance analysis to remove irrelevant and correlated attributes, which makes the use of L1 regularization unnecessary.

Results

- Page 10 Line 16: The expression "of the median" was added to the sentence according to the reviewer suggestion.

Reviewer 1: What is the difference between pcxr and prx described in the supplement? How are these calculated? Is there also a formula for these?

Answer: There is no difference between p_{CXR} and prx, they represent the same variable. We corrected the notation in the Supplement, including the Figures.

Reviewer 1: Is pCT the probability or the logit?

Answer: pCT is the probability of the presence of abnormalities on CT images. The function of the machine learning model was rewritten to include the sigmoid function that restrict pCT between 0 and 1.

Reviewer 1: Can add some chronologic reference to this figure? When were the input variables collected (upon discharge or 6 months later and when was the outcome determined?

Answer: The Figure 1 and its' legend was altered to address the reviewer suggestion.

Reviewer: 2 Dr. Deepak Nag Ayyala, Augusta University

Author: We would like to thank Dr Deepak Nag Ayyala comments' that directly contributed to a better clarification of the question of our study. We included your suggestions in the manuscript and addressed your questions below.

Reviewer 2: Regarding the inclusion of demographic variables along with the variables included in the model, the authors agree that several studies have found demographic and anthropometric measures to be significantly associated with occurrence of lung lesion. If so, there are two concerns:

a. If these variables are significantly associated, then you will observe an improvement in model accuracy by including them. And since there is no additional cost to observe these variables, the authors should provide a stronger rationale (reduction in prediction accuracy) to justify dropping these variables.

b**.** One of the main standing points of the manuscript is the low-cost examinations used in building the predictive model. Consider a model using ONLY the anthropometric and demographic variables (which can be done with no additional cost). The proposed model can be only be considered as significant if it achieves better predictive power than the model using only anthropometric and demographic variables. The authors need to justify how the examinations-based exam compares to the demographics-based model.

Answer: We appreciate the reviewer's suggestion and agree that it would be important to provide results regarding inclusion of demographic and anthropometric variables on the prediction model. Thus, we did experiments using six different combinations of variables (age, gender, BMI, SpO2, mMRC score, FVC and CXR) in the predictive model and the performance of each combination is reported in the table below. These results were also included on the supplemental material.

The model performance with the inclusion of demographic or anthropometric variables does not result in significant improvement. According to our experiments, the combination of SpO2, mMRC score, FVC and CXR presents the best performance as reported in the article. This result reinforces the aim of this study that was to propose a screening clinical protocol to identify COVID-19 survivors that have lung lesions or are at risk to evolve to a chronic lung lesion, focusing on simple, low cost and accessible exams, that could support the clinical decisions on practice in different countries worldwide, reducing radiation exposure and the conduct of costly imaging examinations.

Performance of the predictive model using six combinations of variables (N=257).				
Groups of variables	**Sensitivity**	**Specificity**	**F1-score**	**AUC**
1 **Age, Gender, and BMI**	0.87±0.09	0.40±0.27	0.71±0.03	0.64±0.09
2 **SpO_2, mMRC score, and FVC**	0.87±0.16	0.42±0.33	0.71±0.03	0.68±0.10
3 **Age, Gender, BMI, SpO2, mMRC score, and FVC**	0.95±0.05	0.37±0.30	0.75±0.06	0.71±0.10
4 **CXR**	0.88±0.05	0.52±0.14	0.75±0.04	0.78±0.05
5 **Age, Gender, BMI, SpO_2, mMRC score, FVC, and CXR**	0.87±0.08	0.65±0.16	0.79±0.06	0.79±0.06
6 **SpO_2, mMRC score, FVC, and CXR**	0.85±0.08	0.70±0.14	0.79±0.06	0.80±0.07

Values are presented as means ± standard deviations after five-fold cross validation for each test fold. BMI, body mass index; CXR, chest X-Ray; FVC, forced vital capacity; mMRC, modified Medical Research Council dyspnoea scale.

Reviewer 2: In response to the definition of pCT, the authors note that it is the probability which takes values between 0 and 1. The equation presented in the manuscript and supplementary materials both take a linear form, which is not guaranteed to lie between 0 and 1. Example, consider all CXR probabilities equal to zero. The pCT will be negative. A logistic regression (LR) model studies linear association between the predictor variables and the log-odds, not the actual probability.

Answer: pCT is the probability of the presence of abnormalities on CT images. The function of the machine learning model was rewritten to include the sigmoid function that restrict pCT between 0 and 1.

VERSION 3 – REVIEW

REVIEWER	Patricia Kipnis Kaiser Permanente
REVIEW RETURNED	13-Feb-2022

GENERAL COMMENTS	The revision reads much better. However, there are still some issues that need to be addressed. The response to reviewers page was difficult to follow because the line numbers did not correspons to the original version or the revised version. Also, it wasn't clear to me that the redlined version followed the revised version in the same file. This made it a little more challenging it for me to review the revised version. The abstract and main points pages need clarification. My comments are in the attached document. The reviewer provided a marked copy with additional comments. Please contact the publisher for full details.

REVIEWER	Deepak Nag Ayyala Augusta University, Population Health Sciences
REVIEW RETURNED	24-Feb-2022

GENERAL COMMENTS	Thanks for addressing all the comments/concerns. I have no further comments.

VERSION 3 – AUTHOR RESPONSE

Reviewer: 1 Dr. Patricia Kipnis, Kaiser Permanente

Author: We would like to thank Dr Patricia Kipnis comments' that directly contributed to a better clarification of the question of our study. We included all your suggestions in the manuscript and addressed your questions below.

- Reviewer 1: The revision reads much better. However, there are still some issues that need to be addressed. The response to reviewer's page was difficult to follow because the line numbers did not correspond to the original version or the revised version. Also, it wasn't clear to me that the

redlined version followed the revised version in the same file. This made it a little more challenging it for me to review the revised version.
Answer: In order to clarify the page and line numbers, in our answers below we referred to the page number of the marked copy of the manuscript in the BMJ' PDF, displayed in the top of the page. We hope that it facilitates your revision process.

Abstract
Design, settings and participants:
- Reviewer 1: This description already includes results. Describe the study design before data collection. Put the numbers of patients in the results section.
Answer: This section was reformulated according to the right headings suggested by the BMJOpen guidelines and the Reviewer 1 suggestions. The study design is described in the "Design" heading and it describes how the study was performed. (Manuscript marked copy - Page 24 Line 9) Also, we included the "Participants" heading, according to the the BMJOpen guidelines, where we must describe the number of eligible patients that followed the inclusion criteria. (Manuscript marked copy - Page 24 Line 22)

- Reviewer 1: It is not customary to start a sentence with a numeral.
Answer: This phrase was removed from this section, according to the Reviewer 1 suggestion above. This information was added on the results heading (Manuscript marked copy - Page 24 Line 33).

Outcome Measures:
- Reviewer 1: The metrics described here are predictors and outcomes but the title is outcome measures. Please describe which are predictors and which are outcomes and use the right headings.
Answer: This section was reformulated according to the right headings suggested by the BMJOpen guidelines and the Reviewer 1 suggestions. (Manuscript marked copy - Page 24 Line 25)
Results
- Reviewer 1: Where do the 257 patients described above fit into this section?
Answer: We added a phrase in the results heading in order to clarify where the 257 fit into this section, according to the Reviewer 1 suggestion. (Manuscript marked copy - Page 24 Line 33)

Strengths and limitations of this study
- Phrase 1. Reviewer 1: "a broad assessment of what?" and "This is singular and examinations is plural"
Answer: This phrase was reformulated according to the Reviewer 1 suggestions. (Manuscript marked copy - Page 25 Line 10)

- Phrase 2. Reviewer 1: "How did you show this was the case?"
Answer: Considering the 257 observations, the performance of the model evaluated by AUC is 0.80±0.07 (Table 2) (Manuscript marked copy - Page 33, Table 2). Hosmer and Lemeshow (2000) suggest AUC of 0.70 to 0.80 are 'acceptable', 0.80 to 0.90 'excellent' and 0.9 or above 'outstanding'.

Reference: Hosmer DW and Lemeshow SL (2000). Applied Logistic Regression. 2nd Edition. Wiley:New York. In CBSU library. A third edition is due to be published in 2013.

- Phrase 4. Reviewer 1: "and?? Were less likely to develop lung lesions??"
Answer: This phrase was reformulated according to the Reviewer 1 suggestions. (Manuscript marked copy - Page 25 Line 21)

Introduction

Reviewer 1: Substitute the word “propose” by the word “develop”. (Manuscript marked copy - Page 27 Line 43)
Answer: The word “propose” was changed to “develop” according to the Reviewer 1 suggestion. (Manuscript marked copy - Page 27 Line 43)

Methods. Machine learning (ML) model

- Reviewer 1: “Not clear why L1 was not appropriate. Is it because you had few predictors and wanted to include all of them? If predictors are irrelevant, why do you want them in your model?”

Answer: The key difference between L1 and L2 techniques is that L1 shrinks the less important feature’s coefficient to zero thus, removing some feature altogether. This works well for feature selection in case we have a huge number of features. In this work we have only eight features (SpO2, mMRC score, FVC, pCXR0, pCXR1, pCXR2, pCXR3, and pCXR4) in the predictive model and we have included all of them. (Supplemental Material - Page 51, Figure 1).

- Reviewer 1: “Did you evaluate the model on the same training set? Did you run cross validation? Where are you showing that 257 observations are a robust sample data to build this type of model? What predictors were included in the model?”

Answer: We evaluated the Machine Learning (ML) Model using a five-fold cross-validation strategy. This is described in the performance metrics of the Machine Learning (ML) model (Manuscript marked copy - Page 30 Line 24), in the legends of Table 2 (Manuscript marked copy - Page 33, Table 2) and Table 3 (Supplemental Material - Page 52, Table3). We also included in this revision a sentence to emphasize the use of cross-validation (Supplemental Material - Page 52 Line 3).
The ML model predicts the probability of the presence of abnormalities on CT images based only on the predictors: SpO2, mMRC score, FVC, pCXR0, pCXR1, pCXR2, pCXR3, and pCXR4 (Supplemental Material - Page 51, Figure 1). Considering the 257 observations, the performance of the model evaluated by AUC is 0.80±0.07 (Table 2) (Manuscript marked copy - Page 33, Table 2). Hosmer and Lemeshow (2000) suggest AUC of 0.70 to 0.80 are 'acceptable', 0.80 to 0.90 'excellent' and 0.9 or above 'outstanding'.

Reference: Hosmer DW and Lemeshow SL (2000). Applied Logistic Regression. 2nd Edition. Wiley:New York. In CBSU library. A third edition is due to be published in 2013.

- Reviewer 1: “The results section describes a process of fitting 3 different sets of predictors which is not described in the methods section.”

Answer: In the Supplemental Material (Supplemental Material - Page 49 Line 42) we described the three combinations of input variables (predictors). We also discuss the inclusion of demographic and anthropometric variables in the model (Supplemental Material - Page 53, Table 5).

Methods. Statistical analysis

- Reviewer 1: “What cut offs were used to evaluate sensitivity and specificity?”

Answer: For the DL model, the cut offs for sensitivity and specificity were 0.82 and 0.77, respectively as presented in Table 1 (Supplemental Material - Page 50, Table 1) which resulted in an AUC of 0.89. For the ML Model, the cut offs for sensitivity and specificity were 0.85 and 0.70, respectively as presented in Table 3 (Supplemental Material - Page 52, Table 3), which resulted in an AUC 0.80.

Results

- Reviewer 1: “What are these lower and upper limits? 95% confidence interval or min/max?”

Answer: That is the min/max. This phrase was rewritten in order to clarify it. (Manuscript marked copy - Page 31 Line 34)

- Reviewer 1: “considered? Do you mean observed? Or with the following performance measures….”

Answer: The word “considered” was substituted by the word “observed” according to the Reviewer 1 suggestion. (Manuscript marked copy - Page 33 Line 7)

Discussion

- Reviewer 1: “you only mention cross-validation in the development of the probability that the CXR model not the CT ML model”

Answer: We evaluated the Machine Learning (ML) Model using a five-fold cross-validation strategy. This is described in the performance metrics of the Machine Learning (ML) model (Manuscript marked copy - Page 30 Line 24), in the legends of Table 2 (Manuscript marked copy - Page 33, Table 2) and in the legend of Table 3 (Supplemental Material - Page 52, Table 3). We also included in this revision a sentence to emphasize the use of cross-validation (Supplemental Material - Page 52 Line 3).