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)	Development of a healthcare system COVID hotspotting score in California: an observational study with prospective validation
AUTHORS	Liu, Vincent; Thai, Khanh; Galin, Jessica; Gerstley, Lawrence; Myers, Laura; Parodi, Stephen; Chen, Yi-fen; Goler, Nancy; Escobar, Gabriel; Kipnis, Patricia

VERSION 1 – REVIEW

REVIEWER	Mody, Aaloke Washington University in St Louis, Infectious Diseases
REVIEW RETURNED	26-Jan-2021

GENERAL COMMENTS	Overall comments: This is an interesting study that discusses the development and evaluation of a COVID hotspot score using EHR data from Kaiser. This is definitely a topic of great interest for early identification of upcoming surges. Based on the results, the score the author's develop seems to perform quite well. I have a few main comments that I could help improve the paper. First, I think the authors should attempt to conceptually describe the methods and overall purpose of what is being done at each step. That way a reader who is not familiar with the specifics of the methods and more easily grasp the principles of is being done. I think the authors' should also specify some of decision points more clearly. Second, although the correlation between lagged score and hospital census is very impressive, some more formal analysis on this front akin to more typical evaluation of prediction models would strengthen the paper. Lastly, I think more discussion on the implementation and use of this score would be very helpful. Overall, a good paper with good results. Specific comments: Intro • I would suggest included a brief clarification at the end of discussion discussing that study design: specifically you used early data and then validated in future data. When I first was reading through, I thought this manuscript only included data from the first wave (and was less enthusiastic) until I got to the prospective section in the methods and realized it did indeed include a much more robust design. Methods:

	• I am not familiar with change point analysis. I think a few sentences conceptually describing what is being done to generate the score would be helpful. It seems to me the CPA identifies statistically changes from a baseline and the authors are generating a score based on these changes. A description that is written where a more casual reader can understand what is being done would be helpful. • It seems like only certain indicators ended up making into the score. How were they selected? • How did you select what scores to give the changes also? Results • The correlation between the CHOTS score and relative COVID census are impressive. I am wondering if relative census is the most useful metric though. Wouldn't absolute numbers be more useful to use. • This gets me to I think my main suggestion. I find myself wanting a prediction model and its associated validation metrics. What is presented is fantastic and I think very informative. But as a reader (and I think also user of such an algorithm would want) is for these indicators to be translated into a prediction model that predicted hospital census. Or some other outcome that might trigger different actions. • For example, if having to choose certain lags and indicators, how would this score perform across different facilities and regions where the epidemic dynamics. The correlation tables hint at this but more formal analyses would be very welcome additions. • On the figures, it would be useful to have a marker to indicate which proportions contain the data that were used in developing the score/model (i.e., before May) and which parts were prospective evaluation. Discussion • Was this score implemented at Kaiser? A discussion of its use and early insights it delivered would be fantastic. I think it is clear that such predictive algorithms are in need but I think the question how to use them effectively is also important. A discussion of some of Kaiser's experience could be illuminating. • I would also recommend a section discussing in general how health system would want to implement and utilize this scoring system would be helpful. Should it trigger particular interventions? Only planning at the hospital level? Anything at the community level? Public health messaging about high scores, etc.? This is important. • This score was feasible with Kaiser's EHR. But it also seems that many metrics are likely widely available. How would this score perform say with just COVID tests or other metrics that would be expected to be more widely available. This could potentially be a supplementary analysis (it is mentioned in the discussion). • The authors' discuss not using a machine learning algorithm and that they chose not to develop a full prediction

	model. I think the rational and discussion should be fleshed out more and maybe brought in earlier in the manuscript. If there are aspects of more traditional predictive models and their validation metrics that could be brought into the manuscript, the authors can revisit that per some of the above comments.

REVIEWER	Kia, Arash Icahn School of Medicine at Mount Sinai, Population Health Science and Policy
REVIEW RETURNED	04-Feb-2021

GENERAL COMMENTS	this is great paper. just a few points: 1- Since evaluation is based on including different care facilities it is better to have a table to compare the patient population structure of these facilities 2- different facilities follow different practice styles and they show different resource utilization patterns. I think it is necessary to see how come the different utilization patterns did not have any impact on the performance of CHOTS score

REVIEWER	Jiang, Binyan The Hong Kong Polytechnic University, Applied Mathematics
REVIEW RETURNED	10-Mar-2021

GENERAL COMMENTS	The authors proposed a COVID hotspotting score for identifying emerging COVID activity using data from health systems. It is not clear to me whether similar approaches have been well used/accepted for identifying emerging activities of other diseases such as influenza, and how much real impact would this research bring to our life. I am leaving this for the experts to judge. I am a statistician and my following comments will be mainly focusing on the data analysis part. Overall, the idea of using data from health systems for evaluating/identifying emerging COVID activities is very interesting, and most of the statistical analysis were conducted appropriately. My main concerns are: 1. The scores seem to be rather ad hoc and I am having some doubts on their reliability. 1. 1 On one hand, it is not clear to me why the scores (0, 0.2, 0.4, 0.6, 1 etc.) provided in Appendix table 1 were set in this way, and the Major indicators and minor indicators seem to be selected in a empiricism manner other than data driven. Further, the weights for major and minor indicators are rather ad hoc as well. My main concern here is that any change to these scores/weights might change the results. 1.2 The other concern I am having is that the scores in Appendix Table 1 are based on statistical significance tests. Whether the results would be significant or not is related to the sample size. Sometimes the signal is too weak to be detected when the sample size is too small. If we use the proposed scores in other states/regions, where the population could be potentially very different, the test results could be very differently distributed (for example, hypothetically, if the sample size is too small, one could end up having most of the variables having NS results). I am hence having some doubts on whether it is ok to set the scores based on the significance of the tests.

	Overall, I feel one can develop a relatively more data-driven scoring system. However, as i have pointed out, I am a layman to this area and it is not clear to me whether such kind of setting/approach would make sense to other experts in this area. I believe It would certainly be helpful if the authors can provide further discussions/explanations for their proposed scoring system.

REVIEWER	Faisal, Muhammad University of Bradford
REVIEW RETURNED	16-Mar-2021

GENERAL COMMENTS	Thank you for an opportunity to review this interesting and timely study. My major concern is about generalisability of CHOTS score due to the lack of validation and spurious correlations issues. One can use a couple of hospitals data as validation data to enhanced the generalisability. How the spurious correlation issue has been addressed which raised here Dean, R.T., Dunsmuir, W.T.M. Dangers and uses of cross-correlation in analyzing time series in perception, performance, movement, and neuroscience: The importance of constructing transfer function autoregressive models. Behav Res 48, 783–802 (2016). https://doi.org/10.3758/s13428-015-0611-2 Furthermore, I could not find enough details about statistical model and its performance for developing CHOTS score.

VERSION 1 – AUTHOR RESPONSE

REVIEWER 1 COMMENTS

R1. Overall comments: This is an interesting study that discusses the development and evaluation of a COVID hotspot score using EHR data from Kaiser. This is definitely a topic of great interest for early identification of upcoming surges. Based on the results, the score the author's develop seems to perform quite well. I have a few main comments that I could help improve the paper.

AR1. We thank the Reviewer their supportive comments.

R2. First, I think the authors should attempt to conceptually describe the methods and overall purpose of what is being done at each step. That way a reader who is not familiar with the specifics of the methods and more easily grasp the principles of is being done. I think the authors' should also specify some of decision points more clearly.

AR2. Thank you for this comment. As described above and in the subsequent responses, we have significantly revised and improved the manuscript to give the reader a better understanding of our overall approach.

R3. Second, although the correlation between lagged score and hospital census is very impressive, some more formal analysis on this front akin to more typical evaluation of prediction models would strengthen the paper.

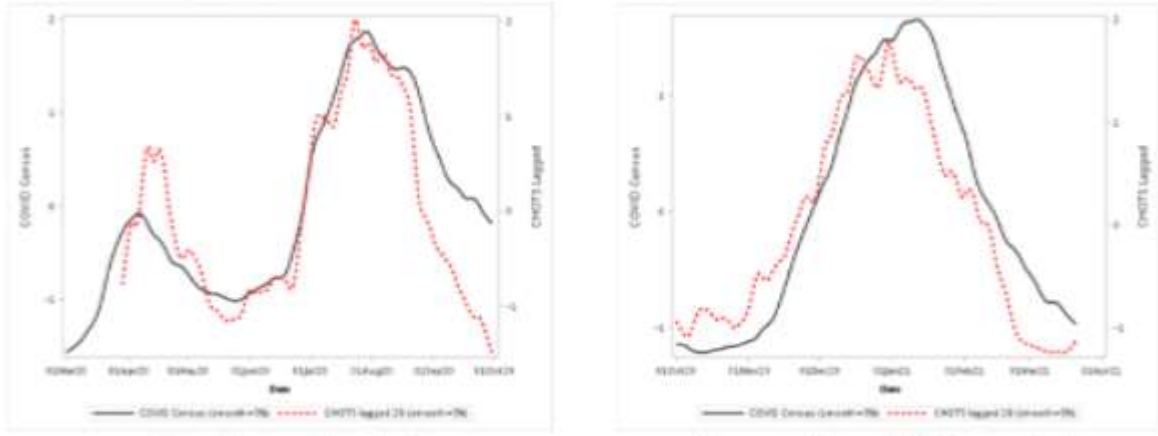
AR3. We thank the Reviewer for these important comments about evaluation that would strengthen the paper that takes into account some approaches used in predictive modeling. A critical concept in assessing performance is prospective validation of a tool once the scoring parameters have been established.

In this case, as our prior data from our initial manuscript only went through September 30, 2020, we have chosen to include data to prospectively validate the CHOTS hotspotting score on a temporally independent and very contemporary sample starting from October 1, 2020 through March 21, 2021. As shown in the Figure and Table below, the CHOTS continued to show excellent correlation with forthcoming COVID19 hospital census during this new period.

The overall maximum correlation with COVID19 hospital census was at a lag of 28 and 25 days (correlation: 0.73) with the maximum correlation by facility ranging from 0.52 to 0.77 with a maximum reached at lags of 21 days (4 facilities), 28 days (9 facilities), 35 days (6 facilities), and 42 days (1 facility).

We have revised the manuscript to reflect these updated prospectively validated correlation metrics, which we think significantly improves the strength of our findings.

Figure. Standardized values of the 28-day lagged COVID19 Hotspotting Score (CHOTS; red dotted line and right y-axis) and COVID19 hospital census (black line and left x-axis) at the regional level in the Development (April 1, 2020 to September 30, 2020; left) and Validation (October 1, 2020 to March 21, 2021; right) periods. Data are smoothed using a LOESS model. The smoothing parameter is the proportion of data in the local neighborhood: a value near 0 results in a curve that nearly interpolates the data whereas a value near 1 is nearly a straight line.



Development period
April – September 2020

Prospective validation period
October 2020 – March 2021

TABLE	**Correlation between lagged COVID Hotspotting Score and forthcoming COVID19-specific hospital census during the Validation period (October 2020 to March 21, 2021), stratified by lag from 7 days to 42 days**					
Location	**7 days**	**14 days**	**21 days**	**28 days**	**35 days**	**42 days**
KPNC Region	**0.38**	**0.54**	**0.66**	**0.73**	**0.73**	**0.66**

Facility A	0.35	0.47	0.58	0.66	0.64	0.59
Facility B	0.38	0.54	0.63	0.66	0.63	0.56
Facility C	0.60	0.72	0.77	0.75	0.68	0.53
Facility D	0.17	0.30	0.43	0.55	0.59	0.59
Facility E	0.48	0.62	0.69	0.67	0.63	0.54
Facility F	0.07	0.27	0.43	0.54	0.62	0.65
Facility G	0.40	0.57	0.67	0.71	0.69	0.60
Facility H	0.62	0.70	0.71	0.63	0.55	0.45
Facility I	0.32	0.51	0.62	0.67	0.67	0.61
Facility J	0.35	0.45	0.58	0.67	0.68	0.61
Facility K	0.04	0.26	0.45	0.59	0.62	0.62
Facility L	0.54	0.67	0.74	0.75	0.69	0.55
Facility M	0.33	0.33	0.42	0.52	0.50	0.40
Facility N	0.40	0.53	0.55	0.52	0.39	0.39
Facility O	0.39	0.52	0.64	0.71	0.68	0.61
Facility P	0.47	0.63	0.72	0.73	0.71	0.61
Facility Q	0.44	0.60	0.72	0.74	0.71	0.62
Facility R	0.18	0.34	0.47	0.58	0.61	0.60
Facility S	0.21	0.36	0.51	0.61	0.65	0.63
Facility T	0.16	0.33	0.48	0.58	0.68	0.67

R4. Lastly, I think more discussion on the implementation and use of this score would be very helpful. Overall, a good paper with good results.

AR4. We have revised the Discussion to discuss more of the implementation and implications for real-time operations in our health system as follows:

> "Implications for clinicians and health system leaders
>
> The CHOTS score has been in use in our health system since June 2020 and is updated on a daily basis in a variety of dashboards that are accessible to our health system and hospital leadership. After KPNC's COVID19 census began to ebb following wave 2, the alarming increase in the trajectory of the CHOTS score before wave 3 was used to inform the reopening of daily Regional COVID19 Command Center operations. The CHOTS tool, along with other predictive models, has also been used to inform decisions about health system staffing and resource allocation as well as clinical care, based on the expected rise,

stabilization, or fall of COVID19 activity across different subregions and individual medical centers. Finally, the CHOTS tools has also informed decisions about the urgency of health system communications with members, communities, and public health agencies, particularly during periods when the easing of social distancing behaviors occurred concurrently with the emergence of increasing COVID19 hotspotting signals."

R5. Intro: I would suggest included a brief clarification at the end of discussion discussing that study design: specifically you used early data and then validated in future data. When I first was reading through, I thought this manuscript only included data from the first wave (and was less enthusiastic) until I got to the prospective section in the methods and realized it did indeed include a much more robust design.

AR5. As suggested by the Reviewer, we have revised the final paragraph of the Discussion to allude to the 2 sets of temporally independent prospective validation data we've used in this study.

> "Numerous efforts are underway to evaluate promising approaches to identify and predict COVID19 hotspots using aggregated social media, viral testing patterns, mobility, biometric, and symptoms data.[6,12-20] In this study, we investigated the development of a composite index to identify and predict emerging hospital COVID19-related activity using passively collected daily electronic health record (EHR) data from a large, regional integrated healthcare system. We further quantified the potential lead time that such data – aggregated as the CHOTS score – might offer to health systems and communities using observational data and 2 periods of prospective validation data across 3 COVID19 waves in Northern California."

R6. Methods: I am not familiar with change point analysis. I think a few sentences conceptually describing what is being done to generate the score would be helpful. It seems to me the CPA identifies statistically changes from a baseline and the authors are generating a score based on these changes. A description that is written where a more casual reader can understand what is being done would be helpful.

AR6. We have added additional clarification about change point analysis to the Methods.

> "CPA algorithms are used to detect changes in the mean values of time-series data and to identify periods marked by a new mean value. They have been used to assess changes in seasonal influenza data."

R7. Methods: It seems like only certain indicators ended up making into the score. How were they selected?

AR7. Thanks for the clarification. Because of the incredible urgency to instantiate a real-time COVID hotspotting score for our health system by June 2020, we did not include all potential leading indicators and we used a series of analyses or prior health system data, heuristics, and clinical judgement to finalize our 10 predictors and the score components. While we would have liked to have the luxury of performing additional statistical analyses to refine a score, COVID did not afford us that opportunity. Fortunately, as we describe in this manuscript, the CHOTS score has proven remarkably robust with 2 prospective waves of COVID in Northern California. We've revised the Methods as follows:

> "Because of the urgent need to establish a hotspotting tool in our health system to prepare for forthcoming COVID19 waves by June 2020, we used visual inspection and association analysis of potential indicators with prior seasonal influenza patterns as well as clinical judgment in our research team to identify the final leading indicators and relevant score components. Final core rules are described in the text below and in **Appendix Tables 1 and 2** significance testing code is available in the Supplement."

R8. Methods: How did you select what scores to give the changes also?

AR8. See answer to R7; while we would have loved the luxury of time to assess scoring regimens, we ultimately used our clinical judgment in selecting the scoring components.

R9. Results: The correlation between the CHOTS score and relative COVID census are impressive. I am wondering if relative census is the most useful metric though. Wouldn't absolute numbers be more useful to use.

AR9. When we set out to develop the CHOTS score, we focused on a metric of correlation with relative census because of historical data indicating that different waves of pandemic disease would result in different absolute values of cases, hospitalizations, and deaths. For example, the 1918 flu pandemic and then more recent H1N1 experience showed how differing waves of infection resulted in tremendous heterogeneity in absolute impact. We focused on making the CHOTS an early indicator of new health system activity, rather than a tool to predict the precise census values. In the process, we developed and used other prediction tools to try to pinpoint short-term census (e.g., SEIR or fitted curve models) which have proven better at predicting actual census. The comment raises a key point and we have revised the manuscript extensively to address this.

> Methods: "Because we designed the CHOTS score to focus on detecting emerging activity rather than on attempting to predict absolute hospital census, our health system also implemented more traditional infectious disease epidemiology and fitted curve models to predict shorter-term absolute hospital census estimates." In this updated text, we cite our work published in BMJ (PIMD: 32444358) that discusses the creation of traditional SEIR models built from Kaiser Permanente data to model absolute hospital census predictions.
>
> Discussion: "In addition, these tools often focus on trying to predict the precise COVID19 hospital census, which has been shown to be highly variable across serial COVID19 waves as well as across waves of other pandemic disease including 1918 influenza and 2009 H1N1 pandemics." And, "Because the CHOTS is designed to inform medium-term decisions, we also chose not to build a model to generate precise predictions of hospital census and instead use other curve-fitting and epidemiologic models to predict absolute hospital census numbers."

R10. Results: This gets me to I think my main suggestion. I find myself wanting a prediction model and its associated validation metrics. What is presented is fantastic and I think very informative. But as a reader (and I think also user of such an algorithm would want) is for these indicators to be translated into a prediction model that predicted hospital census. Or some other outcome that might trigger different actions. For example, if having to choose certain lags and indicators, how would this score perform across different facilities and regions where the epidemic dynamics. The correlation tables hint at this but more formal analyses would be very welcome additions.

AR10. We thank the Reviewer for this comment but point to the other answers and revisions we've made that address this point. While it would be incredibly useful to have a single tool that precisely and perfectly predicts hospital census while also identifying the earliest possible COVID19 activity, we did not attempt to do this with CHOTS, nor did we think it was really possible. As we've stated above, we have strong evidence from COVID19, as well as other prior pandemic disease, that the absolute impact (in terms of quantities like census, visits, infections) varies significantly in successive waves of activity and is likely attributable to a variety of factors that are poorly predictable. For example, in COVID19, this heterogeneity is likely attributable to the social distancing behaviors of individuals and communities, the policies of local and national governments, as well as the changing virulence and biology of the virus itself. Each of these factors has varied across successive waves of the disease which stymie even the best models seeking to precisely predict census. To explain our rationale, we have greatly expanded our Discussion in the following ways:

"This variability is attributable to many factors including the social distancing behaviors of individuals and communities, the policies enacted by local and national governments, the dynamic biology of the virus itself, and other factors that have yet to be elucidated."

"We also did not attempt to develop a predictive model designed to precisely estimate absolute hospital census, instead focusing on a hotspotting approach designed to give the earliest signals of incipient viral activity. We have built other models for precise census prediction and have found that their accuracy is greatest over very short intervals like 1 to 2 weeks."

R11. Figures: On the figures, it would be useful to have a marker to indicate which proportions contain the data that were used in developing the score/model (i.e., before May) and which parts were prospective evaluation.

AR11. We have revised the Figures to specifically denote the period from October 1, 2020 through March 21, 2021 represents temporally distinct prospective validation.

R12. Discussion: Was this score implemented at Kaiser? A discussion of its use and early insights it delivered would be fantastic. I think it is clear that such predictive algorithms are in need but I think the question how to use them effectively is also important. A discussion of some of Kaiser's experience could be illuminating.

AR12. The score was implemented in June 2020 and has been continual use for nearly a year. We have found that it has been extremely useful in our health systems operations. The trajectory of the CHOTS score has been used to restart the standing meetings of our regional COVID19 command centers (after they were stopped following the end of wave 2) as well as in ongoing meetings to help our leadership make decisions about staffing and clinical care. We've expanded this Discussion as follows:

"The CHOTS score has been in use in our health system since June 2020 and is updated on a daily basis in a variety of dashboards that are accessible to our health system and hospital leadership. After KPNC's COVID19 census began to ebb following wave 2, the alarming increase in the trajectory of the CHOTS score before wave 3, was used to inform the reopening of daily Regional COVID19 Command Center operations. The tool, along with other predictive models, has also been used to inform decisions about health system staffing and resource allocation as well as clinical care, based on the expected rise, stabilization, or fall of COVID19 activity across different subregions and individual medical centers."

R13. Discussion: I would also recommend a section discussing in general how health system would want to implement and utilize this scoring system would be helpful. Should it trigger particular interventions? Only planning at the hospital level? Anything at the community level? Public health messaging about high scores, etc.? This is important.

AR13: We have added to the Discussion:

"Finally, the CHOTS tools has also informed decisions about the urgency of health system communications with members, communities, and public health agencies, particularly during periods when the easing of social distancing behaviors occurred concurrently with the emergence of increasing COVID19 hotspotting signals."

R14. Discussion: This score was feasible with Kaiser's EHR. But it also seems that many metrics are likely widely available. How would this score perform say with just COVID tests or other metrics that would be expected to be more widely available. This could potentially be a supplementary analysis (it is mentioned in the discussion).

AR14: Thank you for this important comment. While we designed this score to take advantage of the integrated information systems we have in Kaiser Permanente, it raises an important point about the potential generalizability when implemented in systems without the same type of information. We have added an additional set of analyses that discuss a 'reduced' form of the CHOTS score, which removes data related to calls and e-mail messages. Interestingly, we found that the correlation performance remained robust as shown in the Table below. We have also integrated this into both the Methods, Results and Discussion.

Location	Maximum Correlation Development Period	
	CHOTS	'Reduced' CHOTS
KPNC Region	0.79	0.74
Facility A	0.78	0.71
Facility B	0.59	0.53
Facility C	0.56	0.48
Facility D	0.77	0.67
Facility E	0.65	0.52
Facility F	0.61	0.45
Facility G	0.72	0.66
Facility H	0.55	0.41
Facility I	0.73	0.64
Facility J	0.58	0.49
Facility K	0.62	0.55
Facility L	0.53	0.41
Facility M	0.43	0.19
Facility N	0.64	0.66
Facility O	0.74	0.70
Facility P	0.59	0.54
Facility Q	0.56	0.55
Facility R	0.77	0.68
Facility S	0.52	0.39
Facility T	0.73	0.78

R15. Discussion: The authors' discuss not using a machine learning algorithm and that they chose not to develop a full prediction model. I think the rational and discussion should be fleshed out more and maybe brought in earlier in the manuscript. If there are aspects of more traditional predictive models and their validation metrics that could be brought into the manuscript, the authors can revisit that per some of the above

AR16. As per prior comments addressing R3, R9, and R10, we have significantly revised the manuscript to address this point at an earlier stage in the manuscript.

REVIEWER 2 COMMENTS

R16. This is great paper. just a few points. Since evaluation is based on including different care facilities it is better to have a table to compare the patient population structure of these facilities.

AR16. Thank you for the support and clarification. We now include a Supplemental Table that describes some basic characteristics of the member population attributable to each of 6 subregions within KPNC.

Appendix Table 3. Characteristics of KPNC Adult Population by 6 Main Sub-regional Areas.

Characteristic	Location						
	KPNC Region	A	B	C	D	E	F
N	3,293,223	389,617	575,166	548,614	652,390	640,139	487,297
Age [Mean (SD)]	48 (18.0)	47 (17.8)	49 (18.2)	47 (17.6)	49 (18.4)	49 (18.2)	47 (17.6)
Male [N (%)]	1,564,615 (47.5%)	187,739 (48.2%)	269,400 (46.8%)	257,902 (47.0%)	300,273 (46.0%)	310,066 (48.4%)	239,235 (49.1%)
Race [N (%)]							
Asian	631,943 (19.2%)	42,513 (10.9%)	92,414 (16.1%)	141,987 (25.9%)	85,792 (13.2%)	132,078 (20.6%)	137,159 (28.1%)
Black	217,373 (6.6%)	22,818 (5.9%)	48,462 (8.4%)	67,134 (12.2%)	48,461 (7.4%)	19,053 (3.0%)	11,445 (2.3%)
Hispanic	669,053 (20.3%)	129,643 (33.3%)	101,769 (17.7%)	115,969 (21.1%)	97,577 (15.0%)	105,179 (16.4%)	118,916 (24.4%)
White	1,419,661 (43.1%)	151,582 (38.9%)	274,659 (47.8%)	160,421 (29.2%)	353,256 (54.1%)	315,167 (49.2%)	164,576 (33.8%)
Other	355,193 (10.8%)	43,061 (11.1%)	57,862 (10.1%)	63,103 (11.5%)	67,304 (10.3%)	68,662 (10.7%)	55,201 (11.3%)
COPS2* [Mean (SD)]	14.8 (17.2)	14.7 (17.1)	15.1 (17.8)	14.4 (16.6)	15.6 (18.7)	14.6 (16.7)	14.0 (16.0)

* The COPS2 (COmorbidity Point Score, version 2), described in Escobar et al. (2013) is a score assigned every month to all adults with a Kaiser Permanente Northern California medical record number. Range is from 0 to 1010; higher scores indicate worse mortality risk. The univariate relationship between the COPS2 and 1-year mortality is as follows: 0-39, 0.3%; 40-64, 5.3%; 65+, 17.2%.

R17. Different facilities follow different practice styles and they show different resource utilization patterns. I think it is necessary to see how come the different utilization patterns did not have any impact on the performance of CHOTS score

AR17. Table 2 does show significant variation in the performance of the CHOTS score across facilities, with respect to the strength of the correlation and the maximum lagged value, which we believe reflects some of the differences in case-mix, practice, and resource utilization you allude to. At the same time, many of the COVID19-related practices were also standardized at the regional level, so facilities were not completely independent in each of their actions. We've revised the Discussion as follows:

> "Even at the sub-regional level, where COVID19 infections and hospitalizations have exhibited substantial heterogeneity in timing and size, cross-correlation varied but remained strong at most individual medical centers over similar time frames."

REVIEWER 3 COMMENTS

R18. The authors proposed a COVID hotspotting score for identifying emerging COVID activity using data from health systems. It is not clear to me whether similar approaches have been well used/accepted for identifying emerging activities of other diseases such as influenza, and how much real impact would this research bring to our life. I am leaving this for the experts to judge. I am a statistician and my following comments will be mainly focusing on the data analysis part.

AR18. Thank you for this clarification; similar work using methods like change point analysis have been used to evaluate seasonal influenza activity which we adapted for this tool.

R19. Overall, the idea of using data from health systems for evaluating/identifying emerging COVID activities is very interesting, and most of the statistical analysis were conducted appropriately. My main concerns are: The scores seem to be rather ad hoc and I am having some doubts on their reliability. On one hand, it is not clear to me why the scores (0, 0.2, 0.4, 0.6, 1 etc.) provided in Appendix table 1 were set in this way, and the Major indicators and minor indicators seem to be selected in a empiricism manner other than data driven. Further, the weights for major and minor indicators are rather ad hoc as well. My main concern here is that any change to these scores/weights might change the results.

AR19. Thank you for this clarification. As discussed above in R7 and elsewhere, we were under an extreme urgency to develop a score to assist our health system with its response to COVID19 when little was known about the US experience and we had only experienced what turned out to be a very small wave #1. Thus, by the time we fully implemented this score into operations in June 2020, we had to use prior seasonal influenza data and our clinical judgment and heuristics to select variables and assign scores. If we had had the luxury of time, perhaps the CHOTS score would have differed. Despite this, we have shown in a temporally independent fully prospective validation dataset that the scores have continued to demonstrate very strong performance. We have revised the Discussion as follows in the limitations:

> "We also generated and deployed the CHOTS score during a time of great uncertainty and practice change following the first wave of COVID19 activity in California. As a result of the extreme urgency to prepare our health system, we depended on clinical judgement and

heuristics, in addition to prior health system influenza patterns, to develop our score. With the luxury of time, more advanced machine learning or statistical techniques may have produced a different score. Small sample sizes in each facility may have also impacted statistical significance testing.Nonetheless, the CHOTS continued to show very strong performance through the third wave of COVID in Northern California."

R20. The other concern I am having is that the scores in Appendix Table 1 are based on statistical significance tests. Whether the results would be significant or not is related to the sample size. Sometimes the signal is too weak to be detected when the sample size is too small. If we use the proposed scores in other states/regions, where the population could be potentially very different, the test results could be very differently distributed (for example, hypothetically, if the sample size is too small, one could end up having most of the variables having NS results). I am hence having some doubts on whether it is ok to set the scores based on the significance of the tests.

AR20. Thank you for this comment. It is true that smaller sample sizes at each facility may have impacted the statistical significance of the tests. We revised the limitation section to discuss this:

"Small sample sizes in each facility may have also impacted statistical significance testing."

R21. Overall, I feel one can develop a relatively more data-driven scoring system. However, as i have pointed out, I am a layman to this area and it is not clear to me whether such kind of setting/approach would make sense to other experts in this area. I believe It would certainly be helpful if the authors can provide further discussions/explanations for their proposed scoring system.

AR21. Thank you for your comments and we have done our best to augment the discussion and explanation as described in responses to R5, R10, R12, and others.

REVIEWER 4 COMMENTS

R22. Thank you for an opportunity to review this interesting and timely study. My major concern is about generalisability of CHOTS score due to the lack of validation and spurious correlations issues. One can use a couple of hospitals data as validation data to enhanced the generalisability. How the spurious correlation issue has been addressed which raised here. Dean, R.T., Dunsmuir, W.T.M. Dangers and uses of cross-correlation in analyzing time series in perception, performance, movement, and neuroscience: The importance of constructing transfer function autoregressive models. Behav Res 48, 783–802 (2016). https://urldefense.com/v3/__https://doi.org/10.3758/s13428-015-0611-2__;!!BZ50a36bapWJ!7V_uqI74gTNiAG0Xau1r-livyutUuNAnbPTcUxf8eA1waFd3DVObIZ5t57PKKQ2EwQ$. Furthermore, I could not find enough details about statistical model and its performance for developing CHOTS score.

AR22. Thank you the review and we agree that there are always risks that there can be spurious correlations. We have addressed this concern in two ways: (1) we used a temporally independent prospective period to validate the tool; and (2) we've evaluated generalizability by examining a 'reduced' form of the model. Our data, outlined in responses to R3 and R14, demonstrate the sustained performance of the CHOTS score in these two ways. Also, in responses to R7, R9, R10, and R19, we discuss how we've substantially revised our manuscript to reflect our development process for the CHOTS score.

VERSION 2 – REVIEW

REVIEWER	Mody, Aaloke Washington University in St Louis, Infectious Diseases

REVIEW RETURNED	02-May-2021

GENERAL COMMENTS	I think the author's responded to my comments in the updated manuscript and detailed some of their choices more clearly. No additional comments.

REVIEWER	Jiang, Binyan The Hong Kong Polytechnic University, Applied Mathematics
REVIEW RETURNED	04-May-2021

GENERAL COMMENTS	I thank the authors in addressing all my questions. The main concerns I raised in my previous report have been appropriately discussed or addressed by pointing out the limitations of the proposed SCORE.

REVIEWER	Faisal, Muhammad University of Bradford
REVIEW RETURNED	18-Apr-2021

GENERAL COMMENTS	No further comment