Performance of Existing and Novel Symptom- and Antigen Testing-Based COVID-19 Case Definitions in a Community Setting

Scott Lee3*, Olivia Almendares1*, Jessica L. Prince-Guerra1,2, Charles M. Heilig3, Jacqueline E. Tate1, Hannah L. Kirking1

1COVID-19 Response Team, CDC; 2Laboratory Leadership Service, CDC; 3Center for Surveillance, Epidemiology, and Laboratory Services (CSELS), CDC

*contributed equally to this manuscript

Corresponding Author: Scott Lee, Centers for Disease Control and Prevention, 1600 Clifton Rd NE, Atlanta GA 30322 (yle4@cdc.gov)

Abstract (200 words)

Point-of-care antigen tests are an important tool for SARS-CoV-2 detection, but they are less clinically sensitive than real-time reverse-transcription PCR (RT-PCR), impacting their efficacy as screening procedures. Our goal in this study was to see whether we could improve this sensitivity by considering antigen test results in combination with other relevant information, namely exposure status and reported symptoms. In November of 2020, we collected 3,419 paired upper respiratory specimens tested by RT-PCR and the Abbott BinaxNOW antigen test at two community testing sites in Pima County, Arizona. We used symptom, exposure, and antigen testing data to evaluate the sensitivity and specificity of various symptom definitions in predicting RT-PCR positivity. Our analysis yielded 6 novel multi-symptom case definitions with and without antigen test results, the best of which overall achieved a Youden’s J index of 0.66, as compared with 0.52 for antigen testing alone. Using a random forest as a guide, we show that this definition, along with our others, does not lose the ability to generalize well to new data despite achieving optimal performance in our sample. Our methodology is broadly applicable, and we have made our code publicly available to aid public health practitioners in developing or fine-tuning their own screening rules.
Introduction

Since the emergence of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) in December 2019, the virus has spread globally, causing widespread illness with varying degrees of severity. The emergence of several SARS-CoV-2 variants over time has led to periodic spikes in cases, with the most recent surge in cases related to the B.1.1.529 variant (Omicron). Testing is an important part of the overall strategy to reduce and/or control SARS-CoV-2 spread, and antigen testing in particular has become very common, with tests widely available for both clinical and at-home use. Relative to nucleic acid amplification tests (NAATs) for SARS-CoV2, antigen tests have several benefits: operationally, they are fast, cheap, and relatively easy to administer; and diagnostically, they are very specific, with false positive rates often under 1%. However, the clinical sensitivity of these tests varies substantially, depending, among other things, on the patient’s exposure status, severity of illness, duration of illness, and presence of symptoms (2-8). For the Abbott BinaxNOW COVID-19 Ag Card (BinaxNOW) antigen test, which has been widely adopted in the United States for both community screening and at-home testing, investigators have found sensitivities ranging from as low as approximately 50% (3, 5, 9-12) to as high as 92% (11), with many more in between (e.g. from 74%-77% (6, 7, 13, 14)). These values are much lower than the test’s reported sensitivity of 97%, and their variability makes it difficult for public health practitioners to estimate accurately the performance of their screening programs without relying on routine confirmatory testing with NAATs.

When potential COVID-19 cases are asymptomatic, antigen testing is more feasible and perhaps the better screening option, particularly when epidemiologic risk factors for infection (such as travel history or close contact with a confirmed case) have not been identified. When potential cases are symptomatic, however, the screening scenario changes in two important ways: the sensitivity of the antigen testing tends to improve (Prince-Guerra 2021), and there is more clinical information available for predicting the underlying infection. The current Council of State and Territorial Epidemiologists (CSTE) COVID-19 surveillance case definition (15) follows this line of reasoning for identifying probable cases, requiring them to have either a positive antigen test or to have had close contact with a confirmed case and be symptomatic. To the best of our knowledge, this case definition was not directly derived from case-level clinical and laboratory data (e.g., using the kind of optimization procedure we adopt below), meaning it may not provide an optimal balance of sensitivity and specificity. What balance it does provide is fixed, since it is a single binary rule intended for surveillance, rather than a collection of rules with different operational characteristics intended for a variety of clinical use cases. The question remains, then, of how best to combine symptom information with antigen test results for screening.

In this analysis, our goal was to show how this question might be answered by rigorously exploring the sensitivity-specificity tradeoff in these sorts of hybrid, testing/symptom-based screening case definitions. Building on previous work in improving screening and diagnosis procedures for tuberculosis disease (16) and COVID-19 (17), we show that a brute-force combinatorial search, along with some guidance from a robust machine learning model, can find definitions that improve the sensitivity of antigen testing relative to RT-PCR without sacrificing too much of its specificity. Although the exact definitions we propose may not be appropriate for all COVID-19 screening or diagnosis applications, the discovery procedure itself is adaptable to any dataset with symptom data and a reference standard (e.g., RT-PCR) and will thus be useful for both researchers and practitioners alike in developing definitions tailored to their own scientific and programmatic uses.

Methods

Data collection
We collected paired upper respiratory swabs, demographics, symptom information and exposure history from participants aged 10 years and older at two community testing sites in Pima County, Arizona during November 3-17, 2020, as previously described (3, 10). Here we use symptom data which captured presence of 15 symptoms in the 14 days prior to or on the day of testing. The list of symptoms includes all symptoms per CSTE COVID-19 2020 Interim Case Definition (15). The protocol for this evaluation was reviewed by the Centers for Disease Control and Prevention (CDC) and determined to be non-research and was conducted consistent with applicable federal law and CDC policy as defined in 45 CRF 46.102(l)(2) (18).

All paired swab specimens were tested by RT-PCR and the Abbott BinaxNOW COVID-19 Ag Card (BinaxNOW) antigen test. A healthcare professional collected a bilateral anterior nasal swab first for BinaxNOW testing, immediately followed by a bilateral nasopharyngeal swab for RT-PCR testing. Nasopharyngeal swabs were tested using either the CDC 2019-nCoV RT-PCR Diagnostic Panel (CDC RT-PCR assay) for detection of SARS-CoV-2 (n=2,582) (19) or the Fosun COVID-19 RT-PCR Detection Kit (Fosun RT-PCR assay) (n=837) (20). As described previously (3), validation by the commercial laboratory reported the limit of detection as equivalent at 500 copies/mL for both assays. In the analysis below, both tests were treated equally as reference standards for classification.

**Analytic aims**

Our main goal was to see whether we could improve the performance of antigen testing in predicting RT-PCR positivity by using a hybrid case definition that also considers reported symptoms. To that end, we explored two methods for arriving at such a case definition: heuristics based on combinations of reported symptoms and antigen test results, including an exhaustive combinatorial search of the various predictors; and heuristics based on a tree-based machine learning model trained on the data. We were also interested in determining which of the hybrid case definitions would be most likely to generalize out-of-sample data. Here, we used two heuristics to avoid this kind of overfitting: a preference for parsimony, such that case definitions considering fewer variables are preferable to ones considering more; and a preference for case definitions that do not grossly exceed the generalization performance of the same machine learning model as above.

**Heuristics based on random forests**

To estimate the maximum generalizable classification performance, we might expect from these hybrid case definitions, we turned to the random forest (21), a tree-based machine learning model theoretically guaranteed not to overfit its training data. We trained random forests on two feature sets: one using only the individual symptoms as predictors, and one using both the individual symptoms and antigen test results as predictors. For each set of predictors, we fit five large forests (each with 10,000 trees) to the entire dataset, beginning with trees of depth one for the first forest, and increasing the depth of the trees by one for each successive forest. This process yielded 10 trained forests, one for each unique combination of feature set (n=2) and depth (n=5).

From each of these 10 forests, we retrieved the out-of-bag decision scores for all observations, yielding a forest-derived distribution of predicted probabilities for RT-PCR positivity. We then compared the predicted probabilities with the actual RT-PCR test results to construct a standard receiver operating characteristic (ROC) curve, plotting sensitivity as a function of 1 - specificity. Because this curve contains the random forest’s best-possible scores for Youden’s J index, we treated it as an approximate upper bound on the generalization performance we might expect from naive (i.e., non-learned) case definitions constructed from the same set of variables.

In addition to estimating the upper bound on generalizable classification performance, we also used the random forest as a heuristic for selecting symptoms to include in a combined case definition with the antigen results.
After training, random forests rank their features based on the average amount of information they contributed to the constituent trees’ predictions. Here, we used the feature importances from the symptoms-and-antigen forests to determine which symptoms worked best in combination with test results. We note here that the decision trees in a random forest may split on features “both ways”, such that an important feature may be a lack of that feature and not a presence.

**Heuristics based on combinatorial analysis**

We used a brute-force combinatorial approach to enumerate the full range of classification scores achievable with non-machine-learned case definitions. Here, we considered three sets of predictors: the symptoms alone, the symptoms in combination with antigen test results, and the symptoms plus known exposure to a confirmed case in combination with the antigen test results. For the symptoms alone, we enumerated all possible $m$-of-$n$ combinations of symptoms, where $n$ is the total number of symptoms in the combination, and $m$ is the minimum number of symptoms required to be present for the patient to be considered a case. We used 5 as the maximum value for both $m$ and $n$, mainly to limit the computational complexity of the enumeration, but also to improve the generalizability of the results. For the symptoms and antigen testing together, we combined this enumeration with the binary antigen test results, such that anyone with either a positive antigen test or $m$ of the $n$ symptoms in a given combination would be labeled a case.

**Established case definitions**

To serve as a point of comparison for the combinations, we evaluated the predictive performance of 3 previously established case definitions: the CSTE COVID-19 case definition (15), and two compound symptom combinations found to work well in a published development study (17). We redefine these last two as follows:

- **Reses 1**: loss of taste or smell OR at least 1 of (shortness of breath, fever and/or chills, or myalgia)
- **Reses 2**: loss of taste or smell OR at least 3 of (shortness of breath, difficulty breathing, fever and/or chills, or myalgia)

Both these definitions worked well in the development study; Reses 2 achieved the highest overall performance as measured by F1 score (80%). To our knowledge, this is the first time the definitions have been validated on external data, as well as having their performance examined in combination with antigen test results.

**Classification metrics**

For the purposes of screening and diagnosis in the absence of confirmatory testing, we evaluated prediction methods on several common measures of binary classification performance, including sensitivity (se), specificity (sp), Youden’s J index, and F1 score. For surveillance purposes, we evaluated methods on a single metric, the difference between actual prevalence, defined as the number of RT-PCR positive patients in our dataset, and predicted prevalence, defined as the number of positive calls from the classification method.

**Statistical Analysis**

For inference, we constructed bias-corrected and accelerated bootstrap confidence intervals (22) for all metrics; these were not adjusted for multiplicity. We also used the Wilcoxon signed-rank test to differences in paired classification metrics, where appropriate.

**Software and hardware**
The random forest, combinatorial search, and statistical analysis were implemented in Python 3.8 with the scikit-learn (23) and statsmodels (24) packages. We ran all analyses on a scientific workstation with a 12-core CPU and 128 GB of RAM, and we used parallelization whenever possible to improve runtime. Our code is publicly available on GitHub at https://github.com/scotthlee/az-covid.

Results

We analyzed 3,419 paired specimens from 3,302 total participants aged 10 years and older with complete testing and survey data for which demographic and symptom data has been previously described in detail (3).

Performance of single predictors

Table 1 shows the performance of the single symptoms and antigen test results in predicting RT-PCR positivity. In terms of individual-level performance, congestion was the most sensitive (62%), followed by headache (59%), antigen testing (52%), cough (38%), and chills (37%). Antigen testing was the most specific (99%) and was closely followed by loss of taste or smell (97%), difficulty breathing (97%), rigors (shivers) (96%), and abdominal pain (96%). Overall, antigen testing was the best single predictor, with the highest Youden’s J index (52%) and F1-score (68%), even though its sensitivity was relatively low (53%). In terms of population-level performance, chills had the most accurate prevalence estimate (+4%), followed by diarrhea (-12%) and fever (-20%).

Random forest analysis

Figure 1 shows the out-of-bag Receiver Operating Characteristic (ROC) curves for our five random forests, organized in ascending order by depth (n) and plotted against the full set of ROC points for the m-of-n combinations of equal size n. Depth did not substantially impact performance, with all forests achieving sensitivities of 69% and 74% at specificities of 95% and 90%, respectively. We can also see that the ROC curves generally contain the points for the combinations, suggesting that the performance of most combinations should generalize to new samples from the same population. Regardless of depth, the strongest predictors by feature importance were antigen test result, loss of taste/smell, chills, myalgia, and fever, many of which also appeared in the top-performing combinations.

Combinatorial analysis

Figure 1 also shows the enumeration of symptom combinations in ROC space. Higher values of m make the combinations less sensitive but more specific, and a lower ratio of m to n makes them more sensitive and less specific; increasing n has no substantial impact on performance alone. We see this trend in values of Youden’s J index, as well, which decreases as a function of m (average J was 0.50, 0.46, 0.38, 0.32, 0.28 for combinations with m of 1, 2, 3, 4, and 5, respectively) but not n (average J was 0.39 for all). By these measures, then, 1-of-n rules perform the best. Furthermore, because these rules are comparable to random forests of depth n, we expect those on or near the corresponding forest’s ROC curve both to generalize to new data and to be close to optimal.

The maximum achievable sensitivity for any combination of symptoms without antigen testing was 83% (specificity 61%); with antigen testing, this rose to 89% (specificity 61%). Because these are maxima at low values of m (1) and high values of n (15), and because they are well beyond the corresponding RF ROC curves, we do not expect them to generalize well to other datasets; nonetheless, they serve as useful points of comparison in selecting candidate rules we do expect to generalize.

For combinations near the ROC boundary, we propose three rules that seem to work particularly well in combination with antigen testing. In order of increasing specificity, these are as follows: one of fever, chills, fatigue, myalgia, or loss of taste or smell (82% sensitivity at 80% specificity); one of chills, difficulty breathing, or loss of taste or smell (75% at 91%); and two of rigors (shivers), headache, or loss of taste or smell (68% at 96%).
For ease of reference, we name these *spec80*, *spec90*, and *spec95*, respectively. Table 2 shows their classification performance both with and without antigen testing in comparison to loss of taste or smell alone and the existing case definitions (with the exception of the CSTE definition, none of these rules consider known exposure status).

Of note, the CSTE case definition is the only one that does not perform optimally; all others, including *Reses 1* and *Reses 2*, achieve near-maximal levels of sensitivity at their given levels of specificity. This result is easier to see in Figure 2, which visualizes points for the case definitions against the full enumeration of *m*-of-*n* combinations in ROC space. While no definition is more specific than antigen testing alone, all are more sensitive, and some achieve a better balance of the two metrics, scoring up to 0.14 higher on J (0.52 for antigen alone versus 0.66 for *spec90*).

For surveillance purposes, loss of taste or smell, when combined with antigen results, seems to be a good candidate, since it produces a nearly perfect prevalence estimate (5%) in addition to achieving a relatively high score on J (0.64). This balance between group- and individual-level performance was rare in these data and suggests the rule could be used effectively for both screening and surveillance, especially considering its parsimony (only 2 features evaluated).

**The Benefit of Adding Exposure Status**

Adding exposure status to the combinations generally did not improve their performance. When antigen test results were included alongside the symptoms, adding exposure status to a given *m*-of-*n* combination increased its J index only 43% of the time, lowering it the other 57% of the time (0.2% of the pairs had ties). Although the median J index was higher for rules considering exposure status (0.58) than for those not (0.57), the distribution of pairwise differences suggested the opposite (though very subtle) effect, with the median difference hovering around 0.01 in favor of the latter (Wilcoxon signed-rank test p-value < 0.001). Supplemental Figure 2 shows the distribution of these differences, and Supplemental Figure 3 shows how the distribution changes with respect to the specificity of the original combination (i.e., the combination without exposure status). Adding exposure status always made the original combination more specific, but when the original specificity was already high (more than approximately 0.80), the increase did not outweigh the resulting drop in sensitivity (Supplemental Figure 3).

**Discussion**

In our case, adding reported symptom data to antigen test results boosted the latter’s sensitivity as a screening procedure without incurring a large drop in specificity. Given the basic characteristics of the data in our study, this is perhaps unsurprising: of the 142 study participants who were antigen negative but RT-PCR positive, 110 (78%) were symptomatic, and so the potential gain in sensitivity from considering symptoms was large from the start. Many of these missed cases are recouped by our proposed hybrid definitions (especially *spec80*), but the added sensitivity does come at the cost of reduced specificity, which under some practical circumstances may be non-negligible. This result highlights an important point for public health practice: case definitions should be tailored to their specific use cases, or, perhaps more precisely, to the real-world costs associated with false-positive and false-negative classifications.

From a methodological standpoint, our results suggest that using parsimony as a heuristic for selecting definitions may increase the chances of their performance generalizing to unseen data. *Reses1* and *Reses2* were both developed with a brute-force combinatorial search on different data (as rules with an *m*₁-of-*n*₁ or *m*₂-of-*n*₂ structure, they are also more complex than the simple *m*-of-*n* definitions considered here) [17], and yet they performed well here, skirting the upper boundaries in ROC space for both the random forest and the full
enumeration of symptom combinations. Although this kind of combinatorial search is the most straightforward way of determining which of the candidate combinations work best, it is computationally intensive and may be impractical to run on consumer hardware for very large feature spaces. In these scenarios, a hybrid approach may be sensible, such that a random forest or other machine learning method with a built-in variable-selection/ranking procedure (e.g., LASSO or L1-regularized regression) could be used to whittle down the feature space before performing the combinatorial search.

Relating to SARS-CoV-2 specifically, our results suggest that antigen testing alone may not always be the best strategy for screening or diagnosis in the absence of confirmatory testing (or even in the presence of confirmatory testing, which may vary depending on likelihood of SARS-CoV-2 infection and symptom status), which is in line with current CDC guidance on the use of antigen tests in the community (25). However, given the increase in use of antigen tests and the high demand for rapid test results (26), our analysis here shows the potential benefit for use of antigen tests in the context of specific symptoms to improve sensitivity, albeit at some loss of specificity. Screening scenarios where frequent, repeated antigen testing would be impractical—at border entry points where travelers or migrants may only be tested once before moving on, for example, or in private homes where the cost of repeat testing may be prohibitive—may be good settings for the use of hybrid screening rules that improve on the sensitivity of a single test. Conversely, when confirmatory tests are relatively scarce but required for care, or when the local healthcare system is nearing capacity, a more specific rule, like our spec90 candidate rule or antigen testing alone, may be more appropriate. Finally, case definitions considering exposure status may be less optimal than those that do not, at least in scenarios similar to those in which our data were collected. In our results, these rules were often less sensitive than alternatives achieving the same specificity and less specific than those achieving the same sensitivity. This effect may have been due to the presence of other high-performing predictors in the data, like the antigen test results and loss of taste and/or smell, that made the added specificity from considering exposure less useful than it would have been otherwise.

Our analysis has several limitations. First, the CSTE COVID-19 case definition (27) has been updated since the time of this investigation to add additional symptoms for consideration which were not included in our investigation and cannot be evaluated, although many symptoms added to the most recent definition are more severe in nature (confusion or change in mental status; persistent chest pain/pressure; blue-colored skin, lips, or nail beds, depending on skin tone; and inability to wake or stay awake), and it is unlikely they would have been captured in an outpatient community setting. Second, although SARS-CoV-2 samples were not sequenced as part of this investigation, data collection occurred in November 2020, well before the two variants of concern, Delta (B.1.617.2) and Omicron (B.1.1.529), were documented in the United States. The candidate definitions we propose above, which rely heavily on loss of taste or smell to boost the sensitivity of antigen testing, may not generalize well to populations heavily affected by these variants if the symptoms they cause differ substantially from those cause by the original strain. Additionally, it is unclear whether these results may apply to re-infections or vaccine-breakthrough infections as this study was conducted prior to vaccine availability and widespread concern for re-infection. Because our analysis could easily be rerun with the availability of new data, we hope these limitations will be addressed by future research. Third, the data presented here are cross-sectional and pre-symptomatic individuals could not be identified. Fourth, as mitigation measures have relaxed, such as limits on large gatherings, mask use, etc., the circulation of other non-SARS-CoV-2 respiratory viruses will likely increase and the impact on symptom analysis for SARS-CoV-2 is not fully known. Finally, the FDA recently published a statement (28) suggesting that antigen tests may have reduced sensitivity against the Omicron variant, although there is also some recent data showing antigen test sensitivity remains similar to pre-Omicron levels in a community setting (29).

Conclusion
While confirmatory testing for symptomatic persons with a negative antigen test is recommended (25), this often presents challenges including logistical and cost constraints and longer turnaround times for RT-PCR, thus delaying care and/or isolation of individuals (26). In certain situations, confirmatory testing may not be available at all – especially in low resource settings. Therefore, using the methodology described in this study to better understand combinations of symptoms that are more predictive of SARS-CoV-2 infection in the absence of a positive antigen test may aid clinicians and public health professionals. In situations where confirmatory testing is not available, and immediate public health action is needed, considering the presenting symptoms in antigen negative individuals may provide important information on how to prioritize public health action, especially in congregate living or school settings.

DISCLAIMER The findings and conclusions in this report are those of the authors and do not necessarily represent the official positions of the Centers for Disease Control and Prevention (CDC).

FUNDING: This work was supported by the Centers for Disease Control and Prevention.

COMPETING INTERESTS STATEMENT:
**References**


https://www.fda.gov/media/134922/download, 2020,
https://www.fda.gov/media/137120/download, 2020,
Lancet Infect Dis 2021;21(9):e290-e5.
medRxiv 2022:2022.01.08.22268954.
Table 1. Point estimates and 95% confidence intervals for the classification performance of single-symptom case definitions and for antigen testing.

<table>
<thead>
<tr>
<th>Symptoms</th>
<th>Sensitivity(^1) (95% CI)</th>
<th>Specificity(^2) (95% CI)</th>
<th>PPV(^3) (95% CI)</th>
<th>NPV(^4) (95% CI)</th>
<th>Youden’s J(^5) (95% CI)</th>
<th>F1 score(^6) (95% CI)</th>
<th>Relative Difference in Prevalence(^7) (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fever(^8)</td>
<td>0.3 (0.27, 0.36)</td>
<td>0.95 (0.94, 0.96)</td>
<td>0.38 (0.34, 0.44)</td>
<td>0.93 (0.92, 0.94)</td>
<td>0.26 (0.22, 0.32)</td>
<td>0.34 (0.3, 0.39)</td>
<td>-0.2 (-0.29, -0.1)</td>
</tr>
<tr>
<td>Chills</td>
<td>0.37 (0.31, 0.42)</td>
<td>0.94 (0.92, 0.94)</td>
<td>0.35 (0.3, 0.4)</td>
<td>0.94 (0.93, 0.94)</td>
<td>0.3 (0.25, 0.35)</td>
<td>0.36 (0.31, 0.4)</td>
<td>0.04 (-0.08, 0.18)</td>
</tr>
<tr>
<td>Shivers/Rigors</td>
<td>0.21 (0.17, 0.25)</td>
<td>0.96 (0.96, 0.97)</td>
<td>0.37 (0.31, 0.45)</td>
<td>0.93 (0.92, 0.93)</td>
<td>0.18 (0.14, 0.23)</td>
<td>0.27 (0.22, 0.33)</td>
<td>-0.42 (-0.51, -0.33)</td>
</tr>
<tr>
<td>Muscle Aches</td>
<td>0.47 (0.42, 0.53)</td>
<td>0.89 (0.88, 0.9)</td>
<td>0.29 (0.24, 0.32)</td>
<td>0.95 (0.94, 0.95)</td>
<td>0.36 (0.3, 0.42)</td>
<td>0.36 (0.31, 0.4)</td>
<td>0.63 (0.45, 0.82)</td>
</tr>
<tr>
<td>Congestion(^9)</td>
<td>0.62 (0.58, 0.68)</td>
<td>0.76 (0.75, 0.78)</td>
<td>0.2 (0.18, 0.23)</td>
<td>0.95 (0.95, 0.96)</td>
<td>0.38 (0.34, 0.45)</td>
<td>0.3 (0.28, 0.34)</td>
<td>2.07 (1.71, 2.41)</td>
</tr>
<tr>
<td>Sore Throat</td>
<td>0.41 (0.35, 0.47)</td>
<td>0.85 (0.84, 0.87)</td>
<td>0.21 (0.18, 0.25)</td>
<td>0.94 (0.93, 0.94)</td>
<td>0.27 (0.2, 0.32)</td>
<td>0.28 (0.25, 0.33)</td>
<td>0.97 (0.74, 1.14)</td>
</tr>
<tr>
<td>Cough(^10)</td>
<td>0.38 (0.34, 0.44)</td>
<td>0.9 (0.89, 0.91)</td>
<td>0.27 (0.24, 0.32)</td>
<td>0.94 (0.93, 0.94)</td>
<td>0.28 (0.24, 0.34)</td>
<td>0.32 (0.28, 0.36)</td>
<td>0.41 (0.26, 0.57)</td>
</tr>
<tr>
<td>Shortness of Breath(^10)</td>
<td>0.17 (0.13, 0.21)</td>
<td>0.95 (0.94, 0.96)</td>
<td>0.24 (0.18, 0.3)</td>
<td>0.92 (0.91, 0.93)</td>
<td>0.12 (0.08, 0.16)</td>
<td>0.2 (0.15, 0.25)</td>
<td>-0.31 (-0.41, -0.2)</td>
</tr>
<tr>
<td>Difficulty breathing</td>
<td>0.1 (0.07, 0.13)</td>
<td>0.97 (0.96, 0.98)</td>
<td>0.24 (0.18, 0.33)</td>
<td>0.92 (0.91, 0.93)</td>
<td>0.07 (0.04, 0.11)</td>
<td>0.14 (0.11, 0.19)</td>
<td>-0.58 (-0.7, -0.54)</td>
</tr>
<tr>
<td>Nausea or vomiting</td>
<td>0.15 (0.11, 0.19)</td>
<td>0.94 (0.94, 0.95)</td>
<td>0.21 (0.16, 0.26)</td>
<td>0.92 (0.91, 0.93)</td>
<td>0.1 (0.06, 0.13)</td>
<td>0.17 (0.14, 0.21)</td>
<td>-0.27 (-0.37, -0.17)</td>
</tr>
<tr>
<td>Headache</td>
<td>0.59 (0.55, 0.64)</td>
<td>0.8 (0.78, 0.81)</td>
<td>0.22 (0.2, 0.25)</td>
<td>0.95 (0.94, 0.96)</td>
<td>0.39 (0.36, 0.46)</td>
<td>0.32 (0.29, 0.35)</td>
<td>1.72 (1.46, 1.98)</td>
</tr>
<tr>
<td>Abdominal Pain</td>
<td>0.09 (0.07, 0.15)</td>
<td>0.96 (0.95, 0.96)</td>
<td>0.18 (0.13, 0.25)</td>
<td>0.92 (0.9, 0.92)</td>
<td>0.05 (0.03, 0.09)</td>
<td>0.12 (0.09, 0.19)</td>
<td>-0.49 (-0.55, -0.37)</td>
</tr>
<tr>
<td>Diarrhea</td>
<td>0.17 (0.14, 0.21)</td>
<td>0.93 (0.92, 0.94)</td>
<td>0.19 (0.15, 0.24)</td>
<td>0.92 (0.91, 0.93)</td>
<td>0.1 (0.07, 0.14)</td>
<td>0.18 (0.15, 0.22)</td>
<td>-0.12 (-0.24, 0)</td>
</tr>
<tr>
<td>Loss of Taste or Smell</td>
<td>0.3 (0.26, 0.35)</td>
<td>0.97 (0.97, 0.98)</td>
<td>0.53 (0.48, 0.59)</td>
<td>0.94 (0.93, 0.94)</td>
<td>0.28 (0.23, 0.32)</td>
<td>0.39 (0.34, 0.44)</td>
<td>-0.42 (-0.51, -0.35)</td>
</tr>
<tr>
<td>Fatigue</td>
<td>0.44 (0.4, 0.49)</td>
<td>0.84 (0.83, 0.85)</td>
<td>0.21 (0.19, 0.24)</td>
<td>0.94 (0.93, 0.95)</td>
<td>0.29 (0.24, 0.34)</td>
<td>0.29 (0.26, 0.32)</td>
<td>1.07 (0.86, 1.35)</td>
</tr>
<tr>
<td>Antigen</td>
<td>0.53 (0.47, 0.58)</td>
<td>0.99 (0.99, 1.0)</td>
<td>0.98 (0.95, 0.99)</td>
<td>0.95 (0.95, 0.96)</td>
<td>0.52 (0.47, 0.58)</td>
<td>0.68 (0.64, 0.73)</td>
<td>-0.46 (-0.52, -0.39)</td>
</tr>
</tbody>
</table>

1 Sensitivity, the probability of testing positive when disease is present.
2 Specificity, the probability of testing negative when disease is absent.
3 Positive predictive value, the probability of a patient having disease when test is positive.
4 Negative predictive value, the probability of a patient not having disease when test is negative.
5 Youden’s J index is defined as sensitivity plus specificity minus 1 (perfect score = 1)
6 F1 score is the harmonic mean of sensitivity and PPV (perfect score = 1)
7 Difference in prevalence is the difference in the number of positive symptom screens (i.e. true positive plus false positive) for each combination from the actual number of RT-PCR positive specimens.
8 Includes measured or subjective fever
9 Included nasal congestion or runny nose.
10 Cough, shortness of breath includes only if symptom was new or worsening in the 14 days prior to testing.
### Table 2. Point estimates and 95% bootstrap confidence intervals for the classification performance of multi-symptom case definitions with and without antigen test results. The CSTE definition additionally considers exposure status.

<table>
<thead>
<tr>
<th>Case Definitions</th>
<th>Sensitivity(^1) Value (95% CI)</th>
<th>Specificity(^2) Value (95% CI)</th>
<th>PPV(^3) Value (95% CI)</th>
<th>NPV(^4) Value (95% CI)</th>
<th>Youden’s J(^5) Value (95% CI)</th>
<th>F1 score(^6) Value (95% CI)</th>
<th>Relative Difference in Prevalence(^7) Value (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Without Antigen</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Taste(^9)</td>
<td>0.30 (0.26, 0.35)</td>
<td>0.97 (0.97, 0.98)</td>
<td>0.53 (0.48, 0.59)</td>
<td>0.94 (0.93, 0.94)</td>
<td>0.28 (0.23, 0.32)</td>
<td>0.39 (0.34, 0.44)</td>
<td>-0.42 (-0.51, -0.35)</td>
</tr>
<tr>
<td>Spec95(^10)</td>
<td>0.40 (0.35, 0.45)</td>
<td>0.95 (0.94, 0.96)</td>
<td>0.44 (0.37, 0.48)</td>
<td>0.94 (0.94, 0.95)</td>
<td>0.35 (0.29, 0.40)</td>
<td>0.42 (0.37, 0.46)</td>
<td>-0.08 (-0.16, 0.01)</td>
</tr>
<tr>
<td>Spec90(^11)</td>
<td>0.55 (0.49, 0.61)</td>
<td>0.90 (0.89, 0.91)</td>
<td>0.35 (0.32, 0.39)</td>
<td>0.95 (0.95, 0.96)</td>
<td>0.45 (0.39, 0.51)</td>
<td>0.43 (0.39, 0.47)</td>
<td>0.55 (0.42, 0.73)</td>
</tr>
<tr>
<td>Spec80(^12)</td>
<td>0.69 (0.62, 0.73)</td>
<td>0.81 (0.80, 0.83)</td>
<td>0.26 (0.23, 0.29)</td>
<td>0.96 (0.96, 0.97)</td>
<td>0.50 (0.43, 0.54)</td>
<td>0.38 (0.34, 0.42)</td>
<td>1.64 (1.40, 1.97)</td>
</tr>
<tr>
<td>Reses 1(^13)</td>
<td>0.64 (0.58, 0.69)</td>
<td>0.84 (0.83, 0.86)</td>
<td>0.28 (0.24, 0.31)</td>
<td>0.96 (0.95, 0.97)</td>
<td>0.48 (0.43, 0.53)</td>
<td>0.39 (0.34, 0.42)</td>
<td>1.28 (1.09, 1.58)</td>
</tr>
<tr>
<td>Reses 2(^14)</td>
<td>0.36 (0.31, 0.41)</td>
<td>0.96 (0.95, 0.96)</td>
<td>0.44 (0.37, 0.48)</td>
<td>0.94 (0.93, 0.95)</td>
<td>0.32 (0.26, 0.37)</td>
<td>0.40 (0.35, 0.44)</td>
<td>-0.17 (-0.28, -0.08)</td>
</tr>
<tr>
<td>CSTE(^15)</td>
<td>0.49 (0.44, 0.54)</td>
<td>0.87 (0.86, 0.88)</td>
<td>0.26 (0.23, 0.30)</td>
<td>0.95 (0.94, 0.95)</td>
<td>0.36 (0.30, 0.40)</td>
<td>0.34 (0.31, 0.38)</td>
<td>0.89 (0.68, 1.13)</td>
</tr>
<tr>
<td><strong>With Antigen</strong>(^16)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Taste</td>
<td>0.66 (0.62, 0.72)</td>
<td>0.97 (0.97, 0.98)</td>
<td>0.70 (0.65, 0.74)</td>
<td>0.97 (0.96, 0.97)</td>
<td>0.64 (0.59, 0.69)</td>
<td>0.68 (0.64, 0.72)</td>
<td>-0.05 (-0.13, 0.03)</td>
</tr>
<tr>
<td>Spec95</td>
<td>0.68 (0.63, 0.73)</td>
<td>0.96 (0.95, 0.96)</td>
<td>0.60 (0.56, 0.65)</td>
<td>0.97 (0.96, 0.97)</td>
<td>0.64 (0.60, 0.68)</td>
<td>0.64 (0.60, 0.67)</td>
<td>0.13 (0.05, 0.23)</td>
</tr>
<tr>
<td>Spec90</td>
<td>0.75 (0.70, 0.79)</td>
<td>0.91 (0.90, 0.92)</td>
<td>0.44 (0.39, 0.48)</td>
<td>0.97 (0.97, 0.98)</td>
<td>0.66 (0.61, 0.70)</td>
<td>0.55 (0.50, 0.59)</td>
<td>0.71 (0.56, 0.92)</td>
</tr>
<tr>
<td>Spec80</td>
<td>0.82 (0.79, 0.86)</td>
<td>0.80 (0.79, 0.81)</td>
<td>0.28 (0.25, 0.32)</td>
<td>0.98 (0.97, 0.98)</td>
<td>0.62 (0.58, 0.67)</td>
<td>0.42 (0.38, 0.46)</td>
<td>1.90 (1.60, 2.22)</td>
</tr>
<tr>
<td>Reses 1</td>
<td>0.79 (0.75, 0.84)</td>
<td>0.84 (0.83, 0.85)</td>
<td>0.32 (0.29, 0.35)</td>
<td>0.98 (0.97, 0.98)</td>
<td>0.63 (0.59, 0.68)</td>
<td>0.46 (0.42, 0.50)</td>
<td>1.45 (1.21, 1.71)</td>
</tr>
<tr>
<td>Reses 2</td>
<td>0.68 (0.63, 0.73)</td>
<td>0.95 (0.95, 0.96)</td>
<td>0.59 (0.52, 0.63)</td>
<td>0.97 (0.96, 0.97)</td>
<td>0.63 (0.59, 0.68)</td>
<td>0.63 (0.57, 0.67)</td>
<td>0.15 (0.07, 0.25)</td>
</tr>
<tr>
<td>CSTE</td>
<td>0.70 (0.66, 0.75)</td>
<td>0.90 (0.89, 0.91)</td>
<td>0.40 (0.35, 0.45)</td>
<td>0.97 (0.96, 0.98)</td>
<td>0.61 (0.56, 0.65)</td>
<td>0.51 (0.48, 0.56)</td>
<td>0.78 (0.62, 0.99)</td>
</tr>
<tr>
<td><strong>Antigen Alone</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.53 (0.47, 0.58)</td>
<td>0.99 (0.99, 1.0)</td>
<td>0.98 (0.95, 0.99)</td>
<td>0.95 (0.95, 0.96)</td>
<td>0.52 (0.47, 0.58)</td>
<td>0.68 (0.64, 0.73)</td>
<td>-0.46 (-0.52, -0.39)</td>
</tr>
</tbody>
</table>

1. Sensitivity, the probability of testing positive when disease is present.
2. Specificity, the probability of testing negative when disease is absent.
3. Positive predictive value, the probability of a patient having disease when test is positive.
4. Negative predictive value, the probability of a patient not having disease when test is negative.
5. Youden’s J index is defined as sensitivity plus specificity minus 1 (perfect score = 1)
6. F1 score is the harmonic mean of sensitivity and PPV (perfect score = 1)
7. Difference in the number of positive symptom screens (i.e. true positive plus false positive) for each combination from the actual number of RT-PCR positive specimens.
8. Excludes antigen test results for each combined multi-symptom case definition.
9. Includes only the symptom of loss of taste or smell.
10. Spec95 includes two of rigors (shivers), headache, or loss of taste or smell.
11. Spec90 includes one of chills, difficulty breathing, or loss of taste or smell.
12. Spec80 includes one of fever, chills, fatigue, myalgia, or loss of taste or smell.
13. Reses 1: Loss of taste or smell or at least one of shortness of breath, fever and/or chills, or myalgia.
14 Reses 2: Loss of taste or smell OR at least three of shortness of breath, difficulty breathing, fever and/or chills, or myalgia.
16 Includes antigen test results for each combined multi-symptom case definition.
Figure 1. ROC curves for random forests of depth $n$, along with ROC points for $m$-of-$n$ combinations of symptoms, both excluding (lower regions) and including (higher regions) antigen test results. Darker colors correspond to higher values of $m$, and each subplot specifies a different value for $n$. For both methods, including antigen test results improves sensitivity, and for the $m$-of-$n$ combinations, increasing $m$ increases specificity.

Figure 2. ROC points for preexisting and novel case definitions (red) against the set of points for the full enumeration of $m$-of-$n$ combinations (blue, orange, and green). The colors correspond to the three kinds of predictors that could be included in a combination: symptoms alone (blue), symptoms plus antigen test results (orange), and symptoms plus antigen test results, but with the added requirement that symptoms are only considered with known exposure to a confirmed case (green).