Comparison of Saliva and Nasopharyngeal Swab Nucleic Acid Amplification Testing for Detection of SARS-CoV-2
A Systematic Review and Meta-analysis

Guillaume Butler-Laporte, MD; Alexander Lawandi, MD, MSc; Ian Schiller, MSc; Mandy Yao, BSc; Nandini Dendukuri, PhD; Emily G. McDonald, MD, MSc; Todd C. Lee, MD, MPH

IMPORTANCE Nasopharyngeal swab nucleic acid amplification testing (NAAT) is the noninvasive criterion standard for diagnosis of coronavirus disease 2019 (COVID-19). However, it requires trained personnel, limiting its availability. Saliva NAAT represents an attractive alternative, but its diagnostic performance is unclear.

OBJECTIVE To assess the diagnostic accuracy of saliva NAAT for COVID-19.

DATA SOURCES In this systematic review, a search of the MEDLINE and medRxiv databases was conducted on August 29, 2020, to find studies of diagnostic test accuracy. The final meta-analysis was performed on November 17, 2020.

STUDY SELECTION Studies needed to provide enough data to measure salivary NAAT sensitivity and specificity compared with imperfect nasopharyngeal swab NAAT as a reference test. An imperfect reference test does not perfectly reflect the truth (ie, it can give false results). Studies were excluded if the sample contained fewer than 20 participants or was neither random nor consecutive. The Quality Assessment of Diagnostic Accuracy Studies 2 tool was used to assess the risk of bias.

DATA EXTRACTION AND SYNTHESIS Preferred Reporting Items for Systematic Reviews and Meta-analyses reporting guideline was followed for the systematic review, with multiple authors involved at each stage of the review. To account for the imperfect reference test sensitivity, we used a bayesian latent class bivariate model for the meta-analysis.

MAIN OUTCOMES AND MEASURES The primary outcome was pooled sensitivity and specificity. Two secondary analyses were performed: one restricted to peer-reviewed studies, and a post hoc analysis limited to ambulatory settings.

RESULTS The search strategy yielded 385 references, and 16 unique studies were identified for quantitative synthesis. Eight peer-reviewed studies and 8 preprints were included in the meta-analyses (5922 unique patients). There was significant variability in patient selection, study design, and stage of illness at which patients were enrolled. Fifteen studies included ambulatory patients, and 9 exclusively enrolled from an outpatient population with mild or no symptoms. In the primary analysis, the saliva NAAT pooled sensitivity was 83.2% (95% credible interval [Crl], 74.7%-91.4%) and the pooled specificity was 99.2% (95% Crl, 98.2%-99.8%). The nasopharyngeal swab NAAT had a sensitivity of 84.8% (95% Crl, 76.8%-92.4%) and a specificity of 98.9% (95% Crl, 97.4%-99.8%). Results were similar in secondary analyses.

CONCLUSIONS AND RELEVANCE These results suggest that saliva NAAT diagnostic accuracy is similar to that of nasopharyngeal swab NAAT, especially in the ambulatory setting. These findings support larger-scale research on the use of saliva NAAT as an alternative to nasopharyngeal swabs.
Saliva Nucleic Acid Amplification Testing for Detection of SARS-CoV-2

Key Points

**Question** Is saliva nucleic acid amplification testing (NAAT) comparable to nasopharyngeal NAAT, the current noninvasive criterion standard test for diagnosis of coronavirus disease 2019?

**Findings** In this systematic review and latent class meta-analysis adjusting for the imperfect reference standard, saliva NAAT had a similar sensitivity and specificity to that of nasopharyngeal NAAT.

**Meaning** Given the ease of use and good diagnostic performances, these findings suggest that saliva NAAT represents an attractive alternative to nasopharyngeal swab NAAT and may significantly bolster massive testing efforts.

Testing is the cornerstone of a successful public health response to coronavirus disease 2019 (COVID-19). Nasopharyngeal sampling for nucleic acid amplification testing (NAAT) is the current noninvasive criterion standard diagnostic test. However, nasopharyngeal testing requires trained personnel and handling of a specially designed swab, and the technique cannot be easily or reliably performed in all populations (eg, children and quarantined individuals). A test that can be self-administered would greatly increase the options for case identification in the community. Consequently, alternative sampling sites are being explored to allow broader deployment of testing, increased access in remote regions, and testing among underserviced populations who may have limited access to highly qualified testing personnel. Of the currently available routes for testing, salivary testing likely represents the most practical option because it avoids the inconvenience, discomfort, and required technical expertise of nasopharyngeal sampling and could potentially be broadly deployed in the community. To date, the results of studies on the diagnostic performance of saliva NAAT are conflicting. Furthermore, other studies have relied on the assumption that the sensitivity and specificity of the nasopharyngeal swab is perfect, which is not the case. To summarize the evidence on saliva NAAT performance, we conducted a systematic review of the operating characteristics of saliva NAAT for detection of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and the diagnosis of COVID-19 using nasopharyngeal NAAT as the comparator. We performed a bayesian latent class meta-analysis that adjusts for the imperfect accuracy of the comparator.

**Methods**

**Search Strategy and Review**

This study followed the Preferred Reporting Items for Systematic Reviews and Meta-analyses (PRISMA) reporting guideline for meta-analysis of diagnostic test accuracy. The protocol was registered on PROSPERO. On August 29, 2020, we used the PubMed tool to search the MEDLINE database using the following terms: (novel coronavirus OR SARS-CoV-2 OR COVID19 OR COVID-19) and (PCR OR polymerase chain reaction OR NAAT OR nucleic acid amplification testing OR detection) and (saliva OR oral swabs OR salivary). On the same day, we used the following search terms on MedRxiv for applicable preprints: (SARS-CoV-2 OR COVID-19) and (PCR OR polymerase chain reaction OR NAAT OR nucleic acid amplification testing) and saliva. The MedRxiv search was simplified owing to the limited search functionality compared with PubMed. If preprints were also published in a peer-reviewed journal, we used the peer-reviewed version. We included studies that directly compared saliva and nasopharyngeal NAAT for the detection of SARS-CoV-2 and provided enough data to construct the 2 × 2 diagnostic table. Given that some studies used a combination or nasopharyngeal swab and oropharyngeal or oral swabs for diagnosis in routine clinical activities, we also allowed this composite test as a criterion standard. Allowing for the use of this composite reference test could make saliva NAAT appear less sensitive, hence making our comparison more conservative. Similarly, we allowed for studies using different NAAT platforms so long as the reference was used for routine clinical microbiology diagnostic purposes, ensuring its validity as a comparator. We excluded studies without random or consecutive samples. Owing to well-recognized concerns over publication bias and inflated effect estimates inherent to smaller studies, we further excluded studies with fewer than 20 patients. This threshold was prespecified before our search. Case-control studies were likewise excluded. Finally, we screened references from all included studies for other potential studies to include in the meta-analysis.

Studies were first reviewed by title, then by abstract, and finally by full-text review by 2 authors with experience in systematic reviews (G.B.-L. and T.C.L.). For studies selected after full-text review, we extracted raw data on true-positive, false-positive, true-negative, and false-negative test results for saliva vs nasopharyngeal NAAT. Using a prewritten data extraction chart, we also retrieved information on study design, country, age and sex of patients, proportion of symptomatic patients, and NAAT platforms used. Study selection, review, and data extraction were performed by 2 authors (G.B.-L. and T.C.L.), with disagreements resolved by consensus.

**Quality Assessment**

Two authors (G.B.-L. and A.L.) used the Quality Assessment of Diagnostic Accuracy Studies 2 (QUADAS-2) tool to assess for the risk of bias in each selected study, with disagreements resolved by consensus. The QUADAS-2 tool evaluates 4 bias categories: patient selection, index test, reference test, and flow and timing of testing. It also grades the risk that this bias decreases the applicability of the results in the target population (eg, patients being tested for SARS-CoV-2). Studies that were considered at high risk of bias were not included in the analysis.

**Analysis**

In the primary meta-analysis, we pooled assay sensitivities and specificities from the studies selected. As a secondary prespecified meta-analysis, we limited the analysis to only peer-reviewed studies. Finally, after finishing the data extraction and risk of bias assessment, we decided to perform a post hoc
meta-analysis restricted to studies performed in ambulatory patients and that did not allow repeated testing on multiple days. This analysis was performed to better assess diagnostic accuracy in a setting where saliva sampling is likely to have the greatest effect on public health policy.

To address the fact that the nasopharyngeal NAAT represents an imperfect reference standard, we used a latent class bayesian bivariate random-effects model. This model accounts for the imperfect reference test (nasopharyngeal swab NAAT) by estimating the true disease status of each participant using results from both the index (saliva NAAT) and the reference test. Hence, as opposed to traditional bivariate models, this model does not require that the reference test have a perfect (or near perfect) accuracy. For example, in the unlikely scenario where no tests presented a false-positive result, the latent class model would use a composite test to determine the true SARS-CoV-2 test result. Conditional dependence (ie, dependence that may arise between the tests if they are prone to making the same false-positive or false-negative errors) is also accounted for. The latent class model was also parameterized using a random-effects approach, allowing for correction owing to heterogeneity in sensitivity and specificity between the included studies.

Through the model, sensitivity and specificity of saliva NAAT are then obtained using the true disease estimate as the reference. For each analysis, we present sensitivity and specificity estimates using the latent true disease status as the criterion standard, as well as pooled sensitivity and specificity. We also computed 95% credible intervals (CrIs) for each result, and an overall 95% prediction interval for pooled sensitivity and specificity. Briefly, in a bayesian framework, the 95% CrI for a parameter is the interval within which it is estimated to fall, with a probability of 95%. The 95% prediction interval is the interval within which the parameter obtained in a future (still unmeasured) study, performed in similar populations, would fall with 95% probability. It therefore provides a measure of between-study heterogeneity as well as a measure of uncertainty about the applicability of our results in the studied population. The CrI is therefore always contained within the prediction interval. When between-study heterogeneity is low, the CrI and the prediction interval are expected to be similar. However, when between-study heterogeneity is large, the prediction interval widens. This highlights the fact that unmeasured results from a new study can differ significantly from the pooled estimate, which can suggest additional unmeasured factors explaining differences in diagnostic accuracy that may not be accounted for.

A sample from the posterior distributions of the parameters of interest was obtained using Monte Carlo Markov Chain methods implemented using the JAGS software packages (version 4.10) accessed through the R statistical environment (version 4.0.2 [R Foundation for Statistical Computing]). The programs were accessed through an app on R Shiny, version 1.5.0, written by 3 authors (M.C.Y., I.S., and N.D.) (https://bayesdata.shinyapps.io/meta-analysis/). We used low-information prior distribution on each parameter. For true COVID-19 prevalence in each study, we used a uniform (0, 1) prior distribution. For the hyper prior distribution for the index test sensitivity and specificity (on the logit scale), we used a normal distribution, with mean of 0 and variance of 100. For the between-study variance parameters, we used a gamma prior distribution with a shape parameter of 2 and a rate parameter of 0.5. Finally, we used a uniform (−1, 1) prior distribution for the correlation term between test sensitivity and specificity to account for a threshold effect. Three Gibbs sampling chains were used, with 40 000 iterations each. Pooled estimates were obtained by removing the first 10 000 iterations from each chain and thinning the remaining 30 000 by a factor of 10. Convergence was assessed by visual inspection of each Markov chain, with Brooks-Gelman-Rubin statistics (R-hat <1).14

As a comparator, we also performed a usual bayesian bivariate random-effects model analysis. This approach accounts for both within- and between-study variation in sensitivity and specificity and for the correlation between sensitivity and specificity across studies. However, it does not account for the imperfect reference test.

### Results

#### Review

Our MEDLINE and MedRxiv search yielded 129 and 256 studies, respectively. Only 7 studies were published before 2020, all from the MEDLINE search, suggesting that our search was specific to COVID-19. We initially retained 62 studies for full-text review, including 7 duplicate preprints that were subsequently published in peer-reviewed journals. We identified no further studies from reference parsing. After full-text review, we retained 16 studies for data extraction. The most common reason for exclusion was insufficient information to build the 2 × 2 diagnostic accuracy table (10 of 39 studies), followed by the presence of review articles (7 of 39). Eight of the studies were from the MedRxiv preprint server, whereas the rest were from peer-reviewed journals. However, during our review process, 2 of these studies were published in peer-reviewed literature and were therefore included in the analysis restricted to peer-reviewed literature. The PRISMA flow diagram is shown in Figure 1.

We ultimately retained 16 studies for the quantitative analysis. From these, the mean (SD) number of sample pairs was 370 (616), and the median number of sample pairs was 117 (interquartile range, 88-256). The difference is explained by 2 studies of more than 1900 samples that yielded a high proportion of concordant negative test results (96%) for saliva and nasopharyngeal swab NAAT. Both studies were performed in the context of mass population screening, explaining the low disease prevalence found.

The studies varied by setting and patient population. Although the US was the most highly represented country with 6 studies, we also found studies from Europe,15,17 Asia,3,19,21,26,27 and Australia.20 Patients were selected from drive-through testing clinics or at hospital admission, through a laboratory information system,26 through screening campaigns in returning travelers,21 or participants were...
studies only enrolled patients who fit certain criteria for symptoms or with a known exposure to a contact case. One study enrolled both patients and health care workers, although the exact inclusion and exclusion criteria were unclear. Finally, 1 study enrolled hospitalized patients, whereas 10 studies recruited the general population, and 1 study enrolled patients exclusively tested in an ambulatory setting.

Testing methods also differed between studies. All laboratories used different NAAT assays, but the assays used for the reference standard of nasopharyngeal swab NAAT had all been previously validated by the manufacturer, or in-house by the clinical laboratory in cases of home-brew tests developed by the clinical laboratory in cases of home-brew tests developed by local clinical laboratories. The reference standard used was always that used by the clinical laboratory for diagnosis of COVID-19 in usual clinical practice. Moreover, apart from 2 studies, all cohorts used the same assay to compare the saliva and the nasopharyngeal swab NAATs (other than the different sample types). Furthermore, 1 study used 2 different assays, but each pair of saliva-nasopharyngeal swab samples from each individual patient was analyzed using the same assay. Of note, 2 studies compared results from different NAAT assays on the same patients. For this meta-analysis, we selected the most sensitive of the compared assays (ie, the assay with the highest number of positive results using saliva NAAT). In 1 study, the authors also included an indeterminate category, which we merged with the positive results for the purposes of this meta-analysis. Two studies allowed for repeated tests on the same patients to be performed on separate days, and it was impossible to remove these duplicates from our meta-analysis. These studies also had the 2 highest rates of discordant results between tests with positive saliva findings and negative nasopharyngeal swab findings. Finally, 4 studies used a composite test that included nasopharyngeal swab in combination with either the oral or the oropharyngeal swab as the reference test. The details of the included studies are found in eTable 1 in the Supplement.

Quality Assessment
The QUADAS-2 tool highlights the results mentioned above. Whereas the index and reference tests were unlikely to contribute bias to our analysis, patient selection methods varied between studies, which may have provided biased estimates of sensitivity and specificity. As such, all studies scored a high or unclear risk of bias owing to patient selection, but most of the studies scored as low risk in other categories. The QUADAS-2 graphical summary is shown in Figure 2, and the tabular summary is available in eTable 2 in the Supplement.

Meta-analysis
In total, we included 5922 samples in the primary meta-analysis, of which 941 had a positive result (by saliva NAAT, nasopharyngeal swab NAAT, or both), and 4981 had a concordant negative result (by saliva and nasopharyngeal swab NAAT). In the primary analysis, we obtained a pooled saliva sensitivity of 83.2% (95% CrI, 77.4%-91.4%) and a pooled saliva specificity of 99.2% (95% CrI, 98.2%-99.8%). Individual study-estimated sensitivity and specificity (accounting for each study’s imperfect reference test) can be found in Figure 3, and

SARS-CoV-2 indicates severe acute respiratory syndrome coronavirus 2.
Figure 3. Primary Meta-analysis Results for the Detection of Severe Acute Respiratory Syndrome Coronavirus 2 in Saliva Samples

<table>
<thead>
<tr>
<th>Source</th>
<th>Test finding, No.</th>
<th>Sensitivity (95% CrI), %</th>
<th>Specificity (95% CrI), %</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>TP FP FN TN</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Algun et al14</td>
<td>30  5  25 38</td>
<td>60.6 (46.3-77.0)</td>
<td>98.8 (95.4-99.9)</td>
</tr>
<tr>
<td>Becker et al13</td>
<td>11  2 10 62</td>
<td>65.4 (44.3-89.8)</td>
<td>99.0 (96.5-99.9)</td>
</tr>
<tr>
<td>Byrne et al16</td>
<td>12  0  2 96</td>
<td>82.2 (64.8-95.5)</td>
<td>99.3 (98.0-99.9)</td>
</tr>
<tr>
<td>Caulley et al12</td>
<td>34 14 22 1869</td>
<td>79.5 (58.8-95.8)</td>
<td>99.1 (98.3-99.9)</td>
</tr>
<tr>
<td>Cheuk et al15</td>
<td>104 37 18 70</td>
<td>84.5 (72.4-92.0)</td>
<td>99.2 (97.5-99.9)</td>
</tr>
<tr>
<td>Iwasaki et al16</td>
<td>8  1 1 66</td>
<td>83.8 (65.9-96.1)</td>
<td>99.2 (97.7-99.9)</td>
</tr>
<tr>
<td>Hansson et al13</td>
<td>75 6 5 268</td>
<td>87.3 (74.1-96.5)</td>
<td>99.1 (97.2-99.9)</td>
</tr>
<tr>
<td>Kojima et al17</td>
<td>20  6 3 16</td>
<td>84.6 (69.9-95.0)</td>
<td>99.2 (97.6-99.9)</td>
</tr>
<tr>
<td>Landry et al30</td>
<td>28  2 5 89</td>
<td>81.8 (66.8-93.9)</td>
<td>99.2 (97.8-99.9)</td>
</tr>
<tr>
<td>McCormick-Baw et al29</td>
<td>47  1 2 105</td>
<td>86.0 (71.1-96.0)</td>
<td>99.2 (97.6-99.9)</td>
</tr>
<tr>
<td>Miller et al21</td>
<td>33  1 1 56</td>
<td>86.7 (73.1-96.6)</td>
<td>99.2 (97.6-99.9)</td>
</tr>
<tr>
<td>Pasomsub et al13</td>
<td>16  2 3 179</td>
<td>81.3 (66.1-96.0)</td>
<td>99.2 (97.6-99.9)</td>
</tr>
<tr>
<td>Teo et al18</td>
<td>139 70 11 116</td>
<td>89.4 (77.1-96.3)</td>
<td>99.1 (97.1-99.9)</td>
</tr>
<tr>
<td>Vogels et al13</td>
<td>32  3 2 30</td>
<td>86.4 (72.1-96.1)</td>
<td>99.1 (97.2-99.9)</td>
</tr>
<tr>
<td>Williams et al28</td>
<td>33  1 6 49</td>
<td>79.9 (65.0-91.6)</td>
<td>99.1 (97.5-99.9)</td>
</tr>
<tr>
<td>Yokota et al20</td>
<td>42  6 4 1872</td>
<td>85.9 (71.0-96.6)</td>
<td>99.2 (98.2-99.8)</td>
</tr>
<tr>
<td>Pooled</td>
<td>664 157 120 4981</td>
<td>83.2 (74.7-91.4)</td>
<td>99.2 (98.2-99.8)</td>
</tr>
</tbody>
</table>

Sensitivity and specificity estimates (with their 95% credible intervals [CrI]) are obtained from the latent class bivariate bayesian model. The true-positive (TP), false-positive (FP), false-negative (FN), and true-negative (TN) counts in the sensitivity graph are with respect to the nasopharyngeal swab nucleic acid amplification testing. Note the y-axis scale difference between the 2 forest plots. Diamond indicates pooled data.

Figure 4 shows the resulting summary receiver-operator curve. Nasopharyngeal swab NAAT performed similarly with a pooled sensitivity of 84.8% (95% CrI, 76.8%-92.4%) and a pooled specificity of 98.9% (95% CrI, 97.4%-99.8%).

By comparison, the traditional bivariate model (which assumes a perfect reference test) yielded similar sensitivity for the saliva NAAT (86%; 95% CrI, 80%-92%), but with a slight decrease in specificity (97%; 95% CrI, 93%-99%) for saliva NAAT. The prediction interval was also considerably larger and reached values of specificity as low as 40% (eFigure 1 in the Supplement). This is explained by the fact that cases with positive findings for the saliva NAAT and negative findings for the nasopharyngeal swab NAAT would worsen the estimate of saliva NAAT specificity using traditional methods.

The secondary analysis, restricted to peer-reviewed literature (eFigure 2 in the Supplement), found a pooled sensitivity for saliva NAAT of 85.6% (95% CrI, 77.0%-92.7%) and a pooled saliva specificity of 99.1% (95% CrI, 98.0%-99.8%). The nasopharyngeal swab NAAT had a sensitivity of 85.7% (95% CrI, 76.5%-93.4%) and a specificity of 99.0% (95% CrI, 97.7%-99.7%).

Finally, we performed a post hoc meta-analysis limited to the ambulatory setting, after removal of studies that allowed for repeated testing on multiple days. We retained 9 studies3,10,18,21,24,28,29 for this meta-analysis, for a total of 4851 patients, including 391 (8.1%) with a positive test result (saliva, nasopharyngeal swab, or both). Pooled saliva NAAT estimates remained similar (Figure 5) with a sensitivity of 84.5% (95% CrI, 73.0%-95.3%) and a specificity of 99.0% (95% CrI, 97.7%-99.7%). For nasopharyngeal swab NAAT, the sensitivity was 88.0% (95% CrI, 77.5%-95.8%) and the specificity was 98.7% (95% CrI, 96.2%-99.8%). The summary receiver-operator curve was similar to that of the primary analysis (eFigure 3 in the Supplement).

Detailed results and the summary receiver-operator curves for the meta-analysis on peer-reviewed studies can be found in eFigure 2 in the Supplement. Forest plots for the diagnostic accuracy of the nasopharyngeal swabs can be found in eFigure 4 in the Supplement. Gibbs sample chains and posterior densities for all analyses can be found in eFigures 5 to 7 in the Supplement.

Discussion

Our meta-analysis found that the diagnostic sensitivity for saliva NAAT is approximately 83.2% (95% CrI, 74.7%-91.4%), which is comparable to that reported for nasopharyngeal swab NAAT4 and to the result obtained using our latent class model analysis (84.8%; 95% CrI, 76.8%-92.4%). Given the ease of sample procurement and increased patient comfort, testing centers should strongly consider adopting saliva as their first sample choice, especially in community mass screening programs.

Strengths and Limitations

Our study has several strengths. Our search strategy was sensitive, because no additional studies were identified from reference screening or by other means than our primary search strategy. We also addressed a highly clinically relevant question that could potentially rapidly affect global public health policy for testing strategies. It is incredibly important to validate this sampling technique for possible deployment in countries or communities with continually high case rates and es-
especially for those with developing health care systems and less access to specialized care. An increasingly large number of published studies on the subject suggest that a reliable alternative to nasopharyngeal swab NAAT is clearly being sought.30 Finally, we used a meta-analysis that accounts for the imperfect reference standard. The standard meta-analysis methods would have provided negatively biased estimates of the true specificity of the saliva NAAT, with a larger amount of uncertainty (as can be seen from the prediction intervals). Accounting for this, we were able to obtain a more precise estimate of saliva NAAT diagnostic accuracy.

Our results are limited by the heterogeneity of the included studies, which varied in terms of the study population and the timing of testing. For example, hospitalized and critically ill patients were significantly underrepresented compared with studies performed in the community, and so our results may not apply in this setting. Similarly, many studies compared saliva and nasopharyngeal swabs later in the disease course, and only a few were able to obtain paired samples on the first day of presentation. Two studies23,26 also allowed for repeated samples on multiple days, and correlation between samples may have been imperfectly accounted for, which might have biased our results. However, after removal of these studies in the post hoc analysis, we obtained similar pooled sensitivity and specificity estimates. Furthermore, although each study used assays with similar technology that were all used in clinical practice, small differences between the assays and their targets could have contributed to heterogeneity.

Of note, most studies did not provide details on patient symptoms, and it remains unclear whether certain clinical syndromes (eg, predominantly lower respiratory tract disease or critical illness) may warrant a specific specimen type for optimal NAAT diagnostic accuracy. Such differences may explain the wide prediction intervals obtained in the traditional meta-analysis (without adjustment for the imperfect reference test), and some caution is warranted in applying the findings of our study to populations that were underrepresented in our meta-analysis. Nevertheless, all latent class meta-analyses showed much less uncertainty in the pooled diagnostic accuracy, including the post hoc analysis in the ambulatory patient population. Hence, despite these limitations, we believe that saliva sampling is a reasonable alternative to nasopharyngeal swabs, especially in community testing centers.

### Figure 4. Summary Receiver-Operator Curve for the Primary Meta-analysis of Detection of Severe Acute Respiratory Syndrome Coronavirus 2 in Saliva Samples

Each study is represented by a box of size proportional to the number of samples tested. Some boxes are small and overlap. The orange line is the saliva nucleic acid amplification testing (NAAT) 95% credible interval. The dark solid line is the summary receiver-operator curve. The dark circle is the pooled sensitivity and specificity estimate. The dark solid line is the summary receiver-operating characteristic curve.

### Figure 5. Post hoc Meta-analysis Results for the Detection of Severe Acute Respiratory Syndrome Coronavirus 2 in Saliva Samples From Ambulatory Patients

<table>
<thead>
<tr>
<th>Source</th>
<th>Test finding, No.</th>
<th>Sensitivity (95% CrI), %</th>
<th>Specificity (95% CrI), %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Becker et al15</td>
<td>11 2 10 62</td>
<td>70.7 (46.1-96.1)</td>
<td>98.7 (96.1-99.8)</td>
</tr>
<tr>
<td>Caulley et al22</td>
<td>34 14 22 1869</td>
<td>81.6 (58.0-97.6)</td>
<td>99.1 (97.9-99.8)</td>
</tr>
<tr>
<td>Hanson et al23</td>
<td>75 6 5 268</td>
<td>88.6 (75.7-97.8)</td>
<td>98.9 (97.0-99.8)</td>
</tr>
<tr>
<td>Kojima et al17</td>
<td>20 6 3 16</td>
<td>85.9 (71.6-96.6)</td>
<td>98.7 (95.6-99.8)</td>
</tr>
<tr>
<td>Landry et al27</td>
<td>28 2 5 89</td>
<td>83.7 (69.0-96.4)</td>
<td>98.9 (97.0-99.8)</td>
</tr>
<tr>
<td>Miller et al21</td>
<td>33 1 1 56</td>
<td>88.4 (75.6-97.9)</td>
<td>98.9 (96.9-99.9)</td>
</tr>
<tr>
<td>Pasomsub et al3</td>
<td>16 2 3 179</td>
<td>84.9 (68.6-97.4)</td>
<td>98.9 (97.2-99.8)</td>
</tr>
<tr>
<td>Williams et al18</td>
<td>33 1 6 49</td>
<td>82.3 (67.9-95.1)</td>
<td>98.9 (96.8-99.8)</td>
</tr>
<tr>
<td>Yokata et al32</td>
<td>42 6 4 1872</td>
<td>87.1 (72.9-97.7)</td>
<td>99.0 (97.7-99.8)</td>
</tr>
<tr>
<td>Pooled</td>
<td>292 40 59 4460</td>
<td>84.6 (73.0-95.3)</td>
<td>99.0 (97.7-99.7)</td>
</tr>
</tbody>
</table>

Sensitivity and specificity estimates (with their 95% credible intervals [CrI]) are obtained from the latent class bivariate bayesian model. The true-positive (TP), false-positive (FP), false-negative (FN), and true-negative (TN) counts in the sensitivity graph are with respect to the nasopharyngeal swab nucleic acid amplification testing. Note the y-axis scale difference between the 2 forest plots. Diamond indicates pooled data.
The timing of testing also deserves careful consideration; it is plausible that the timing of testing may have affected the diagnostic accuracy of saliva and nasopharyngeal swab testing differentially. In this review, studies with the highest discordant rates of positive findings for saliva and negative findings for nasopharyngeal swab paired samples included patients with repeated tests performed on multiple days. This finding suggests that saliva NAAT results may remain positive for longer than nasopharyngeal swab NAAT results. In fact, the 1 study we excluded from our meta-analysis, performed exclusively in patients who had clinically recovered from proven COVID-19, found a high rate of positive findings for saliva and negative findings for nasopharyngeal swab paired samples. This hypothesis has been supported by other studies that compared viral load in saliva and nasopharyngeal specimens and found that more viral RNA could be amplified from saliva than nasopharyngeal swabs and for longer periods. Given that NAAT does not necessarily imply presence of live virus, more research is needed to determine the role of saliva NAAT in following viral shedding time, and whether this approach overestimates the duration of infectiousness.

Last, other limitations of saliva NAAT may emerge with time and more widespread use. We therefore caution that any large-scale deployment of saliva NAAT should be accompanied by an ongoing rigorous quality control program. Despite these limitations, community testing of symptomatic or mildly symptomatic persons was well represented in our analyses, and a subgroup analysis limited to this population also demonstrated excellent diagnostic accuracy, implying that saliva NAAT could be deployed in a community setting with a fair level of confidence. Saliva testing could be particularly useful for larger screening efforts or in patients who are reliable enough to self-report to a testing center (e.g., health care workers). Our findings open the door for patient self-collected tests, which would make self-isolation easier to implement and increase community research capacity. In theory, because saliva samples are self-collected, this could also reduce occupational exposures for personnel collecting samples. This sampling strategy could also conceivably facilitate community prophylaxis and early treatment studies of COVID-19. Importantly, saliva NAAT allows for large-scale testing in high-risk populations that are difficult to reach using hospital-based approaches, an approach that has been deemed to be cost-effective when using nasopharyngeal NAAT. The use of techniques to address inefficiencies, such as saliva testing or pooling of test results, could reduce costs by as much as 40% and personnel by 20%.

The duration of infectiousness.

During a pandemic, when the pretest probability of infection is elevated and lack of exposure is unreliable in ruling out COVID-19, large community-based testing programs are required for an effective public health response. Although some questions remain about saliva NAAT diagnostic accuracy in certain populations and settings, especially hospitalized and critically ill patients, saliva NAAT yields a sensitivity and specificity comparable to nasopharyngeal swab NAAT in ambulatory patients presenting with minimal or mild symptoms. Saliva NAAT should therefore be prioritized for larger-scale deployment with prospective studies conducted by clinical microbiology laboratories and public health authorities.

ARTICLE INFORMATION

Accepted for Publication: December 5, 2020.
Published Online: January 15, 2021.

Correction: This article was corrected on March 1, 2021, to fix Ms Yao’s name and academic degree in the byline.

Author Affiliations: Division of Infectious Diseases, Department of Medicine, McGill University Health Centre, Royal Victoria Hospital, Montréal, Québec, Canada (Butler-Laporte, Lee); Department of Critical Care Medicine, National Institutes of Health, Clinical Center, Bethesda, Maryland (Lawandi); Centre for Outcomes Research, McGill University Health Centre, Montréal, Québec, Canada (Schiller, Yao, Dendukuri, McDonald, Lee); Clinical Practice Assessment Unit, Department of Medicine, McGill University, Montréal, Québec, Canada (McDonald, Lee); Division of General Internal Medicine, Department of Medicine, McGill University, Montréal, Québec, Canada (McDonald, Lee); Division of General Internal Medicine, McGill University, Montréal, Québec, Canada (McDonald, Lee);

Author Contributions: For the purposes of authorship, Drs McDonald and Lee contributed equally. Dr Butler-Laporte had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Conflict of Interest Disclosures: Dr Dendukuri reported receiving grants from the Canadian Institutes of Health Research for development of the Shiny app to perform bayesian diagnostic meta-analysis during the conduct of the study. No other disclosures were reported.

Funding/Support: This study was supported by a joint scholarship from the Fonds the Recherche du Québec-Santé and Quebec’s Ministry of Health and Social Services (Dr Butler-Laporte), by the Intramural Research Program of the National Institutes of Health, Clinical Center (Dr Lawandi), by the Canadian Institutes of Health Research grant PJT-156039 (Dr Dendukuri), and by research salary funds from the Fonds de Recherche du Québec-Santé (Drs McDonald and Lee).

Role of the Funder/Sponsor: The sponsors had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Disclaimer: The findings and conclusions in this study are those of the authors and do not necessarily represent the official position of the National Institutes of Health.

REFERENCES

© 2021 American Medical Association. All rights reserved.


12. Riley RD, Higgins JPT, Deeks JJ. Interpretation of random effects meta-analyses. BMJ. 2011;342:d549. doi:10.1136/bmj.d549
