Johann Holzmann - Comment posted on *medRxiv* on June 8, 2020

Thank you for making the pre-print accessible, I read it with great interest.

How do your findings regarding the presumptive false-positive rate of SARS CoV2 detection using RT-PCR relate with the very low RT-PCR positive rate as currently seen in many countries or regions with a very low prevalence of SARS CoV2?

For example Australia runs between 30.000 to 35.000 PCR test daily for the last month and only gets around 10 positive assays per day.

Other examples with a ratio of PCR assays per day to positive assays of around 600-2000:1 are Iceland, Greece, Croatia, Thailand and certain parts of Germany (eg Sachsen-Anhalt, Mecklenburg Vorpommern) or Austria (eg Tirol).

Wouldn't these data indicate a much lower false-positive rate than the one suggested in your manuscript?

Thank you again for making your research accessible

kind regards

Response from Andrew Cohen, June 15, 2020

Johann,

It's a good question, one that I've been pondering in regard to the recent mass testing conducted by China. In two cities, China reported 0.003% testing positive, an order of magnitude lower than your lowest example of about 0.03% positive in Australia.

I'll start with some general observations, discuss issues specific to China and Australia, and then review hypotheses that could explain the large gaps between the calculated positivity rates in these two countries and the false positive rates (FPRs) we reported from external quality assessments (EQAs).

<u>General observations:</u>
• In general, national data on testing and test results is not entirely reliable, both (1) because of errors and inaccuracies that relate to the complexity and variation in testing programs and in the determination of test results, and errors and delays in local results being reported and aggregated at the national level; and (2) because data on both testing and results have become highly politicized. The politicization has mostly worked in the direction of authorities wanting to overstate the extent of testing and understate the number of infected individuals. While these influences may sometimes lead to reporting of a lower than actual positivity rate, in most cases I doubt it could amount to an order of magnitude difference (but see comments below on China's data).

• We used the EQA data on other RT-PCR tests to suggest a range where we might expect the mean FPR for SARS-CoV-2 RT-PCR tests to fall. The FPR that we reported for each EQA was the mean of the results for individual participating laboratories. With an average of 85 laboratories participating in each EQA, some individual laboratories had a considerably higher and some a considerably lower FPR than the mean for that EQA. Similarly, some countries will have a higher and some a lower FPR than the global average, and we shouldn't be surprised if there are some individual countries with markedly different FPRs than an estimate derived from mean values in EQAs. Could some countries be an order of magnitude or more off of the global average FPR? That seems like a lot to me, but I don't have any data with which to assess this.

China:

China recently conducted mass testing in two cities, Wuhan and Mudanjiang, that had suffered large COVID-19 outbreaks. CCTV (Chinese state TV) reported that 9,899,828 people were tested in China, with 300 persons testing positive, all of them asymptomatic. In Mudanjiang it was reported that 650,000 people were tested with 19 persons, all asymptomatic, testing positive. In both cases the test positivity rate is 0.003%. China's COVID-19 data have often been questioned—questions that were sometimes raised with political motivations but often, I think, with a legitimate basis. There is good evidence that authorities denied, failed to report, and tried to squelch evidence of the disease during its early spread. A recent study argues that deaths from COVID-19 in Wuhan in February and March may have been more than an order of magnitude greater than the official figures. The Wuhan testing in particular was attended by a lot of publicity: the Chinese authorities presented it as testing everyone in a city of 10 million people within 10 days—in other words, presented it as a feat that the Chinese government would accomplish. Given the political/publicity context, I think it's possible that some effort was made—either in the design of the testing, its execution, or its reporting—to keep the number of positive results and the positivity rate low.

From press reports we know that a substantial amount of the testing was done with pooled or batch testing, with 5-10 samples per pool. This procedure involves taking an aliquot from each sample, combining each aliquot with 4-9 others in a pooled sample, and if the pooled sample tests positive then individually testing second aliquots taken from the 5-10 samples represented in the pool. This introduces an element of second confirmatory testing—because both the pooled sample and the subsequent individual aliquot have to test positive for the sample to be reported as positive—which would tend to reduce the effective FPR relative to the single-test rate. How much it would be reduced depends on what causes the false positives. If, for example, all false positives are due to contamination that occurs just before or during the amplification stage (that is, after the pooled sample is pooled and after the subsequent individual aliquot is taken from the initial sample), such that an overall false positive result requires two independent contamination events, then the effective FPR would be equal to the square of the single-test FPR. If the single test rate was 0.8% (the rate we used in our modeling), the effective rate would be 0.0064%, which is still larger than but within hailing distance of the 0.003% positivity rate from the mass testing of the two Chinese

cities. If, however, some contamination occurred at earlier stages in the test process, or was due to causes other than contamination, the reduction from the single-test rate would be less. Also, one news report stated that pooled samples were used in up to 25% of the tests, which would further limit the effective reduction in the overall FPR.

Another possible factor, if false positives are primarily caused by contamination, is that use of a less sensitive test would yield fewer false positives. The news articles don't report on what assay was used in the mass testing events, but there might have been logistical or expense reasons to use a less sensitive test, or, if the purpose of the mass-testing was to produce good news rather than accurate information, authorities might have been tempted to go with a less sensitive test.

Finally, there is some information in the EQAs that we reviewed that suggests that Chinese laboratories are simply better than most at avoiding false positives. Three of the four EQAs that involved more than 25 negative samples and reported no false positives were EQAs of Chinese laboratories (Zhou & Luo 2018, zero false positives in 49 negative samples; Wang et al. 2015, zero in 60 negative samples; and Zhang et al. 2016, zero in 168 negative samples). However, none of the EQAs provided data that allow us to compare the performance of Chinese laboratories to laboratories in other countries when analyzing the same panel.

Australia:
The Australian data you reference indicates a positivity rate of 0.029%-0.033%, an order of magnitude higher than the Chinese mass-testing rates, but still nearly two orders of magnitude lower than the median FPR value from the EQAs (2.3%). The 7-day average test positivity, calculated from daily data on the Our World in Data website, stays between around 0.03% and 0.05% from May 13 to the most recent data posted, June 11. This is consistent with your numbers. However, the Australian data on number of tests on the website is for "tests performed," not "persons tested"; the latter would be lower (because some people are tested more than once), yielding a higher test positivity. The limited data I have on this from other countries suggest that there are usually 1.2-1.5 tests conducted per person tested. Using the higher number to adjust the number of Australian tests yields positivities since May 13 of 0.04% to 0.08%.

According to the Australian case definition (as in most other countries) a single positive result from a nucleic acid test without any accompanying clinical signs or epidemiological indicators is sufficient to confirm a case of SARS-CoV-2 infection. However, early newspaper accounts suggest that at least some laboratories routinely checked positive results by sequencing, by culturing the virus, or by electron microscopy. It's not clear how widespread this was, whether it continued to be done, and whether an unsuccessful attempt at sequencing, culturing the virus, or finding the virus with electron microscopy meant that an individual with a positive PCR result was not counted as a confirmed case. Sometime between May 13 and June 4, a section on asymptomatic persons that test positive was added to Australia's guidance to Public Health Units on COVID-19, stating that the "veracity of the test" should be confirmed, which in some cases should include retesting. From my review, the guidance from most

national health ministries (with the exception of Norway's guidance after May 27) and from international health organizations like WHO do not contain recommendations to retest asymptomatic positive results for SARS-CoV-2. Also, on June 9 Australia's Public Health Laboratory Network issued a detailed statement on false positive PCR results in SARS-CoV-2 testing, recommending that asymptomatic individuals that test positive, any weak positive results, and any unexpected positive results should be retested with a second assay. Again (with the exception of Norway), I've not seen similar statements from health ministries or international health organizations addressing false positive PCRs in SARS-CoV-2 testing. All of this suggests that in Australia, more than in most countries, laboratories may have made efforts to confirm positive results, especially for asymptomatic persons, with either a second PCR test or an independent method of verification. If there is a significant amount of such confirmatory work it should produce a national FPR that is lower than the average FPR from a single PCR-based test.

Hypotheses:

1. The EQA test conditions weren't equivalent to normal laboratory working conditions, and produced higher FPRs. Thus could happen, for example, if several of the EQA managers produced and sent out for testing negative samples that had been contaminated with the RNA target. This seems unlikely. A possibly significant difference is that the panels usually contained more positive than negative samples, sometimes in a fairly high ratio (e.g. 4-6 positive samples with 1 negative sample). This may be higher than the usual ratio of positive to negative samples during normal work in a diagnostic laboratory, and if so might provide more potential for a negative sample to become contaminated. On the other hand, if laboratory staff knew which samples were from EQA panels they might have, consciously or not, paid closer attention when working on those samples, which would tend to improve performance. Overall, it doesn't seem likely that the EQAs would have produced systematically higher FPRs than would other diagnostic testing conducted in the same laboratories.

2. We calculated the FPRs in the EQAs incorrectly. In some cases information in the EQA reports was unclear and had to be interpreted, and this could have affected some numbers used in our calculations of FPRs. We were aware of this and consciously tried to avoid interpreting ambiguous reporting in a manner that consistently or significantly raised the calculated FPR. In most cases the numbers were clear enough, and I don't believe that the few cases where the information might reasonably be interpreted differently would significantly alter the calculated range of FPRs. At any rate, our work on this can be checked by comparing the EQA reports to our counts of negative samples and false positive results from those reports.

3. For some reason, the RT-PCR assays for SARS-CoV-2 produce a much lower (1-2 orders of magnitude lower) average FPR than the average FPRs from RT-PCR assays used to test for other viruses from 2004-2019. For the subset of 37 EQAs with more than 100 negative samples we found a statistically weak downward correlation between FPR and Year, but no correlation for the full set of 43 EQAs. From this we might anticipate a slightly lower average FPR for RT-PCR assays conducted in 2020 than for

RT-PCR assays conducted in 2004-2019, but nothing like an order of magnitude difference. Other than that, since the design of modern PCR-based tests has eliminated false positives due to cross-reactivity except in rare cases, the likely sources of false positives in PCR-based testing (contamination, and sample or data handling errors) are more directly connected to sampling and laboratory practices and layouts than to which particular assay is used. It's therefore unlikely that there would be a systematic difference between the FPRs among the scores of different assays used to detect SARS-CoV-2 and the hundreds of different assays in the EQAs.

4. The Chinese and Australian test data are incorrect. I think this is a possible explanation, or partial explanation, for the Chinese results, given the questions that have been raised about China's COVID-19 data, the Chinese government's ability to control the reporting of data, and the highly politicized nature of the two mass-testing programs. I think it's much less likely for Australia.

5. The SARS-CoV-2 test data used to calculate the positivity rate have different baselines, so that the calculated positivity rate is below the actual value. This appears to be the case in Australia, where the number of tests are reported on a per sample basis while the number of positive results are reported on a per person basis. Correcting for this could have a significant impact—potentially raising the calculated positivity rate by as much as 50% based on the limited data that I have—though probably not anything close to an order-of-magnitude impact. As near as I can tell this factor does not appear to an issue with the Wuhan and Mudanjiang data (both numbers of tests and numbers of positive results appear to be reported on a per person basis).

6. The SARS-CoV-2 test data are not each based on a single positive result from a nucleic acid assay. As discussed, from news reports this appears to be at least a significant partial explanation for the difference between the EQA FPRs and the reported Wuhan FPR, where there was significant use of pooled samples; and might explain the Mudanjiang and Australia results. More complete information on how testing was done, and how results were counted as positive cases, would resolve this.

7. The SARS-CoV-2 test data are based on non-PCR-based nucleic acid assays with inherently lower false positive rates. It's possible that some part of the testing in Wuhan, Mudanjiang or Australia was done with non-PCR-based methods, such as NEAR or LAMP assays. It's also possible that some of these non-PCR techniques are less likely to produce false; I believe I've seen some claims to that effect, but haven't investigated this question. More complete information on how testing was done, and on different assays' propensity to produce false positives, could resolve this.

8. Laboratories in China or Australia follow practices that produce much fewer false positives than the average laboratory elsewhere, or for other undetermined reasons are outliers in terms of their FPRs during the relevant periods. We should expect that some laboratories will have substantially lower FPRs than the average, and possibly that some countries' laboratories will have substantially lower FPRs than the average, and this could explain at least part of the difference between the average range of FPRs

suggested by the EQA results and the positivity rates in particular regions over a particular period. There are some EQA data that suggest this might be true for Chinese laboratories.

To sum up, hypotheses 1-3 seem unlikely to be true or at least unlikely to produce any significant part of the differences between the countries' reported positivity rates and the EQAs' FPRs. Hypothesis 4 seems a possible explanation or substantial part of the explanation for China, but not likely for Australia, and Hypothesis 5 for Australia (in part) but not China. Hypothesis 6 is probably at least a significant part of the explanation for China (pooled sampling in Wuhan). It might be an explanation for Australia, depending on the actual testing practices there; further investigation could resolve this. Hypothesis 7 could be investigated and resolved; I don't have enough information to venture an opinion about it. Hypothesis 8 could be a significant part of the explanation; it could be investigated, perhaps through EQAs, though they are not currently structured to investigate this sort of question. If true, then there might be something useful that the rest of the world's laboratories could learn from Chinese and Australian laboratories.

If readers of this reply have information that would help resolve the open issues with regard to hypotheses 6 (how testing was done and how test results were counted as positive cases in Wuhan, Mudanjiang and Australia) and 7 (the extent to which testing for viral RNA was conducted with non-PCR-based assays in Wuhan, Mudanjiang and Australia; and whether these non-PCR methods have inherently lower FPRs), please share.

Andrew Cohen