Estimating the number of undetected COVID-19 cases among travellers from mainland China [version 1; peer review: 3 approved with reservations]

Sangeeta Bhatia¹, Natsuko Imai¹, Gina Cuomo-Dannenburg¹, Marc Baguelin¹, Adhiratha Boonyasiri¹, Anne Cori¹, Zulma Cucunubá¹, Ilaria Dorigatti¹, Rich FitzJohn¹, Han Fu¹, Katy Gaythorpe¹, Azra Ghani¹, Arran Hamlet¹, Wes Hinsley¹, Daniel Laydon¹, Gemma Nedjati-Gilani¹, Lucy Okell¹, Steven Riley¹, Hayley Thompson¹, Sabine van Elsland¹, Erik Volz¹, Haowei Wang¹, Yuanrong Wang¹, Charles Whittaker¹, Xiaoyue Xi¹, Christl A. Donnelly¹,², Neil M. Ferguson¹

¹MRC Centre for Global Infectious Disease Analysis, Department of Infectious Disease Epidemiology, Imperial College London, London, UK
²Department of Statistics, University of Oxford, Oxford, UK

Abstract
Background: Since the start of the COVID-19 epidemic in late 2019, there have been more than 152 affected regions and countries with over 110,000 confirmed cases outside mainland China.

Methods: We analysed COVID-19 cases among travellers from mainland China to different regions and countries, comparing the region- and country-specific rates of detected and confirmed cases per flight volume to estimate the relative sensitivity of surveillance in different regions and countries.

Results: Although travel restrictions from Wuhan City and other cities across China may have reduced the absolute number of travellers to and from China, we estimated that more than two thirds (70%, 95% CI: 54% - 80%, compared to Singapore; 75%, 95% CI: 66% - 82%, compared to multiple countries) of cases exported from mainland China have remained undetected.

Conclusions: These undetected cases potentially resulted in multiple chains of human-to-human transmission outside mainland China.

Keywords
Epidemiology, COVID-19, novel coronavirus, SARS-CoV-2, international open peer review
This article is included in the Coronavirus (COVID-19) collection.
Background
As of 18th March 2020, over 80,000 cases of COVID-19 (formerly 2019-nCoV) have been reported in China (with 3231 deaths), and more than 110,000 cases have been detected in 152 regions and countries outside mainland China (including Hong Kong SAR and Macau SAR)\(^1\). Several analyses have been undertaken to predict or estimate the risk of exported cases by country on the basis of flight connections between Wuhan City, China or mainland China as a whole and other regions and countries\(^2\). In this analysis we built on published work\(^3\) to analyse COVID-19 cases reported and confirmed in different countries that were exported from mainland China, comparing the region- and country-specific rates of detected cases per flight volume to estimate the relative sensitivity of surveillance in different countries. We then estimate the number of COVID-19 cases exported from mainland China that have remained undetected worldwide.

Methods
Data sources
Air traffic volume. Air travel data for the months of January, February, and March 2016 were obtained from the International Air Travel Association (IATA), with the sum divided by three to get destination-region- and destination-country-specific monthly averages. These numbers were not scaled up to reflect recent growth in air travel because any constant scaling of the monthly averages would simply be absorbed into the \(\lambda\) estimate and not affect other results. Flows of passengers within mainland China were excluded from this analysis.

Number of cases detected outside mainland China. We collated data on 3276 cases in international travelers from media reports and provincial and national department of health press releases up until 27 February 2020\(^1\).\(^4\).\(^5\). We defined a local transmission as any transmission that occurred outside mainland China (Hong Kong SAR and Macau SAR are considered outside mainland China for this analysis). We only consider cases that were not transmitted locally. That is, we only considered cases detected outside mainland China that had a travel history to China and arrived outside mainland China by air, excluding repatriation flights (Table 1). Based upon these inclusion criteria, a total of 173 cases were included in our analysis. The earliest date of travel for the cases included in the analysis is 1 January 2020, and the latest date of travel is 25 February 2020.

Analysis
We assume that the number of exported cases in a country \(i\) is Poisson distributed with a mean that depends on the air traffic from Wuhan to \(i\), and the sensitivity of surveillance in \(i\) relative to a country \(j\), denoted by \(s_{ij}\). For each country \(i\), let \(X_i\) be the number of exported cases (a count) and let \(F_i\) be the volume of air traffic from Wuhan to country \(i\). We can then write a joint log likelihood for the data from countries \(i\) and \(j\):\[ l = X_i \ln(s_{ij} \lambda F_i) - s_{ij} \lambda F_i + X_j \ln(\lambda F_j) - \lambda F_j \]
ignoring additive constants. Thus, the maximum likelihood estimates for \(\lambda\) and \(s_{ij}\) are:
\[
\hat{\lambda} = \frac{X_i}{F_i} \quad \text{and} \quad \hat{s}_{ij} = \frac{X_j F_j}{X_i F_i}
\]
The likelihood-based confidence intervals are obtained by calculating the maximum log likelihood (over values of \(\lambda\)) for each value of \(s_{ij}\). Then the 95% confidence interval includes all those values of \(s_{ij}\) such that \(2 \hat{\lambda} X_i - X_j \ln(\hat{\lambda} F_j) - \lambda F_j \leq 3.84\) (the 95th centile of the chi-squared distribution with 1 degree of freedom). These calculations were all performed using R version 3.6.0.

The relative sensitivities can also be estimated relative to \(J\) countries simultaneously using a method similar to above but with the log likelihood:
\[
l = X_i \ln(s_{ij} \lambda F_i) - s_{ij} \lambda F_i + \sum_{j \neq i} X_j \ln(\hat{\lambda} F_j) - \lambda F_j
\]
Expected values can then be calculated for every country \(i\) as simply \(\hat{s}_{ij} F_i\), and the expected value for all countries \(\sum_{i \neq j} F_i\), where \(N\) is the total number of countries with air traffic from Wuhan Tianhe International Airport (\(N = 119\)).

Results
The number of exported cases by country was plotted as a function of the average monthly passenger volume originating from Wuhan Tianhe International Airport on international flights (Figure 1). This showed Singapore to be an outlier in terms of having relatively many exported cases compared to the measure of air traffic volume.

The relative sensitivity of surveillance in individual countries was estimated compared to Singapore, Finland, Nepal, Philippines, Sweden, India, Sri Lanka, and Canada were all found to have relative sensitivity estimates greater than 1 (i.e. more cases were detected per passenger flight than in Singapore). Thus, a second set of relative sensitivity estimates was obtained for all other individual countries compared simultaneously to Singapore, Finland, Nepal, Philippines, Sweden, India, Sri Lanka, and Canada.

The region- and country-specific expected numbers of exported COVID-19 cases were in several cases substantially higher than the numbers detected (Figure 2). The sum of the expected numbers of exported COVID-19 cases for all regions and countries other than mainland China was 576.8 (95% CI: 372.2 - 845.4), based on the analysis relative to Singapore only, and 704.4 (95% CI: 510.3 - 942.3), based on the analysis relative to Singapore, Finland, Nepal, Philippines, Sweden, India, Sri Lanka, and Canada. Given that 173 such cases were detected, these central estimates suggest that between 70% (95% CI: 54% - 80%, relative to Singapore only) and 75%...
Figure 1. Exported COVID-19 cases vs average air traffic from Wuhan Tianhe International Airport by destination. The number of exported COVID-19 cases detected by region and country plotted against the average monthly international air traffic volume from Wuhan Tianhe International Airport aggregated by destination country. The colour of the points denotes the continent of the destination country (Asia - orange, Europe - light blue, Africa - green, North America - dark blue, South America - pink, and Oceania - dark orange).

Table 1. Number of cases detected outside mainland China with travel history to China.

<table>
<thead>
<tr>
<th>Country</th>
<th>Travel History to Hubei</th>
<th>No Travel History to Hubei</th>
<th>Unknown Travel History within China</th>
<th>Total (cases with a travel history to China)</th>
<th>Travelled by air (not repatriation flight)</th>
<th>Travelled on repatriation flight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Australia</td>
<td>15</td>
<td>0</td>
<td>0</td>
<td>15</td>
<td>15</td>
<td>0</td>
</tr>
<tr>
<td>Belgium</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Cambodia</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Canada</td>
<td>7</td>
<td>1</td>
<td>0</td>
<td>8</td>
<td>8</td>
<td>0</td>
</tr>
<tr>
<td>Finland</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>France</td>
<td>5</td>
<td>0</td>
<td>1</td>
<td>6</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>Germany</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Hong Kong SAR</td>
<td>12</td>
<td>3</td>
<td>0</td>
<td>15</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>India</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>Italy</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>Japan</td>
<td>24</td>
<td>4</td>
<td>0</td>
<td>28</td>
<td>28</td>
<td>0</td>
</tr>
<tr>
<td>Macau SAR</td>
<td>5</td>
<td>3</td>
<td>0</td>
<td>8</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Malaysia</td>
<td>9</td>
<td>4</td>
<td>0</td>
<td>13</td>
<td>9</td>
<td>0</td>
</tr>
<tr>
<td>Nepal</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Philippines</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>Singapore</td>
<td>21</td>
<td>3</td>
<td>0</td>
<td>24</td>
<td>23</td>
<td>1</td>
</tr>
<tr>
<td>South Korea</td>
<td>12</td>
<td>2</td>
<td>0</td>
<td>14</td>
<td>12</td>
<td>1</td>
</tr>
<tr>
<td>Sri Lanka</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Sweden</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Taiwan</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>8</td>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>Thailand</td>
<td>14</td>
<td>3</td>
<td>5</td>
<td>22</td>
<td>19</td>
<td>0</td>
</tr>
<tr>
<td>United Arab Emirates</td>
<td>0</td>
<td>0</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>United Kingdom</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>United States of America</td>
<td>10</td>
<td>2</td>
<td>1</td>
<td>13</td>
<td>13</td>
<td>0</td>
</tr>
<tr>
<td>Vietnam</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>162</td>
<td>26</td>
<td>16</td>
<td>204</td>
<td>173</td>
<td>3</td>
</tr>
</tbody>
</table>
(95% CI: 66% - 82%, relative to Singapore, Finland, Nepal, Philippines, Sweden, India, Sri Lanka, and Canada) remained undetected.

Discussion
Consistent with similar analyses\(^4\), we estimated that more than two thirds of COVID-19 cases exported from mainland China have remained undetected worldwide, potentially leaving sources of human-to-human transmission unchecked (70%, 95% CI: 54% - 80% and 75%, 95% CI: 66% - 82%, undetected, based on comparisons to Singapore only and to Singapore, Finland, Nepal, Philippines, Sweden, India, Sri Lanka, and Canada, respectively). Undoubtedly, the exported cases vary in the severity of their clinical symptoms, making some cases more difficult to detect than others. However, some countries have detected significantly fewer than would have been expected based on the volume of flight passengers arriving from Wuhan City, China. These undetected cases potentially resulted in multiple chains of human-to-human transmission outside mainland China.

Data availability
Source data
The air travel data used in this analysis can be purchased from International Air Transport Association (IATA) via the following link: https://www.iata.org/en/services/statistics/air-transport-stats/.

Underlying data

This project contains the following underlying data:
- exported_cases.csv (information on the date of report, country of report and travel history of 3,276 cases outside mainland China)

Extended data

This project contains the following extended data:
- data_processing.R (R code to post-process international case data)

Data are available under the terms of the Creative Commons Zero “No rights reserved” data waiver (CC0 1.0 Public domain dedication).

References
This is a well-done analysis that estimates the number of unreported cases that were imported from China early on in the pandemic. It does this using data on air travel volume between China and other countries, and the reported numbers of cases in the places that reported the highest numbers of cases given their air traffic volume. This analysis was highly relevant early in the pandemic as infections spread from China.

I have a few comments on points for clarification below.

Abstract:
The statement of the results about 70/75%... in the abstract is confusing. Suggest rephrasing.

Conclusion: I wonder if this is a conclusion from the paper. I would suggest the addition here that the analysis leads to estimates that there were many unreported imported infections, and that potentially lead to transmission.

Main text:
Background: It would be helpful to have a statement about what the previous analysis in reference 4 did/showed.

Methods: Please add more detail on how airports in China were used in the flow calculation, and also on the definition of a destination region. In the results section initially Wuhan is focused on but the general conclusions seem to be from all of mainland China. Please clarify throughout. Please add more detail from where the collated data on imported cases was obtained. Were all excluded cases excluded because they were defined as local or due to missing information on this? Was data available on which location was travelled from within China for the imported cases? If not, how was this dealt with in the analysis?
Results:
Figure 2 legend, are the numbers shown relative to Singapore numbers, or is the analysis done relative to Singapore and then the estimates of imported cases shown? At the moment, the legend reads as the former, but my understanding of the analysis is that is it the latter. Please clarify.

Discussion: Please add on limitations of the analysis, in particular how this relates to the available data including classification as imported vs local, and that this data needed to be publicly available.

Is the work clearly and accurately presented and does it cite the current literature?  
Partly

Is the study design appropriate and is the work technically sound?  
Yes

Are sufficient details of methods and analysis provided to allow replication by others?  
Partly

If applicable, is the statistical analysis and its interpretation appropriate?  
Yes

Are all the source data underlying the results available to ensure full reproducibility?  
Yes

Are the conclusions drawn adequately supported by the results?  
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Infectious disease epidemiology and dynamics

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 12 Aug 2021

Christl Donnelly, Imperial College London, London, UK

1. The statement in the abstract has been rephrased for clarity.
2. The conclusion in the abstract has been edited as suggested.
3. We have added a sentence about the analysis and conclusions of Reference 4.
4. Destination-regions refer to Hong Kong SAR and Macau SAR, this clarification has been added to the text.
5. We have now edited the Results and the Discussion sections to emphasise that we are estimating the number of exported cases from Wuhan (rather than China).
6. We have expanded the section on data collation to clarify the points raised by the
reviewer. The inclusion criteria have been elaborated to clarify when a case was excluded.

7. As the reviewer has rightly noted, the numbers in figure 2 are the estimates of imported cases. The legend has been edited to reflect this.

8. We have expanded the discussion to highlight the limitations and potential biases of the analysis.

**Competing Interests:** No competing interests were disclosed.

---

**Reviewer Report 31 July 2020**

https://doi.org/10.21956/wellcomeopenres.17332.r39133

© 2020 Funk S. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

**Sebastian Funk**
Centre for Mathematical Modelling of Infectious Diseases, London School of Hygiene & Tropical Medicine, London, UK

This manuscript aims to quantify the amount of underdetection of cases outside of China during the early phase of the COVID-19 pandemic. It assumes that the numbers of cases are centred around the product of the number of flights from Wuhan, an estimated per-country parameter $\lambda$, and an estimated relative per-country surveillance capacity.

Overall the methodology is sensible and the findings of interest at the time.

Comments:
- The manuscript is missing detail on how the 3276 cases were collated. Was this done systematically and if so how? Were non-English reports collated and was there any attempt to correct for language biases? Have you got references for the cases and, if yes, could they be added, e.g., to the csv file?

- The index could be more consistent, e.g. it seems $\lambda = X_j/F_j$ should have an index $j$, and $s$ should have indices $i$ and $j$ (and no $e$).

- It would be interesting to see the results here compared to Golding et al., https://www.medrxiv.org/content/10.1101/2020.07.07.20148460v1, which had the same aim but used a different methodology/data - also, to compare to other estimates of underdetection in the relevant countries.

- A thorough discussion of potential biases/limitations is missing.
References

Is the work clearly and accurately presented and does it cite the current literature?
Partly

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Partly

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Partly

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: infectious disease dynamics

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 12 Aug 2021

Christl Donnelly, Imperial College London, London, UK

1. The section on data collation has been expanded to clarify the points raised by the reviewer.
2. Thanks, we have edited the indices.
3. A direct comparison with the results of Golding et al is difficult because they present country-specific ascertainment estimates while the goal of our analysis was to estimate the number of undetected COVID-19 cases globally and we have therefore not provided per-country estimates of surveillance sensitivity.
4. We have expanded the discussion to highlight the limitations and potential biases of the analysis.

Competing Interests: No competing interests were disclosed.
Summary: Bhatia et al. estimate the numbers of undetected COVID-19 cases exported from Mainland China based on historical air traffic patterns and public data on COVID-19 cases imported to other countries that originated in China. The observed number of cases imported to each country is assumed to be Poisson distributed, and the expected number of imported cases relative to Singapore (used as a reference country because of its high ratio of detected cases per flight volume) is derived using standard maximum likelihood estimation methods. A similar procedure is used to compare the expected number of imported cases relative to a set of reference countries which had estimated sensitivities of surveillance higher than that of Singapore. The main finding is that at least 2/3 of cases exported from China remained undetected worldwide.

Issues and comments:

**Major**

The code provided does not run out-of-the-box. In data_processing.R `exported_cases_paper_test.csv` should be `exported_cases.csv` and the line defining the total number of exported cases should be moved after the definition of `exported`.

Furthermore, only the data processing script for cases in international travelers is provided, not the code for data analysis or visualization of results.

While we understand that the authors cannot share the IATA data, they should at least provide results sufficient to reproduce Figure 2, and preferably all code for analysis and visualization such that anyone with access to the data could reproduce the results.

**Minor**

There is a disconnect between the use of passenger flow data from only WTIA and analysis of cases thought to have acquired infection in mainland China, regardless of whether they travelled through Wuhan or Hubei; some explanation of the decision to use these inconsistent definitions is warranted.
Please clarify why repatriation flights have been excluded when selecting cases for analysis.

At the beginning of the analysis section, “We assume that the number of exported cases in a country $i$...” should read “We assume that the observed number of exported cases in a country $i$...”. This distinction should be clarified throughout.

Some mention could be made of the reason for using 2016 flight data rather than more recent data.

Some mention of potential biases and pitfalls in methods and the data would be appropriate.

There are a small number of relevant, recently published works not mentioned - e.g. https://www.medrxiv.org/content/10.1101/2020.03.23.20038331v2 and https://doi.org/10.1016/S0140-6736(20)30411-6.

The background in the abstract is very out of date; suggest updating numbers and adding a date marker (‘as of...’ or similar)

Additional labels for the figures would be useful (beyond the select countries labeled in Figure 1). Alternatively, per-country estimates could be made available as a table.

In the methods section (‘Data Sources’) the parameter lambda is referred to before being introduced.

In the discussion “Consistent with similar analyses” has only one citation.

References

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Partly

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Partly

Are the conclusions drawn adequately supported by the results?
Yes

**Competing Interests:** No competing interests were disclosed.

**Reviewer Expertise:** infectious disease epidemiology and modelling

We confirm that we have read this submission and believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however we have significant reservations, as outlined above.

Author Response 12 Aug 2021

Christl Donnelly, Imperial College London, London, UK

1. We apologise for the omission of relevant code. We have now added all relevant code so that someone with relevant data can reproduce the results. We have also included a file with dummy data on travel volume to help the readers. The README file has also been updated to include instructions on running the code.

2. Everyone we classified as cases detected overseas had travel by air either explicitly mentioned, or implied as the most probable mode of travel from mainland China to the destination (e.g. from China to Italy). Where multiple modes of travel are possible e.g. from mainland China to Hong Kong, we have only classified individuals as cases detected overseas where the mode of travel was explicitly mentioned as air. The text has been updated to clarify this.

3. In most instances, all passengers on repatriation flights had been tested for the presence of SARS-CoV2. The cases detected through surveillance of repatriation flights are therefore not representative of the typical sensitivity of surveillance in a country. We have therefore excluded these from the analysis. The text has been updated to clarify this.

4. This sentence has now been edited and the distinction has been emphasised in the rest of the text.

5. The data from 2016 were the most recent data to which we had access when undertaking our analysis. Further, any constant scaling of the volume of passengers would not affect the estimates of model parameters (lambda and s_e). This has been emphasised in the text.

6. We have included the limitations of the method and data sources in the discussion.

7. Thanks for highlighting these relevant references. We have now included reference to these in the section Background.

8. The abstract and the reference were updated as of August 2021.

9. The reference to lambda has been removed from the methods section and a reference added to the appropriate section.

10. We have added reference to other studies conducted at the time which provide estimate surveillance sensitivity globally.

**Competing Interests:** No competing interests were disclosed.