Infectivity, susceptibility, and risk factors associated with SARS-CoV-2 transmission under intensive contact tracing in Hunan, China

Shixiong Hu, Wei Wang, Yan Wang, Maria Litvinova, Kaiwei Luo, Lingshuang Ren, Qianlai Sun, Xinghui Chen, Ge Zeng, Jing Li, Lu Liang, Zhihong Deng, Wen Zheng, Mei Li, Hao Yang, Jinxin Guo, Kai Wang, Xinhua Chen, Ziyan Liu, Han Yan, Huilin Shi, Zhiyuan Chen, Yonghong Zhou, Kaiyuan Sun, Alessandro Vespignani, Cécile Viboud, Lidong Gao, Marco Ajelli, Hongjie Yu

- ## Getting started

The R scripts are written assuming you have the following folder structure:

```
NatComm_COVID-19_Hunan
│       0. README.pdf
└─── 1. input
└─── 2. code
└─── 3. output
```

Where
- All data needed to perform the analyses is reported in folder '1. input';
- The code is reported in folder '2. code';
- Tables and Figures produced by the code will be stored in folder '3. output'.

Note that 'data_Table S3', 'data_Table S4', 'data_Table S5', 'data_Figure 1', 'data_Figure S2', 'data_Figure S4', and 'data_Figure S6' contain the aggregated data from the official line list and contact tracing data of COVID-19 surveillance system in Hunan province. However, as the original database contains confidential patient information it cannot be made public. Therefore, we generated a mock dataset ('data_model') that can be used to test the code. Please note that no aggregated statistics of the original dataset were used to generate the artificial one, and the

results obtained by running the analysis generated on the mock dataset will differ from those presented in the manuscript.

• Software and packages

The code was written in R (version 3.6.3). The following packages are needed to run the scripts:

- readxl
- data.table
- dplyr
- tidyr
- reshape2
- openxlsx
- fitdistrplus
- lme4
- DHARMa
- broom.mixed
- ggeffects
- splines
- mgcViz
- mgcv
- oddsratio
- itsadug
- ggplot2

• Analysis

1. Sample description

Aggregated data used for descriptive analyses are presented in Figure 1a/1d, Figure S2 and Figure S4, while Figure 1b to 1c, Table 1, Table S2, Figure S1, Figure S3 represent data

including confidential patient information which cannot be made public. Aggregated data are provided in files "data_Figure_1.xlsx", "data_Figure_S2.xlsx", and "data_Figure_S4.zip". Table 1, Figure S1 and Table S2 present descriptive analysis generated simply by using base R function "table". The code to reproduce Figure 1, and Figures S2 - S4 is provided in scripts "Code_Figure_1.R", "Code_Figure_S2.R", "Code_Figure_S3.R" and "Code_Figure_S4.R".

Variable list in file "data_Figure_1.xlsx":

- onset_date: date of illness onset , format: YYYY/MM/DD
- type_a: case type. Values: 1="asymptomatic cases";2="locally-acquired cases"; 3="travel-related cases"
- total: number of cases with data on the date of illness onset or lab-confirmation

Variable list in file "data_Figure_S2.xlsx":

- onset_date: date of illness onset or lab-confirmation, format: YYYY/MM/DD
- type_a: case type. Values: 1="asymptomatic cases";2="clustering cases"; 3="sporadic cases"
- total: number of cases

Variable list in file "data_Figure_S4.zip":

- age: age of infectors. Values: 1="0-14";2="15-64"; 3="65+"
- clinic: clinical severity. Values: 1="mild and general cases";2="severe and critical cases"
- common: whether cases share the same exposure or a unequivocal human-to-human transmission can be identified. Values: 1="shared same exposure"; 2="human-to-human transmission"
- gender: gender. Values: 1="Male"; 2="Female";
- group: Type of the infected individual with whom a close contact occurred. Values: 1="close contacts of symptomatic cases"; 2="close contacts of asymptomatic subjects"
- cluster: Identifiability of the transmission chain a case belongs to. Values: 1="sporadic/cluster index cases"; 2="cluster successive cases"

- Travel: travel relatedness of the infection. Values: 1="travel-related cases";2="locally-acquired cases"
- no: number of cases
- percent: proportion of cases in %

2. Time-to-key-event distribution and pre-symptomatic transmission

2.1 Incubation period

The data and code used to fit the distribution of incubation period are available in "data_Table S3.xlsx" and "Code_Table_S3.R", respectively. There are three sheets in "data_Table S3.xlsx". Sheet "Main analysis" contains 268 records for the main analysis. Sheet "Sensitivity analysis 1" and "Sensitivity analysis 2" contains 258 and 251 records for the sensitivity analyses (Supporting Information Tab. S3).

Variable list in file "data_Table S3.xlsx":

- left: the lower bound of the incubation period, which was calculated by subtracting the date of the last possible exposure from the date of symptom onset of the cases
- right: the upper bound of the incubation period, which was calculated by subtracting the date of their first possible exposure from the date of symptom onset of the cases

2.2 Serial interval

The data and code used to fit the distribution of serial interval are available in "data_Table S4.csv" and "Code_Table_S4.R", respectively.

Variable list in file "data_Table S4.csv":

- id: the identification number of the COVID-19 case
- x.lb: the dates of symptom onset of the first infector with respect to a reference date ("2020-01-01")

- x.ub: the dates of symptom onset of the last infector with respect to a reference date ("2020-01-01")
- y: the date of symptom onset of the infectee with respect to a reference date ("2020-01-01")
- prd: the period of the epidemic, where "prd=1" refers to the interval from January 5 to January 23 and "prd=2" refers to the interval from January 24 to April 2

2.3 Infectiousness profile and pre-symptomatic transmissions

The data and code used to fit the infectiousness profile are available in "data_Table S4.csv" and "Code_Infectiousness profile.R"

2.4 Generation time

The data and code used to fit the distribution of generation time are available in "data_Table S3.xlsx" and "Code_Generation time.R"

2.5 Other key time-to-event intervals

The data used to fit the distribution of time interval from symptom onset to the date of collection of the sample for PCR testing is available in "data_Table_S5.xlsx" (Sheet="data_onset_smp"). The data used to fit the distribution of time interval from symptom onset to laboratory confirmation is available in "data_Table_S5.xlsx" (Sheet="data_onset_diag"). The code to fit these two intervals is shown in "Code_Table_S5.R"

Variable list in file "data_Table_S5.xlsx":

- id: the identification number of the COVID-19 case
- int_onset_samp: time interval from symptom onset to the date of collection of the sample for PCR testing (in days)
- int_onset_diag: time interval from symptom onset to laboratory confirmation (in days)

3. Transmission chain in all the clusters showing evidence of symptomatic and asymptomatic SARS-CoV-2 transmission

The original database containing confidential patient information cannot be made public, but an illustration of the deidentified data subsamples is provided in Figure 3 and Figure S5.

4. SARS-CoV-2 risk factors

4.1 age-specific matrix

The data and code to reproduce Figure S6 are available in "data_Figure_S6.csv" and "Code_Figure_S6.R".

Varible list in file "data_Figure_S6.csv":

- Age_index: age of infector
- Age_contact: age of contact
- case: the total number of infections caused by an infector of a given age among the contacts of a given age
- mean: the mean number of infections caused by an infector of a given age among the contacts of a given age

4.2 Summary of contact tracing data by age of infectors/contacts and generation of transmission

The original database containing confidential patient information cannot be made public, but description of the data is provided in tables S6-7 (generated by using base R function "table").

4.3 GLMM and GAMM models

As the original database containing confidential patient information cannot be made public, we have created artificially generated mock database (which do not reflect the reality) to demonstrate the underlying data structure. All code and mock database to reproduce Table 2 in

the main text, and Figures S7-S10 and Tables S8-S13 in the supplementary information are provided in "Code_Figure_model.R" and "data_model.csv" files, respectively.

Varible list in file "data_model.csv":

- id_case: case identification number
- cluster_id: cluster identification number
- Infectee: whether a contact has been infected by a given infector.
- agegroup_index: age of infector
- agegroup: age of contacts
- contact_type1: type of contacts
- generation_y: generation of infector in the transmission chain
- no.persons: number of close contacts made by such infector
- gender_index: gender of infector
- gender: gender of contacts
- clinic_index2: clinical severity of infector
- logage_index: log-transformed age of infector
- logage: log-transformed age of contacts
- observ_period2: period of observation

Code authors

– Wei Wang
– Yan Wang
– Maria Litvinova
– Marco Ajelli

Acknowledgments

He X., et al (https://github.com/ehylau/COVID-19/blob/master/Fig1c_Rscript.R)