# Supplementary Information for “Using social media data to assess the impact of COVID-19 on mental health in China”

Yongjian Zhu1 Liqing Cao2 Jingui Xie3 Yugang Yu1 Anfan Chen4 Fengming Huang2

1School of Management, University of Science and Technology of China, Hefei, China.

2The First Affiliated Hospital of USTC, University of Science and Technology of China, Hefei, China

3School of Management, Technical University of Munich, Heilbronn, Germany.

4School of Humanity and Social Science, University of Science and Technology of China, Hefei, China.

**Email addresses:**

ustczyj@mail.ustc.edu.cn (Yongjian Zhu)

caoliqing@ustc.edu.cn (Liqing Cao)

jingui.xie@tum.de (Jingui Xie)

ygyu@ustc.edu.cn (Yugang Yu)

caf16@ustc.edu.cn (Anfan Chen)

3234119867@qq.com (Fengming Huang)

**Corresponding author:** Liqing Cao

**Contents:**

Supplementary Figure 1~Figure2

Supplementary Table 1 ~ Table 5



**Supplementary Figure 1 | The** **time trend of the cumulative number of treated cities.** The cumulative number of cities that have reported COVID-19 cases from January 1, 2020 to March 1, 2020 are shown. Human-to-human transmission was declared on January 20, 2020.



**Supplementary Figure 2 | The** **effect of COVID-19 on mental health over time using mean sentiment value as the dependent variable.** The estimated coefficients from equation (2) and their 95% confidence intervals (error bars) are shown. The dummy variable indicating one week before the occurrence of COVID-19 is omitted from the regression (see Methods). Thus, the difference in mental health status between treated and control cities one week before the treatment is set to be zero and serves as the reference point. The estimation signifies the difference in mental health status in each period relative to the difference one week before the treatment.

|  |
| --- |
| **Supplementary Table 1. Summary statistics** |
|  | All cities(1) | Treatment group | Control group(4) |
| Before treatment(2) | After treatment(3) |
| *Dependent variable* |  |  |  |  |
| City-level mental health status | 0.6397(0.0684) | 0.6535(0.0633) | 0.6311(0.0618) | 0.6363(0.1043) |
| *Independent variable* |  |  |  |  |
| COVID-19 | 0.5441(0.4981) | 0(0) | 1(0) | 0(0) |
| *Control variables* |  |  |  |  |
| AQI | 78.1637(52.2649) | 94.8988(61.7197) | 68.3883(40.0593) | 71.3907(60.0207) |
| Mean temperature (°C) | 4.4893(9.9080) | 3.2688(10.4081) | 6.2043(8.9934) | -0.7453(10.4503) |
| Wind speed (m/s) | 1.6052(1.1122) | 1.4402(1.0115) | 1.7470(1.1816) | 1.4145(0.9365) |
| Rainfall (mm) | 1.3692(4.8751) | 1.3008(4.4224) | 1.5826(5.4289) | 0.4021(2.3889) |
| Cloud (%) | 61.8355(36.1747) | 66.9818(34.8240) | 59.6237(37.0007) | 54.6302(33.7333) |
| Each column summarizes the mean values and standard deviations of different variables in our panel data. |

|  |
| --- |
| **Supplementary Table 2. The effect of COVID-19 on mental health after excluding cities in Hubei province** |
|  | (1) | (2) |
| COVID-19 | -0.0082\*\*(0.0028) | -0.0090\*\*(0.0028) |
| Air pollution and weather conditions |  | Yes |
| City fixed effects | Yes | Yes |
| Date fixed effects | Yes | Yes |
| Observations | 20,906 | 18771 |
| R2 | 0.4662 | 0.5541 |
| Due to some missing values of air pollution and weather data, the numbers of observations in the two columns are not the same. Standard errors are clustered at the city level and shown in parentheses. \**P* < 0.05; \*\**P* < 0.01; \*\*\**P* < 0.001. |

|  |
| --- |
| **Supplementary Table 3. The effect of COVID-19 on mental health using mean sentiment value as the dependent variable** |
|  | (1) | (2) |
| COVID-19 | -0.0071\*\*(0.0027) | -0.0084\*\*\*(0.0027) |
| Air pollution and weather conditions |  | Yes |
| City fixed effects | Yes | Yes |
| Date fixed effects | Yes | Yes |
| Observations | 21,882 | 19381 |
| R2 | 0.5070 | 0.5947 |
| Due to some missing values of air pollution and weather data, the numbers of observations in the two columns are not the same. Standard errors are clustered at the city level and shown in parentheses. \**P* < 0.05; \*\**P* < 0.01; \*\*\**P* < 0.001. |

|  |
| --- |
| **Supplementary Table 4. Assessing the effect of COVID-19 on mental health at the tweet level** |
|  | (1) | (2) | (3) |
| COVID-19 | -0.0084\*\*\*(0.0014) | -0.0087\*\*\*(0.0014) | -0.0089\*\*\*(0.0011) |
| Air pollution and weather conditions |  | Yes | Yes |
| User fixed effects |  |  | Yes |
| City fixed effects | Yes | Yes | Yes |
| Date fixed effects | Yes | Yes | Yes |
| Observations | 13,478,142 | 12,983,153 | 11,895,215 |
| R2 | 0.0165 | 0.0168 | 0.2958 |
| The numbers of observations in columns (2) and (3) are different since the users only posting one tweet were dropped in column (3). Standard errors are clustered at the city level and shown in parentheses. \**P* < 0.05; \*\**P* < 0.01; \*\*\**P* < 0.001. |

|  |
| --- |
| **Supplementary Table 5. Randomized occurrence time of COVID-19 (Placebo Test)** |
|  | Randomization  |
|  | On treated cities(1) | On all cities(2) |
| Mean of $β\_{pseudo}$ | -0.0006 | 0.0000 |
| standard deviation of $β\_{pseudo}$ | 0.0014 | 0.0015 |
| Estimated $β$ | -0.0097 | -0.0097 |
| Replication times | 1000 | 1000 |
| Z-score | 211.4830 | 203.2985 |
| *p*-value | <0.0001 | <0.0001 |
| The randomization refers to the procedure of randomly assigning COVID-19’s pseudo presence to treated cities and all cities, respectively, with 1,000 times of repetition. $β\_{pseudo}$ is the coefficient for COVID-19’s pseudo presence, and $β$ is the true coefficient for COVID-19’s occurrence reported in Table 1; both were estimated using equation (1). |