Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
To avoid the noncausal association between environmental factor and COVID-19 when using aggregated data: Simulation-based counterexamples for demonstration

Shi Zhao *
JC School of Public Health and Primary Care, Chinese University of Hong Kong, Hong Kong, China
CUHK Shenzhen Research Institute, Shenzhen, China

A B S T R A C T
In the infectious disease epidemiology, the association between an independent factor and disease incidence (or death) counts may fail to infer the association with disease transmission (or mortality risk). To explore the underlying role of environmental factors in the course of COVID-19 epidemic, the importance of following the epidemiological metric’s definition and systematic analytical procedures are highlighted. Cautiousness needs to be taken when understanding the outcome association based on the aggregated data, and overinterpretation should be avoided. The existing analytical approaches to address the inferential failure mentioned in this study are also discussed.

© 2020 Elsevier B.V. All rights reserved.

1. Background and problem

Identifying environmental factors, e.g., meteorological factors and air pollutants, affecting the transmission and mortality risk of COVID-19 is of importance in understanding the features of the ongoing COVID-19 pandemic (Daraei et al., 2020; Ran et al., 2020). Pani et al found the associations between several meteorological factors and COVID-19 incidences under a time series study design, and they implied similar associations with the COVID-19’s transmission (Pani et al., 2020). Frontera et al found the positive associations between levels of several air pollutants and COVID-19 cases and deaths under a cross-region ecological study design (Frontera et al., 2020). They concluded that “air pollution may have a strong impact on the high rate of infection and mortality”. Similar analytical approaches as well as main findings appeared in other recent studies (Şahin, 2020; Suhaimi et al., 2020; Shahzad et al., 2020).

In the infectious disease epidemiology, the number of incidences, i.e., commonly known as the epidemic curve, is driven by the disease’s transmission process that is strongly determined by (i) the transmissibility, and (ii) the seed cases in the latest few days (Bauch et al., 2005; Cori et al., 2013; Lin et al., 2020; Riou and Althaus, 2020; Wallinga and Lipsitch, 2007; Wang et al., 2020; Wu et al., 2020a; Zhao et al., 2020a; Zhao et al., 2020b). Thus, the autocorrelation is highly likely undermine the statistical inference of the role of independent factors when using the aggregated number of COVID-19 incidences time series directly. Time series and ecological studies are two commonly used study designs aiming at exploring the relationship between an independent factor and the course of a disease, e.g., (Pani et al., 2020; Frontera et al., 2020). Both of these two classic study designs adopt the aggregated (instead of individual level) dataset. Hence, cautiousness needs to be taken when understanding the outcome association based on the aggregated data, and overinterpretation should be avoided.

In this study, it is demonstrated that the association between an independent factor and disease incidence (or death) counts fails to infer the association with disease transmission (or mortality risk) by using simple simulation-based counterexamples. The existing analytical approaches to address this inferential failure are also discussed.

2. Simulation framework and construction of counterexamples

To mimic the real-world COVID-19 epidemic growing process, the simple modelling framework in Tuite and Fisman (2020) is adopted
with the mean serial interval, $\tau$, at 5 days, referring to previous studies (Zhao et al., 2020c; Ferretti et al., 2020; Ganyani et al., 2020; He et al., 2020; Zhao, 2020), and population size, $N$, at 10 million. Then, for the $t$-th day since the first COVID-19 case, the daily number of new cases $c_t = c_0 - r_t^{1/\tau}$, where $c_0 = 1$ for first seed case at the start of the outbreak. Here, $r_t$ denotes effective reproduction number, and thus $r_t = R_t [1 - C_t / N]$. The $R_t$ is the reproduction number, a well-accepted metric quantifying the instantaneous transmissibility of an infectious disease (Zhao et al., 2020a). The $C_t$ is the cumulative number of cases at the $t$-th day, and obviously $C_t \geq C_{t-1} > 0$. The values of $R_t$ time series will be directly assumed in our counterexamples, which could also be a constant in some scenarios.

2.1. Counterexample #1: a time series study

In this counterexample, the time series study design as in Pani et al. (2020) is considered. An independent factor $X$ was considered as a determinant of the reproduction number ($R_t$) of COVID-19. Three pre-defined scenarios are considered that included

- scenario (I): the correlation between $X$ and $R_t$ was negative;
- scenario (II): the correlation between $X$ and $R_t$ was positive; and
- scenario (III): the correlation between $X$ and $R_t$ was zero.

In Fig. 1A, the factor $X$ time series are generated by using the arithmetic sequence with additive random noise terms. For COVID-19, $R_t$ largely ranges from 1.5 to 3 (Riou and Althaus, 2020; Wang et al., 2020; Wu et al., 2020a; Tuite and Fisman, 2020; Ferretti et al., 2020; He et al., 2020; Li et al., 2020; Zhao et al., 2020d), and the range from 1.5 to 2.5 is used in the simulation. With the same fashion, in Fig. 1B, three types of $R_t$ time series are generated in decreasing, increasing and constant trends matching with scenarios (I)–(III), respectively. In Fig. 1C, the COVID-19 epidemic curves are simulated in terms of the daily number of cases by using the three $R_t$ time series in Fig. 1B. Similarly, Fig. 1D shows the cumulative number of COVID-19 cases. The consistency between two kinds of pairwise associations (or correlations) are examined, and they include

- the association between $X$ and $R_t$,
- the association between $X$ and number of COVID-19 cases, i.e., daily or cumulative.

The inconsistency of these two associations indicates an inferential failure that the association between factor $X$ and number of COVID-19 cases failed to infer the underlying true relationship between $X$ and COVID-19 transmissibility.

2.2. Counterexample #2: an ecological study

In this counterexample, the same cross-region ecological study design is considered as in Frontera et al. (2020). As model setting, $n = 51$ regions are considered with different region-specific reproduction parameters.
numbers ($R$), which can be regarded by setting $R_t$s as constants, and case fatality ratios (CFR), and thus there were 51 COVID-19 outbreaks to be simulated. For the $i$-th region, the reproduction number was denoted by $R_i$, the case fatality ratio was denoted by CFR$_i$, and both of them are constants for a specific region. For COVID-19, $R$ was set from 1.5 to 3 (Riou and Althaus, 2020; Wang et al., 2020; Wu et al., 2020a; Tuite and Fisman, 2020; Ferretti et al., 2020; He et al., 2020; Li et al., 2020; Zhao et al., 2020d), and CFR ranged from 1.5% to 3% (Verity et al., 2020; Russell et al., 2020; Wu et al., 2020b) across different regions. The CFR$_i = (4.5 - R_i)/100$ is formulated for simple demonstration, as shown in the color code bar in Fig. 2. In the simulation, the time lag between onset of illness and death was ignored for simplicity, which should not affect our main conclusions. Fig. 1A shows the simulated COVID-19 epidemic curve of 51 regions, i.e., each curve represents the outbreak in one region. Fig. 1B shows the trends of the daily number of COVID-19 deaths for each region.

For a factor, $X$, negatively correlated with CFR, see Fig. 2C, the association between $X$ and the cumulative number of deaths is examined across all 51 regions. Similar as in counterexample #1, the consistency between two pairwise associations (or correlations) is examined, and they include:

- the pre-defined negative association between $X$ and CFR, and
- the association between $X$ and cumulative number of COVID-19 deaths.

In same fashion, the inconsistency of these two associations indicates an inferential failure that the association between factor $X$ and number of COVID-19 deaths failed to infer the underlying true relationship between $X$ and COVID-19 mortality risk.

3. Results and discussion

For the counterexample #1, Fig. 1E shows the relationship between factor $X$ and three types of $R$ series, which matches the scenarios (I)–(III) introduced above. However, in Fig. 1F and G, factor $X$ was found positively associated with the daily or cumulative number of COVID-19 cases in all scenarios. Therefore, the proposed inferential failure was evidently demonstrated. Furthermore, as remark, even though this counterexample merely presented the scenarios when $R > 1$ for demonstration, the conclusions may still hold when $R < 1$ as well as in more complex contexts.

For the counterexample #2, the predefined negative correlation between the factor $X$ and CFR was shown in Fig. 1C. However, the association between $X$ and the cumulative number of deaths are obviously positive, see Fig. 1D, which was opposite the association in Fig. 1C. Thus, it was clear that the positive association between factor $X$ and cumulative number of deaths failed to infer the underlying true relationship between $X$ and mortality risk of COVID-19 (in terms of CFR). A similar inferential failure also occurred when using the daily number of deaths (data not shown). Furthermore, as remark, that even though this counterexample presents the situation when CFR was negatively correlated with $R$ for demonstration, similar kinds of inferential failure may still occur in alternative situations and even more complex contexts.

The epidemic curve of an infectious disease is driven by its transmission process that can be measured by several metrics including $R$. The daily number of cases is strongly determined by the strength of transmissibility and the number of seed cases in the latest few days, which is determined by the serial interval. Hence, the autocorrelation is highly
likely undermine the statistical inference of transmission driven factors when using the number of cases time series directly, i.e., without defining quantifying the disease transmissibility. The cautiousness needs to be taken in avoiding this kind of noncausal association between independent factors and COVID-19 incidences. It is important to transform the incidence data to transmissibility by using plausible analytical approaches (Cori et al., 2013; Leung et al., 2020; Wallinga and Teunis, 2004; Ali et al., 2018), and check the associate between transmission rate and external factors thereafter, e.g., (Ran et al., 2020; Yao et al., 2020a).

The mortality risk of COVID-19 is reflected by both the number of disease-induced deaths and the number of COVID-19 cases. Thus, the COVID-19 CFR is of importance to properly quantify the mortality risk. It is important to transform the cases and deaths data to CFR by using analytical frameworks in previous studies (Verity et al., 2020; Russell et al., 2020; Wu et al., 2020b; Leung et al., 2020; Yang et al., 2020; Jung et al., 2020). One may then examine the association between CFR and external factors directly, e.g., (Yao et al., 2020b; Yao et al., 2020c).

Recently, an increasing number of studies present ‘associations’ between environmental factors and COVID-19 transmission, which is mainly due to two reasons. They include

• in previous literatures, weather parameters are commonly believed as critical drivers of respiratory virus’ transmission (Pica and Bouvier, 2012; Kutter et al., 2018), e.g., SARS-CoV (Yuan et al., 2006; Chan et al., 2011), MERS-CoV (Van Doremalen et al., 2013), and influenza (Ali et al., 2018; Dalziel et al., 2018); and
• exposing to air pollution may damage the immune system and thus lower the resistance to viral infections.

Hence, the associations between environmental factors and COVID-19 transmission or related mortality risks are possible, and consistent with previous findings in other respiratory viruses. However, the counterexamples demonstrate that implausible conclusions may be reached due to flawed analytical procedures. This study highlights the importance of following systematic analytical procedures to explore the role of environmental factors in the infectious disease epidemiology with aggregated dataset.

4. Conclusive remarks

It is important to follow the systematic analytical procedures to explore the role of environmental factors in the course of COVID-19 pandemic. To avoid noncausal association as well as over-conclusion, the definition of epidemiological metrics for infectious disease should be introduced and clarified before the analytical procedures. Although some simplifications have been made, the two counterexamples proposed in this study apply to other infectious diseases in alternative contexts.

Ethics approval and consent to participate

Not applicable.

Data availability

No real-world data was used in this work.

Authors’ contributions

S Zhao conceived the study, carried out the analysis, drafted the letter, discussed the results, critically read and revised the letter.

Funding

This work was not funded.

Disclaimer

The funding agencies 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; or decision to submit the manuscript for publication.

Declaration of competing interest

None declared.

Acknowledgements

The author thanks the insightful discussion with two anonymous colleagues, and helpful comments from two anonymous reviewers and the handling editor, which improves this manuscript.

References

Ali, S.T., Wu, P., Cauchemez, S., He, D., Fang, V.J., Cowling, B.J., et al., 2018. Ambient ozone and influenza transmissibility in Hong Kong. Eur. Respir. J. 51 (5).
Bauch, C.T., Lloyd-Smith, J.O., Coffee, M.P., Galvani, A.P., 2005. Dynamically modeling SARS and other newly emerging respiratory illnesses: past, present, and future. Epidemiology (Cambridge, Mass) 16 (6), 791–801.
Chan, K.H., Peiris, J.S.M., Lam, S.Y., Poon, L.L.M., Yuen, K.Y., Seto, W.H., 2011. The effects of temperature and relative humidity on the viability of the SARS coronavirus. Adv. Virol. 2011, 754069.
Cori, A., Ferguson, N.M., Fraser, C., Cauchemez, S., 2013. A new framework and software to estimate time-varying reproduction numbers during epidemics. Am. J. Epidemiol. 178 (9), 1505–1512.
Daraei, B.D., Kessler, S., Gog, JR., Viboud, C., Bjernadst, O.N., Metcalfe, C.J.E., et al., 2018. Urbanization and humidity shape the intensity of influenza epidemics in U.S. cities. Science 362 (6410), 75.
Daraei, H., Toolabian, K., Kazempour, M., Javanbakht, M., 2020. The role of the environment and its pollution in the prevalence of COVID-19. J. Inf. 81, E168–E169.
Ferretti, L., Wyman, C., Kendall, M., Zhao, L., Nurtay, A., Abeler-Dörner, L., et al., 2020. Quantifying SARS-CoV-2 transmission suggests epidemic control with digital contact tracing. Science 368, eabb9336.
Frontera, A., Gianfani, L., Vlahoi, K., Landoni, G., Cremona, G., 2020. Severe air pollution links to higher mortality in COVID-19 patients: the “double-hit” hypothesis. J. Inf. 81, 255–259.
Gananyi, T., Kremer, C., Chen, D., Torreri, A., Faes, C., Wallinga, J., et al., 2020. Estimating the generation interval for coronavirus disease (COVID-19) based on symptom onset data, March 2020. Eurosurveillance 25 (17), 2000257.
He, X., Lai, E.H.Y., Wu, P., Deng, X., Wang, J., Hao, X., et al., 2020. Temporal dynamics in viral shedding and transmissibility of COVID-19. Nat. Med. 1–4.
Jung, S.M., Akhmetzhanov, A.R., Hayashi, K., Linton, N.M., Yang, Y., Yuan, B., et al., 2020. Real-time estimation of the risk of death from novel coronavirus (COVID-19) infection: inference using exported cases. J. Clin. Med. 9 (2).
Kutter, J.S., Fraij, P.L., Fouchier, R.A.M., Herfst, S., 2018. Transmission routes of respiratory viruses among humans. Curr. Opin. Virol. 28, 142–151.
Leung, K., Wu, J.T., Liu, D., Leung, G.M., 2020. First-wave COVID-19 transmissibility and severity in China outside Hubei after control measures, and second-wave scenario planning: a modelling impact assessment. Lancet 395, 1382–1393.
Li, Q., Guan, X., Wu, P., Wang, X., Zhou, L., Tong, Y., et al., 2020. Early transmission dynamics in Wuhan, China, of novel coronavirus–infected pneumonia. N. Engl. J. Med. 382, 1199–1207.
Lin, Q., Zhao, S., Gao, D., Lou, Y., Yang, S., Musa, S.S., et al., 2020. A conceptual model for the coronavirus disease 2019 (COVID-19) outbreak in Wuhan, China with individual reduction and governmental actions. Int. J. Infect. Dis. 93, 211–216.
Pani, S.K., Lin, N.-H., Ravindrababu, S., 2020. Association of COVID-19 pandemic with meteorological parameters over Singapore. Sci. Total Environ. 140112.
Pica, N., Bouvier, N.M., 2012. Environmental factors affecting the transmission of respiratory viruses. Curr. Opin. Virol. 2 (1), 90–95.
Ran, J., Zhao, S., Han, L., Chen, D., Yang, Z., Yang, L., et al., 2020. The ambient ozone and COVID-19 transmissibility in China: a data-driven ecological study of 154 cities. J. Inf. 81, e9–e11.
Riou, J., Althaus, C.L., 2020. Pattern of early human-to-human transmission of Wuhan 2019 novel coronavirus (2019-nCoV). December 2019 to January 2020. Eurosurveillance 25 (4).
Russell, T.W., Hellewell, J., Jarvis, C.I., Van Zandvoort, K., Abbott, S., Ratnayake, R., et al., 2020. Estimating the infection and case fatality ratio for coronavirus disease (COVID-19) using age-adjusted data from the outbreak on the Diamond Princess cruise ship, February 2020. Eurosurveillance 25 (12), 2000256.
Şahin, M., 2020. Impact of weather on COVID-19 pandemic in Turkey. Sci. Total Environ. 728, 138810.
Shahzad, F., Shahzad, U., Fareed, Z., Iqbal, N., Hashmi, S.H., Ahmad, F., 2020. Asymmetric nexus between temperature and COVID-19 in the top ten affected provinces of China: a current application of quantile–on–quantile approach. Sci. Total Environ. 736, 139115.
Suhaimi, N.F., Jalaludin, J., Latif, M.T., 2020. Demystifying a possible relationship between COVID-19, air quality and meteorological factors: evidence from Kuala Lumpur, Malaysia. Aerosol Air Qual. Res. 20 (7), 1520–1529.

Tuite, A.R., Fisman, D.N., 2020. Reporting, epidemic growth, and reproduction numbers for the 2019 novel coronavirus (2019-nCoV) epidemic. Ann. Intern. Med. 172, 567–568.

Van Doremalen, N., Bushmaker, T., Munster, V.J., 2013. Stability of Middle East respiratory syndrome coronavirus (MERS-CoV) under different environmental conditions. Eurosurveillance 18 (38), 20590.

Verity, R., Okell, L.C., Dorogovtsev, S.N., Gog, J., Imai, N., et al., 2020. Estimates of the severity of coronavirus disease 2019: a model-based analysis. Lancet Infect. Dis. 20, 659–677.

Wallinga, J., Lipsitch, M., 2007. How generation intervals shape the relationship between growth rates and reproductive numbers. Proc. R. Soc. B Biol. Sci. 274 (1609), 599–604.

Wallinga, J., Teunis, P., 2004. Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures. Am. J. Epidemiol. 160 (6), 509–516.

Wang, K., Zhao, S., Liao, Y., Zhao, T., Wang, X., Zhang, X., et al., 2020. Estimating the serial interval of the novel coronavirus disease (COVID-19) based on the public surveillance data in Shenzhen, China from January 19 to February 22, 2020. Transbound. Emerg. Dis. https://doi.org/10.1111/tbed.13647.

Wu, J.T., Leung, K., Leung, G.M., 2020a. Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study. Lancet 395, 689–697.

Wu, J.T., Leung, K., Bushman, M., Kishore, N., Niehus, R., de Salazar, P.M., et al., 2020b. Estimating clinical severity of COVID-19 from the transmission dynamics in Wuhan, China. Nat. Med. 26 (4), 506–510.

Yang, S., Cao, P., Du, P., Wu, Z., Zhuang, Z., Yang, L., et al., 2020. Early estimation of the case fatality rate of COVID-19 in mainland China: a data-driven analysis. Ann. Transl. Med. 8 (4).

Yao, Y., Pan, J., Liu, Z., Meng, X., Wang, W., Kan, H., et al., 2020a. No association of COVID-19 transmission with temperature or UV radiation in Chinese cities. Eur. Respir. J. 55 (5).

Yao, Y., Pan, J., Liu, Z., Meng, X., Wang, W., Kan, H., et al., 2020b. Temporal association between particulate matter pollution and case fatality rate of COVID-19 in Wuhan. Environ. Res. 109941.

Yao, Y., Pan, J., Wang, W., Liu, Z., Kan, H., Qi, Y., et al., 2020c. Association of particulate matter pollution and case fatality rate of COVID-19 in 49 Chinese cities. Sci. Total Environ. 140396.

Yuan, J., Yun, H., Lan, W., Wang, W., Sullivan, S.G., Jia, S., et al., 2006. A climatologic investigation of the SARS-CoV outbreak in Beijing, China. Am. J. Infect. Control 34 (4), 234–236.

Zhao, S., 2020. Estimating the time interval between transmission generations when negative values occur in the serial interval data: using COVID-19 as an example. Math. Biosci. Eng. 17 (4), 3512–3519.

Zhao, S., Musa, S.S., Hebert, J.T., Cao, P., Ran, J., Meng, J., et al., 2020a. Modelling the effective reproduction number of vector-borne diseases: the yellow fever outbreak in Luanda, Angola 2015–2016 as an example. PeerJ 8, e8601.

Zhao, S., Stone, L., Gao, D., Musa, S.S., Chong, M.K.C., He, D., et al., 2020b. Imitation dynamics in the mitigation of the novel coronavirus disease (COVID-19) outbreak in Wuhan, China from 2019 to 2020. Ann. Transl. Med. 8 (7).

Zhao, S., Cao, P., Gao, D., Zhuang, Z., Cai, Y., Ran, J., et al., 2020c. Serial interval in determining the estimation of reproduction number of the novel coronavirus disease (COVID-19) during the early outbreak. J. Travel Med. 27, taaa033.

Zhao, S., Musa, S.S., Lin, Q., Ran, J., Yang, G., Wang, W., et al., 2020d. Estimating the unreported number of novel coronavirus (2019-nCoV) cases in China in the first half of January 2020: a data-driven modelling analysis of the early outbreak. J. Clin. Med. 9 (2).