REVISTA ESPAÑOLA DE COMUNICACIÓN EN SALUD
2024, Suplemento 1, S54-S59
https://doi.org/10.20318/recs.2024.7951

Modelización estadística para la estimación y predicción de la incidencia de COVID-19 en España

Statistical modelling for estimating and forecasting COVID-19 incidence in Spain

David Moriña[a], Alessandra Ybargüen[a]

[a] Departamento de Econometría, Estadística y Economía Aplicada - Riskcenter-IREA, Universitat de Barcelona, España

Resumen

Introducción: Basar procesos de toma de decisiones en datos que contienen errores e imprecisiones es inevitable en muchos de contextos por diferentes razones. La situación derivada de la pandemia mundial de COVID-19 es un claro ejemplo, donde los datos proporcionados por fuentes oficiales no siempre fueron fiables debido a problemas de recopilación de datos y a la alta proporción de casos asintomáticos. **Objetivos:** Cuantificar la gravedad de la información errónea en una serie temporal y reconstruir la evolución más probable del proceso, así como una discusión sobre los métodos estadísticos más adecuados para obtener predicciones en este contexto. **Métodos:** Se propone el uso de un modelo autoregresivo con heterocedasticidad condicional y estimación de los parámetros mediante Bayesian synthetic likelihood. **Resultados:** Solo alrededor del 51% de los casos de COVID-19 en el período 23 de febrero de 2020 al 27 de febrero de 2022 se notificaron en España, observándose también diferencias relevantes en la intensidad del subregistro entre comunidades autónomas. **Conclusión:** La metodología propuesta proporciona a los tomadores de decisiones en salud pública una valiosa herramienta para mejorar la evaluación de la evolución de una enfermedad bajo diferentes escenarios, ya que permite generar predicciones realistas en este contexto.

Palabras clave: modelización estadística; series temporales; datos infraregistrados; enfermedades infecciosas; COVID-19

Abstract

Introduction: Basing decision-making processes on data containing errors and inaccuracies is unavoidable in many situations. The COVID-19 pandemic related data is a clear example, where the information provided by official sources was often unreliable due to data collection mechanisms and the amount of asymptomatic cases. **Objectives:** To estimate the amount of misreported data in a time series and reconstructing the most probable evolution of the process and provides a discussion on the more appropriate statistical methods able to yield reliable forecasts in this context. **Methods:** The usage of a model based on autoregressive conditional heteroskedastic time series is proposed, estimating the parameters by Bayesian synthetic likelihood. **Results:** Only around 51% of the cases of COVID-19 in the period from February 23rd, 2020 to February 27th, 2022 were observed in Spain, also detecting remarkable differences in the reporting issues between Autonomous communities. **Conclusion:** The presented method allows generating realistic predictions under different possible scenarios, and therefore it represents a valuable tool for policy makers in order to improve the evaluation of the evolution of a situation.

Keywords: statistical modelling; time series; underregistered data; infectious diseases; COVID-19.

CONTACTO: David Moriña dmorina@ub.edu

Fecha de envío:04/08/2023
Fecha de aceptación:08/10/2023

Introducción

La situación provocada por el virus SARS-CoV2, desde finales de 2019, ha puesto de manifiesto que es fundamental contar con datos de calidad en la cadena de toma de decisiones, tanto en epidemiología como en muchos otros campos. Existe una enorme preocupación mundial en torno a esta enfermedad, lo que llevó, en marzo de 2020, a la declaración de emergencia de salud pública por parte de la Organización Mundial de la Salud (OMS) (Sohrabi et al., 2020).

La modelización matemática permite, mediante la utilización de herramientas matemáticas y la información disponible sobre una determinada enfermedad, representar y predecir una situación epidémica, estimar situaciones futuras y evaluar actuaciones ya realizadas. Desde el inicio de la pandemia se hizo evidente no solo la necesidad de disponer de información epidemiológica fiable, sino también la importancia de utilizar estas técnicas como apoyo en la gestión de la crisis por parte de las autoridades sanitarias.

La modelización matemática permite representar y predecir una situación epidémica, estimar situaciones futuras y evaluar actuaciones ya realizadas

Hasta la fecha, se han realizado muchos esfuerzos metodológicos para lidiar con los datos de COVID-19 mal informados, siguiendo las ideas introducidas en la literatura desde finales de los noventa (Alfonso, Løvseth, Samant & Holm, 2015; Arendt et al., 2013; Bernard et al., 2014; Gibbons et al., 2014; Rosenman et al., 2006; Winkelmann, 1996). Estas propuestas van desde el uso de factores de multiplicación (Stocks et al., 2018) hasta modelos basados en cadenas de Markov (Azmon et al., 2014; Magal & Webb, 2018) o modelos espacio-temporales (Stoner, Economou & Drummond Marques da Silva, 2019). Recientemente se han propuesto varios enfoques basados en series de tiempo discretas (Fernández-Fontelo, Cabaña, Joe, Puig & Moriña, 2019; Fernández-Fontelo, Cabaña, Puig & Moriña, 2016; Fernández-Fontelo, Moriña, Cabaña, Arratia & Puig, 2020) y continuas, una característica de los datos de COVID-19 y típicamente presente en el modelado de enfermedades infecciosas. En este sentido, en Moriña, Fernández-Fontelo, Cabaña, Puig, Monfil, Brotons & Diaz (2021) se introduce un nuevo enfoque para datos longitudinales que no presentan correlaciones temporales y en Moriña, Fernández-Fontelo, Cabaña & Puig (2021) se presenta un modelo capaz de manejar estructuras temporales usando un enfoque diferente. Una limitación típica de este tipo de modelos es el esfuerzo computacional necesario para estimar correctamente los parámetros.

Métodos

La estimación de la incidencia de COVID-19 que se utiliza para generar las predicciones descritas en este trabajo se basan en un modelo de series temporales autorregresivas con errores heterocedásticos (ARCH), cuyos detalles se pueden encontrar en Moriña, Fernández-Fontelo, Cabaña, Arratia & Puig (2023). Una vez estimada la incidencia de COVID-19, incluyendo casos no reportados, se estudian las predicciones generadas por diferentes métodos tanto a largo plazo (predicciones para todo un año) como a corto plazo (a una y dos semanas en el futuro). Los diferentes modelos predictivos considerados incluyen smoothing splines (Hastie & Tibshirani, 1990) para las predicciones a corto plazo, y una red neuronal (Hochreiter & Schmidhuber, 1997) para generar las predicciones a largo plazo.

Resultados

Este apartado reporta los resultados más reseñables producidos por la metodología propuesta, en relación con la estimación de la incidencia real de COVID-19 en cada comunidad autónoma en el período 23 de febrero de 2020 al 27 de febrero de 2022 y a las predicciones generadas para la Comunidad de Madrid en todo el año 2022.

Estimación de la incidencia

Los primeros casos de neumonía provocados por el betacoronavirus SARS-CoV-2 se detectaron en diciembre de 2019 en la ciudad de Wuhan (China) (Sohrabi et al., 2020), posteriormente diagnosticados como COVID-19. Teniendo en cuenta que muchos casos cursan sin desarrollar síntomas o solo con síntomas muy leves, es razonable suponer que la incidencia de esta enfermedad ha sido subregistrada. En Moriña et al., 2023 se analiza la incidencia semanal de COVID-19 registrada en España en el periodo del 23 de febrero de 2020 hasta el 27 de febrero del 2022, viendo que las fuentes oficiales reportaron 11.056.797 casos de COVID-19 en España, mientras que el modelo predice un

total de 21.639.627 casos (solo se reportaron el 51,10% de los casos reales). Este trabajo también reveló que, si bien la frecuencia de subregistro es extremadamente alta para todas las regiones, la intensidad de este subregistro no es uniforme entre las regiones consideradas: Aragón y Ceuta son las comunidades autónomas con mayor intensidad de subregistro mientras que Región de Murcia y Comunidad Valenciana son las regiones donde los valores estimados son los más cercanos al número de casos notificados.

El modelo permite estudiar también el impacto de covariables que potencialmente estén relacionadas con la incidencia de la enfermedad que se está estudiando. En este caso se analiza el efecto del periodo de estado de alarma declarado entre el 15 de marzo y el 21 de junio de 2020 y el efecto de la vacunación, considerando como periodo post-vacunación el momento en el cual el 50% de la población española había recibido al menos una dosis de alguna vacuna contra el COVID-19, a partir de mayo de 2021. Aunque el principal impacto de los programas de vacunación se aprecia en los datos de mortalidad, puede apreciarse también un descenso significativo en el número de casos semanales en todas las comunidades autónomas excepto en Euskadi (Moriña et al., 2023).

Generación de predicciones a corto plazo

Una vez reconstruida la evolución más probable de la incidencia de COVID-19, se pueden obtener pronósticos a corto plazo a una y dos semanas en adelante. Estimando la incidencia semanal real en el período del 29 de octubre de 2021 al 31 de diciembre de 2021, se han producido pronósticos para cada semana de 2022, utilizando el último valor de incidencia conocido de 2021, a una y dos semanas en el futuro, basadas en la técnica de smoothing splines. Los valores pronosticados para las 52 semanas de 2022 en la Comunidad de Madrid se muestran en la Figura 1.

Generación de predicciones a largo plazo

La Figura 2 muestra las predicciones a largo plazo para todo un año (53 semanas) en base a una red neuronal entrenada con los datos del período del 1 de enero de 2020 al 31 de diciembre de 2021, comparadas con los casos observados en 2022 y las primeras semanas de 2023.

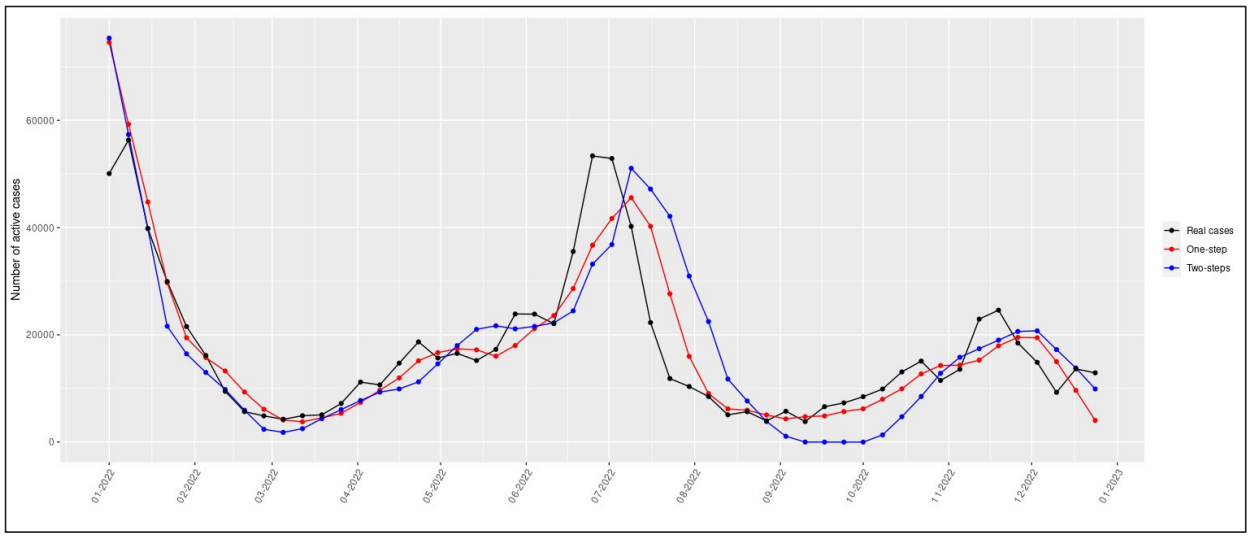

Figura 1. Predicción del número de casos semanales activos según el método de *smoothing splines* de COVID-19 en la Comunidad de Madrid en 2022.

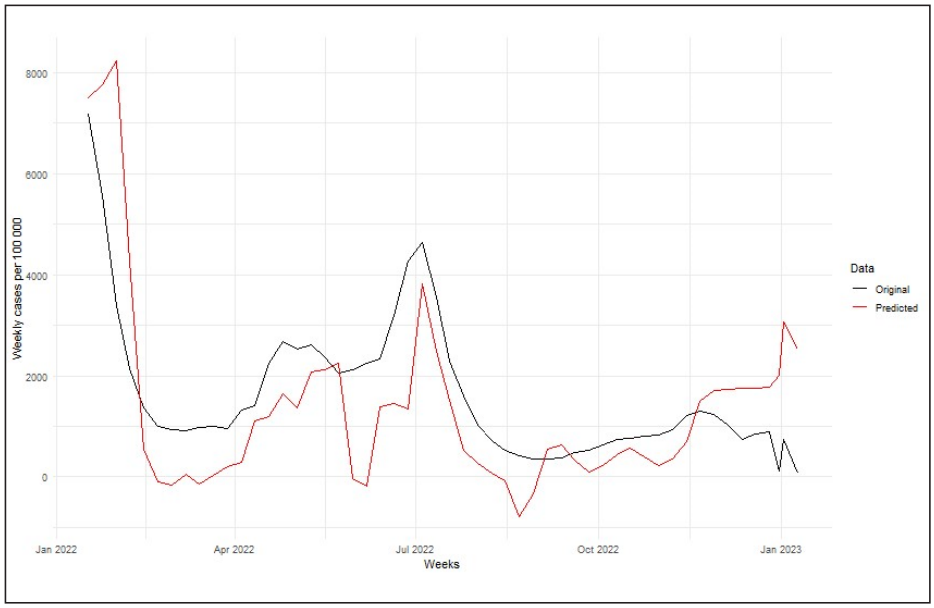

Figura 2. Predicción de la incidencia de COVID-19 por 100,000 individuos según el método de red neuronal de COVID-19 en España en 2022 e inicio de 2023.

Conclusiones

Aunque es muy común en la investigación biomédica y epidemiológica utilizar datos de los registros oficiales, recientemente se ha hecho evidente una preocupación acerca de su fiabilidad, y se han realizado algunos esfuerzos para estandarizar los protocolos a fin de mejorar la precisión de los registros de información de salud (ver por ejemplo Harkener et al., 2019; Kodra et al., 2018). Sin embargo, como ha puesto en evidencia la pandemia de COVID-19, no siempre es posible implementar estas recomendaciones de manera adecuada.

Otro trabajo que analizó la carga acumulada de COVID-19 en España estimó que solo alrededor del 21% de los casos se notificaron en el período del 1 de enero de 2020 al 1 de junio de 2020 (Moriña, Fernández-Fontelo, Cabaña, Arratia, Ávalos & Puig, 2021). Pero debemos tener en cuenta el fenómeno de la subnotificación se dió con mayor intensidad en las primeras etapas de la pandemia y, por lo tanto, se espera un menor subregistro general en el período más largo considerado en Moriña et al. (2023). Adicionalmente, la metodología presentada permite un seguimiento en tiempo real y no solo el análisis de la incidencia acumulada en un periodo de tiempo. Disponer de datos fiables es clave para abordar cualquier proceso de toma de decisiones en salud pública de forma óptima, y para mejorar la precisión de los modelos dinámicos destinados a estimar la propagación de la enfermedad (Zhao et al., 2020) y predecir su comportamiento.

El análisis de los datos españoles de COVID-19 muestra que, en promedio, solo se notificó alrededor del 51% de los casos en el período del 23 de febrero de 2020 al 27 de febrero de 2022

La metodología propuesta por Moriña et al. (2023) puede tratar los datos mal informados de una manera muy natural y directa. Además, es capaz de reconstruir el proceso oculto más probable, proporcionando a los tomadores de decisiones de salud pública una herramienta valiosa para predecir la evolución de la enfermedad de interés bajo diferentes escenarios.

El análisis de los datos españoles de COVID-19 muestra que, en promedio, solo se notificó alrededor del 51% de los casos en el período del 23 de febrero de 2020 al 27 de febrero de 2022, y que existen diferencias importantes en la severidad del subregistro entre las distintas comunidades autónomas españolas. También es posible evaluar el impacto del programa de vacunación sobre la dinámica de la enfermedad, logrando una disminución significativa de la incidencia de COVID-19 en casi todas las comunidades autónomas después de que el 50% de la población tuviera al menos una dosis de la vacuna (aunque estos resultados probablemente serían notablemente diferentes si en el período de estudio se incluyeran variantes como la BA.4 o BA.5, con un escape de inmunidad mayor y que en el momento de escribir este trabajo son predominantes en muchos países), mientras que el modelo solo pudo detectar el impacto del confinamiento obligatorio en siete de las 19 comunidades autónomas analizadas.

Los resultados de este trabajo muestran también la dificultad de obtener predicciones a largo plazo para series temporales no estacionarias y con un comportamiento altamente volátil como la considerada, así como la importancia

de combinar información proveniente de diversas fuentes para obtener predicciones a largo plazo fiables. La baja capacidad predictiva observada también hace altamente recomendable el uso de combinaciones de modelos predictivos o ensemble, cuya popularidad ha aumentado de forma notable recientemente, como consecuencia de la mejora de la capacidad computacional (Sherratt et al., 2023).

Contribuciones de los autores

Los autores participaron igualmente en la elaboración del manuscrito y aprobaron la versión final presentada.

Financiación

Esta investigación no recibió financiación.

Declaración de disposición de datos

Los datos presentados en este estudio pueden ser solicitados al autor de correspondencia.

Agradecimientos

Agradecemos al proyecto PredCov (Multi-source and multi-method prediction to support COVID-19 policy decision making) la oportunidad de presentar los resultados de nuestra investigación en este número especial.

Conflicto de interés

Los autores declaran que no hay conflicto de interés.

Referencias bibliográficas

Alfonso, J. H., Løvseth, E. K., Samant, Y. and Holm, J. (2015). Work-related skin diseases in Norway may be underreported: Data from 2000 to 2013. *Contact Dermatitis*, *72*(6), 409-412. https://doi.org/10.1111/cod.12355

Arendt, S., Rajagopal, L., Strohbehn, C., Stokes, N., Meyer, J., & Mandernach, S. (2013). Reporting of foodborne illness by U.S. consumers and healthcare professionals. *International journal of environmental research and public health*, *10*(8), 3684-3714. https://doi.org/10.3390/ijerph10083684

Azmon, A., Faes, C., & Hens, N. (2014). On the estimation of the reproduction number based on misreported epidemic data. *Statistics in medicine*, *33*(7), 1176-1192. https://doi.org/10.1002/sim.6015

Bernard, H., Werber, D., & Höhle, M. (2014). Estimating the under-reporting of norovirus illness in Germany utilizing enhanced awareness of diarrhoea during a large outbreak of Shiga toxin-producing E. coli O104: H4 in 2011—A time series analysis. *BMC Infectious Diseases*, *14*(1), 116-116. https://doi.org/10.1186/1471-2334-14-116

Fernández-Fontelo, A., Moriña, D., Cabaña, A., Arratia, A. and Puig, P. (2020). Estimating the real burden of disease under a pandemic situation: The SARS-CoV2 case. *PLoS ONE*, *15*, e0242956. https://doi.org/10.1371/journal.pone.0242956

Fernández-Fontelo, A., Cabaña, A., Joe, H., Puig, P. and Moriña, D. (2019). Untangling serially dependent underreported count data for gender-based violence. *Statistics in Medicine*, *38*(22), 4404-4422. https://doi.org/10.1002/sim.8306

Fernández-Fontelo, A., Cabaña, A., Puig, P. and Moriña, D. (2016). Under-reported data analysis with INAR-hidden Markov chains. *Statistics in Medicine*, *35*(26), 4875-4890. https://doi.org/10.1002/sim.7026

Gibbons, C. L., Mangen, M.-J. J., Plass, D., Havelaar, A. H., Brooke, R. J., Kramarz, P., Peterson, K. L., Stuurman, A. L., Cassini, A., Fèvre, E. M., Kretzschmar, M. E. E., & Burden of Communicable diseases in Europe (BCoDE) consortium. (2014). Measuring underreporting and under-ascertainment in infectious disease datasets: A comparison of methods. *BMC public health*, *14*(1), 147. https://doi.org/10.1186/1471-2458-14-147

Harkener, S., Stausberg, J., Hagel, C., & Siddiqui, R. (2019). Towards a Core Set of Indicators for Data Quality of Registries. *Studies in health technology and informatics*, *267*, 39-45. https://doi.org/10.3233/SHTI190803

Hastie, T. J., & Tibshirani, R. J. (1990). *Generalized Additive Models*. CRC Press.

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, *9*(8), 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735

Kodra, Y., Weinbach, J., Posada-De-La-Paz, M., Coi, A., Lemonnier, S. L., van Enckevort, D., Roos, M., Jacobsen, A., Cornet, R., Ahmed, S. F., Bros-Facer, V., Popa, V., van Meel, M., Renault, D., von Gizycki, R., Santoro, M., Landais, P., Torreri, P., Carta, C., … Taruscio, D. (2018). Recommendations for improving the quality of rare disease registries. *International Journal of Environmental Research and Public Health*, *15*(8). https://doi.org/10.3390/ijerph15081644

Magal, P., & Webb, G. (2018). The parameter identification problem for SIR epidemic models: Identifying unreported cases. *Journal of Mathematical Biology*, *77*(6-7), 1629-1648. https://doi.org/10.1007/s00285-017-1203-9

Moriña, D., Fernández-Fontelo, A., Cabaña, A., Arratia, A., & Puig, P. (2023). Estimated Covid-19 burden in Spain: ARCH underreported non-stationary time series. *BMC Medical Research Methodology*, *23*, 75. https://doi.org/10.1186/s12874-023-01894-9

Moriña, D., Fernández-Fontelo, A., Cabaña, A., Arratia, A., Ávalos, G., & Puig, P. (2021). Cumulated burden of Covid-19 in Spain from a Bayesian perspective. *European Journal of Public Health*, *31*(4), 917-920. https://doi.org/10.1093/eurpub/ckab118

Moriña, D., Fernández-Fontelo, A., Cabaña, A., & Puig, P. (2021). New statistical model for misreported data with application to current public health challenges. *Scientific Reports*, *11*(1), 23321. https://doi.org/10.1038/s41598-021-02620-5

Moriña, D., Fernández-Fontelo, A., Cabaña, A., Puig, P., Monfil, L., Brotons, M., & Diaz, M. (2021). Quantifying the under-reporting of uncorrelated longitudal data: The genital warts example. *BMC Medical Research Methodology*, *21*(1), 6-6. https://doi.org/10.1186/s12874-020-01188-4

Rosenman, K. D., Kalush, A., Reilly, M. J., Gardiner, J. C., Reeves, M., & Luo, Z. (2006). How much work-related injury and illness is missed by the current national surveillance system? *Journal of occupational and environmental medicine / American College of Occupational and Environmental Medicine*, *48*(4), 357-365. https://doi.org/10.1097/01.jom.0000205864.81970.63

Sherratt, K., Gruson, H., Grah, R., Johnson, H., Niehus, R., Prasse, B., Sandmann, F., Deuschel, J., Wolffram, D., Abbott, S., Ullrich, A., Gibson, G., Ray, E. L., Reich, N. G., Sheldon, D., Wang, Y., Wattanachit, N., Wang, L., Trnka, J., … Funk, S. (2023). Predictive performance of multi-model ensemble forecasts of Covid-19 across European nations. *eLife*, *12*, e81916. https://doi.org/10.7554/eLife.81916

Sohrabi, C., Alsafi, Z., O'Neill, N., Khan, M., Kerwan, A., Al-Jabir, A., Iosifidis, C., & Agha, R. (2020). World Health Organization declares Global Emergency: A review of the 2019 Novel Coronavirus (Covid-19). *International journal of surgery (London, England)*, *76*, 71-76. https://doi.org/10.1016/j.ijsu.2020.02.034

Stocks, T., Britton, T., & Höhle, M. (2018). Model selection and parameter estimation for dynamic epidemic models via iterated filtering: Application to rotavirus in Germany. *Biostatistics*. https://doi.org/10.1093/biostatistics/kxy057

Stoner, O., Economou, T. and Drummond Marques da Silva, G. (2019). A Hierarchical Framework for Correcting Under-Reporting in Count Data. *Journal of the American Statistical Association*, *114*(528), 1481-1492. https://doi.org/10.1080/01621459.2019.1573732

Taylor, S. J., & Letham, B. (2018). Forecasting at Scale. *The American Statistician*, *72*(1), 37-45. https://doi.org/10.1080/00031305.2017.1380080

Winkelmann, R. (1996). Markov chain Monte Carlo analysis of underreported count data with an application to worker absenteeism. *Empirical Economics*, *21*(4), 575-587. https://doi.org/10.1007/BF01180702

Zhao, S., Musa, S. S., Lin, Q., Ran, J., Yang, G., Wang, W., Lou, Y., Yang, L., Gao, D., He, D., & Wang, M. H. (2020). 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. *Journal of Clinical Medicine*, *9*(2), 388. https://doi.org/10.3390/jcm9020388