Challenges in Interpreting Epidemiological Surveillance Data – Experiences from Germany

Cornelius Fritz,
Giacomo De Nicola,
Felix Günther,
David Rügamer,
Martje Rave,
Marc Schneble,
Andreas Bender,
Maximilian Weigert,
Ralph Brinks,
Annika Hoyer,
Ursula Berger,
Helmut Küchenhoff,
Göran Kauermann

COVID-19 Data Analysis Group (CODAG@LMU)
Ludwig-Maximilians-Universität Munich

September 21, 2022

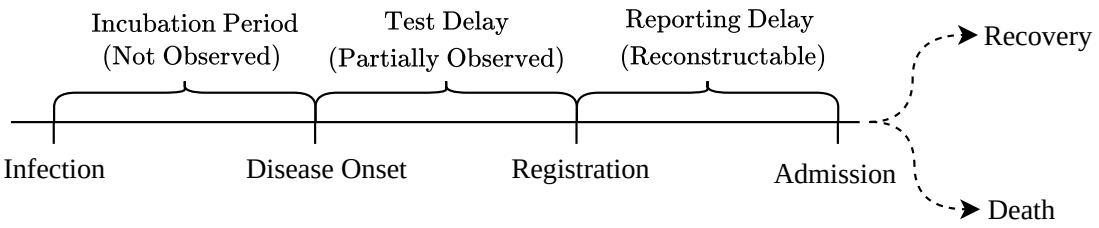

Figure 1: Illustrative temporal path of a COVID-19 infection, collected surveillance data and resulting delays.

As early as March 2020, the authors of this letter started to work on surveillance data to obtain a clearer picture of the pandemic's dynamic. This letter outlines the lessons learned during this peculiar time, emphasizing the benefits that better data collection, management, and communication processes would bring to the table. We further want to promote nuanced data analyses as a vital element of general political discussion as opposed to drawing conclusions from raw data, which are often flawed in epidemiological surveillance data, and therefore underline the overall need for statistics to play a more central role in public discourse.

Structure of COVID-19 Surveillance Data To better convey the lessons we learned, we start with a short introduction of the underlying mechanisms of surveillance data collection in Germany. For SARS-CoV-2, each infected person goes through different stages, as illustrated in Figure 1 for symptomatic cases. After infection and an incubation period, there is a disease onset, followed by recovery or severe disease progression, potentially leading to death. In Germany's case, each infected case in the respective surveillance data hence corresponds to multiple timestamps. Each record starts with the *registration date* on day t when a positive PCR test result is reported to the local health authorities (at the patient's place of residence). Due to the German healthcare system's federal structure, it is then passed on to the respective state counterpart and finally to the federal authority, the Robert Koch Institute (RKI), lead by the Federal Ministry of Health. Unfortunately, this reporting chain is not generally digitized in Germany (Sachverständigenrat Gesundheit, 2020). Therefore, all daily cases enter the dataset with a delay of d days, that is usually in the order of a couple of days. We thus call day $t + d$ the *admission date*. When testing is symptom-based, the registration date occurs after symptoms onset. However, as part of contact tracing and available screening tests, some cases may also be identified in the pre- or asymptomatic stage. Information on the *disease onset date* is also recorded in the German surveillance data based retrospective personal communication between health authorities and patients. Still, this date is only available in about 70% of all cases, either because no information was collected or because the corresponding infections are identified in the pre- or asymptomatic stage. Due to privacy issues, the date of death is not directly available. We are only provided with the number of fatal infections with a given local registration date and disease onset date.

Data Management and Transmission One should account for this specific data structure in any statistical analysis of the pandemic dynamics. Generally, reporting delays in surveillance data are not new, and *nowcasting* methods are suitable for accounting for occurred-but-not-yet reported cases (Lawless, 1994). We tackle this problem in various ways that enable a better assessment of the current pandemic's state (Günther et al., 2020; Schneble et al., 2020; De Nicola et al., 2022). Predicting future deaths and registrations allows, for instance, local management of the healthcare facilities in hospitals. At the same time, we can utilize the nowcasted epidemic curve to study, e.g., the effectiveness of pandemic containment measures (Küchenhoff et al., 2021) or estimate the R-value (Günther et al., 2020). In contrast, reported raw numbers of registered new infections and deaths only provide incomplete information, which may also structurally deviate in their temporal development from the actual infection dynamics and hence lead to misjudgments regarding the current situation.

Conceptually, implementing such nowcasting approaches would be relatively straightforward if the available data contained the relevant information. To nowcast deaths, information on the case's admission date would be needed which the RKI does not report for Germany. A possible workaround, which we implemented, is to automatically download the entire database daily, consecutively matching the inclusion date and the reporting date across all available downloads to obtain the admission dates.

Data Analysis and Policymaking With the start of the second wave, the incidence rate became the central measure on which policy decisions are based on. However, this number is biased by the occurrence of non detected cases, which depends on test capacities and test strategies (Brinks et al., 2020).

The issue of reporting delays also translates to the calculation of the 7-day-incidences, which in Germany the RKI reports daily because containment measures are carried out contingent on meeting a specific goal,i.e., a threshold of 50 infections per 100.000 capita. On day T, all infections with registration date in the last 7 days are then added up, i.e. infections with registration date $T-1$, $T-2$, $\ldots$,$T-7$. This causes a bias as infections with registration date T and reporting delay d are included in the central database on day $T+d$, and thus affect the 7-day-incidence only for $7-d$ days, leading to a downward bias in the order of 10%. Therefore, the 7-day incidence calculated by the RKI underestimates the true incidence. Improved estimates can be obtained using the *nowcasting* approaches mentioned above. Note that the ones just mentioned are instances of what we believe to be a broader problem: Too much focus is put on raw numbers, which are oftentimes intrinsically flawed.

Impact and Lessons Learned Having laid out shortcomings in currently available data, we were fortunate enough to have had access, through the RKI and other institutions, to coherent data in the first place. Though incomplete, these data have allowed us to conduct analyses, point out possible improvements, and design solutions through scientifically challenging work. Still, our data-driven approach seems to have previously had a limited bearing on public debate and policies; However, this changed when our group started issu-

ing biweekly reports[1]. These reports have been cited with increasing frequency in the local and national press, which led to media exposure. Despite having seen some progress in this direction in the general discussion, a long way remains to convince policymakers that statistics and nuanced data analysis should play a much more central role in policymaking and public discourse.

Acknowledgments

The authors gratefully acknowledge support from the German Federal Ministry of Education and Research (BMBF) under Grant No. 01IS18036A and the German Research Foundation (DFG) for the project KA 1188/13-1.

Conflict of interest

None.

References

Brinks, R., H. Küchenhoff, J. Timm, T. Kurth, and A. Hoyer (2020). Epidemiological measures for informing the general public during the SARS-CoV-2-outbreak: simulation study about bias by incomplete case-detection. *medRxiv*. 10.1101/2020.09.23.20200089.

De Nicola, G., M. Schneble, G. Kauermann, and U. Berger (2022). Regional now- and forecasting for data reported with delay: toward surveillance of COVID-19 infections. *AStA Advances in Statistical Analysis*. 10.1007/s10182-021-00433-5.

Günther, F., A. Bender, K. Katz, H. Küchenhoff, and M. Höhle (2020). Nowcasting the COVID-19 pandemic in Bavaria. *Biometrical Journal 63*(3), 490–502. 10.1002/bimj.202000112.

Küchenhoff, H., F. Günther, M. Höhle, and A. Bender (2021). Analysis of the early covid-19 epidemic curve in germany by regression models with change points. *Epidemiology & Infection 149*, e68. 10.1017/S0950268821000558.

Lawless, J. (1994). Adjustments for reporting delays and the prediction of occurred but not reported events. *Canadian Journal of Statistics 22*(1). 10.2307/3315826.n1.

Sachverständigenrat Gesundheit (2020). Corona: Daten teilen, besser heilen - Sachverständigenrat Gesundheit. *Der Spiegel*. https://www.spiegel.de/wirtschaft/soziales/corona-daten-teilen-besser-heilen-sachverstaendigenrat-gesundheit-a-ed21193d-84cf-4765-a085-cca5de840078 (visited 2-12-2021).

[1] https://www.covid19.statistik.uni-muenchen.de/newsletter/index.html

Schneble, M., G. De Nicola, G. Kauermann, and U. Berger (2020). Nowcasting fatal COVID-19 infections on a regional level in Germany. *Biometrical Journal 63*(3), 471–489. 10.1002/bimj.202000143.