Stochastic Environmental Research and Risk Assessment (2020) 34:1281–1283
https://doi.org/10.1007/s00477-020-01821-0

SHORT COMMUNICATION

Undersampling in action and at scale: application to the COVID-19 pandemic

Andreas Langousis[1] · Alin Andrei Carsteanu[2]

Published online: 2 June 2020

Abstract

It is the purpose of this short communication to analyze the possible caveats in the statistical interpretation of collected data, particularly in the light of decision-making concerning the current COVID-19 coronavirus pandemic. A mitigation of undersampling is proposed, based on re-scaling of statistics that can be considered reliable, such as deaths, and epidemic properties like mortality, that may be considered comparable between countries with similar levels of health care, which would not have reached a saturation level.

Keywords Susceptible-infectious-removed (SIR) models · COVID-19 pandemic · Asymptomatic carriers · Undersampling

1 Introduction

Compartmental models have been used in epidemiology since the basic Susceptible-Infectious-Removed (SIR) model was proposed in the seminal work of Kermack and McKendrick (1927). The 3 compartments represent the 3 fundamental states in which a member of the population can be found, assuming that the removed category lumps together the deceased and recovered individuals, of which the latter are immune, and therefore not susceptible. Depending on the type of disease, further categories may (or may not) be necessary for a more accurate representation of its epidemic evolution. For instance, a disease with high mortality may require a distinction between deceased and recovered (SIRD formulation, which is being used herein) instead of the lumped-up category of removed individuals (see e.g. Osemwinyen and Diakhaby 2015), given that generally the deceased have less contact with the susceptible, than do the recovered. Introducing a compartment of exposed individuals (SEIR end SEIRD models), allows to distinguish between infectious individuals and those who have been infected, but are not yet infectious. However, since infectiousness most often comes along with symptoms, it is unclear whether exposed non-infectious individuals would make a considerable difference when distinguished from the susceptible compartment. A more delicate problem is raised by the asymptomatic carriers: aside from the difficulty to identify those individuals among the general population, a fact that also complicates statistical parametrization, it is generally little known for how long an asymptomatic carrier remains in that state, or whether a sudden drop in immune response could make the asymptomatic carrier develop the illness. The actual process, which describes the epidemic infection caused by random interaction between members of a population of individuals, deals with discrete-valued random variables, representing the counting-measure (or cardinality) of the 3 subsets in question. Necessarily, the stochastic process described by those random variables is characterized at any point in time, by an expected value, and a standard deviation around it.

It is the purpose of this communication to analyze the possible caveats in the statistical interpretation of collected data, particularly in the light of decision-making concerning the current COVID-19 pandemic, caused by SARS-CoV-2 (severe acute respiratory syndrome coronavirus 2). In the following sections, statistical considerations with

✉ Andreas Langousis
andlag@alum.mit.edu

[1] Department of Civil Engineering, University of Patras, Patras, Greece

[2] ESFM, Instituto Politécnico Nacional, Ciudad de México, México

emphasis on undersampling are being analyzed, in terms of their consequences for decision making. Conclusions are being summarized in the corresponding section.

2 Statistical estimation

The possibility of constructing unbiased estimators, e.g. for the true number of infected individuals at some point in time, or for the mortality rate in an epidemic, based on the data that are being routinely collected under the circumstances, is a feasible task, but it's not a trivial one. Given its importance in decision making, however, it is essential to understand what the estimation challenges are.

One concern, during an epidemic, is the estimation of the mortality rate. At the end of the epidemic, the quotient $Deceased/(Deceased + Recovered)$ represents a realization of the random variable whose expected value is sought, and thereby an unbiased estimator thereof. However, while the epidemic develops, it is unclear what the reference number ought to be, against which the number of casualties should be compared, since the daily number of infections evolves, and the transition rate from infectious to recovered is generally quite different from the transition rate from infectious to deceased. Such being the case, the above quotient should be estimated for the subset of individuals who were infected before a given date, once all cases of that subset are closed, in order to obtain an unbiased estimator of mortality—but such a statistic is unobtainable. The closest one can come is to consider the individuals whose illness was confirmed before a certain date, once all their cases are closed. Such a statistic is feasible, but generally not readily available. It should be noted that making it available, would offer an unbiased estimator of the mortality rate, within the intrinsic standard deviation of that variable. In its absence, an upper and a lower limit of mortality rate estimates can be established from current-time data, considering the two extreme situations; i.e. that the number of infected individuals $I(t)$ existing at the present time t could evolve either to the recovered (R) or deceased (D) states. In the first case, the lower estimation limit is calculated to be $D(t)/[D(t) + I(t) + R(t)]$, whereas in the second case the upper estimation limit becomes $[D(t) + I(t)]/[D(t) + I(t) + R(t)]$, leaving a rather wide margin of uncertainty during the epidemic. It should be noted that when $I(t) \rightarrow 0$ at the end of the epidemic, both limits tend to the unbiased estimator, but during the epidemic, we shall use the estimator with respect to the closed cases to date, in spite of the aforementioned difference in transition rates.

Taking into account the above considerations, the effects of undersampling, particularly during the period of initial exponential growth of an epidemic, can be evidenced. The following example, taken from the current pandemic, shows how exhaustive sampling in a given community or country, can be used to mitigate undersampling elsewhere. At $t =$ April 27, 2020, the reported situation in the USA was the following (https://www.worldometers.info/coronavirus/country/us/): total recorded cases $D(t) + I(t) + R(t) = 987{,}322$, closed cases $D(t) + R(t) = 174, 196$, of which deaths $D(t) = 55, 415$ and recovered $R(t) = 118{,}781$. At the same time the reported situation in Germany was (https://www.worldometers.info/coronavirus/country/germany/): total cases $D(t) + I(t) + R(t) = 157{,}770$, closed cases $D(t) + R(t) = 120{,}476$, of which deaths $D(t) = 5, 976$ and recovered $R(t) = 114{,}500$. One can notice the discrepancy in $D(t)/[D(t) + R(t)]$ ratios of ≈ 0.32 in the USA vs. ≈ 0.05 in Germany, with the latter being very close to the worldwide mortality rate that has been reported so far, i.e. 3–4% [see e.g. Wang et al. (2020), and https://www.worldometers.info/coronavirus/coronavirus-death-rate/].

A similar discrepancy appears also in the $[D(t) + I(t)]/[D(t) + I(t) + R(t)]$ ratios. More precisely $[D(t) + I(t)]/[D(t) + I(t) + R(t)] \approx 0.88$ in USA vs. ≈ 0.27 in Germany, signifying that the COVID-19 outbreak in the USA is at early stages. Also, the much higher $D(t)/[D(t) + R(t)]$ ratio in the USA (by a factor of approximately 6.5), relative to Germany where the outbreak decelerates, means that there has been significant undersampling, as testing has been taking place at medical facilities, health care units, hospitals etc., where symptomatic cases are sampled at higher frequency. Applying the $D(t)/[D(t) + R(t)]$ ratio of Germany to the USA, one can obtain a rough estimate of the number of closed cases based on the number of deaths. In this case, $D(t) + \widehat{R(t)} = 55{,}415/0.05 \approx 1{,}100{,}000$, and one can use the current $[D(t) + R(t)]/[D(t) + I(t) + R(t)]$ ratio of the USA (i.e. $= 174{,}196/987{,}322 \approx 0.18$) to estimate the actual number of total cases. This calculation gives $D(t) + \widehat{I(t)} + \widehat{R(t)} \approx 6{,}100{,}000$, which corresponds to an estimated value of $\widehat{I(t)} \approx 6{,}100{,}000 - 1{,}100{,}000 = 5{,}000{,}000$ active cases, from which only $987{,}322 - 174{,}196 = 813{,}126$ are detected and tracked. Under this setting, if we assume that the US economy opens in an unconstrained setting, and each undetected case infects one additional individual every 2.776 days on the average (which is the value witnessed before the shelter-at-home order was issued in the USA), then within 10 days the total number of cases will be $D(t) + \widehat{I(t)} + \widehat{R(t)} = (5{,}000{,}000 - 813{,}126) \times 2^{(10/2.776)} + 987{,}322 \approx 50{,}000, 000,$ resulting to approximately $3\% \times 50{,}000{,}000 \approx 1{,}500{,}000$ deaths in total, from which only 55, 145 have already been witnessed. If we assume that social distancing accompanied by extraordinary protection measures is effectively applied, and each undetected case infects one additional individual every 23.52 days on the average (which is the value witnessed after the shelter-at-home order was issued in the USA), then within 10 days the total number of cases will be $D(t) + \widehat{I(t)} + \widehat{R(t)} = (5{,}000{,}000 - 813{,}126) \times 2^{(10/23.52)} + 987{,}322 \approx 6{,}600{,}000$, resulting to approximately $3\% \times 6{,}600{,}000 \approx 200{,}000$ deaths in total, from which only 55, 145 have already been witnessed.

3 Implications and conclusions

According to the comparison of mortality statistics, it appears that in the USA there has been a significant undersampling of active cases, most probably because testing has been conducted at medical facilities, healthcare units, hospitals etc., where symptomatic cases are sampled at higher frequency. Under the current circumstances, it is suggested by the descriptive statistics of data and simple mathematical calculations that as of April 27, 2020 the active cases in the USA are of the order of 5 million, which may include patients with lighter symptoms and/or asymptomatic carriers, who may very effectively transmit the virus. Therefore, if/when economic activity is resumed, extraordinary protection measures should be taken for everyone, similar to those taken presently for medical personnel.

While testing hospitalized patients for COVID-19 whenever they present any symptom thereof is a medical necessity, in order to determine the treatment, our results show that testing lighter cases, as well as random samples among the population at large could enormously benefit the accuracy of statistics concerning the epidemic. It is important to note that random samples need not be large, but their importance for informed decision-making cannot be overstated, as well as for determining the fraction of asymptomatic carriers, who may influence the evolution of an epidemic in an essential way.

While this paper focuses on the statistics of collected data, trying to highlight the important risks associated to incomplete information, a brief comment should be added in terms of related risks, as those emerge from medical literature. More precisely, a considerable number of serious cases have been linked to causes beyond the mere lung infection caused by the virus, such as the formation of blood clots (Lillicrap 2020), heart problems (Bangalore et al. 2020), kidney failure (Xiao et al. 2020), some of which may be related to the overproduction of proinflammatory cytokines (Lillicrap 2020), but all of which can favor the development of serious cases in patients with no previous immune deficiency. Under this setting, relying on the option of developing a sufficient number of immune individuals in the population (usually referred to as “herd” immunity), may lead to saturation of medical services, with its inherent consequences, well before such a state is reached. However, if social distancing is carefully enforced, even after the lockout, saturation of the medical system can be avoided, as the number of infected cases reduces significantly.

Acknowledgements The authors would like to thank the Editorial Team of SERRA and involved reviewing parties, who made possible this work to have an impact during the current pandemic. Unfortunately, during this same process, one of the Authors has developed COVID-19 symptoms, but hopes to recover.

References

Bangalore S, Sharma A, Slotwiner A, Yatskar L, Harari R, Shah B, Ibrahim H, Friedman GH, Thompson C, Alviar CL, Chadow HL, Fishman GI, Reynolds HR, Keller N, Hochman JS (2020) ST-segment elevation in patients with covid-19: a case series. N Engl J Med. https://doi.org/10.1056/NEJMc2009020

https://www.worldometers.info/coronavirus/coronavirus-death-rate/. Retrieved: April 27, 2020

https://www.worldometers.info/coronavirus/country/germany/. Retrieved April 27, 2020

https://www.worldometers.info/coronavirus/country/us/. Retrieved April 27, 2020

Kermack WO, McKendrick AG (1927) A contribution to the mathematical theory of epidemics. Proc R Soc A 115(772):700–721

Lillicrap D (2020) Disseminated intravascular coagulation in patients with 2019-nCoV pneumonia. J Thromb Haemost. https://doi.org/10.1111/jth.14768

Osemwinyen A, Diakhaby A (2015) Mathematical modelling of the transmission dynamics of ebola virus. Appl Comput Math 4:313–320

Wang C, Horby PW, Hayden FG, Gao GF (2020) A novel coronavirus outbreak of global health concern. Lancet 395:470–473. https://doi.org/10.1016/S0140-6736(20)30185-9

Xiao G, Hu H, Wu F, Sha T, Huang Q, Li H, Han J, Song W, Chen Z, Zeng Z (2020) Acute kidney injury in patients hospitalized with COVID-19 in Wuhan, China: a single-center retrospective observational study. The Preprint Server or Health Sciences. https://doi.org/10.1101/2020.04.06.20055194