DATA AND STATISTICS AS BASIS FOR POLITICAL DECISIONS: LESSONS TO BE LEARNT FROM THE COVID-19 PANDEMIC

Joachim Engel and Adalbert Wilhelm
Ludwigsburg University of Education, Germany
Jacobs University Bremen, Germany
engel@ph-ludwigsburg.de

The Covid-19 crisis has impressively raised the general awareness that our social coexistence and political decisions are essentially based on data, the weighing of risks and thus on probability estimates. This places high demands on the ability of health authorities, policy makers and the media to communicate statistical information as well as on the ability of citizens to understand these messages. In this paper we reflect on the role of scientific evidence in democratic societies and analyze selected illustrative examples of communicating evidence via visualizations and simulation, media reports, and expert's statements. We identify venues and formats of communicating statistical information about the pandemics to the public that seems to be effective contrasting less helpful formats. We conclude by presenting recommendations for stakeholders in politics, media and statistics agencies on how to communicate empirical evidence to the public efficiently, released by the Deutsche Arbeitsgemeinschaft Statistik, an umbrella organization of statistical associations in Germany.

INTRODUCTION

The global COVID-19 pandemic provides a vivid example for the need to use evidence to inform policy and demonstrates the importance of, and need for, public understanding and reasoning with data. Every country affected by Covid-19 faces the threat of widespread deaths, economic damage, and social disruption. Citizens, policy makers, public health officials and governments need to take account of existing and emerging evidence, in order to decide on effective action. Democratic societies thrive on transparency, and transparency is enhanced by good communication of information (Martignon et al., 2021). For measures to be effective in democratic societies, governments need to give transparent and convincing explanations for their decisions. This puts high demands on skills associated with communicating statistical evidence on the side of governments and media, and a citizenry able to understand statistical messages. Statistical data on public health and other socially burning issues often have specific characteristics, and understanding them requires competencies for which our educational institutions and curricula inadequately prepare (Engel et al., 2020; ProCivicStat Partners 2018; Ridgway, 2021). Understanding these issues is essential for civic engagement in modern societies, but it is often based on complex multivariate data, the interpretation and development of which requires knowledge that is not taught in ordinary undergraduate education in mathematics and statistics - either in schools or universities - let alone in school subjects such as politics or civics. Few high school teachers in mathematics or social science receive any training on how to teach statistics. As a result, when reaching the statistics parts of a national curriculum, teachers stay within their comfort zone and overemphasize a narrow range of statistical techniques and computations (mathematics) or fail to engage with statistical ideas at all (social science). In particular, they pay too little attention to working with and understanding multivariate data that are characteristic of social trends, or to the analysis and interpretation of and communication about the meaning of such data (Ridgway, 2021).

In the following we showcase along examples in the context of the COVID-19 pandemic some barriers and obstacles to understanding, indicate venues to overcome these difficulties and end with recommendations for communicating evidence in times of a pandemic.

CRITICAL APPRECIATION OF DATA QUALITY

The basis for evidence-informed decisions are high quality up-to-date data that are of relevance for the questions of interest. Regarding quality, the conflicting goals of timeliness and accuracy are particularly relevant in times of a pandemic. In the context of data collection, the following questions are of high relevance for policy makers as well as for engaged citizen: What are the data used for? Have suitable data been obtained, do they contain the appropriate variables for

In: R Helenius, E Falck (Eds.), *Statistics Education in the Era of Data Science*
Proceedings of the Satellite conference of the International Association for Statistical Education (IASE),
Aug-Sept 2021, Online conference. ©2021 ISI/IASE iase-web.org/Conference_Proceedings.php

answering the research question of interest? How were the data obtained? Much of the data during a pandemic come from observational studies, which makes it difficult to make robust causal attribution. Decisions are made about measurements and operationalizations, the backgrounds of which are to be questioned and discussed: How can cases or deaths be measured accurately enough to be a basis for good decisions? What does 7-day incidence measure? What is the significance of the R_0 reproduction figure?

Data are often presented in aggregate form, and the underlying spatial structures result from administrative frameworks that should not be directly related to infection events. For which questions and under which circumstances is it appropriate to add incidences from different regions? How valid are comparisons of incidence values of different groups such as school children and people in nursing homes, vulnerable populations, and people without relevant preexisting conditions? There are many difficult issues with data quality: first of all, only reported cases are registered. Can we estimate the unreported cases? An emerging disease that spreads globally at a rapid pace leads to heterogeneity in data and collection methods, and thus to time-dependent and geographically dispersed data from different data sources: Data on the course of epidemics are collected over longer periods of time and in different locations. The same methods are not used everywhere and at all times. Disease modeling requires estimates of the number of susceptible individuals in the population and the chances of cure for those who have contracted the disease. Both parameters change over time as more people become immune and treatments improve.

THE DYNAMICS OF A PANDEMIC: EXPONENTIAL GROWTH

Understanding the dynamics of a pandemic is quite a challenge: Considered in isolation of any other accompanying factors, the spread of an epidemic may well be modeled exponentially, captured by the basic reproduction rate R_0 which represents the expected number of new cases of infected people directly generated by one infected person. Understanding exponential growth is essential. Those, who are familiar with the characteristics of exponential growth can easily refute arguments that COVID-19 deaths in the early stages of an epidemic are not a cause for concern when, for example, the numbers are low compared to road traffic deaths. But most people have severe difficulties understanding the nature of exponential relationships. And many of those, who have a basic theoretical understanding, often underestimate its powerful dynamics in practice (Vittert and Podkul, 2020). While the exponential model is helpful, the dynamics of a pandemic are more complex because the parameters such as R_0 are not fix but influenced by policy decision and changing human behavior leading to effects, that are being realized with time delay. There is a strong impact of feedback-loops: today's behavior influences tomorrows spread of the disease. This is why epidemiological models go far beyond simple exponential or logistic relationships.

However, there is a transparent way, to communicate the dynamics of a pandemic to the public in an understandable way: Simulations. Some media used simulations in a particularly illustrative way to visualize the spread of the virus, under various scenarios. An inspiring example illustrating the spread of the epidemic appeared as early as March 14, 2020, in the Washington Post[1], entitled "Why outbreaks like coronavirus spread exponentially, and how to flatten the curve." The Washington Post made this simulation available free of charge and in all major languages, which resulted in it being distributed worldwide, including repeatedly on German television[2]. The New York Times[3] published a dynamic graphic titled "How the Virus Won" that maps the spread of COVID-19 cases from February to June 2020 in the United States. It shows how an analysis of the associations between different COVID-19 strains and travel patterns can help understand the spread of the disease.

UNDERSTANDING TEST RESULTS

In testing situations, the difference between the sensitivity of a diagnostic test and the positive-predictive value is not always present (Gigerenzer et al, 2007; McDowell at al., 2019). In particular, the prevalence (or base rate) is often neglected. When applied to a broad swath of the population, a test's performance can be surprisingly counterintuitive. It can perform worse than expected, producing a potentially large proportion of false positives in populations less likely to have the disease. Consider a scenario with Covid-19 testing in a population with 1 in 101 people infected. Assume the test is always positive in individuals with the disease (100% sensitivity) but falsely positive 5% of the time (i.e., 95% specificity) which would be superior to many medical tests in use.

As shown in Figure 1, the chance that someone with a positive test result is actually infected is about 17% (1 in 6). If the prevalence rate (more realistically) is 4 in 1000 as (roughly) in the Germany in November of 2020, then the probability of being infected after a positive test result is less then 7.5%.

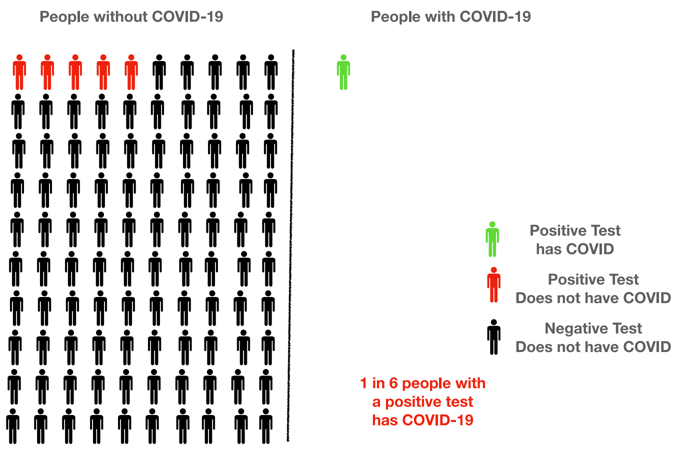


Figure 1. Under an incidence rate of 1% and a test with 100% sensitivity and 95% specificity, only 1 in 6 positively tested persons actually has COVID.

UNDERSTANDING RISK: THE CASE OF ASTRAZENECA

Unforeseen safety issues routinely emerge after any new medicine or vaccine goes from testing in tens of thousands of volunteers to actual public use on tens of millions. A very small percentage developed a strange blood clot after receiving the vaccine of AstraZeneca. As consequence, in spring of 2021 eighteen mostly European countries suspended AstraZeneca over possible side effects either completely or for some portions of its population. After much media attention to this problem, a substantial number of people who were eligible to receiving AstraZeneca rejected this vaccine, preferring going unvaccinated, at least for an undetermined period of time. The issue at stake here is the rational balancing of two risks (while the broader lesson to be learnt is that an absolute risk free life is an illusion). All medical treatments have potential harms as well as potential benefits, and it is important to be able to weigh these against each other. The life-threatening blood clots, accompanied by an oddly low count of clot-promoting platelets, appear to strike about one per 100,000 receiving AstraZeneca's (Ledford, 2021) while age is a factor modifying the risk. Figure 2, from the Winton Centre for Risk and Evidence Communication at Cambridge University[4], provides transparent information allowing a rational weighing of the two risks involved. Notice, that the potential benefits of a vaccination vary with the incidence rate of the virus. The Winton Centre website provides also information for scenarios with other incidence rates of 6 per 10000 and 20 per 10000.

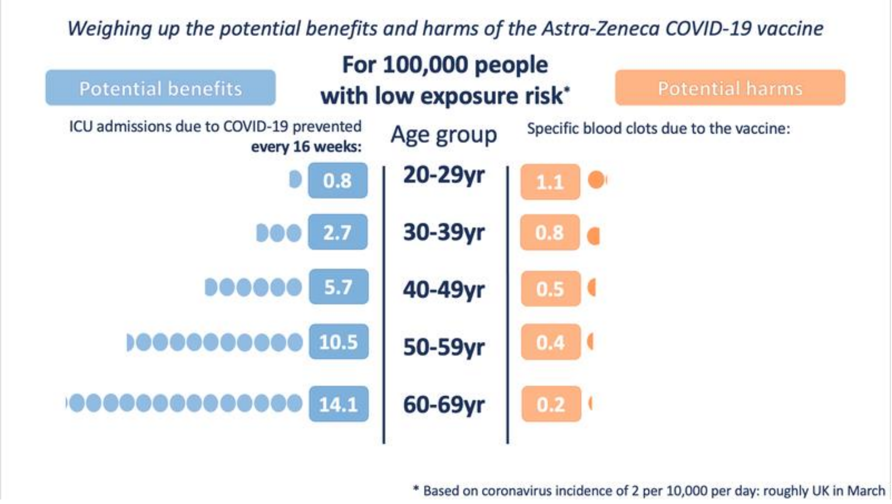


Figure 2. Illustration of the potential harms and benefits at a low exposure, incidence of 2 in 10,000 per day (Source: Winton Centre for Risk and Evidence Communication at Cambridge University).

VISUAL COMMUNICATION

New developments in computer graphics and animation are shaping the way data are presented in media as illustrations or as animated simulations. Visual representations occupy a central position in public communication and aim to represent the relevant dynamics and content in a way that can be quickly understood. Mainly, either time-dependent parameters or data with a spatial reference are visualized. For the spatially distributed data, choropleth maps are preferred, in which ideally the local authorities responsible for health issues are colored according to the distribution density of infection figures or derived variables (see Figure 3). Almost exclusively, ordinance thresholds are used as the basis for color scaling in medial practice, so that no color distinction between areas is realized in the case of area-wide over- and under-achievement of the valid warning values. This ignores the real spatial distribution differences and minimizes the information content of the choropleth map. For a sensible visual representation, color scales that use different basic colors based on limit values and then vary them depending on intensity would be an appropriate means for better information communication. means of better information communication.

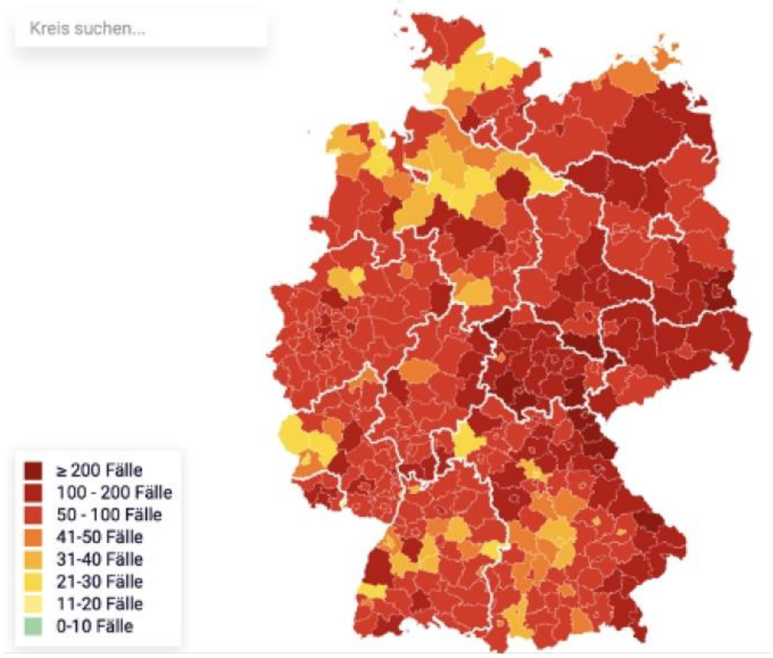
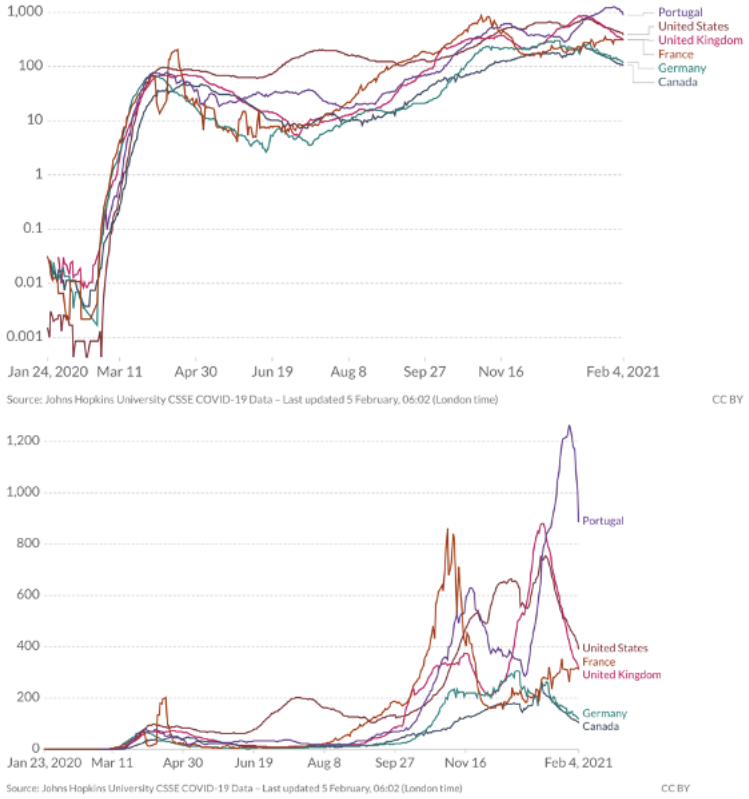


Figure 3. Chloropleth map of incidence figures for Germany by county
(Source: Robert Koch Institute)

For the time-dependent parameters, different variants of time series diagrams are used, mainly line and column diagrams. Classical errors of the graphical representation such as an overemphasis of temporal variability by reducing the vertical axis to a small section, have now largely disappeared from the media. Switching between lines and bar charts for purely design considerations in order to produce corresponding graphical diversity nevertheless seems questionable. The use of logarithmic scales in time series charts should be evaluated with caution. On the one hand, they tempt superficial readers to underestimate dynamic growth processes, and on the other hand, they increase the mathematical and statistical literacy requirements of the readership without corresponding advantages of visual representation. Figure 4 shows the time course of 7-day incidence per 100,000 population between January 24 and February 4, 2021, for some selected countries. While on the logarithmic scale the differences appear relatively small, the linear scale shows substantial differences.

Some media use innovative visualizations for illustration. In the Financial Times[5] , for example, under the headline "COVID-19's soaring death toll dwarfs figures from the first wave" which gives an overview of the development of deaths attributable to Corona, broken down by world by regions of the world. Also illustrative is a simulation from ZEIT Online[6] that - based on models developed by a group of researchers at the Max Planck Institute for Chemistry - estimates probabilities in different scenarios of an infected person infecting other people indoors. While the visualizations present the simulated infection processes in a catchy way, the dependence of the simulations on parameter assumptions and settings is usually not addressed. Simulations should also always make transparent on which model assumptions and which data basis the simulations were created.



Source: Johns Hopkins University CSSE COVID-19 Data – Last updated 5 February, 06:02 (London time) CC BY

Source: Johns Hopkins University CSSE COVID-19 Data – Last updated 5 February, 06:02 (London time) CC BY

Figure 4. The time course of 7-day incidence for different countries. On the logarithmic scale (upper graph), differences appear smaller. The linear scale (lower graph) shows considerable differences. Source: Our World in Data.

RECOMMENDATIONS AND CONCLUSIONS

The Corona virus can only be overcome through joint efforts, both internationally and through the participation of all members of society. Furthermore, society's future acceptance of scientific evidence communicated through visualizations and simulations may pave the way for modern societies resilience to many more future challenges such as global warming, war or the problems identified by the United Nations Sustainable Development Goals (SDGs). The willingness to comply with decided measures requires citizens to understand the context, facts and rules. For measures to be effective, decision-makers need to provide transparent and convincing explanations for their decisions. The Deutsche Arbeitsgemeinschaft Statistik, an umbrella organization of 18 academic statistics associations in Germany (DagStat, 2021) recently released a declaration on data and statistics as basis for political decision making. Among several other aspects, this declaration contained recommendations for stakeholders in politics, media and statistics agencies on how to communicate empirical evidence to the public efficiently. We wholeheartedly endorse these recommendations.

1. Existing uncertainties due to the novel situation and the still insufficient data basis should be communicated consistently.
2. Transparent communication of decision models and risks and uncertainties can promote active endorsement of recommended preventive measures and responsible health-related behavior. The data basis including the origin of the data, the scope, and the collection methods, should be transparent.

3. Simulations are a very effective means of illustrating the dynamics of infection events. When they are used, underlying model assumptions should be made transparent.
4. Figures should be related to a unifying benchmark (e.g. X deaths per 100,000 population). For better classification, they should be compared with other health risks and illustrated by graphics. Especially for percentages, it is important that the reference value is mentioned and that it is relevant to the question.
5. To promote statistical literacy at all levels, collaboration of statisticians with all stakeholders involved in statistical education in higher or secondary education or promoting statistical literacy among citizens is needed. These stakeholders include educators at all levels, school and university administrators, policy makers, official statistical providers, researchers, media professionals, teacher educators, software developers, and many others.

REFERENCES

Deutsche Arbeitsgemeinschaft Statistik (2021). Stellungnahme der DAGStat Daten und Statistik als Grundlage für Entscheidungen: Eine Diskussion am Beispiel der Corona-Pandemie. *https://www.dagstat.de/fileadmin/dagstat/documents/DAGStat_Covid_Stellungnahme.pdf*

Engel, J., Biehler, R., Frischemeier, D., Podworny, S., Schiller, A. & Martignon, L. (2019). Zivilstatistik: Konzept einer neuen Perspektive auf Data Literacy und Statistical Literacy. *AStA Wirtsch Sozialstat Arch* 13. https://link.springer.com/article/10.1007/s11943-019-00260-w.

Gigerenzer, G., Gaissmaier, W., Kurz-Milcke, E., Schwartz, L. M. & Woloshin, S. (2007). Helping doctors and patients make sense of health statistics. *Psychological Science in the Public Interest, Supplement*, *8*(2), 53–96. https://doi.org/10.1111/j.1539-6053.2008.00033.x

Ledford, H. (2021). How Could a COVID Vaccine Cause Blood Clots? *Nature 592*, 334-335. https://doi.org/10.1038/d41586-021-00940-0

Martignon, L, Mousavi, S., & Engel, J. (2021*).* Democratic societies defeat (COVID-19) disasters by boosting shared knowledge. *Mind & Society.* https://doi.org/10.1007/s11299-021-00278-0

McDowell, M. E., Gigerenzer, G., Wegwarth, O. & Rebitschek, F. G. (2019). Effect of tabular and icon fact box formats on comprehension of benefits and harms of prostate cancer screening: A randomized trial. *Medical Decision Making*, *39*, 41-56. https://doi.org/10.1177/0272989X18818166

ProCivicStat Partners (2018). Engaging civic statistics: a call for action and recommendations. A product of the ProCivicStat project. *http://IASE-web.org/ISLP/PCS*

Ridgway, J. ed. (Forthcoming 2021). *Statistics for Empowerment and Social Engagement –teaching Civic Statistics to develop informed citizens*. Springer.

Vittert, L., & Podkul, A. (2020). The Coronavirus Exponential: A Preliminary Investigation into the Public’s Understanding. *Harvard Data Science Review*. https://doi.org/10.1162/99608f92.fec69745

[1] https://www.washingtonpost.com/graphics/2020/world/corona-simulator/

[2] https://web.br.de/interaktiv/corona-simulation/

[3] ttps://www.nytimes.com/interactive/2020/us/coronavirus-spread.html

[4] https://wintoncentre.maths.cam.ac.uk/news/communicating-potential-benefits-and-harms-astra-zeneca-covid-19-vaccine/

[5] https://www.ft.com/content/a2901ce8-5eb7-4633-b89c-cbdf5b386938

[6] https://www.zeit.de/wissen/gesundheit/2020-11/coronavirus-aerosole-ansteckungsgefahrinfektion-hotspot-innenraeume