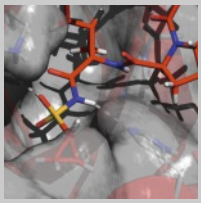
doi:10.2533/chimia.2020.613 Chimia 74 (2020) 613–614



Medicinal Chemistry and Chemical Biology Highlights

Division of Medicinal Chemistry and Chemical Biology
A Division of the Swiss Chemical Society

Communicating Near Real-Time Data During the COVID-19 Pandemic

Daniel Probst*

*Correspondence: Dr. D. Probst, E-mail: daniel.probst@dcb.unibe.ch; Department of Chemistry and Biochemistry, University of Bern

Keywords: COVID-19 · Data analysis · Data visualization · Public health · Scientific communication

Introduction

During the ongoing COVID-19 pandemic caused by SARS-CoV-2, near real-time population-scale health data, which usually evokes a relatively low degree of public interest compared to other near real-time data sources such as sports events, have become a mainstay in journalistic publications. Meanwhile, models and extrapolations based on these data, including those of non-peer-reviewed studies, are being readily disseminated to the public. While visualizing and communicating data within the scientific community is almost exclusively limited to finite data, great care must be taken when visualizations based on near real-time data are published to a broader audience. Near real-time data only allows for limited analysis with increased uncertainty and, in the current situation, highly depends on local or national policies and unclear diagnostic rules, among other strong confounders. Also, conclusions and projections based on the available data can change daily. Nevertheless, the data visualization must be interpretable by a person without scientific training while avoiding being unintentionally misleading due to simplifications. When communicating data to the public, available data visualization methods have to be selected with care, adapted, and well explained to be interpretable by a broad audience while the creator has to uphold scientific rigor during the creation of the visualization. In addition, methods widely applied in a scientific context such as logarithmic scales can easily be misleading to members of the general public. Here I discuss the web-based dashboard *corona-data.ch*, with which I made data gathered by cantonal authorities on COVID-19 accessible to the public through visualizations.[1]

Aggregating and Evaluating Near Real-Time Data on COVID-19

During the first weeks of the COVID-19 pandemic, the access to case numbers in Switzerland was limited, as there was no system in place that transparently and efficiently aggregated cantonal case numbers. The Federal Office of Public Health (FOPH) started to release cantonal numbers for COVID-19 cases on its website on the 8. March 2020 (with a total of 281 cases) as a daily updated PDF document. However, for three days between the 15. and 17. March, the data released by the FOPH no longer contained cantonal numbers, prompting me to manually check the cantonal websites for their respective case numbers and running web crawlers where applicable (a task that has been taken over by the Specialist Unit for Open Government Data Canton of Zurich as of mid-March).[2] At this point, it became evident that the data reported by the FOPH was different from those reported by the cantons, with the FOPH reporting significantly lower numbers; however, no information fully explaining the discrepancies between the two data sources has been released by the cantons nor by the FOPH. Maximizing the overlap of the cumulative epidemic curves defined by the historical data sets reported by the cantons and the FOPH (FOPH archive) and the data set that is currently being made available by the FOPH (which is assumed to be based on test dates rather than the case number publication dates), it can be shown that the cantonal data was reported approximately one day after testing, constituting a timely release of the information by the cantons. In contrast, the FOPH data was published approximately three days after testing during the recent peak of the pandemic in Switzerland (Fig. 1a).

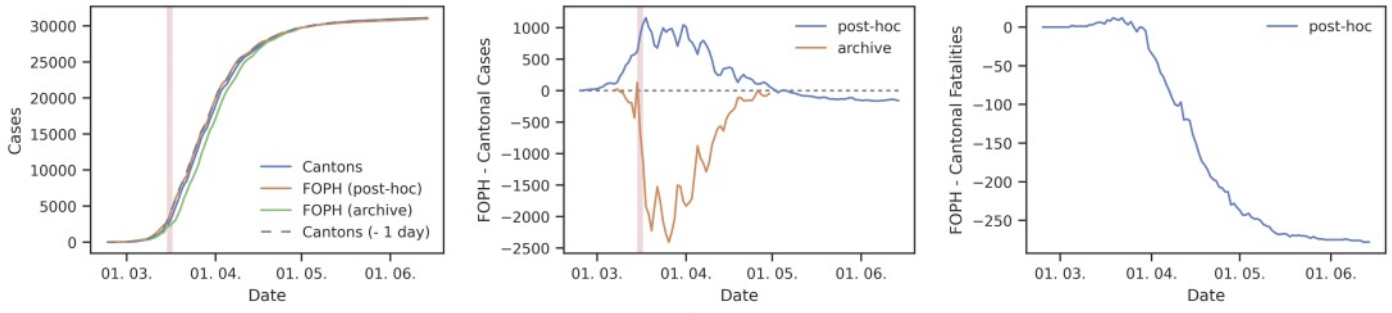


Fig. 1. Visualizing the discrepancies between FOPH and cantonal data sources during the peak of the recent COVID-19 pandemic in Switzerland. (a) By maximizing the overlap of the cumulative epidemic curves based on cases reported by the FOPH (archive) and the cantons with the curve derived from the later released true date-of-test FOPH (post-hoc) data, it can be shown that the curve based on cantonal data was shifted to the right by one day while the curve based on the FOPH was shifted to the right by three days. (b) Deviations of the case numbers reported by the FOPH (archive) and the FOPH (post-hoc) data from the data reported by the cantons during the recent pandemic peak. (c) As yet unexplained gap in fatalities between data reported from the FOPH and the cantons.

Can you show us your Medicinal Chemistry and Chemical Biology Highlight?
Please contact: Dr. Cornelia Zumbrunn, Idorsia Pharmaceuticals Ltd., E-mail: cornelia.zumbrunn@idorsia.com

This delay in reporting led to an apparent underreporting by the FOPH during March and April as compared to the cantonal data (Fig. 1b, orange); in contrast, the cantonal data was significantly closer to the actual case numbers (Figure 1b, blue). While the difference in reported case numbers between the FOPH and the cantons can be explained by this three-fold comparison of the released data, the reported number of COVID-19-related fatalities still exhibits unexplained discrepancies, with cantonal sources reporting more than 200 additional deaths compared to the FOPH as of 16 June (Fig. 1c).

Data Visualization Methods used for COVID-19 Data

Starting in early March, I created static line plots, choropleth maps, as well as a heatmap visualizing the absolute day-over-day growth per canton based on FOPH case numbers and made them available online (*http://corona-ch.surge.sh/*).[3] While the line plots and choropleth maps provided valuable insights into the geographic distribution and the initial exponential growth, respectively, the heatmap visualizing the day-over-day growth rate was highly problematic as it would quickly approach zero in the absence of exponential growth, thus hiding any sub-exponential growth behind a misleading decrease in growth rates. Moving the visualizations to the interactive dashboard *corona-data.ch* (Fig. 2a),[1] based on the open-source project Dash and a Python backend, the problematic heatmap was dropped. Compared to many other COVID-19-related websites at the time, I refrained from publishing predictions; instead, I focused on visualizing the available data as understandable as possible. A common mistake of many available dashboards and media outlets was to use logarithmic scales. While non-linear scales are a common occurrence in scientific literature, they were shown to have the potential to be highly misleading to the general public (Fig. 2b,c).[4] In order to minimize the potential for misleading the public, graphs shown on *corona-data.ch* applied a linear scale by default while keeping the option to switch to a logarithmic scale. An exception were the graphs showing a log-log plot displaying weekly new cases vs. the total number of cases developed by Bhatia *et al.* (*https://github.com/aatishb/covidtrends*),[5] which were used as a supplemental visualization of Swiss and cantonal cases. In addition, all plots were augmented with captions and annotations, describing the reasoning behind the plots as well as any caveats where applicable, including warnings for incomplete data.

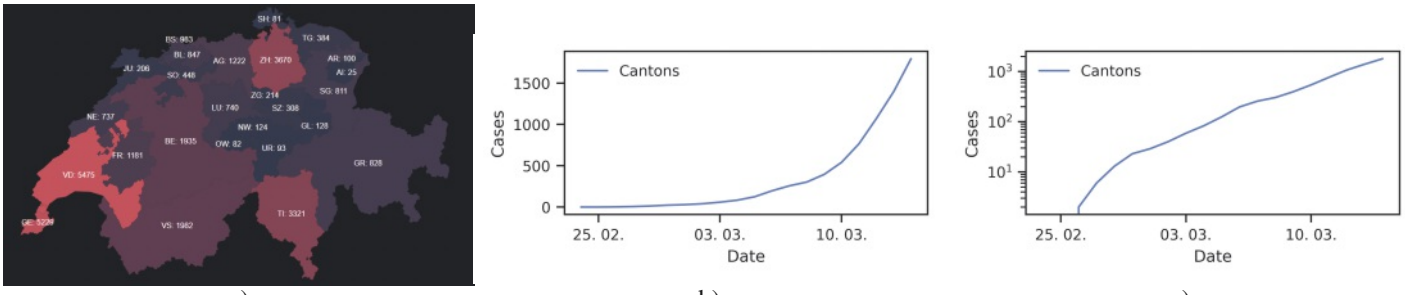
Conclusion

Based on the overall positive feedback by the public and more than 14 million visitors between the 26 March 2020 and 15 June 2020, a case can be made that aside from timeliness, transparency regarding data source and quality, as well as simplicity and interpretability of the data representation are of high importance when communicating near real-time population-scale health data to the public. As the public interest shifts from case numbers to emerging therapeutics and vaccines, a similar approach towards the reported outcomes could be immensely beneficial to the public's understanding of the current state and the overall process of drug development regarding COVID-19.

The data discussed in this article is available online (*https://github.com/daenuprobst/corona_data_retrospective*).

Received: June 17, 2020

[1] COVID-19 Information for Switzerland, Latest updates of COVID-19 development in Switzerland, accessed 16 June, **2020**.
[2] *opendata.swiss, https://opendata.swiss/en/*, accessed June 16, **2020**.
[3] Corona Figure Dump, *http://corona-ch.surge.sh/*, accessed 16 June, **2020**.
[4] A. Romano, C. Sotis, G. Dominioni, S. Guidi, 'COVID-19 Data: The Logarithmic Scale Misinforms the Public and Affects Policy Preferences', Social Science Research Network, Rochester, NY, **2020**, DOI: 10.2139/ssrn.3588511.
[5] A. Bhatia, 'aatishb/covidtrends', **2020**.



a) b) c)

Fig. 2. Visualizing near real-time data on *corona-data.ch.* (a) The choropleth map displaying the current case counts per canton as shown on top of *corona-data.ch*. The epidemic curve in an early stage (b, linear; c, logarithmic; first 20 days of the pandemic in Switzerland) is an example of how logarithmic scales can be misleading to many readers unfamiliar with the method.