Data Science in a Time of Crisis: Lessons from the Pandemic

Chiara Sabatti and John M. Chambers

The exceptional shock of the COVID-19 pandemic has brought about an equally exceptional scientific response, over a wide range of disciplines and with a spirit of collaboration and mutual support.

This issue of *Statistical Science* contains five reports on research related to the pandemic and four perspectives on lessons learned and thoughts for the future. We hope the issue will contribute to ongoing exploration and discussion from the special perspective of the *Statistical Science* community, particularly in the context of avoiding or managing future crises.

As with any substantial data science challenge, three fundamental aspects are involved: *data*—obtaining, organizing and validating the information relevant to the study; *analysis*—models, summaries and other computations motivated by the scientific questions; and *communication*—valid information derived from the study in a form helpful for the scientific community, for decision makers or for the public.

The COVID-19 pandemic has challenged the scientific response in all aspects. Data on the prevalence and effects of the disease fall short because of the rapidity of the spread, the wide range of symptoms and especially from the global context. The accuracy, consistency and simple availability of the relevant data suffer. Future fixes will need improvements in policy but statistical techniques can to some extent compensate. Pooled testing can use models to improve accuracy (Comess et al. [3]) and analytic techniques can compensate for time delays (Jahja, Chin and Tibshirani [4]).

The relevant science ranges from the most basic genomic understanding of the virus to all the social, economic and other human effects of the pandemic. At the genomic level, powerful models are available but with challenges at the limits of analysis and computing, as discussed by Cappello et al. [1].

Communication of the scientific insights is arguably the challenge needing the greatest novelty of approach. For the pandemic, this has to encompass a wide variety of listeners: those directly involved in the fight; those involved in formulating policy; and the whole population affected by the results. Wang et al. [8] present analysis and graphical techniques to aid patient monitoring in the crucial hospital context. Nicholson et al. [7] introduce the concept of *interoperability* as a framework for communication between policy makers and statistical analysis.

The technical papers are supplemented by four perspectives. Yu and Singh [9] distill seven principles from their experience with the pandemic. Mukherjee [6] and Lin [5] recount their experiences, in India and with the early Wuhan data respectively, and go on to present reflections with implications for dealing with future epidemics. Chambers [2] considers how science might respond to a broad range of still graver threats facing us, using the COVID-19 experience and the example of Bell Labs research as reference points.

Any single journal issue can only report a very small fraction of the dramatic surge in activity. To the credit of the profession, many statisticians engaged directly with the analysis and modeling of data related to the pandemic. They provided support for local, national and international health organizations. Presentation of their analysis helped the public to interpret information and assisted authorities in designing policies. The vast majority of the researchers involved in these efforts had not specialized in such data before spring 2020.

Our goal for this special issue was to highlight a few of these contributions. In no sense could we hope to find a "representative" sample, let alone a "best" selection. Given the desire to produce the issue in a reasonable time frame, and considering the heavy commitments of potential authors, we used a thoroughly *ad hoc* search process. A first targeted request for contributions was sent out in early autumn 2020. We also sought suggestions for additional contributions, leading to further invitations. This very unsystematic sampling is inevitably biased, notably towards United States activities, for which we apologize. The international perspectives in the papers by Nicholson et al., Mukherjee and Lin are particularly appreciated.

Inevitably also a number of technical topics are conspicuously absent; for example, studies related to vaccines or to the health disparities exposed by the pandemic. Much other research has taken place and still more challenges exist for future investigation. The perspective by Mukherjee has a valuable summary of important areas of statistical research.

Chiara Sabatti is Professor, Biomedical Data Science and of Statistics, Stanford University, Stanford, California 94305-2070, USA (e-mail: sabatti@stanford.edu). John M. Chambers is Adjunct Professor in Statistics and Senior Advisor, Data Science, Stanford University, Stanford, California 94305-2070, USA (e-mail: jmc4@stanford.edu).

In the event, we are impressed by the range and quality of the papers included and thankful to all the authors for their efforts, given their pressing commitments. Our special thanks to the Editor, Sonia Petrone, for her generous contribution of time and advice throughout the process. Thanks also to the IMS and its technical staff for assistance and particularly for agreeing to make the issue open access.

Lastly, one historical coincidence gives cause for reflection. Our subtitle "Lessons from the Pandemic" corresponds intriguingly to an article in the May 25, 1919 issue of *Science*, "The Lessons of the Pandemic". The influenza pandemic of 1918-20 was indeed devastating at a level beyond the current experience: some 20 to 50 million deaths in a human population four times smaller than in 2020. The *Science* article made some notable points: many of the preventive steps suggested are still relevant; data on individuals and communities is needed to understand relative risk; and perhaps most surprisingly, a guess that the pandemic might have been triggered by zoonotic transmission from another species.

The article's first and most emphatic observation, however, was science's ignorance of the fundamental mechanism leading to the illness: the mysteries of "what the disease was, where it came from, how to stop it". A century later, science has much to say on these questions, as the articles in this issue illustrate, but as usual the advances have posed new questions. More fundamentally, the exponential growth of humanity and its influence raise the complexity and urgency of the scientific response.

Most of all, let's hope that the spirit of openness, collaboration and support of science displayed in fighting the pandemic can be exported and augmented to respond to future, even more serious challenges.

REFERENCES

[1] CAPPELLO, L., KIM, J., LIU, S. and PALACIOS, J. A. (2022). Statistical challenges in tracking the evolution of SARS-CoV-2. *Statist. Sci.* **37** 162–182.

[2] CHAMBERS, J. M. (2022). Data, science, and global disasters. *Statist. Sci.* **37** 284–288.

[3] COMESS, S., WANG, H., HOLMES, S. and DONNAT, C. (2022). Statistical modeling for practical pooled testing during the COVID-19 pandemic. *Statist. Sci.* **37** 229–250.

[4] JAHJA, M., CHIN, A. and TIBSHIRANI, R. J. (2022). Real-time estimation of COVID-19 infections: Deconvolution and sensor fusion. *Statist. Sci.* **37** 207–228.

[5] LIN, X. (2022). Lessons learned from the COVID-19 pandemic: A statistician's reflection. *Statist. Sci.* **37** 278–283.

[6] MUKHERJEE, B. (2022). Being a public health statistician during a global pandemic. *Statist. Sci.* **37** 270–277.

[7] NICHOLSON, G., BLANGIARDO, M., BRIERS, M., DIGGLE, P. J., FJELDE, T. E., GE, H., GOUDIE, R. J. B., JERSAKOVA, R., KING, R. E. et al. (2022). Interoperability of statistical models in pandemic preparedness: Principles and reality. *Statist. Sci.* **37** 183–206.

[8] WANG, Z., BOWRING, M. G., ROSEN, A., GARIBALDI, B., ZEGER, S. and NISHIMURA, A. (2022). Learning and predicting from dynamic models for COVID-19 patient monitoring. *Statist. Sci.* **37** 251–265.

[9] YU, B. and SINGH, C. (2022). Seven principles for rapid-response data science: Lessons learned from Covid-19 forecasting. *Statist. Sci.* **37** 266–269.