OPINION PAPER

Two years of COVID-19 pandemic: the Italian experience of Statgroup-19

Giovanna Jona Lasinio[1] | Fabio Divino[2] | Gianfranco Lovison[3] | Marco Mingione[6] | Pierfrancesco Alaimo Di Loro[4] | Alessio Farcomeni[5] | Antonello Maruotti[4]

[1]Dpt. of Statistical Sciences, "La Sapienza" University of Rome, Rome, Italy

[2]Dpt. of Bio-Sciences, University of Molise, Italy

[3]Dpt. of Economics, Management and Statistics, University of Palermo, Palermo, Italy

[4]Dpt. GEPLI, Libera Univerità Maria Ss. Assunta (LUMSA), Rome, Italy

[5]Dpt. of Economics and Finance, University of Rome "Tor Vergata", Rome, Italy

[6]Dpt. of Political Sciences, University of Roma Tre, Rome, Italy

Correspondence
Giovanna Jona Lasinio
Email: giovanna.jonalasinio@uniroma1.it

Funding information
FISR?

The amount and poor quality of available data and the need of appropriate modelling of the main epidemic indicators require specific skills. In this context, the statistician plays a key role in the process that leads to policy decisions, starting with monitoring changes and evaluating risks. The "what" and the "why" of these changes represent fundamental research questions to provide timely and effective tools to manage the evolution of the epidemic. Answers to such questions need appropriate statistical models and visualization tools. Here, we give an overview of the role played by Statgroup-19, an independent Italian research group born in March 2020. The group includes seven statisticians from different Italian universities, each with different backgrounds but with a shared interest in data analysis, statistical modelling, and biostatistics. Since the beginning of the COVID-19 pandemic the group has interacted with authorities and journalists to support policy decisions and inform the general public about the evolution of the epidemic. This collaboration led to several scientific papers and an accrued visibility across various media, all made possible by the continuous interaction across the group members that shared their unique expertise.

KEYWORDS
COVID-19, Epidemic data, Data quality

1 | INTRODUCTION

The scientific method provides humanity with rules to distil knowledge from the observed reality, draw inferences about its roots, and anticipate its evolution. Its effective operation requires careful observation of our surroundings. The latter can be converted into data for analysis and a trusting relationship between the scientific community and the institutions it interfaces with. While the worldwide struggle against COVID-19 has favoured scientific progress in several fields, it also highlighted generalised frailties in the data collection systems and a lack of adequate management and communication tools. A large part of the scientific literature on the matter foresaw this event as early as fifty years ago, but it was left unheard, so the abrupt spread of the virus caught us unprepared. Spillover events will likely increase in frequency because of global warming and bad hygiene practices in human-animal and animal-animal interactions. The hard lessons learned during the past two challenging years should be treasured and not forgotten. They can make the difference in managing the current pandemic and preventing and controlling future ones. This paper discusses our experience as a working group of statisticians collaborating on analysing, monitoring, and forecasting the COVID-19 pandemic. The statistical community has never been under the spotlight as much as during these proving times. Here, we briefly highlight some of the most relevant challenges and dangerous pitfalls the community had to face, by sharing the story of four lessons we learnt along the way.

2 | DATA AVAILABILITY, QUALITY AND MANAGEMENT ARE CRUCIAL ISSUES

The COVID-19 pandemic taught us that data collection, management, and availability are more of a key-issue than the statistical methods per-se in the context of statistics-for-policy. Publicly available data were often scarce and inconsistent, especially during the first year of the pandemic. Sadly enough, this problem still persists two years and a half after it all began. On top of that, access to the limited sources of more accurate information was hindered by data collection issues, unclear privacy policies, and institutional unwillingness to share data. All of the above delayed the scientific community's understanding of the pandemic dynamics and the possibility of adequately implementing data-driven policy-making.

Most of the modelling efforts of our group, as those of most of the statistical community, were indeed addressed toward the management of unreliable, spatio-temporally misaligned, incoherent and incomplete data. Data quality issues called for strategies considering relevant uncertainty in the estimation process. Poor quality hampered the application of highly sophisticated statistical models. These models require timely and high-quality data to produce accurate and reliable results. We strongly believe that health monitoring systems, not limited to the COVID-19 situation, should be set up locally and nationally standardised. Anonymised data from real-time surveillance systems for privacy reasons should become freely available to all researchers. That would be in the spirit of the open-data community driving the fast scientific progress of our era.

3 | COMMUNICATING STATISTICAL RESULTS EFFECTIVELY IS AS IMPORTANT AS PRODUCING THEM APPROPRIATELY

Until recently, the need for harmonised data collection systems and other data quality issues have been of minor concern to the general public. It is the COVID-19 pandemic that pointed the public attention to the need for high-quality data and their role in guiding decision-making. Such awareness led to the involvement of eminent scientists in the governments' decision process and the national media. The critical role of the statistician in this apparatus has been debated across the world, with different fortunes in different nations. For example, the Guardian in the UK set the column The weekly stats uncovered edited by the well known statisticians David Spiegelhalter and Anthony Masters. Other countries, such as Italy, followed a very fuzzy path. Scientific committees were set up, involving a variety of scientists but none with a statistical background. Most of the media of these countries gave space to more appealing than more competent scientists, often neglecting their statistical skills. That choice despite the Italian health data, like those of many other countries, require a solid statistical background to be adequately analysed. Indeed, the country is divided into 19 administrative regions and two autonomous provinces, each with its health system and data handling procedures. This stream of information flows to Rome at the National Health Institute, where it is tentatively cleaned and made available. The system works reasonably well when time is not an issue; otherwise, a huge amount of noise flows together with the authentic information. That issue soon highlighted the need to involve statisticians in the day-to-day monitoring of the pandemic, especially in its early stages. That is why the Italian statistical community, together with a large opinion group (more the 50,000 people signed the petition `https://www.datibenecomune.it/`), took action and asked for freely available data and more competent committees to be involved in the decision-making process. The public action encouraged a nationwide discussion on the matter, but the requests ultimately fell on deaf ears.

4 | THE NATURE AND COMPLEXITY OF THE CHOSEN STATISTICAL MODELS SHOULD TAKE THE QUALITY OF DATA INTO ACCOUNT

A computational temptation is often driving researchers attention away from data-quality issues. The almost unlimited computational possibilities of our time point toward methodological developments while overlooking data quality. Again, the pitfalls of data quality affected studies that were not backed by a solid statistical background. Sophisticated mathematical models were developed and proved to be ineffective in analysing noisy data like the COVID-19 surveillance data. During epidemiological emergencies, the data cannot be collected under a stable collection scheme because the surveillance conditions change as the epidemiological process evolves. That may be due to: public health authorities' interventions, such as introducing diagnostic tests with different sensitivities/specificities or restrictions; population behavioral adjustments; changes in the spread dynamics, etc.

Therefore, the data generating mechanism changed during the pandemic and introduced a large amount of noise in the collected data. Any model attempting to describe this phenomenon should be defined coherently to the surveillance conditions and account for those issues not to bias the estimates and to adequately quantify the overall uncertainty. For instance, Statgroup-19 and other research groups attempted to provide solutions based on empirical models that could still embed some epidemiological notion in their definition (Mingione et al., 2022; Girardi et al., 2022; Alaimo Di Loro et al., 2021; Farcomeni et al., 2021; Chowell et al., 2016). These data-driven models proved particularly suitable for the COVID-19 *dirty* data, as opposed to compartmental models (Diekmann et al., 2013) that often provided inaccurate inferences (Ioannidis et al., 2020). While theoretically most suitable for modelling epidemic dynamics, the

latter relies on the accurate initial estimates of several key quantities and can greatly suffer the consequences of poor data input. As a general guide, when the poor quality of the data does not allow for correctly estimating sophisticated models, the researcher should prefer robustness over precision. Statisticians know well how to perform this task successfully. They know the dangers and difficulties of dealing with dirty data and fitting miss-specified models. They know how to correct such miss-specifications and properly model random components to account for unobserved heterogeneity.

5 | WHEN IT COMES TO ASSESSING THE EFFICACY OF POLICIES, DISTINGUISHING BETWEEN CORRELATION AND CAUSATION IS PARAMOUNT

Much research has been devoted to assessing the effects of various pharmaceutical and non-pharmaceutical interventions (PI and NPI), especially in 2020, when very little was known about this new disease. With a few noticeable exceptions (see, for example, Mader and Rüttenauer, 2022; Steiger et al., 2021), most of these analyses were carried out in correlational terms, that is, by observing the magnitude and direction of changes in relevant outcomes (number of cases, number of hospitalisations, number of intensive care admissions, etc.) corresponding to changes in policies: closing/opening of schools, enforcement/relaxation of legal requirements in terms of mask-wearing or social distancing, etc. (see, for example, Brauner et al., 2021; Haug et al., 2020). Although often helpful in supporting public decisions, these studies can never be conclusive. In a complex setting like a pandemic, it is paramount to try a causal approach, taking the many confounders that can bias the conclusions of correlational studies into explicit account. For example, in Italy, there has been a lively debate about the claim that the re-opening of schools in Autumn 2020 was the leading cause of the resurgence of virus circulation, which gave rise to the second wave of the pandemic. However, little effort has been devoted to a genuinely causal approach. That should have considered some of the main confounders, such as: (i) meteorology - the re-opening of schools at the end of September corresponds to the end of Summer and the beginning of the colder season, a well-known risk factor for respiratory viruses; (ii) simultaneous re-opening of most economic activities, e.g. shops, factories, etc. and the related traffic and increased people circulation, as the re-opening of both schools and economic activities increase the use of public transportation; (iii) the arrival of new and more aggressive variants, as the *alpha* variant in October 2020.

On the one hand, it is crucial to warn against using purely correlational studies to support decision-making, as they expose to the risk of making decisions on the grounds of spurious relationships. On the other hand, rigorous causal studies on the efficacy of pharmaceutical and not pharmaceutical interventions require disaggregated or individual data for estimating counterfactual models. These data were not available. That reinforces the request for better data discussed in Sec. 3.

6 | CONCLUSIONS

From a purely political point of view, not much has changed for the Italian statistical community. However, something changed in the media perception of our profession. Newspapers and TV talk shows interested in gathering and sharing factual information started inviting data scientists and statisticians to discuss the flow of numbers. On the contrary, other means of communication preferred sensational headlines, blending accurate information and fake news altogether. That partly contributed to the considerable confusion about vaccines and their effectiveness, especially during the first vaccination campaign in Italy. This problem persists as about 4 million Italians never got a vaccine dose: how is this possible? The scientific community must take responsibility for part of this unfortunate result. Scientists

are creative people, a bit like artists. As artists, they live in very competitive environments and develop great determination during their working lives. Many of them develop a hypertrophic ego, which sometimes lets scientific rigour slip into the backseat; professional ethic is bent for the benefit of oneself, and the headline in a newspaper becomes a better and faster reward than a paper in a scientific journal. Let us be honest: scientists are human beings subject to desires and passions and can forget their responsibilities as anybody else. However, the damage may be massive if they forget their role in highly uncertain periods like the one we are still living in. The pandemic took a huge toll on us, and since February 2022 the Ukrainian war is adding more insecurity to our daily lives.

Our international community should speak loud, insisting on the request for open data and transparency in the decision-making process. We need to genuinely understand the world around us more than ever. **Ask a statistician!**, this should be our future motto.

references

Alaimo Di Loro, P., Divino, F., Farcomeni, A., Jona Lasinio, G., Lovison, G., Maruotti, A. and Mingione, M. (2021) Nowcasting covid-19 incidence indicators during the italian first outbreak. *Statistics in Medicine*, **40**, 3843–3864.

Brauner, J. M., Mindermann, S., Sharma, M. and *et al.* (2021) Inferring the effectiveness of government interventions against covid-19. *Science*, **371**, eabd9338.

Chowell, G., Hincapie-Palacio, D., Ospina, J., Pell, B., Tariq, A., Dahal, S., Moghadas, S., Smirnova, A., Simonsen, L. and Viboud, C. (2016) Using phenomenological models to characterize transmissibility and forecast patterns and final burden of zika epidemics. *PLoS currents*, **8**.

Diekmann, O., Heesterbeek, H. and Britton, T. (2013) *Mathematical tools for understanding infectious disease dynamics*, vol. 7. Princeton University Press.

Farcomeni, A., Maruotti, A., Divino, F., Jona-Lasinio, G. and Lovison, G. (2021) An ensemble approach to short-term forecast of covid-19 intensive care occupancy in italian regions. *Biometrical journal*, **63**, 503–513.

Girardi, P., Greco, L. and Ventura, L. (2022) Misspecified modeling of subsequent waves during covid-19 outbreak: a change-point growth model. *Biometrical Journal*, **64**, 523–538.

Haug, N., Geyrhofer, L., Londei, A. and *et al.* (2020) Ranking the effectiveness of worldwide covid-19 government interventions. *Nature Human Behaviour*, **4**, 1303–1312.

Ioannidis, J. P., Cripps, S. and Tanner, M. A. (2020) Forecasting for covid-19 has failed. *International journal of forecasting*.

Mader, S. and Rüttenauer, T. (2022) The effects of non-pharmaceutical interventions on covid-19 mortality: A generalized synthetic control approach across 169 countries. *Frontiers in Public Health*, 740.

Mingione, M., Alaimo Di Loro, P., Farcomeni, A., Divino, F., Lovison, G., Maruotti, A. and Lasinio, G. J. (2022) Spatio-temporal modelling of covid-19 incident cases using richards' curve: An application to the italian regions. *Spatial Statistics*, **49**, 100544.

Steiger, E., Mussgnug, T. and Kroll, L. E. (2021) Causal graph analysis of covid-19 observational data in german districts reveals effects of determining factors on reported case numbers. *PloS one*, **16**, e0237277.