SFI TRANSMISSION

COMPLEXITY SCIENCE FOR COVID-19

STRATEGIC INSIGHT: **We must use a modeling approach to COVID-19 data that will yield the least biased inference and prediction.**

FROM: **Amos Golan, American University; Santa Fe Institute; and Pembroke College, Oxford University**

DATE: 6 July 2020 **NO:** 035.1

As the world faces the possibility of recurring waves of the current novel coronavirus pandemic, it is critical to identify patterns and dynamics that could be leveraged to decrease future transmission, infection, and death rates. At this stage in the pandemic, data on disease patterns and dynamics are emerging from almost all countries in the world. Variations across countries with respect to coronavirus infection rates, public health policies, social structure, norms, health conditions, environmental policy, climate, and other factors provide us with the data to investigate the impact of different underlying factors and governmental policies on COVID-19 transmission, infection, and death rates.

Despite the fact that millions have been infected and hundreds of thousands have died from COVID-19, the available information is still insufficient for reaching precise inferences and predictions. This is because the available data on each patient are very limited, the variables of interest are highly correlated, and great uncertainty surrounds the underlying process. In addition, though the death rate from COVID-19 is high relative to other infectious diseases, from an inferential point of view, it is still very small since the number of deaths relative to those who did not die is extremely small. As a result, the observations are in the tail of the survival probability distribution. In short, the available data for analysis of COVID-19 are complex, constantly evolving, and ill-behaved. Inferring and modeling with such data results in a continuum of explanations and predictions. We need to use a modeling and inferential approach that will yield the least biased inference and prediction. Unfortunately, traditional approaches impose strong assumptions and structures — most of which are incorrect or cannot be verified — leading to biased, unstable, and misguided inferences and predictions. Information theory offers a solution. It provides a rational inference framework for dealing with mathematically underdetermined problems, allowing us to achieve the least biased inferences.

An information-theoretic approach — specifically, info-metrics — is situated at the intersection of information theory, statistical inference, decision-making under

uncertainty, and modeling. In this framework, all information enters as constraints plus added uncertainty within a constrained optimization setup, and the decision function is an information-theoretic one. That decision function is defined simultaneously as the entities of interest — say, patients' survival probabilities — and the uncertainty surrounding the constraints. That framework extends the maximum entropy principle of Jaynes,[1] which uses Shannon's entropy[2] as the decision function for problems that are surrounded with much uncertainty.[3] Info-metrics has clear parallels with more traditional approaches, where the joint choice of the information used (within the optimization setting) and a particular decision function will determine a likelihood function. The encompassing role of constrained optimization ensures that the info-metrics framework is suitable for constructing and validating new theories and models, using all types of information. It also enables us to test hypotheses about competing theories or causal mechanisms. For certain problems, the traditional maximum likelihood is a special case of info-metrics.

The info-metrics approach is well suited to dealing with the complex and uncertain cross-country COVID-19 pandemic data, specifically the relatively small sample size of detailed data, high correlations in the data, and the observations in the tail of the distribution. For this analysis, we developed a discrete-choice, binary (recovered/died) model to infer the association between the underlying country-level factors and death. The model controls for age, sex, and whether the country had universal vaccination for measles and Hepatitis B. This information-theoretic approach also allows us to complement existing data with priors constructed from the death frequency (by age and sex) of individuals who were infected with Severe Acute Respiratory Syndrome (SARS). For the detailed study, see Golan et al.[4]

Using data from twenty countries published on the public server on April 24, 2020, our study fund a number of country-level factors with a significant impact on the survival rate of COVID-19. One of these is a country's past or present universal TB (BCG) vaccination. Another one is the air-pollution death rate in the country. Some quantified results (by age — the x-axis — and sex) are presented in the figure below. The left panel shows the predicted death probability conditional on a universal BCG vaccination. There are three universal vaccination possibilities: countries that never had it (say, the United States), that currently have it (say, the Philippines), or that had it in the past (say, Australia). The huge impact on survival rates, across ages, of a universal BCG vaccination, is clear. The right panel demonstrates the probability of dying conditional on air-pollution death — the number of deaths attributable to the joint effects of household and ambient air pollution in a year per 100,000 population. The continuous line reflects the 90th percentile of pollution. The dashed line reflects the 10th percentile of pollution.

The same framework can be used for modeling all other pandemic-related problems, even under much uncertainty and evolving, complex data. Examples include conditional Markov processes, dynamical systems, and systems that evolve simultaneously. The info-

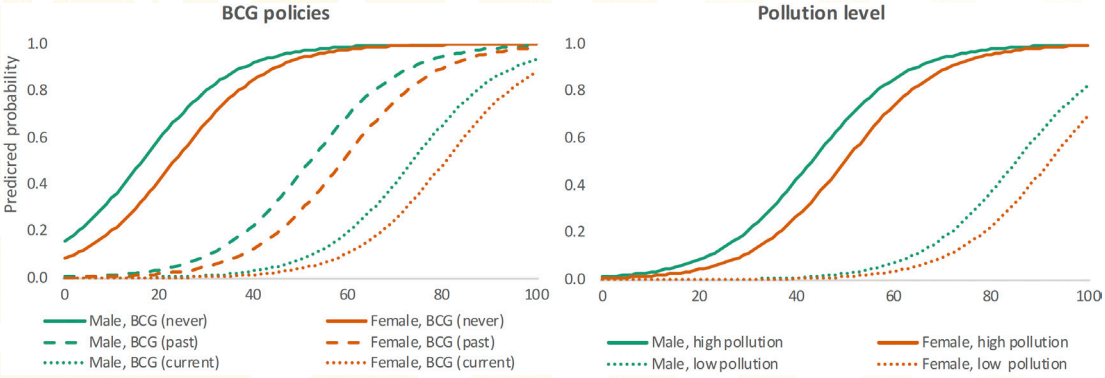

Figure. *The probability of dying conditional on BCG vaccination policies (left). The probability of dying conditional on Pollution (right) show the death rate in the 10th percentile (dots) vs. those at the 90th percentile (continuous). The x-axis is the patients' age.*

metrics framework allows us to construct theories and models and to perform consistent inferences and predictions with all types of information and uncertainty. Naturally, each problem is different and demands its own information and structure, but the info-metrics framework provides us with the general logical foundations and tools for approaching all inferential problems. It also allows us to incorporate priors and guides us toward a correct specification of the constraints — the information we have and use — which is a nontrivial problem.

So, should we always use info-metrics? To answer this, it is necessary to compare info-metrics with other methods used for policy analysis and causal inference. All inferential methods force choices, impose structures, and require assumptions. With complex and ill-behaved pandemic data, more assumptions are needed. Together with the data used, these imposed assumptions determine the inferred solutions. The assumptions and structures include the likelihood function, the decision function, and other parametric (or even nonparametric) assumptions on the functional form or constraints used. The reason for that is, without this additional information, all problems are under-determined. A logical way to compare different inferential approaches (classical and Bayesian), especially in relation to complex and ill-behaved pandemic data, is within a constrained optimization setup. That way, the comparison is on a fair basis as we can account for the information used in each approach.[3] But such a detailed comparison, including other approaches like agent-based models (ABM), deserves its own paper and is outside to scope of this essay. Here, I point toward two basic choices we need to make when using the info-metrics approach. First, the choice of the constraints; the constraints are chosen based on the symmetry conditions or the theory we know (or hypothesize) about the problem. They capture the rules that govern the system we study. Mathematically, they must be satisfied within the optimization. Statistically, if specified correctly, they are sufficient statistics. In the more classical and Bayesian approaches, the constraints are directly related to the parametric functional form used (say, linear, nonlinear, etc.). But specifying the constraints within info-metrics, or the functional

forms in other approaches, is far from trivial and affects the inferred solution. Info-metrics provides us with a way to falsify the constraints and points us in the direction of improving them. That choice, together with the decision function used, determines the exact functional form of the solution, or the inference.

The second choice we make in the info-metrics framework is constructing the constraints as stochastic. This is different than the classical maximum entropy approach where the constraints must be perfectly satisfied. This is also different than classical approaches where the likelihood and functional forms must be perfectly specified. But there is no free lunch. To achieve this more generalized framework, which allows us to model and infer a larger class of problems, we must bear the cost of specifying the bounds on the uncertainty. These bounds are theoretically or empirically derived. But, regardless of that derivation, it implies that what we give up is the assurance that our solution is first-best; rather, it may be a second-best solution, a solution describing an approximate theory, or the evolution of a complex theory derived from a mix of different underlying elements and distributions. The benefit is that whenever we deal with insufficient and uncertain information, it allows us to account for all types of uncertainties and to handle ill-behaved data. It provides us with a way to make inferences even under much uncertainty and ill-behaved data. Out of all possible methods, it is the one that uses the least amount of information and therefore tends to produce the least biased inference.

Whether it is more convenient or appropriate to choose a likelihood function or to determine the structure of the constraints from symmetry conditions and other information is a decision faced by each researcher. When approaching this decision, we should keep in mind that the constraints are only one part of the decision. The choice, however, of what method to use, depends on the problem we try to solve, the information we have, and the researcher's preference.

REFERENCES

1 Jaynes, E. T. (1957). "Information Theory and Statistical Mechanics." *Physical Review* 106 (4): 620–630. https://doi.org/10.1103/PhysRev.106.620.

2 Shannon, C. E. (1948). "A Mathematical Theory of Communication." *Bell System Technical Journal* 27: 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x.

3 Golan, A. (2018). *Foundations of Info-Metrics: Modeling, Inference, and Imperfect Information*. Oxford University Press. http://info-metrics.org.

4 Golan, A. et al (2020). "Effect of Universal TB Vaccination and Other Policy-Relevant Factors on the Probability of Patient Death from COVID-19," Working Paper 2020-041, Human Capital and Economic Opportunity Working Group (U Chicago).

Read more posts in the Transmission series, dedicated to sharing SFI insights on the coronavirus pandemic: santafe.edu/covid19