An Information-Theoretic Method for Identifying Effective Treatments and Policies at the Beginning of a Pandemic

Amos Golan[12]; Tinatin Mumladze[1]; Jeffrey M. Perloff [3]; Danielle Wilson[1]

Abstract

Background: Identifying effective treatments and policies early in a pandemic is challenging because only limited and noisy data are available, and biological processes are unknown or uncertain. Consequently, classical statistical procedures may not work or require strong structural assumptions. An information-theoretic approach can overcome these problems and identify effective treatments and policies. The efficacy of this approach is illustrated using a study conducted at the beginning of the COVID-19 pandemic.

Methods: An information-theoretic inferential approach with and without prior information was applied to the limited data available in the second month (April 24, 2020) of the COVID-19 pandemic. For comparison, a second statistical analysis used a large sample with millions of observations available at the end of the pandemic's pre-vaccination period (mid-December 2020).

Results: Even with limited data, the information-theoretic estimates performed well in identifying influential factors and helped explain why death rates varied across nations. Later experiments and statistical analyses based on more recent, richer data confirm that these factors contribute to survival.

Conclusions: An information-theoretic statistical technique is a robust method that can overcome the challenges of under-identified estimation problems in the early stages of medical emergencies. It can easily incorporate prior information from theory, logic, or previously observed emergencies.

[1]Department of Economics, American University, Washington, D.C. 20016
[2]Santa Fe Institute, Santa Fe, NM, 87501
[3]Department of Agricultural & Resource Economics, University of California Berkeley, Berkeley CA, 94720

Keywords: Information-theoretic inference, COVID-19, identification, mortality

Introduction

It's not if but when the next pandemic will strike.

When the next pandemic strikes, how can we choose treatments and policies to reduce deaths before a new vaccine is available? Eventually, we will have a plethora of data and an understanding of the relevant biology, so we can use standard statistical techniques to determine what we should have done using 20-20 hindsight. Unfortunately, at the start of a pandemic, we have few observations and a limited understanding of a disease's process, so typical statistical methods are infeasible or require strong, possibly inaccurate assumptions. We show that an information-theoretic inferential model using few observations works well without imposing heroic conjectures.

We demonstrated the efficacy of this approach in an analysis conducted in May 2020 [1] using data from the first few months of the COVID-19 pandemic, when only 485 individual observations from 20 countries were available. The study identified two factors, including an existing vaccination, associated with lower COVID-19 death rates. Later experiments and statistical analyses based on more recent, richer data confirm that these factors contribute to survival.

Methods

Statistical inference with uncertainty and little information results in multiple possible solutions, each consistent with the observed information because the problem is underdetermined. The principle of Maximum Entropy [2–4] uses the available information as constraints in an optimization problem to select a solution using Shannon entropy [5] as the

decision function. The maximum entropy solution is the least-biased approach. It is not biased by structural modeling assumptions. It is the flattest, and therefore least informative, probability distribution compatible with the information captured in the constraints [6–8].

The classical Maximum Entropy (ME) formalism may not work in the presence of model ambiguity and insufficient, noisy, and complex information. However, an information-theoretic approach, which generalizes the ME, accommodates these challenges (see the Supplement). In the absence of these complications, the solution of this information-theoretic approach converges to that of the ME.

This approach incorporates each piece of information as a flexible constraint with additive mean-zero uncertainty. It maximizes the Shannon entropy decision function defined over the probabilities of interest (here, the survival rate), accounting for the uncertainties in the constraints. It can be applied even with few observations and little or no knowledge of the underlying biological model.

The binary choice information-theoretic approach we used dominates the classical maximum likelihood for finite samples and allows us to use informative priors, significantly improving the inference [9]. The priors may reflect fundamental principles, logical reasoning, or empirical observations. Empirical priors must be independent of the data used for the analysis but capture the universal characteristics and features of the population of interest [10]. In our application, the priors are observed death frequencies by age and sex for individuals previously infected with SARS because different coronaviruses with similar characteristics cause SARS and COVID-19 [11]. Our application demonstrates that this choice of priors improves the model's in- and out-of-sample predictions.

At the beginning of the COVID-19 pandemic, we applied this approach to identify existing treatments and policies that could reduce death rates. The data came from the Open COVID-19 Data Curation Group [12], which had only a small amount of publicly available patient-level data as of April 24, 2020. Although the disease had infected millions, the dataset contained only 485 individuals from twenty countries with the age, sex, and survival information necessary for our analysis. We supplemented this dataset with country-specific information on BCG–tuberculosis and polio vaccination policies, public health policies, pollution levels, education, and economic characteristics. (Because the polio vaccination, education, and economic variables were not statistically significant, we do not report them in the following results.) The polio and BCG vaccinations are used because these are well-studied and known to positively affect the immune system (especially the BCG) beyond their original purpose (e.g., Rivas et al.).

We use two binary BCG policy variables. The first equals one if a country never had a universal vaccination program (e.g., the United States) and zero otherwise. The second equals one if a country's former BCG policy ended before the pandemic (e.g., Australia). The base case is a current BCG policy (e.g., the Philippines).

Our environmental variable is the air pollution death rate. The World Bank estimates the annual deaths attributable to household and ambient air pollution.

We used three health variables for each country: the World Health Organization's estimate of the domestic private health expenditure per capita (in international dollars at the purchasing power parity) and the measles and hepatitis B immunization rates [13]. We did not expect those vaccinations to affect COVID-19 outcomes directly but viewed them as proxies for

health policies in general. Table S1 in the supplement summarizes the dataset used and its resources.

We estimated (in early 2020) an information-theoretic, binomial model [1, 14] to infer the survival probability of an infected individual, conditional on age, sex, and country-specific factors with and without priors.

Results

The models with and without priors fit the data well and give similar qualitative results. In the model with priors, all estimated coefficients except for health expenditures and females are statistically significant at the 0.05 level. The asymptotic t-statistics are 4.60 for age, 4.23 for BCG never, 2.91 for BCG past, 6.67 for air pollution, 1.42 for females, 1.53 for health expenditure, 1.85 for measles, and 3.57 for hepatitis B. The pseudo-R^2 is 0.54. Table 1 show our estimated coefficients and confidence intervals (CI). The last column shows the marginal effects: the change in death probability as a change of some explanatory variable. For example, if a country never had a BCG program, the death rate would be 57.4% (with a 95% confidence interval of [25.5%, 89.2%]) higher than if it currently has such a program, holding other variables constant.

The model with priors better predicts outcomes. Its in- and out-of-sample predictions were about 90% correct. Table 2 shows the out-of-sample predictions for the models with and without priors. We estimated the model using 200 randomly selected individuals (41% of the sample) and predicted the outcomes for the 285 others. The model without priors correctly predicted 222 (= 50 + 172) or 78% of the out-of-sample people. The model with priors accurately predicted 264 or 93%. Henceforth, we discuss the results for the model with priors.

We focus on the pollution and BCG vaccination effects on COVID-19 patients' survival probabilities. Pollution substantially affected COVID-19 patients' survival probabilities. A one percentage point increase in the air pollution death rate raised the probability of death from the virus by about 2.6 percentage points evaluated at the other explanatory variables' means other than age.

Figure 1 shows the death probabilities controlling for all other factors and vaccines. The death probability curves rise sharply with age. The curves evaluated at the ninetieth pollution percentile lie substantially above those at the median and the tenth. The women's curves lie below the men's.

Panel A of Figure 2 shows that current or past BCG vaccine policies substantially increased survival probabilities. At the margin for the middle of the age distribution, men and women from countries that never had a universal BCG vaccination policy were about 50 percentage points more likely to die from COVID-19 than individuals from countries with a current universal policy, and about 30 percentage points more likely than those with a previous vaccination policy. The death probability rises substantially with age, and the curves for women lie below those for men.

We used a much larger publicly available dataset from a longer period to determine whether these results based on limited and imperfect information available at the beginning of the pandemic are plausible and qualitatively accurate. Panel B of Figure 2 shows the cumulative frequencies from the pandemic's beginning through mid-December 2020, the end of the pre-vaccination period for 12,654,066 individuals in eleven countries. To illustrate the effects of BCG policies while controlling for pollution, this figure includes only countries with low (10%) pollution levels. The death rates are much lower in Panel B than in Panel A because these data

are primarily from later in the pandemic. However, the qualitative results are the same as our small sample estimates from early in the pandemic: The BCG policies reduced death rates.

For example, we estimated that a 55-year-old man's death probability was 7.6 times greater in a country that never had a BCG policy relative to one with a current policy in the initial analysis. This ratio for the cumulative frequencies is 7.4. The corresponding ratio for a country with a past BCG policy to a current one was 5.5 using predicted probabilities and 4.5 using frequencies.

Discussion

Our early, small-data-set study identified two factors that reduce death rates. Our study using a larger dataset was consistent with these results.

Moreover, well-designed experimental evidence also supports these results. For example, a randomized, double-blinded, placebo-controlled trial to test the efficacy of the BCG vaccine against COVID-19 found that BCG is safe and is approximately 92% efficacious relative to a placebo group [15]. See also [16-18]. Two studies [19-20] confirm the negative impact of pollution on COVID patients.

Conclusions

An information-theoretic approach can identify factors affecting patients' survival probabilities in the face of great uncertainty stemming from limited information about a complex system and few collinear observations at the beginning of a pandemic. Thus, it can allow policymakers to respond before more reliable experimental studies and data are available early in pandemics and before a new vaccine is available.

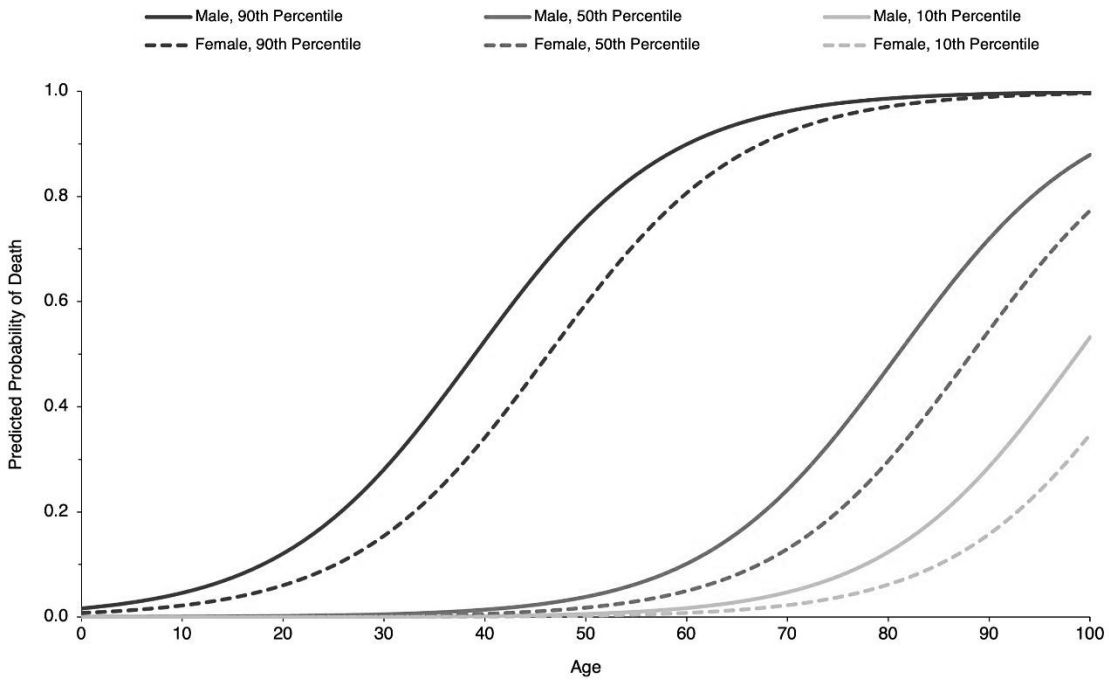
The same information-theoretic approach can be used in other scenarios where the data are limited and imperfect, and we are uncertain about the underlying physiological process, such as with emerging diseases. It could also be used for imperfect experiments (with attritions or imperfect protocols or that include a small number of individuals) and for initial study of rare diseases.

References

1. Golan A, Mumladze T, Wilson D et al., *Effect of universal TB vaccination and other policy-relevant factors on the probability of patient death from COVID-19*. Human Capital and Economic Opportunity Global Working Group at the University of Chicago; 2020. https://hceconomics.uchicago.edu/research/working-paper/effect-universal-tb-vaccination-and-other-policy-relevant-factors
2. Jaynes ET. Information theory and statistical mechanics. *Phys Rev*. 1957;106(4):620–630. doi: https://doi.org/10.1103/PhysRev.106.620
3. Levine RD, Tribus M, eds. *The Maximum Entropy Formalism*. MIT Press; 1979.
4. Skilling J. Data analysis: The maximum entropy method. *Nature*. 1984;309:748–749. doi: https://doi.org/10.1038/309748a0
5. Shannon CE. A mathematical theory of communication. *BSTJ*. 1948;27:379–423.
6. Skilling J. The axioms of maximum entropy. In Erickson GJ, Smith CR, eds. *Maximum-Entropy and Bayesian Methods in Science and Engineering*. Springer Netherlands;1988:173–187.
7. Shore JE, Johnson RW, Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. *IEEE Trans Inf Theory*. 1980;26(1):26–37. doi: 10.1109/TIT.1980.1056144
8. Golan A, Harte J. Information theory: A foundation for complexity science. *Proceedings of the National Academy of Sciences*. 2022;119(33). doi: https://doi.org/10.1073/pnas.2119089119
9. Golan A, Judge G, Perloff JM. A maximum entropy approach to recovering information from multinomial response data. *JASA*. 1996; 91(434):841-853. doi: https://doi.org/10.2307/2291679
10. Golan, A. Prior information. In *Foundations of Info-Metrics: Modeling, Inference, and Imperfect Information*. Oxford University Press; 2018:194-230.
11. Karlberg J. Do men have a higher case fatality rate of severe acute respiratory syndrome than women do? *Am J Epidemiol*. 2004;159(3):229–231. doi: https://doi.org/10.1093/aje/kwh056
12. Open access epidemiological data from the COVID-19 outbreak. Accessed April 24, 2020. https://github.com/beoutbreakprepared/nCoV2019
13. World Bank open data. Accessed May 18, 2020. https://data.worldbank.org
14. A. Golan. *Foundations of Info-Metrics: Modeling, Inference, and Imperfect Information*. Oxford University Press; 2018.
15. Petrosillo N, Viceconte G, Ergonul O, Ippolito G, Petersen E. COVID-19, SARS and MERS: are they closely related? *CMI*. 2020;26(6):729–734. doi: 10.1016/j.cmi.2020.03.026
16. Berg MK, Yu Q, Salvador CE, Melani I, Kitayama S. Mandated Bacillus Calmette-Guérin (BCG) vaccination predicts flattened curves for the spread of COVID-19. *Sci. Adv.* 2020:6(32). doi: 10.1126/sciadv.abc1463
17. Rivas MN, Ebinger JE, Wu M et. al. BCG vaccination history associates with decreased SARS-CoV-2 seroprevalence across a diverse cohort of health care workers, *J Clin Invest.* 2021;131(2). doi: https://doi.org/10.1172/JCI145157

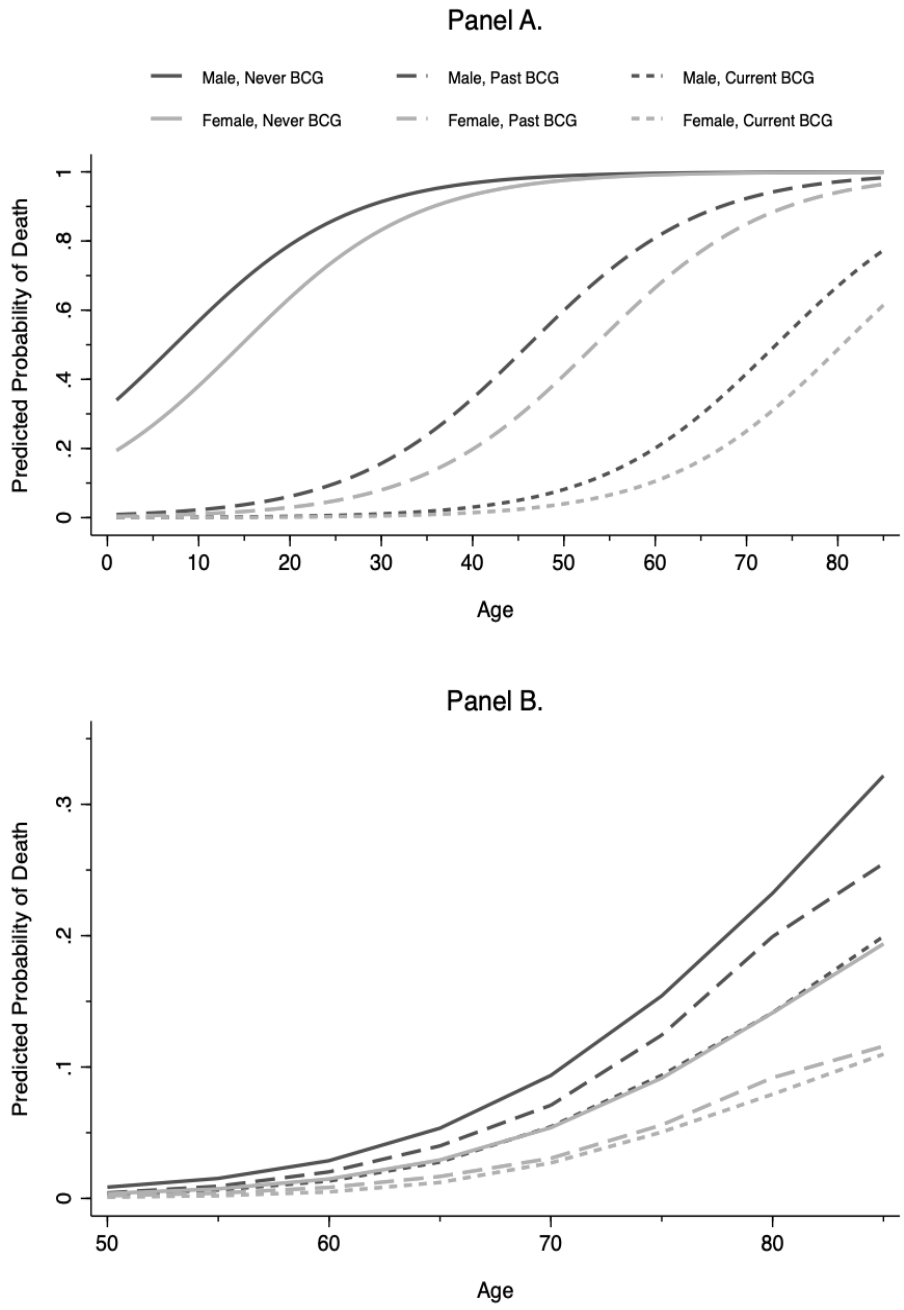
18. Faustman DL, Lee A, Hostetter ER et al. Multiple BCG vaccinations for the prevention of COVID-19 and other infectious diseases in type 1 diabetes. *Cell Rep Med*. 2022;3(9). doi: 10.1016/j.xcrm.2022.100728
19. Yu Z, Bellander T, Bergström A et al. Association of short-term air pollution exposure with SARS-CoV-2 infection among young adults in Sweden. *JAMA Netw Open*. 2022;5(4). doi: 10.1001/jamanetworkopen.2022.8109
20. Wu X, Nethery RC, Sabath MB, Braun D, Dominici F. Air pollution and COVID-19 mortality in the United States: Strengths and limitations of an ecological regression analysis. *Sci Adv*.2020; 6(45). doi: 10.1126/sciadv.abd4049

Figure 1. Pollution-Death Probability Relations by Sex and Age



Note: Figure shows the probability an infected COVID-19 patient died by age and sex (Females, dashed lines) conditional on pollution at low (10th percentile, light gray), median (dark gray line) and high (90th percentile, black line) levels.

Figure 2. BCG-Death Rate by Age and Sex



Note: Panel A shows the estimated probability an infected COVID-19 patient died by age and sex (male – dark gray, female – light gray) conditional on BCG vaccination policies using data through April 24, 2020. Panel B illustrates the cumulative frequency of death of infected individuals over 50 years old using data from the beginning of the pandemic through mid-December 2020, the pre-vaccine period, in countries with a low pollution level.

Table 1. Estimated Coefficients and Marginal Effects of the Model with Priors.

	Coefficients	Marginal Effects
Female	–0.760	-0.063
	(–1.44, –0.079)	(-0.123, -0.004)
Age	0.104	0.009
	(0.082, 0.126)	(0.006, 0.012)
BCG never	6.868	0.574
	(3.582, 10.154)	(0.255, 0.892)
BCG past	2.825	0.236
	(0.638, 5.012)	(0.043, 0.430)
Health expenditure	0.000	0.000
	(–0.001, 0.000)	(0.000, 0.000)
Air pollution	0.036	0.003
	(0.025, 0.048)	(0.002, 0.004)
Measles	–0.111	-0.009
	(–0.216, –0.006)	(-0.018, -0.0001)
Hepatitis– B	0.193	0.016
	(0.089, 0.297)	(0.006, 0.026)
Constant	–17.098	-1.428
	(–21.994, –12.203)	(-0.647, -0.188)
Observations	485	
Entropy	132.386	
Normalized Entropy	0.394	
Entropy Ratio Statistic	407.580	
P–Vale for LR	0.000	
Pseudo R–squared	0.54	

Note: 95% confidence interval in parentheses.

Table 2. Out-of-Sample Prediction Table (Counts)

		Observed Death		Total
		Yes	No	
Predicted Death	Yes	50	24	74
		(86)	(18)	(104)
	No	39	172	211
		(3)	(178)	(181)
Total		89	196	285
		(89)	(196)	(285)

Note: Out-of-sample prediction: estimates of randomly chosen 200 observations are used to predict the other 285 observations. Comparison of the information-theoretic model with priors and that without priors. Results not in parenthesis are from the model without priors, while results in parenthesis are from the model with priors.