## Appendix: Data, Methodology, and Limitations

April 22, 2020

### Terms of Use

*All users of this information must acknowledge the purpose, limitations, and lack of warranty for this analysis and the related data, and accept the requirement to disclose these terms to any person to whom they provide this information:*

**1. Purpose.** Our purpose is to provide information useful to others who may face a decision regarding their activities in the coming months. The purpose is not to diagnose or treat any disease; provide health advice; provide legal advice; or even to make unconditional forecasts of the number of cases or patients.

**2. Known Limitations and Errors in Data.** This analysis relies upon data that have known errors, omissions, and limitations. These include:

- The diagnosis and classification of those afflicted with Covid-19 has been inconsistent across jurisdictions, and across time.
- Daily case counts are subject to reporting and compilation errors.
- The proper location for a case is often unclear, particularly with travelers and people who live near borders.
- In some situations, governments have retroactively re-classified cases.
- In some areas of the world, governments have censored, suppressed, or otherwise failed to report accurately the relevant data.
- Given the volume of data and the frequency with which it is revised, some inadvertent errors are certain to arise.
- The underlying data, and therefore the estimated parameters and the estimated and projected paths, will change over time.

**3. Known Limitations of Epidemiological Models.** Every model is an approximation of reality. Like every other model, the one used here will never completely represent the underlying behavior, and will often distort some portion of it.

**4. No Warranty on Information.** There is NO WARRANTY provided for any of this information, including no warranty of merchantability or fitness.

**5. Requirement to Disclose these Terms.** If you provide this information to any other person, you must disclose to them these terms and provide a copy of this statement.

*If you do not acknowledge and accept these terms, then you should not read, share, or rely upon this information.*

**Table of Contents**

*Data*

**Data.** The state-level data we used in this analysis were collected by the *New York Times* organization from state and local sources. We selected this data source after reviewing multiple alternatives, and recognizing the benefits of the consistency and data checking this organization brought to the effort.

- The NYT website for these data is: https://www.nytimes.com/interactive/2020/us/coronavirus-us-cases.html.
- The NYT data repository on GitHub is: https://github.com/nytimes/covid-19-data.

**Discrepancies in Data Sources; Revisions over Time.** There are discrepancies among sources for data, even for the same areas. These discrepancies arise for a number of reasons, including lapses in reporting; differences in classification; and ambiguities regarding the location of the persons. Furthermore, government entities routinely revise these data, and in many cases have retroactively reclassified patients. For these reasons, users of these data should expect that analyses prepared on different dates will often show different historical data, as well as different projected data.

*Comparison and Alternative Sources of Data and Analysis*

There are multiple other sources of data, as well as alternative sources of analysis. These have widely-varying quality, and some are sponsored by organizations that advocate for specific policies as well as report data.

We urge readers of our analyses to compare the information they have obtained from us with information from other sources. Here are an example of comparison sources of information, for the State of Michigan in the United States:

- https://covid19.healthdata.org/united-states-of-america/michigan
- https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/cases-in-us.html#epi-curve
- https://www.uofmhealth.org/covid-19-update
- https://www.michigan.gov/coronavirus

*Model*

We used a SIR model to forecast the impacts of COVID-19 on states across the U.S. We chose this model after conducting an extensive review of available epidemiological models.

Our review confirmed that the standard SIR model, when consistent daily data are available, often provides a reliable and understandable assessment. This judgement is based partially on the decades of experience across a wide number of countries and epidemics.

It is also based on our assessment of how this model and the available data had performed in assessing the COVID-19 outbreak in Italy, Germany, Lombardy (a province in Italy), the United States, New York City, Ohio, Michigan, Illinois, the Detroit metropolitan area and counties, and other areas.

The SIR model is a “compartmentalized” model, meaning that it presumes the population can be classified as if they are in different “compartments,” and can move among them when they become infected with a disease and when they recover or perish from it. The acronym comes from “Susceptible-Infected-Removed.”

We use the data from various states and a specific implementation of the SIR model to fit the data, and estimate parameters such as R0 (discussed further below). We use these parameters to estimate a future curve, which represents the prediction based on past data and the assumptions of the model. As noted below, our implementation of the SIR model provides for iterative re-estimating of R0 values and estimates of the epidemic size and susceptible population.

**History of the Use of the SIR Model.** The SIR model has been recognized and used as a basis for the diffusion of ideas and other transmissions, as well as for epidemiological studies, since at least 1964.[1] The model was originally published in 1927, in an article that helped establish the basis for modern studies of epidemiology.[2]

[1] .An influential article in this development was Goffman W, Newill V. “Generalization of epidemic theory.” *Nature*. 1964;204(4955):225–228. doi: 10.1038/204225a0. This article outlined the now standard nomenclature of “susceptible-infected-removed” subgroups of a population, and stated the standard assumptions such as a homogenous population and regular “mixing” of it. (The “R” portion is now commonly called “recovered.”)
The intellectual roots of this model date back to the early 1900s, in which the first “compartmented” mathematical models, separating population cohorts, were developed. See, e.g. Sooknanan, J., Comissiong, D.M.G. “When behaviour turns contagious: the use of deterministic epidemiological models in modeling social contagion phenomena.” *Int. J. Dynam. Control* 5, 1046–1050 (2017). https://doi.org/10.1007/s40435-016-0271-9.

It is also widely described in academic and professional settings, including having many easily-accessible resources, such as an online teaching version provided by the Mathematics Association of America.[3]

**Alternatives and Limitations of the SIR Model.** Every model is an approximation to reality, and all have limitations. In the case of the SIR model, known limitations include:

- The model only focuses on one wave of an epidemic.
- The model assumes that "removed" individuals attain immunity from the disease. For some novel diseases, this has not been demonstrated.
- The model relies on certain assumptions (such as homogenous population and consistent mixing of the population) that are rarely accurate.

Furthermore, all models rely upon data—and the available data are never perfect in epidemiological studies.

There are alternatives to this model, many of which are variations on it. As with alternative sources of data, readers should consider whether an alternative model should be also be considered.

**Inclusion of Diagnostic Information.** Our implementation of the standard SIR model includes a number of features that allow for inspection of the results, and for a better understanding of the reliability of the model's estimated parameters. These include:

- Our complete results allow for the inspection of prediction errors, allowing readers to compare our historical estimates against historical data.
- Using innovations developed by researchers cited below, we also iteratively re-estimate the model over time, capturing the evolution of model parameters as more data become available.
- We fully disclose the model data and parameters, allowing the results to be replicated by other researchers.
- We explicitly note the known limitations of the model and data.

**Elements of the SIR Model.** The typical curve generated by an SIR model that estimates the number of people infected per day is an "S" shaped curve in which the number of new

[2] The 1927 article by W.O.Kermack and A.G. McKendrick, "A Contribution to the Mathematical Theory of Epidemics," Proceedings of the Royal Society A, vol. 115 no. 772, outlined the SIR model, is often credited as the first to identify the underlying mathematics of the path of an epidemic. In this 1927 article, they plot the deaths from an epidemic in "the island of Bombay" from December 1905 to July 1906. It followed a bell-shaped curve of the shape shown in SIR and other models a century later.

[3] The MAA teaching version, authored by David Smith and Lang Moore, is available at: https://www.maa.org/press/periodicals/loci/joma/the-sir-model-for-spread-of-disease-the-differential-equation-model. An overview of this and many variants is presented in H.W. Hethcote, "The Mathematics of Infectious Diseases," *SIAM Review* 42 (4) (2000).

infections per day begins slowly, then grow exponentially larger before reaching a “peak” point. Following the peak, number of cases per day decrease and slowly reach zero, or the neutral level.

**Parameters.** A key element of the model is known as “R0,” (sometimes called “R nought”) or the reproduction rate of the infection. In the SIR model, if R0 equals 2, that means for each time period of the outbreak, 1 infected person is expected to infect 2 others.

In actual epidemics, we expect the reproduction rate to be a dynamic value that changes over time. In its initial stages, the R0 value can grow quickly. As the epidemic slows, the R0 value declines and ultimately decreases to near zero.

- See “Exhibit: Key Variables in SIR Model ” on page 6.

*Exhibit: Key Variables in SIR Model*

### Reproduction Rate

- R = Reproduction number, number of people infected by each infected person
- R0 = Basic reproduction number ("R nought")
  calculated by Beta/Gamma (with scale factors)
- Beta = Average contact frequency
- Gamma = Average removal frequency

### Population and Epidemic Size

- N = Population size (approximate initial size of susceptible population)
- $C_{end}$ = Epidemic size (total recovered population)
- $S_{end}$ = Final number of susceptible individuals left

### Implied Timing

- Day = epidemy day number
- date0 = start day

### Errors, Uncertainty, and Model Fit

- RMSE = Root mean square error
- AdjR2 -- adjusted R2 ("R squared") statistic

### Data

- C = reported case data
- Ce = projected case data, given the model, data, and parameters
- Diff = difference between reported and projected cases

### *Known Limitations and Common Assumptions in Epidemiological Models*

There are known limitations of this and other epidemiological models. We state the most important ones here:

1. This and other epidemiological models presume that, after some time, the population acquires "herd immunity" that limits further spread of the same strain of the disease.

2. There is a likelihood of subsequent waves of the same disease. These are typically of lesser magnitude. For example, numerous variations of influenza cause illness and deaths each year, and are related to similar prior outbreaks. In general, only one wave is represented by this type of model.

3. "Comorbidity" causes difficulty in assigning properly both cases and the number recovering and perishing from a disease. In particular, the number of deaths due to influenza each year is very large, and this cannot always be distinguished (especially in the early stages of the epidemic) from the specific disease being studied.

4. In general, these models assume a certain amount of "mixing" of the population. The mathematical treatment of this interaction is never accurate in small settings or specific short time periods, and usable models require significant data and enough time to allow for patterns across entire societies to develop. The amount of mixing also varies by age, and varies across social groups and among areas within a country.

5. In addition, all models rely upon data that are collected with reporting and measurement errors. In this epidemic, differences in testing, and in some countries pressure to mis-report or fail to report results, have clearly caused some difficulties in obtaining accurate data.

6. All statistical models involve uncertainty, and all projections resulting from these models carry with them this uncertainty.

7. This is a new disease. It will react in unpredictable ways with different populations.

8. The population itself changes its behavior over time. This is perhaps the largest unknown at the early stage of a severe epidemic in which a strong change in behavior, and new forms of treatment, occur.

Because of these inherent limitations, we present the model and data along with the available diagnostic and related information, to allow an informed decision. We urge policymakers to consider this information, and additional information, before making their decisions.

### *Implementation of Model*

Supported Intelligence adapted an SIR model authored by Milan Batista, Professor at the University of Ljubljana in Slovenia. Mr. Batista and his colleagues describe this model, provide a public file exchange for the code, and apply it to the Coronavirus outbreak in the following publication and websites:

- Batista, Milan. (2020). *Estimation of the final size of the coronavirus epidemic by the SIR model*. Monograph found at ResearchGate.
- A file exchange version is maintained for review by other researchers; it is here: https://www.mathworks.com/matlabcentral/fileexchange/74658-fitviruscovid19.
- Related work using this same model for other countries and with different data sources is available here: https://www.fpp.uni-lj.si/en/research/researh-laboratories-and-the-programme-team/research-programme-team/

The version used by Supported Intelligence is modified to include the following improvements for the purposes of our work:

1. Supported Intelligence built a data assembly routine that gathers live data on cases reported per day from the *New York Times*.

2. We provide additional diagnostic information, and have created different data visualizations for the purpose of presenting the results.

3. We clarify a number of terms, and adjust the presentation to date formats common in the United States, to avoid ambiguity and misunderstanding.

**Acknowledgements.** Contributors to this methodology statement include:

- Patrick L. Anderson
- Brian Peterson
- Sarp Mertdogan
- Anderson Economic Group LLC staff.