**Data Science Quarterly: Manuscript for Opinions Section**
Volume 1, Issue 1, January 2021
ISSN 2693-8871 (print) ISSN 2693-8863 (online)

# Protect Public Health Data Now

**Thomas W. Miller**

Revised 24 August 2020

**Abstract** Early in July 2020, the Trump administration required hospital COVID-19 data to be submitted to HHS rather than the CDC. By the end of July 2020, data reporting delays and inaccuracies were already being observed. Changes in data reporting have been worrisome to many, including the author, a data scientist who builds predictive models. Data reporting changes made by the Trump administration could lead to data reporting delays, inaccurate data and models, and public policy mistakes.

**Keywords** CDC · COVID-19 · Data provenance · Epidemiology · Public health

## 1 Opinion

The way to get ahead of a pandemic is to understand where it is going, to make accurate predictions about the progress of the disease, and to use those predictions to drive public policy. And the way to make accurate predictions is to have complete and accurate data.

With more than one thousand people dying in a single day, the possibility of more than one million new cases in the next two weeks, and fears that infections may be many times higher than reported, we have a critical need for timely and accurate data about the current pandemic.

Think about what you would want for your personal health data. Suppose you are a woman who feels a lump in her breast or a man who hears about a blood test with unusually high prostate screening levels. You visit a doctor. The doctor sends samples to a lab, and you await test results.

What outcome do you want? Of course, you hope for negative test results. You want to hear that you are healthy and can live life to its fullest. But whatever the results, you want the test to be accurate. You want assurance that tissue and blood samples are properly cared for and analyzed. You want to know the truth about the data so you can make informed decisions about your life.

Hospitals are now required to submit COVID-19 patient data to the Department of Health and Human Services (HHS) rather than the Centers for Disease Control and Prevention (CDC). Over the years, health care workers and epidemiologists have relied on the CDC for ready and reliable access to public health data. It is unclear why the administration called for a data reporting change in the middle of a pandemic. Data reporting delays and inaccuracies are already being detected.

Drawing on scientists' experience with climate change data at the Environmental Protection Agency, we have reason for concern regarding pandemic data reporting. The current administration has not been fully transparent in its past data practices.

I am a data scientist with more than thirty years of experience working with data and building models. Recently I have been building models of epidemics to share with students. Model building exercises show students that accurate predictions about epidemics depend on the data used to drive the models.

A model building exercise might begin by assuming that each member of the population is in one of four disease states: susceptible to infection, infected with the disease, immune, or dead. Even the simplest of models, which are called finite Markov chains, can forecast the proportion of the population in each disease state at any future point in time. The models provide a sobering picture of how epidemics develop.

Data scientists and epidemiologists obtain initial data for models from the CDC, a trusted resource. We see no reason to change data collection and distribution processes that have worked well for many years.

When COVID-19 was just entering the United States, we had much uncertainty about the disease. We relied on data from other countries with little assurance of their completeness or accuracy. Does the current administration want to renew fears about missing or distorted data?

Data provenance, an important concept in the data science community, refers to the lineage or life cycle of data. Where do data come from? Has there been data filtering or selection? What about data aggregation or transformations? And most importantly, who has touched the data? In addition to making HHS the locus for COVID-19 patient data reporting, the administration is entrusting public health data to a private firm for analysis, a firm outside the purview of the CDC.

Data provenance is key, especially when the data in question relate to matters of life and death. The administration should seek CDC involvement or scrutiny at every stage of the data pipeline, from initial data collection to distribution and reporting.

Scientists need complete, up-to-date data for predictive models of disease. Attempts to hide public health data from scientists would be a mistake. Attempts to distort or destroy data would be a mistake. Delays in data reporting would be a mistake. Any of these could affect the predictive accuracy of models. Public policy mistakes could follow.

Just as we seek assurance that personal tissue and blood samples provided to doctors and labs are properly cared for and analyzed, we must demand that all COVID-19 data are protected and made available to health care workers and scientists shortly after being collected.

The pandemic is our shared disease. We need to know the truth about it so we can make informed decisions in our lives. Public health data are not the property of those in power to do with as they please. Public health data belong to the people. Protect our data now.

## 2 About the Author

Thomas W. Miller is faculty director of the data science program at Northwestern University, author of six books about data science, and owner of Research Publishers LLC.