ABOUT THESE GRAPHS AND MAPS - EPIDEMIOLOGY AND TRACKING OF THE COVID-19 EPIDEMIC (Updated 2020-06-13)

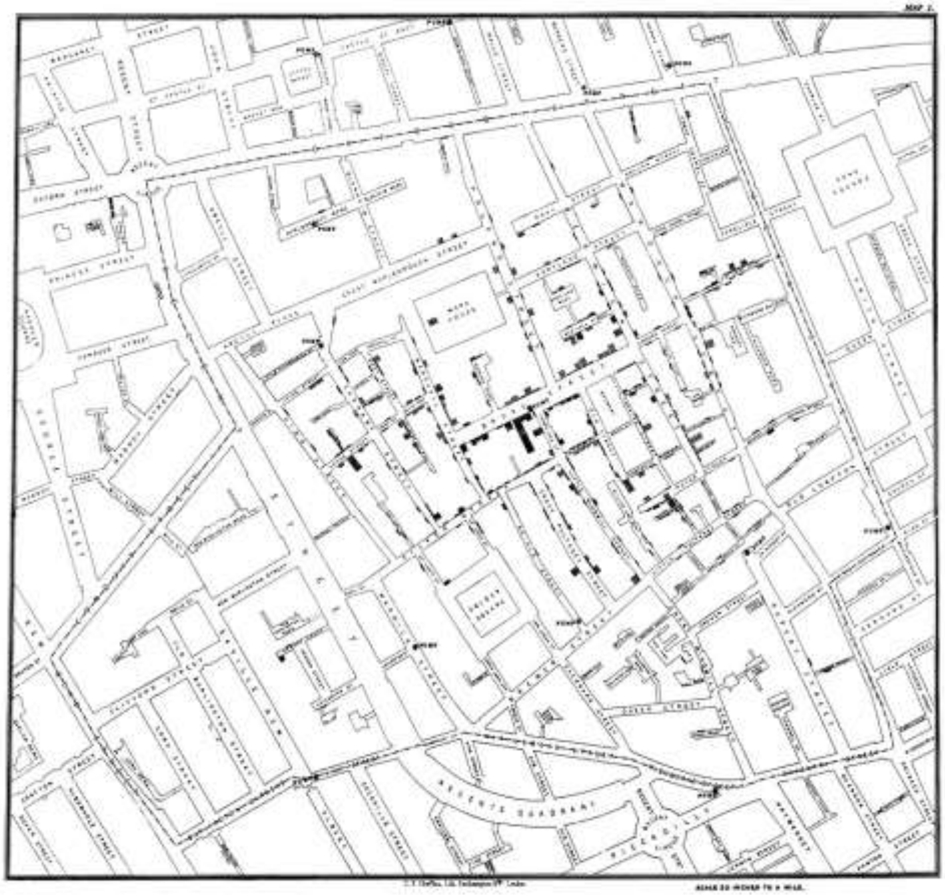
William L. Salomon, MD MS MPH
Clinical Metrics, LLC
covid-19@clinical-metrics.com

NOTE: The Multiple Graphs and Maps come from only 4 Workbooks (US-, State-, and County-level graphs, and one map). They are posted as examples of what you will find, and you can navigate freely in each Workbook to show what interests you.

Each workbook has about a 8-22 Panes and are arranged in the same order for Cases and Deaths - having learned one series, you have learned them all. They show how the various measures are related and illustrate that each Pane is insufficient to gain a thorough overview of the epidemic. The on-line version is fully functional - use the checkboxes on the right to add States and Counties for comparison, or un-check boxes if the provided starting point is too "busy". Consider jotting down your settings for your next session.

This is an Epidemiologist's tool meant to illustrate how each measure contributes to the whole.

The science of Epidemiology (the tracing of outbreaks of disease) is attributed to the London physician John Snow who established the source of repeated outbreaks of cholera (a bacterial diarrheal affliction) to a single contaminated well in London in 1854. In a poor area (SoHo), the well sourced water from the heavily polluted River Thames. The epidemics were stopped by removing the handle of the pump and obtaining clean water elsewhere.



John Snow's Spot-Plot of London (1854)

The Fundamental Epidemiologic Statistics

INCIDENCE is the **occurrence** of a disease or affliction within a certain area or population. It is measured over some period of time. For COVID-19, this is the occurrence of the disease clinically, or a “positive” test result each day. (For other disorders, it could be longer – per week, month, year, etc.). On a graph, this is a line that starts at 0 (usually) peaks in a “hump” and trends back down towards 0 (hopefully).

CUMULATIVE INCIDENCE (TOTAL INCIDENCE) is the **ongoing accumulation of Incidence over time**. On a graph, this is a line that continually trends upwards. As an epidemic is controlled, the curve flattens out, indication of no new cases (that is, or an Incidence of zero (0)).

Where multiple outbreaks occur, it is helpful to “adjust” these "raw" numbers for the number of people involved. A small town may be suffering as much as a big city. For that reason, the Incidence and Cumulative Incidence are divided by the population involved, giving a “rate” per person (or some arbitrary measure, such as per 10,000 people). It measures the chance that any person, in a city/town large or small will be afflicted.

To explain the concept further - suppose every county (large and small) has a local "Super Store" that is usually "packed". Suppose further that each store is the same size, and can accommodate only 100 shoppers at a time due to "social distancing" (with lines out the door regardless of location). The "per 10,000" measure is the probability that someone in the store has tested "positive" up to and including the day you are shopping, or has tested positive on that day. Said differently, the probability is the relative chance that you will directory encounter a “positive” person if you encounter every person in the store; go to a different store, you chance encounter will be more or less likely. The same applies to business, restaurants, etc. that have a random sample of that County.

ADJUSTED INCIDENCE RATE is the Incidence divided by the population multiplied by 10,000 (or some other number (100,000, 1 million, etc.). In these graphs and maps, 10,000 is used, making it easy to relate to the size of a town or small city. Mathematically, this is:

Adjusted Incidence Rate = (Incidence * 10,000)/Population

(This is also called the **ATTACK RATE.)**

ADJUSTED CUMULATIVE INCIDENCE RATE, similarly, is:

Adjusted Cumulative Incidence Rate = (Cumulative Incidence * 10,000)/Population

To show the relationship between the Incidence and Cumulative Incidence, some Panes both have both curves, the Incidence curve is usually above the Cumulative Incidence. Tool-tip numbers are **Bolded** in accordance with the graph you are “mousing” over. Both numbers will be in enlarged 14 pt font to emphasize the relationship between the two.

PREVALENCE is the existence of a disease or “positive” test at any given point in time, and varies from day to day, week to week, etc. So far, given that many individuals with COVID-19 are asymptomatic (estimates vary greatly – 50-80%), and the lack of universal or even targeted statistical sampling, the actual Prevalence of COVID-19 is a matter of huge speculation. It is not addressed in these graphs and maps.

Why These Numbers Matter

The matter of when to “re-open” (or perhaps to “clamp-down” again), is a “political” decision (“politics” coming from the Latin “*populus*” – of the people). It is a decision of at what level of Incidence is it “acceptable” to “re-open” and how to structure degrees of re-opening. To re-open means risking a higher rate of sickness, death, and expense for medical care, but with the benefit increased of economic activity, resumption of business, and “normal” daily affairs. To remain “closed” results in lower rates of illness and death, increased psychological stress, lack of income, and lower health care costs until a vaccine and/or treatment is developed.

In John Snow's time, disease travelled at the speed of walking, on horseback, or sailing ship. It could take years to spread a thousand miles. It should be remembered that smallpox first came to the New World in the early 1500's and killed an estimated 90% of indigenous peoples.

Four hundred years later, the population in much of the world is now highly "mobile" - disease can cross an ocean in 12-18 hours; and on land, people may travel from a High Incidence area to a Low Incidence area in an hour or two. Hence, looking at your own vicinity in conjunction with those two hours away will inform your decision-making,

Rational, justifiable Public Policy is made upon these kinds of numbers which are “real” and explainable to all. And, as was stated, these are decisions made on what is “acceptable” and to whom, not whether it is “safe”. “Safe” occurs only when the Incidence falls to zero (0).

Technical Notes

In respects, the methods used in Epidemiology have changed little since 1854. Case tracing and mapping are clearly easier - what was done with push-pins in map on a wall is now done with a computer. However, case tracing is still labor-intensive, done by on-site field work, telephone, and laboratory testing.

1. These figures are from the NY Times Repository of COVID-19 Cases and Deaths. See: https://github.com/nytimes/covid-19-data/ .

2. These data are "refreshed" daily at 11:00 AM EDT. It is dependent on when the NY Times posts it, and the author downloading it and processing. On some days this can be hours later as this is done manually - not by an automatic "scrape".

3. These graphs and maps come with a complete set of data for the US and Territories. They have been “preset” for a given State and the top 7 Counties of Incidence. It is suggested that you take at least 30 minutes to explore your own State, and others of interest. There are multiple panels, so explore “raw” and “adjusted” Cases and Deaths.

4. At this point, trying to calculate a “case-fatality rate” (the chance of death for a given case) is futile as the data were never collected in a way to calculate that (the death may have occurred in a person counted as a case the month before).

5. There is Federal repository of such information at the CDC; however, in the past several days questions have arisen as to just how the data was "massaged" and "reaggregated". Until this is clarified, the author’s use of the NY Times data will continue.

6. This particular work was generated by the lack of such information being available publicly at the state level. It was done to answer the author's own questions, and those presented to him by physicians and other health professionals, legislators, and neighbors and acquaintances. Daily isolated numbers without historical context are not a substitute. It is the obligation Public Health Departments and Centers for Disease Control (CDCs) to provide this information daily in “de-identified” digital download, graphical and map form so the Public can be made aware of current risks to its health. Only by releasing such data can Government formulate current and rational

Public Policy. To date, the author's own state continues to issue only "raw" numbers at the County level, and not adjusted to the County Population (Adjusted Incidence Rate or Attack Rate); it has just released ZIP-Code day on line at its Website, but only a "click-on-it" map without Rates /10,000 and without historical data (no downloadable data sets for anything).

7. Originally, this work was done on a spreadsheet (Excel). The NY Times data have now grown to 200,000 rows of data, and grows 3,000 rows/day for every State and County in the US and US Territories. Clearly, this outgrew "spreadsheet size" long ago.

8. All tools used are in the Public Domain or available at no charge. The database is MySQL, the graphics are done with Tableau Public. (Note, this is not an endorsement of Tableau *per se* but it is frequently used for such purposes. The Public free version is awkward, but it works. (If you have access to the full Desktop version, by all means, use it.) In other words, *anyone*, *any agency*, or *any business* can do this with the right tools (free), and some persistence with nothing more than a laptop computer and Internet connection.

9. No attempt has been made to "clean-up the data". The "spikes and valleys" in the Incidence data are what they are. There are techniques to "curve smooth". The author has used a Moving Average of 6 days - the day of interest, -3 days previously, and +2 days ahead; this determined by "inspection" with alterations of the "window" and knowledge of when "outbreaks" occurred – this provided the best smoothing of cyclic variations while allowing known "spikes" to show themselves. Clearly there is a relationship to day of the week when testing, workloads, etc. may be "stressed". More sophisticated modeling techniques such as regressions and AI "clustering techniques would no doubt yield better results. This may be an interesting project for those looking at the quality of data from various jurisdictions.

10. There are more "elegant" ways of presenting this data; however, the clarity of the graph and map approach is time-tested and allows the viewer to make their own interpretation of what they see. The map coloration is arbitrary – it can be made to look as "hot" or "cool" as one wants. The underlying Tableau file can be downloaded and "played" with. It is complete with data for that *day only* (unless you "refresh" with your own data).

11. If "porting" this to another data source (such as a relational database instead CSV files), it is *crucial* that the column names of the database be *identical* to those in Data Pane (the left-hand-most Pane) of each Worksheet. Without doing so, you will find yourself rebuilding every formula (a somewhat arduous task given the under of Panes for each Workbook – up to 22). The one exception is if a column name, such as **`Date`** and/or **`State`** is a "reserved" word in your database (a word that the system uses) – in the that case you much use an "escape" character to tell the system that this is word is a data element name, not a system word. (In MySQL, **`Date`** is a reserved word, and a backwards single quote (**`` ` ``**) is the escape character. Hence **`Date`** becomes **`` `Date` ``** which is "legal". When printed as the header in a SQL query **`` `Date` ``** looks like just **`Date`** (the backward single quotes do not appear).

12. The NY Times Data are Cumulative Incidence data, not Daily Incidence data. Programmatically, the Daily Incidence was organized by State, County and Date, then obtained by subtraction in a SQL "self-join" displaced by one row. (This might be a useful "hack" for the student of Epidemiology – see the SQL scripts.)

13. The demographic data is from: https://www2.census.gov/programs-surveys/popest/datasets/2010-2019/counties/totals/co-est2019-alldata.csv . It has been modified by:

    a. Adding a complete 4-5 FIPS (Federal Information Processing Standards) code for each jurisdiction
    b. Adding an entry for "New York City, NY" composed of the 5 boroughs (3651000)

c. Adding an entry for “Kansas City, MO “(2938000)
d. Adding US Territories
e. Adding FIPS codes for “Unknown” counties with a Population = 1 to prevent “divide-by-zero” errors. The codes were “synthesized” by concatenating the 1-2 State FIPS code with “999”. (e.g. for “Unknown, NY” this is 36999)
(Note: This is a “hack” for “Unknown” counties. It will “graph” as an extra County. However, it will not “map” as there is no associated Latitude and Longitude. In the case of “New York City”, the Cases and Deaths were apportioned to each borough based on population. Hence, “New York City” will “graph” along with 5 boroughs; only the 5 boroughs will “map” as they have valid FIPS 4-5 digits FIPS codes

14. With a “batch file” download script and MySQL database scripts, the processing time to get the data “Tableau-ready” is about 5-10 minutes. The final version of the scripts will have 1. directly load to MYSQL, from which Tableau Desktop can "auto-refresh", and a second version for Tableau Public which requires CSV files to be produced at the end of the MySQL script

15. The author considers this work to be in the Public Domain, available at no charge, on a “time as available” basis (requests to data@clinical-metrics.com). The Source Code is available at https://github.com/Clinical-Metrics-LLC/COVID-19-Viz . This includes:

    a. Windows batch file for downloading and extracting the NT Times Data
    b. Modified Census 2019 .CSV for direct loading
    c. SQL scripts for MySQL for creating the data set at US, State, and County levels (including calculating the Daily Cases/Deaths from Cumulative Cases/Deaths, "parceling out" NY City to the 5 boroughs for mapping) and the modified Census files for 2019)
    d. Tableau .twbx and .twb files for use with a MySQL server via an ODBC Connector
    e. Tableau .twbx files with data for use with .CSV text files.

16. Comments and errata should be addressed to the author at data@clinical-metrics.com.

Acknowledgements:

Many organizations and individuals made this work possible...

NIH / National Library of Medicine
Medical Informatics Trainee Grant - LM-07037

Decision Systems Group (Brigham and Women's Hospital, Department of Radiology)
Robert Greenes, MD, PhD - Director (now at Arizona State University)
Aziz Boxwala, MD PhD (now at UCSD)
Omolola Ogunyemi PhD (now at Charles R. Drew University / UCLA)
Qing Zeng PhD (now at Washington University)

Harvard School of Public Health - in particular:
James Ware, PhD (course - Biostatistics - Rates and Proportions)
Jennifer Leaning, MD SMH (course - Disaster Management)
Marc Roberts, PhD (course - Medical Ethics)

Massachusetts Institute of Technology (Laboratory of Computer Science and Artificial Intelligence)
Peter Szolovits, PhD

Maine Health
Dora Mills, MD MPH (formerly Director of the Maine CDC)

Department of Defense SBIR grant to Scientific Systems, Inc. (Contract No. DAMD17-03-C-0054)
Problem-Oriented Information Networking Tool (POINT) – A Cognitive Integrated Medical Data Display System
Lead Research Engineer at SSI - Lingi Chen, PhD - (now Senior Research Scientist at nuTonomy/Aptiv)

Members of the Tableau User Community, in particular:
Don Wise