Harnessing the Power of Visual Analytics & Data Science in Health Trends

Practical Guidance and Insights from TIBCO Spotfire® and PerkinElmer

Live Presentation with:

Michael O'Connell, Chief Analytics Officer, TIBCO Software® Inc.
Dan Weaver, Senior Product Manager and Solutions Architect, Research Informatics

April 21, 11 AM ET

Register Now
Agenda

COVID-19 Basics
- Reproduction Numbers
- Case Fatality Rates
- Interventions : NPIs

Genomics
- COVID19 Spike protein genomics

Demonstrations
- Data Science
- Genomics

Rejoinder
- Current controversies / analyses
- Genomics directions
- Back to work directions
The COVID-19 Webinar Panel

TIBCO & PerkinElmer

Michael O'Connell - chief analytics officer at TIBCO.

Neil Kanungo - Spotfire visual analytics ninja.

Dan Weaver – PerkinElmer Bioinformatics.

Gerard Conway – PerkinElmer Spotfire guru.
TIBCO Data Science and COVID-19

Provide data-driven review of issues & analytics work in the space
• Epidemiology, Data Science, Data Journalism

Contribute data and analytics to the community
• Data, Spotfire apps, R and Python code and global intervention data timelines

Bring a visual analytics and data science perspective
• Recognize the many sources of variability; potential for wide array of outcomes

Show TIBCO Predict & Unify products at work
• Visual Data Science
• GeoSpatial Data Science
• Unified Data and Data Wrangling
• Real-time data updates
**Latent Period** = time between occurrence of infection and onset of infectiousness (when the infected individual becomes infectious).

**Serial Interval** = duration of time between the onset of symptoms in primary case and onset of symptoms in secondary case infected by primary case.

**Incubation Period** = time period between the occurrence of infection (or transmission) and the onset of disease symptoms.

**Common epidemiology model: SEIR**

- **Susceptible** → **Exposed** → **Infectious** → **Removed**

Removed = recovered or died
**R0**: the reproduction number; and Re, Rt

**R0 = Basic Reproduction Number**
The number of people infected from an infected person, in usual behavior (no interventions e.g. distancing)

**Re = Effective Reproduction Number**
Shows the effects of interventions

**Re < 1 and virus stops spreading**
*Re is a leading indicator of case curve bending and epidemic curve flattening*

**COVID-19 Reproduction Numbers**
- **R0** is in the range 2-3

**Non-Pharmaceutical Interventions (NPIs) can drop the Re dramatically**
- China: 2.4 to 1.1 with severe travel / lockdown restrictions
Time Lags & Reproduction Numbers

**Time lags**
Cases & fatalities that we see now, were initiated a few weeks ago
The world we are in today, doesn’t show its COVID status for some weeks in the future

slow the transmission (social distancing) ... and the case counts drop
CFR: the case fatality rate

**CFR - risk that someone who develops symptoms will eventually die from infection**

Best way to calculate CFR would be to track a large group of people from the point when they develop symptoms until they later die or recover

- Then calculate the proportion of all these cases who had died.
- This is not possible in the real world.

**It is incorrect to just divide the total number of deaths by total number of cases as this does not account for unreported cases or the delay from illness to death.**

- *Many unreported cases* eg due to unavailable test kits. In the US, Bedford estimates ~10-20X under-reporting of cases as of March 13.
- *Time delay*, consider 20 new people admitted to a hospital with confirmed COVID-19 infection on a given day -- that doesn’t mean the CFR is zero! We need to wait to see what happens to them.
- Conversely any deaths that occur now, are people who showed symptoms some weeks before.

**Early estimates of CFR in epidemics is typically high as focus is on the sickest of the sick.**

- [Early CDC estimates](https://www.cdc.gov) suggested a wide range of 0.25%-3.0% back in January.
CFR: the case fatality rate

Data

- **Wu et al.** : CFR of COVID-19 in Wuhan of **1.4% (0.9–2.1%)**
  - Lower than naïve confirmed case fatality risk of $2,169/48,557 = 4.5%$
  - Lower than the approximator of deaths/(deaths + recoveries): $2,169/(2,169 + 17,572) = 11%$
  - Risk of symptomatic infection increased with age
  - Those above 59 years were 5.1X (4.2–6.1) more likely to die after developing symptoms, compared to those aged 30–59.
- Other recent estimates support 0.5 – 1.5% CFR (Wuhan, International)

Experts

- **Kucharski**: COVID-19 is ~1% ie ~10X+ more deadly than Flu
- **Fauci et al.** "if one assumes that the number of asymptomatic or minimally symptomatic cases is several times as high as the number of reported cases, the case fatality rate may be considerably less than 1%. This suggests that the overall clinical consequences of Covid-19 may ultimately be more akin to those of a severe seasonal influenza (which has a case fatality rate of approximately 0.1%) or a pandemic influenza (similar to those in 1957 and 1968) rather than a disease similar to SARS or MERS, which have had case fatality rates of 9 to 10% and 36%, respectively."
- Some others posit that CFR is lower than this eg **Gupta** (U Oxford), **Bendavid and Bhattacharya** (Stanford); they suggest are larger proportion of the population is already infected.
Genomics

The world research community has move at remarkable speed:

- >10,000 SARS-CoV-2 genomes sequence
- Crystal structure solved
- Spike protein mutations that correlate with severity being identified

1. https://www.epicov.org/epi3/frontend#
Anatomy of a virus – focus on the “Spike”

- The Spike protein allows the virus to bind to the ACE2 receptor on the cell, stimulating endocytosis.
- The Spike is cleaved as part of releasing the virus RNA into the cell.

- The spike glycoprotein is a key determinant of host specificity and pathogenicity.
- The spike is what gives the coronavirus its crown (it’s corona) in electron microscope.
- One emergent finding suggestions that the D614G mutation after the virus arrived in Europe.

---

2. Sequences in this analysis were obtained from NCBI Virus. We gratefully acknowledge the availability of these sequences and the many researchers that have isolated and submitted them to the NCBI.

© Copyright 2000-2020 TIBCO Software Inc.
The structure of the Spike

- The top of the Spike contains the Receptor Binding Domain (RBD) in 2 parts NTD and CTD.
- The bottom of the Spike mediate membrane fusion.
- The top has to be cleaved off for the bottom to work.

Structural overlay of mutations found:
- The red region is where most of the known mutations suspected to impact severity have been found.
- The orange region is where there seems to be mutations of unknown impact.
- The blue amino acid is the D614G mutation

Demonstrations

**Data Science & Epidemiology**
- Reproduction Number Predictions
- Cases and Fatalities
- Interventions
- Healthcare locators

**Genomics**
- COVID19 Genome
As of April 20, there are at least 2,471,408 cases confirmed worldwide.

Using local county level data, lasso select different regions of the United States to find the biggest outbreaks in a region. The calculations are relative to your selection so highlighting different regions will emphasize outbreaks in that selection.
The basic reproduction number (R0) is the number of people one person will infect if they have a disease, assuming no interventions. The effective reproduction number (Re) is how many people the disease actually infects, including interventions such as social distancing. The time between infections is the serial interval (D0) which can be adjusted here. Read the Page Info for more details.

**Re Estimate by Country**

- **Slovakia**, **Ukraine**, **San Marino**, **Republic of Serbia**, **Finland**, **Estonia**, **Kosovo**, **Romania**, **Macedonia**, **Denmark**, **Ireland**, **Italy**, **Bosnia and Herzegovina**, **Czech Republic**, **Italy**, **Switzerland**, **Belarus**, **Iceland**, **Austria**, **France**
COVID-19 Case Trajectory by Country

Compare Growth Rates

Confirmed Cases are growing exponentially around the world. Here, each country is shown time normalized after its 100th case where reporting is assumed to be more consistent.

Choose countries to compare from the checklist.

Cumulative Cases over Time

It's important to remember many countries report differently, and there is non-uniform testing across populations. Therefore, one should not project infection spread from this chart.
Demonstrations

**Data Science & Epidemiology**
- Reproduction Number Predictions
- Cases and Fatalities
- Interventions
- Healthcare locators

**Genomics**
- COVID19 Genome
CFR: the case fatality rate

Issue: How to assess the total number of COVID-19 infections?

Gupta (Oxford)

Fundamental principles of epidemic spread highlight the immediate need for large-scale serological surveys to assess the stage of the SARS-CoV-2 epidemic.

Jose Lourenco, Robert Paton, Mahan Ghafuri, Moritz Kremer, Craig Thompson, Peter Simmonds, Paul Kleiman, Sanjiv Gupta

doi: https://doi.org/10.1101/2020.03.24.2004291

This article is a preprint and has not been peer-reviewed [what does this mean?]. It reports new medical research that has yet to be evaluated and so should not be used to guide clinical practice.

Gupta described a number of scenarios. The media picked up on the most extreme, which implied the majority of the UK had already been infected. The Gupta analysis analysis focused solely on deaths.

Kucharski: Based on multiple studies it seems about 20–80% of people infected with Covid-19 could show symptoms. If this range turns out to be correct – and we combine it with our estimate that one in 15 people with symptoms are being reported – it would mean that hundreds of thousands of people in the UK have probably been infected with Covid-19 already, but not tens of millions.

Bendavid and Bhattacharya (Stanford)

Is the Coronavirus as Deadly as They Say?

Current estimates about the Covid-19 fatality rate may be too high by orders of magnitude.

Bendavid and Bhattacharya suggest CFR could be more like 0.01 to 0.1% ie more in line with seasonal Flu.

Argument based on under-reporting, lack of diagnostics

Existing evidence suggests that the virus is highly transmissible and that the number of infections doubles roughly every three days. An epidemic seed on Jan 1 implies that by March 9 about six million people in the U.S. would have been infected.

We don’t know the true infection rate in the U.S. Antibody testing of representative samples to measure disease prevalence (including the recovered) is crucial.
COVID-19 Antibody Seroprevalence in Santa Clara County, California

Eran Bendavid, Bianca Mulaney, Neeraj Sood, Soleil Shah, Emilia Ling, Rebecca Bromley-Dulflano, Cara Lai, Zoe Weissberg, Rodrigo Saavedra, James Tedrow, Dona Tversky, Andrew Bogan, Thomas Kupiec, Daniel Eichner, Ribhav Gupta, John Ioannidis, Jay Bhattacharya


This article is a preprint and has not been certified by peer review [what does this mean?]. It reports new medical research that has yet to be evaluated and so should not be used to guide clinical practice.

Results

The unadjusted prevalence of antibodies to SARS-CoV-2 in Santa Clara County was 1.5% (exact binomial 95 CI 1.11-1.97%), and the population-weighted prevalence was 2.81% (95 CI 2.24-3.37%). Under the three scenarios for test performance characteristics, the population prevalence of COVID-19 in Santa Clara ranged from 2.49% (95 CI 1.80-3.17%) to 4.16% (2.58-5.70%). These prevalence estimates represent a range between 48,000 and 81,000 people infected in Santa Clara County by early April, 50-85-fold more than the number of confirmed cases. Conclusions The population prevalence of SARS-CoV-2 antibodies in Santa Clara County implies that the infection is much more widespread than indicated by the number of confirmed cases. Population prevalence estimates can now be used to calibrate epidemic and mortality projections.

Assay Sensitivity and Specificity

A combination of both data sources provides us with a combined sensitivity of 80.3% (95 Cl 72.1-87.0%) and a specificity of 99.5% (95 CI 98.3-99.9%).

The total number of positive cases by either IgG or IgM in our unadjusted sample was 50, a crude prevalence rate of 1.50% (exact binomial 95% CI 1.11-1.97%). After weighting our sample to match Santa Clara County by zip, race, and sex, the prevalence was 2.81% (95% CI 2.24-3.37 without clustering the standard errors for members of the same household, and 1.45-4.16 with clustering). We further improved our estimation using the available data on test kit sensitivity and specificity, using the three scenarios noted above. The estimated prevalence was 2.49% (95CI 1.80%-3.17%) under the S1 scenario, 4.16% (95CI 2.58%-5.70%) under the S2 scenario, and 2.75% (95CI 2.01%-3.49%) under the S3 scenario.
Bendavid and Bhattacharya - Issues

- B&B publish a confidence interval on the specificity of the test of 98.3% - 99.9%.
- But only 1.5% of all the tests are positive!
- Thus, if true specificity is close to 98.3%, then ALL of the positive results may be accounted as false positives.
- B&B then report a 95% confidence interval for the prevalence of COVID-19 in Santa Clara County of 2.01% to 3.49% (using delta method, adjusting for patient sample demographics and test specificity).

Concerns with the B&B paper
- unstable population weights, adjustment, error propagation
- can specificity be this high?
- consent bias
- consistency with other serosurvey data?

Statistical Appendix

In this appendix, we describe our statistical methods in more detail. Section 1 describes our approach to calculating population weights. Section 2 describes our approach to adjusting our population prevalence estimate for the sensitivity and specificity properties of the LFA test kit we are using. Finally, Section 3 describes our approach to incorporating three separate sources of uncertainty in our prevalence estimates: sampling variability, error in the sensitivity estimate, and error in the specificity estimate.

We solve for \( \pi \) as a function of the sample prevalence, sensitivity, and specificity:

\[
\pi = \frac{q + s - 1}{r + s - 1}
\]

There is one important caveat to this formula: it only holds as long as (one minus) the specificity of the test is higher than the sample prevalence. If it is lower, all the observed positives in the sample could be due to false-positive test results, and we cannot exclude zero prevalence as a possibility. As long as the specificity is high relative to the sample prevalence, this expression allows us to recover population prevalence from sample prevalence, despite using a noisy test.

See: John Cherian (@jjcherian) for alternate bootstrap analysis; avoiding asymptotic assumptions of delta method
https://github.com/jjcherian/medrxiv_experiment
Sensitivity and Specificity

- **SENSITIVITY measures how well we can detect patients with disease.** Imagine a nasal swab from someone with SARS-CoV-2 infection. Sensitivity is the probability this sample will test positive.

- **With RT-PCR,** the specimen is declared positive if viral RNA is detected. Patients with COVID19 often have high viral loads in their throats. Thus it is relatively easy to detect virus, and the **test is typically highly sensitive.**

- **When sensitivity is low,** we get **FALSE NEGATIVES.** These are swabs from which no virus is detected even though the person is infected. This could occur if the specimen was degraded, virus didn't amplify well, etc.

See: Natalie Dean (@nataliexdean)
Sensitivity and Specificity

- **Sensitivity** measures how well we can detect patients with disease. Imagine a nasal swab from someone with SARS-CoV-2 infection. Sensitivity is the probability this sample will test positive.

- **With RT-PCR**, the specimen is declared positive if viral RNA is detected. Patients with COVID19 often have high viral loads in their throats. Thus it is relatively easy to detect virus, and the test is typically highly sensitive.

- **When sensitivity is low, we get FALSE NEGATIVES.** These are swabs from which no virus is detected even though the person is infected. This could occur if the specimen was degraded, virus didn't amplify well, etc.

- **Specificity** measures how well we rule out infection for people who are tested but aren't infected. Imagine a nasal swab from someone without SARS-CoV-2 infection. This should come back negative on RT-PCR. Specificity measures how often these come back negative.

- **Because RT-PCR measures viral RNA, it is rare for someone to test positive if they aren't infected.** It can happen, though e.g. if there is random contamination in the lab. In general, RT-PCR is usually highly specific and there are few FALSE POSITIVES.

See: Natalie Dean (@nataliexdean)
So what is % population infected?

What is the under-reporting of cases?

Table 1: Posterior model estimates of percentage of total population infected as of 28th March 2020.

<table>
<thead>
<tr>
<th>Country</th>
<th>% of total population infected (mean [95% credible interval])</th>
</tr>
</thead>
<tbody>
<tr>
<td>Austria</td>
<td>1.1% [0.36%-3.1%]</td>
</tr>
<tr>
<td>Belgium</td>
<td>3.7% [1.3%-9.7%]</td>
</tr>
<tr>
<td>Denmark</td>
<td>1.1% [0.40%-3.1%]</td>
</tr>
<tr>
<td>France</td>
<td>3.0% [1.1%-7.4%]</td>
</tr>
<tr>
<td>Germany</td>
<td>0.72% [0.28%-1.8%]</td>
</tr>
<tr>
<td>Italy</td>
<td>9.8% [3.2%-26%]</td>
</tr>
<tr>
<td>Norway</td>
<td>0.41% [0.09%-1.2%]</td>
</tr>
<tr>
<td>Spain</td>
<td>15% [3.7%-41%]</td>
</tr>
<tr>
<td>Sweden</td>
<td>3.1% [0.85%-8.4%]</td>
</tr>
<tr>
<td>Switzerland</td>
<td>3.2% [1.3%-7.6%]</td>
</tr>
<tr>
<td>United Kingdom</td>
<td>2.7% [1.2%-5.4%]</td>
</tr>
</tbody>
</table>

Source: report by @MRC_Outbreak, March 30

Trevor Bedford (@trvrb) on the US undercount: predicts 10-20X more people in the US are infected c.f. reported cases; based on ~20 sparks that caught between Jan 15 and Feb 15 and have grown into outbreaks.

So, we have ~760k cases as of Apr 20 in US. With 10-20% under-reporting, this means we may have 7.5-15M infected people in the US at this time, or around 2-4% of the US population.

This is consistent with these other WW results.

Source: @rivm (Dutch CDC) presentation for members of parliament.
The structure of the Spike

Spike in context of domain of full genome

We gratefully acknowledge these Data provided by William Buscher and Jack Bramley of Washington University
Preliminary Coupling Analysis near Spike RBD Domain

- Researchers are beginning to look for more complex genetic trends.
- “Coupling Analysis” shows statistical signals of sequences that co-vary and thus may be structurally important.

We gratefully acknowledge these Data provided by William Buscher and Jack Bramley of Washington University.
Wrap-Up

Common Misconceptions

Back to Work Planning
Don’t divide case curves by population without %pop-normalizing

Logarithmic scale on the vertical axis with start at same number of cases is best
- x-axis is days since first <n> cases
- slopes reflect growth rates
- heights reflect prevalence
- relative timing is maintained

If you divide by population
- Need to start at same % of population infected
- x-axis is days since first <%> population infected
Don’t divide case curves by population

without %pop-normalizing

Logarithmic scale on the vertical axis with start at same number of cases is best
• x-axis is days since first \(<n>\) cases
• slopes reflect growth rates
• heights reflect prevalence
• relative timing is maintained

If you divide by population
• Need to start at same % of population infected
• x-axis is days since first \(<\%>\) population infected

Exponential Growth is alarming
1 and 2 weeks growth
• \(2^7 = 128\)
• \(2^{14} = 16,384\)
Don’t divide case curves by population

without %pop-normalizing

**Logarithmic scale on the vertical axis with start at same number of cases is best**
- x-axis is days since first \(<n>\) cases
- slopes reflect growth rates
- heights reflect prevalence
- relative timing is maintained

If you divide by population
- Need to start at same % of population infected
- x-axis is days since first \(<\%>\) population infected

**Pay close attention to axis scales**
- Especially in the media
- the logarithmic scale !! (Yann LeCun, Twitter)
**Effects of Age on COVID Fatalities**

- Risk (death) increases exponentially with age in COVID and non-COVID people
- COVID packs ~1 years risk into 3 weeks
- COVID multiplies baseline risk similarly across age bands
- Males: 5.8% of deaths are COVID, Females: 3.8% are COVID

---

**Risk of Dying vs Age for non-COVID infected people. UK ONS**
- proportion increases ~9% per year
- average risk of death doubles every ~8 years

**UK REGISTERED deaths: COVID and non-COVID. March 21 - 27 (week 13)**

---

**Analysis: David Spiegelhalter**
Back to Work Planning

Positive Test Rates remain high – epidemic still underway

Learnings from 1918 Spanish Flu

Social distancing in Philadelphia avoided outbreak in St Louis

Denver double-hump when reopened early
## Back to Work Planning

### Digital Surveillance

<table>
<thead>
<tr>
<th>Category/Study Name</th>
<th>1st Data source(s)</th>
<th>Institution(s)</th>
<th>Country</th>
</tr>
</thead>
<tbody>
<tr>
<td>DETECT Study and COVID-19 Wearables</td>
<td>Smartwatch, smart ring, EHR</td>
<td>Scripps, Stanford</td>
<td>US</td>
</tr>
<tr>
<td>Kinsa</td>
<td>Smart thermometer</td>
<td>NA</td>
<td>US</td>
</tr>
<tr>
<td>COVID-19 Citizen Science</td>
<td>Smartphone +</td>
<td>UCSF</td>
<td>US</td>
</tr>
<tr>
<td>Corona-Datenspende-App</td>
<td>Smartphone, body temp</td>
<td>Robert Koch Institute</td>
<td>Germany</td>
</tr>
<tr>
<td>Symptom checkers</td>
<td>Multiple smartphone apps with daily inputs</td>
<td>NA</td>
<td>Israel, UK, Canada, Switzerland, Germany</td>
</tr>
<tr>
<td>Aggregated Mobility/Geolocation</td>
<td>Cellphone</td>
<td>NA</td>
<td>Germany, China and other Asia, EU countries</td>
</tr>
</tbody>
</table>

### Contact Tracing

<table>
<thead>
<tr>
<th>Category/Study Name</th>
<th>1st Data source(s)</th>
<th>Institution(s)</th>
<th>Country</th>
</tr>
</thead>
<tbody>
<tr>
<td>Google/Apple</td>
<td>Smartphone</td>
<td>NA</td>
<td>Global</td>
</tr>
<tr>
<td>PACT (Private Automatic Contact Tracing)</td>
<td>Smartphone</td>
<td>MIT, CSAIL</td>
<td>US</td>
</tr>
<tr>
<td>Trace Together</td>
<td>Smartphone</td>
<td>NA</td>
<td>Singapore</td>
</tr>
</tbody>
</table>

---

**Apple, Google Bring Covid-19 Contact-Tracing to 3 Billion People**

By Mark Gurman
April 10, 2020, 10:00 AM PDT Updated on April 10, 2020, 4:53 PM PDT
Test Trace Isolate

Alice infected at day 0

Alice infectious at day 3
Alice shows symptoms at day 5

Alice infects Bob at day 4

Alice is tested and gets results at day 7 and Bob is alerted to quarantine

Transmission chain is cut and Carol stays uninfected

Day 0 1 2 3 4 5 6 7 8 9 10
Resources

1) TIBCO COVID-19 Visual Analysis Hub
   tibco.com/covid19

2) Community Headquarters
   community.tibco.com/covid19

Community Tech Blogs
   community.tibco.com/blog

3) Live Spotfire App
   Live Spotfire App

4) Social
   LinkedIn
   linkedin.com/in/michaelo15/

Twitter: Daily update
   @MichOConnell

April 29: Hands-on building session
If you need a Spotfire license we will provide
Go to TIBCO Community HQ to get
- Spotfire
- Spotfire starter app
- Data
- R and Python scripts
Resources

1) TIBCO COVID-19 Visual Analysis Hub
   tibco.com/covid19

2) Community Headquarters
   community.tibco.com/covid19

Community Tech Blogs
   community.tibco.com/blog

3) Live Spotfire App
   Live Spotfire App

4) Social
   LinkedIn
   linkedin.com/in/michaelo15/

Twitter: Daily update
   @MichOConnell

April 29: Hands-on building session
If you need a Spotfire license we will provide
Go to TIBCO Community HQ to get
- Spotfire
- Spotfire starter app
- Data
- R and Python scripts

TIBCO Community COVID-19 Visual Data Science Headquarters

By: Mike Alperin
Last updated: 2:49pm Apr 17, 2020