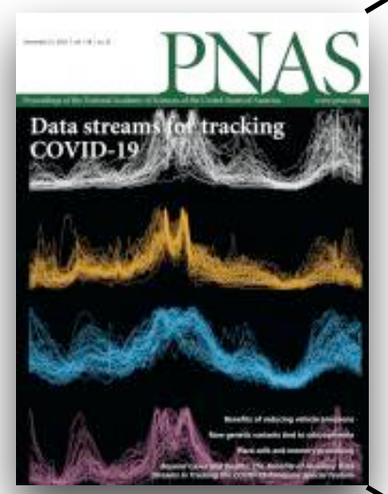
Data source agreement analysis

MSP project intro

Jacob Bien, USC (CMU Delphi) Feb 2, 2022

So much COVID data…



An open repository of real-time COVID-19 indicators

Alex Reinhart, Logan Brooks, Maria Jahja, Aaron Rumack, Jingjing Tang, Sumit Agrawal, Wael Al Saeed, Taylor Arnold, Amartya Basu, Jacob Bien, Ángel A. Cabrera, Andrew Chin, Eu Jing Chua, Brian Clark, Sarah Colquhoun, Nat DeFries, David C. Farrow, Jodi Forlizzi, Jed Grabman, Samuel Gratzl, Alden Green, George Haff, Robin Han, Kate Harwood, Addison J. Hu, Raphael Hyde, Sangwon Hyun, Ananya Joshi, Jimi Kim, Andrew Kuznetsov, Wichada La Motte-Kerr, Yeon Jin Lee, Kenneth Lee, Zachary C. Lipton, Michael X. Liu, Lester Mackey, Kathryn Mazaitis, Daniel J. McDonald, Phillip McGuinness, Balasubramanian Narasimhan, Michael P. O'Brien, Natalia L. Oliveira, Pratik Patil, Adam Perer, Collin A. Politsch, Samyak Rajanala, Dawn Rucker, Chris Scott, Nigam H. Shah, Vishnu Shankar, James Sharpnack, Dmitry Shemetov, Noah Simon, Benjamin Y. Smith, Vishakha Srivastava, Shuyi Tan, Robert Tibshirani, Elena Tuzhilina, Ana Karina Van Nortwick, Valérie Ventura, Larry Wasserman, Benjamin Weaver, Jeremy C. Weiss, Spencer Whitman, Kristin Williams, Roni Rosenfeld, and Ryan J. Tibshirani

PNAS December 21, 2021 118 (51) e2111452118; https://doi.org/10.1073/pnas.2111452118

Can auxiliary indicators improve COVID-19 forecasting and hotspot prediction?

Daniel J. McDonald, Jacob Bien, Alden Green, Addison J. Hu, Nat DeFries, Sangwon Hyun, Natalia L. Oliveira, James Sharpnack, Jingjing Tang, Robert Tibshirani, Valérie Ventura, Larry Wasserman, and Ryan J. Tibshirani

PNAS December 21, 2021 118 (51) e2111453118; https://doi.org/10.1073/pnas.2111453118

The US COVID-19 Trends and Impact Survey: Continuous real-time measurement of COVID-19 symptoms, risks, protective behaviors, testing, and vaccination

Joshua A. Salomon, Alex Reinhart, Alyssa Bilinski, Eu Jing Chua, Wichada La Motte-Kerr, Minttu M. Rönn, Marissa B. Reitsma, Katherine A. Morris, Sarah LaRocca, Tamer H. Farag, Frauke Kreuter, Roni Rosenfeld, and Ryan J. Tibshirani

PNAS December 21, 2021 118 (51) e2111454118; https://doi.org/10.1073/pnas.2111454118

So much COVID data…

Table 1. Data sources available in the COVIDcast API (19), as of date of publication

Data source	Signals available	First date	Resolution
Change healthcare	Percentage of outpatient visits with COVID diagnostic codes or codes indicating COVID-like symptoms; based on deidentified claims data processed by Change Healthcare	1 February 2020	County*
Doctor visits	Percentage of outpatient visits primarily about COVID-like symptoms, based on deidentified claims data provided by health system partners	1 February 2020	County*
Hospital admissions	Percentage of new hospital admissions with COVID diagnostic codes, based on deidentified claims data provided by health system partners	1 February 2020	County**
Quidel	Test positivity rates for COVID-19 antigen tests produced by Quidel	26 May 2020	County**
SafeGraph	Mobility metrics, such as time away from home or visits to bars and restaurants, based on cell phone mobility data collected by SafeGraph (27, 28)	1 January 2019	County
COVID-19 Trends and Impact Survey	COVID symptoms, social distancing behaviors, mental health, economic impact, behavior (e.g., mask wearing, vaccination attitudes), and COVID testing signals based on daily surveys conducted nationally by Delphi through Facebook (24, 25)	6 April 2020	County**
Health and Human Services	Counts of hospital admissions due to confirmed or suspected COVID-19, as reported by the Department of Health and Human Services	31 December 2019	State
CovidActNow	COVID-19 testing results, such as positivity rate and number of tests, compiled by CovidActNow from CDC reporting	2 March 2020	County*
Google symptoms	Trends in Google search volume for terms related to anosmia and ageusia (loss of smell or taste), which correlate with COVID activity, based on data shared by Google (26)	13 February 2020	County***
Cases and deaths	Confirmed COVID-19 cases and deaths, compiled by JHU CSSE (1) and by USAFacts (3)	22 January 2020	County
NCHS mortality	Weekly totals of deaths broken down by cause, such as COVID, flu, or pneumonia, compiled by the National Center for Health Statistics (29)	26 January 2020	State

The above are accessible using `library(covidcast)` in R.

Source: https://www.pnas.org/content/pnas/118/51/e2111452118.full.pdf

Main question

When two different data signals are supposedly getting at the same concept, do they agree?

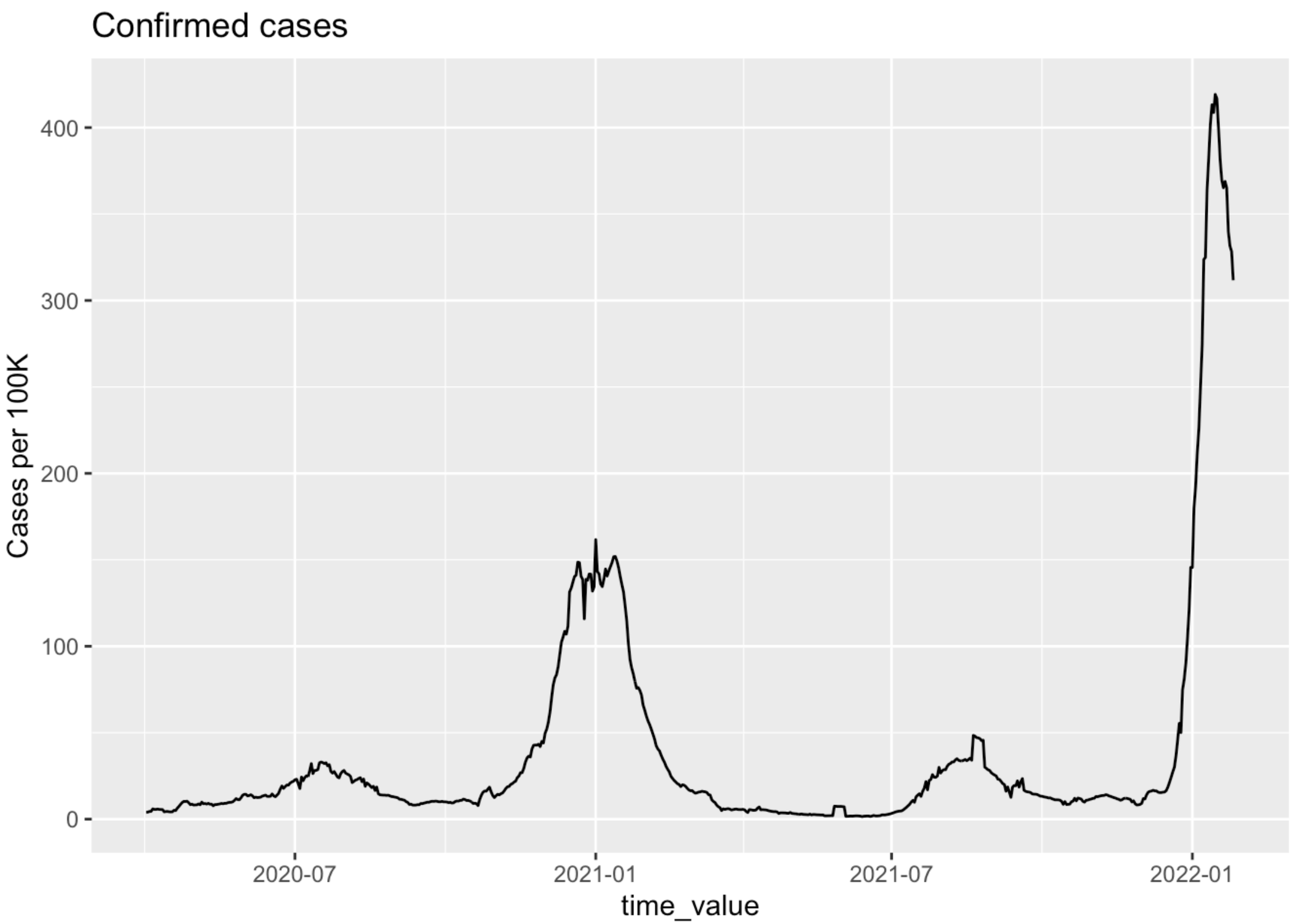
Why is this interesting?

- *There is often no “ground truth,” so it is difficult to assess data quality.*
- *When two sources happen to measure the same underlying concept, it is a unique opportunity to assess quality via data consistency.*
- *Caveat: When two sources disagree, this isn’t necessarily bad. Understanding why they are different can give us deeper insight into the data signals themselves, their strengths and weaknesses. Nuance is useful.*

A look at Los Angeles

with `library(covidcast)`

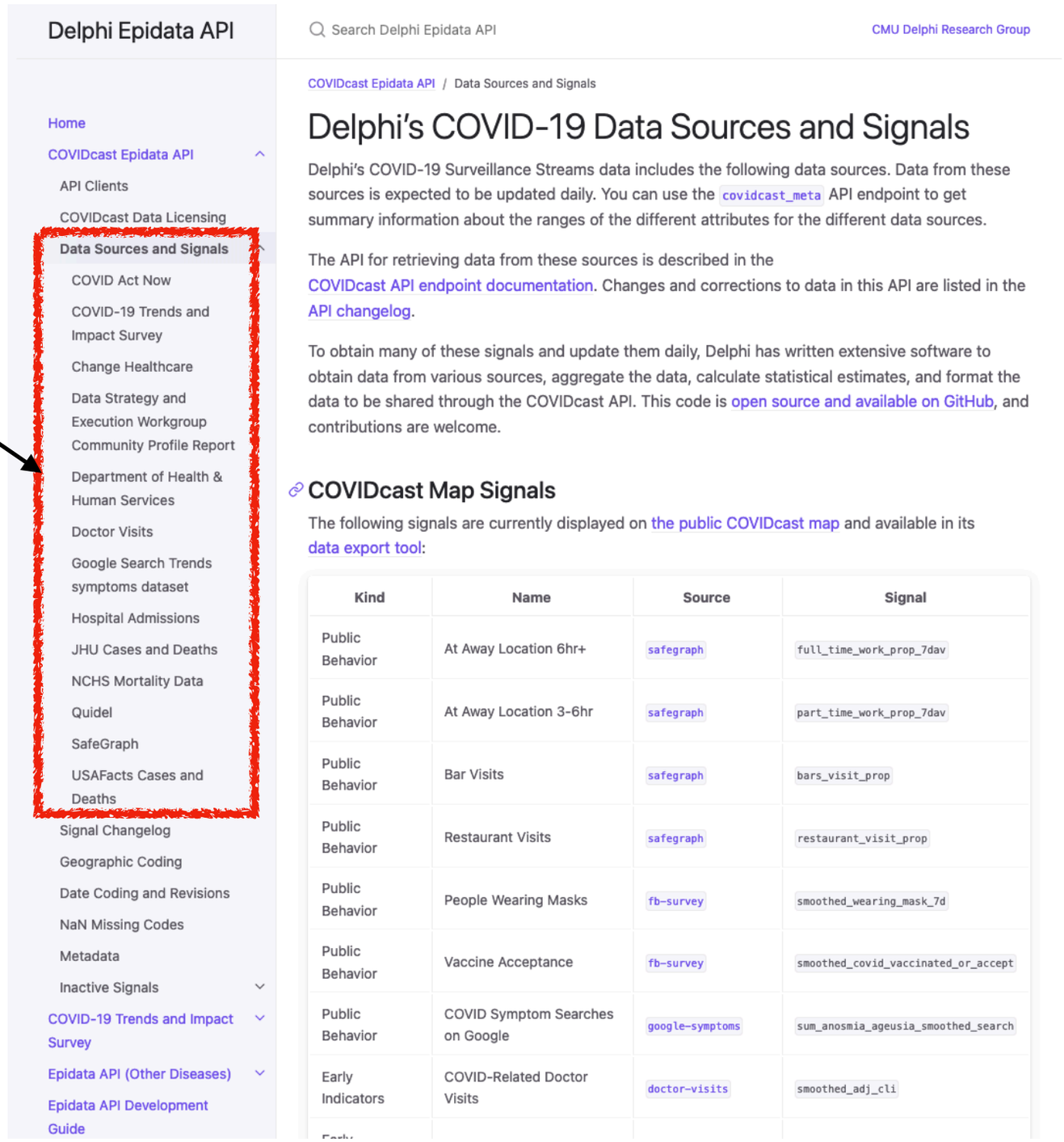
(I can share an R markdown notebook to reproduce the plots that follow)


Confirmed cases
400
300
200
100
0
Cases per 100K
2020-07
2021-01
2021-07
2022-01
time_value

Covid-like Illness (CTIS)

Where to find data…

Click on these links to find many data signals accessible via the **covidcast** R package



https://cmu-delphi.github.io/delphi-epidata/api/covidcast_signals.html

Examples of signals getting at the same idea

(But please look for more examples than this)

- Test positivity rate:
  - **CTIS** - *Estimated test positivity rate (percent) among people tested for COVID-19 in the past 14 days*
  - **Quidel** - *Percentage of antigen tests that were positive for COVID-19*
  - **NAAT from Community Profile Report**

Examples of signals getting at the same idea

(But please look for more examples than this)

- Symptoms:
  - **CTIS** - *Estimated % of people reporting that they or someone in their household experienced a sore throat in the past 24 hours*

    Source: https://cmu.app.box.com/s/ymnmu3i125go4aue0qxosi3rbcae20bj/file/714076275364
  - **Google search** - *Measures search volume for "sore throat" symptom*

    Source: https://github.com/GoogleCloudPlatform/covid-19-open-data/blob/main/docs/table-search-trends.md

(And likewise for a **large number** of symptoms)

What does it mean for two sources to agree?

- Exact numerical agreement, while great, may be too high of a bar unless they are measuring exactly the same quantity.

- Spatial / temporal correlations

- Perhaps rank correlation rather than Pearson correlation

- See https://delphi.cmu.edu/blog/2020/08/26/covid-19-symptom-surveys-through-facebook/ for discussion of correlations “sliced by county” and “sliced by time.”

- Open ended: Can you think of new effective ways for getting at whether two signals are in agreement?

Questions?