Measuring Icebergs @CoronaSurveys

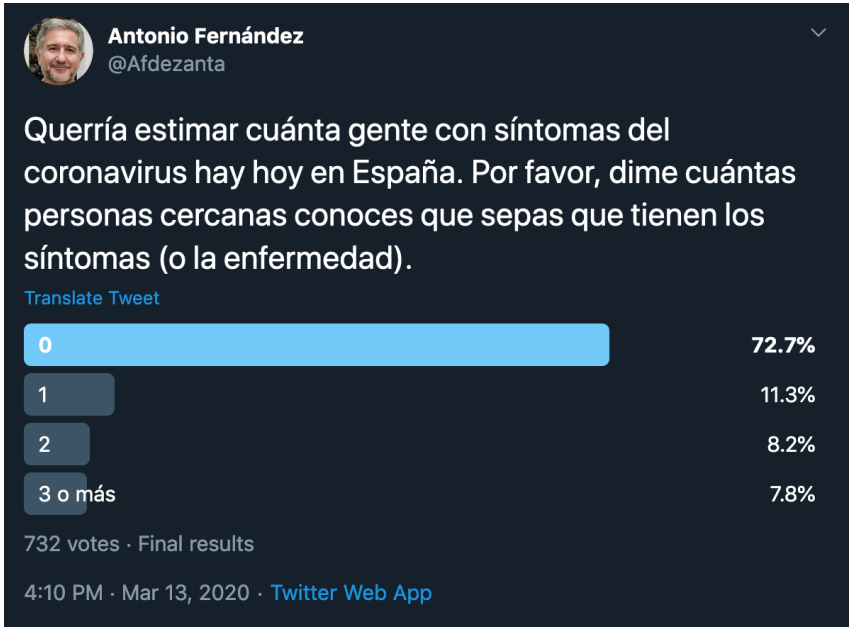
Davide Frey & @CoronaSurveys Team

WIDE Team, Inria Rennes

4 May 2020

Spain, March 13th 2020

Antonio Fernandez Anta, IMDEA Networks, Spain



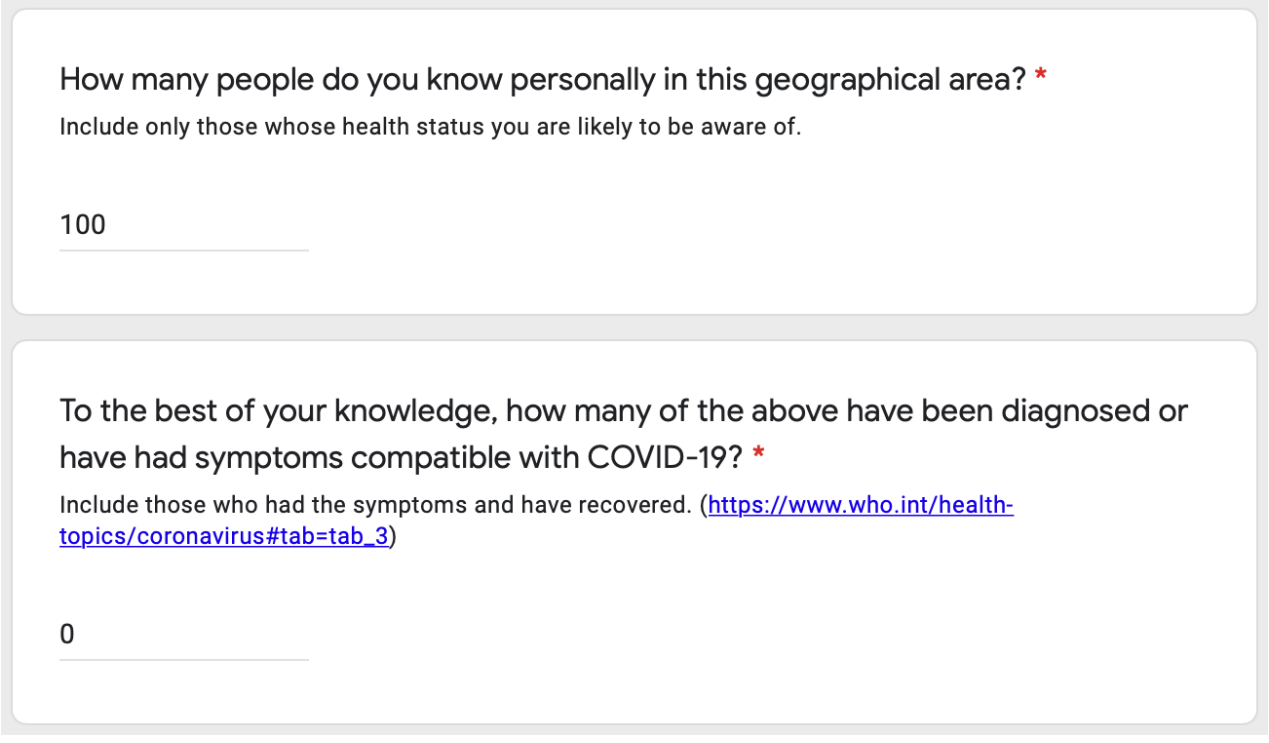
- $cases = 374,\ reach = 732 * 150$ (Dunbar number is 150)
- $total = \frac{cases}{reach} * ESpop \approx 153000$ ($\approx 30*$ official)

From Twitter to Google Forms

- Spread the survey beyond Spain
- Initial countries:
  Spain, Italy, Portugal, UK, USA, Germany, Cyprus
- French Survey Timeline
  - March 20: Antonio Fernandez contacted me and several other researchers
  - March 22: First "daily" survey in France on Google Forms
  - March 23-27: COERLE's feedback
  - March 28-today: Survey Resumes: Google Forms, always open.
- Currently 24 countries + Arab countries + rest of the world
- Migrating survey off Google Forms.

Two “simple” questions

For any given country/region and after informed consent we ask:


How many people do you know personally in this geographical area? *

Include only those whose health status you are likely to be aware of.

100

To the best of your knowledge, how many of the above have been diagnosed or have had symptoms compatible with COVID-19? *

Include those who had the symptoms and have recovered. (https://www.who.int/health-topics/coronavirus#tab=tab_3)

0

A few optional questions to be added in new version

- estimate network bias, and other parameters
- without collecting personal information

Trade-offs in Open Surveys

Positive Points:

- No personal data: health status of respondent is not asked
- No GDPR issues (but Ethical Board approval)
- Number of answers and reach size

Negative Points:

- Unorthodox approach: no panel selection, no stratification
- Not possible to track answers from users across time
- Fault injections

@CoronaSurveys at https://coronasurveys.org

https://www.instagram.com/coronasurveys/
https://tinyurl.com/coronasurveysfrance

CoronaSurveys Project Summary

The CoronaSurveys project is a collaborative endeavour from several universities and research institutions (team members). Data about COVID-19 cases is collected via anonymous open surveys (all the data collected is openly available). The results below present estimations on the incidence of COVID-19 from this and other available data. You can help by regularly completing the anonymous survey.

Media Presence

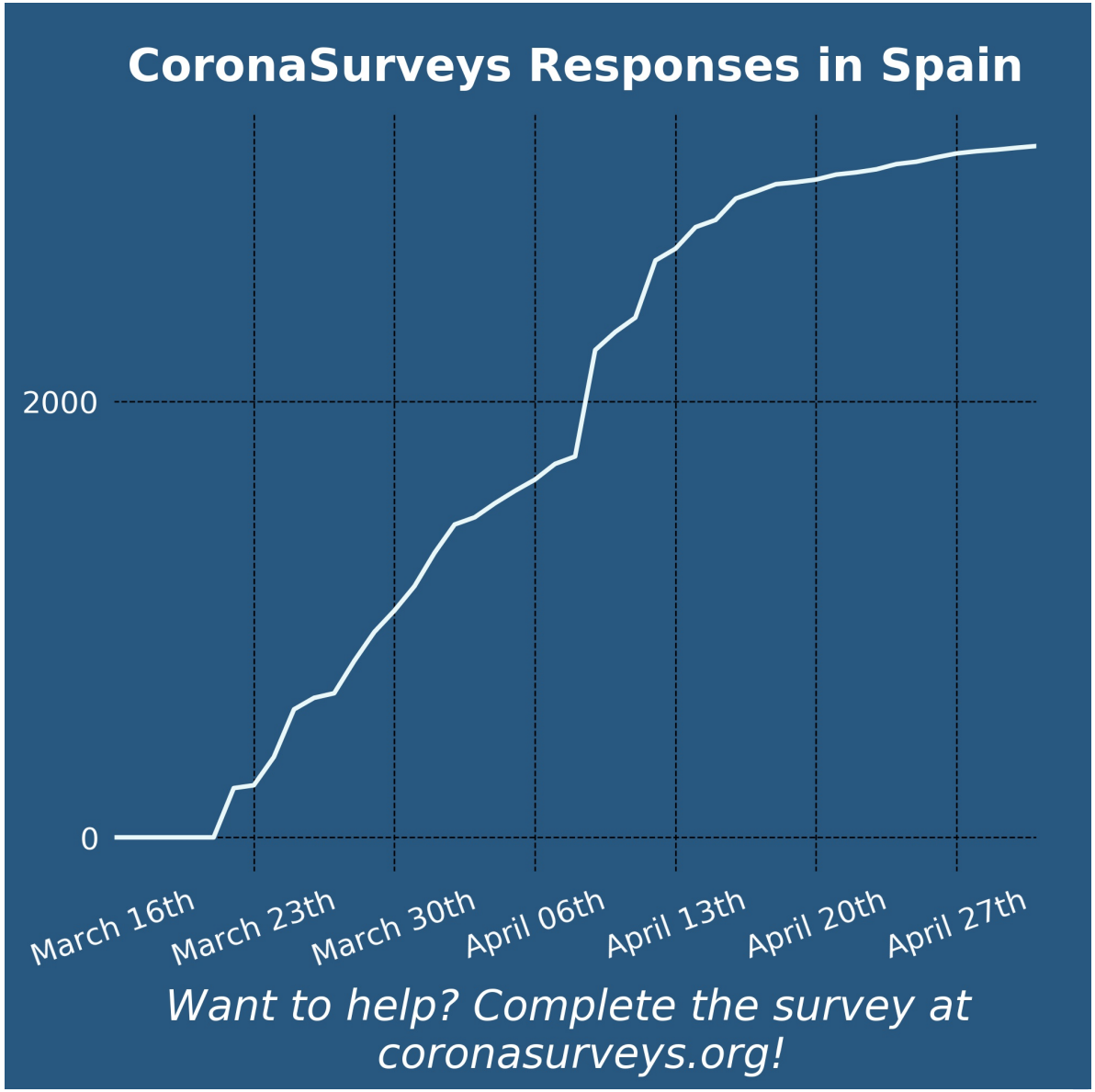
elpais.com/politica/2020/03/24/actualidad/1585077503_994849.html

EL PAÍS

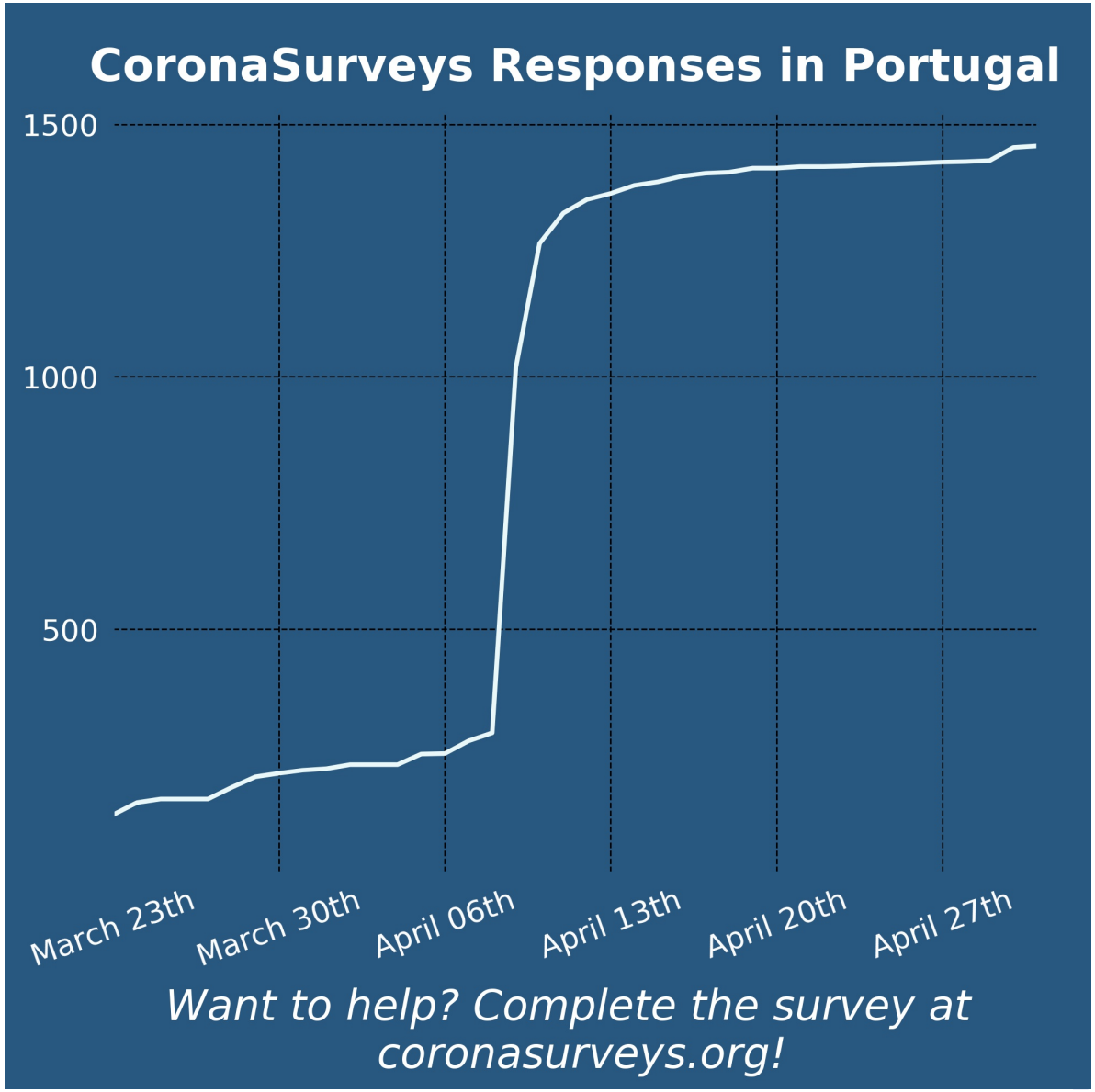
- El cálculo anterior lo podemos hacer también en Madrid: daría más de 200.000 infectados solo allí.

Cálculo C: Unos 900.000 infectados. Antonio Fernández Anta y otros investigadores están usando encuestas (Twitter y Facebook) para estimar los infectados en España. Sus cálculos del martes iban desde 900.000 hasta más de 2 millones.

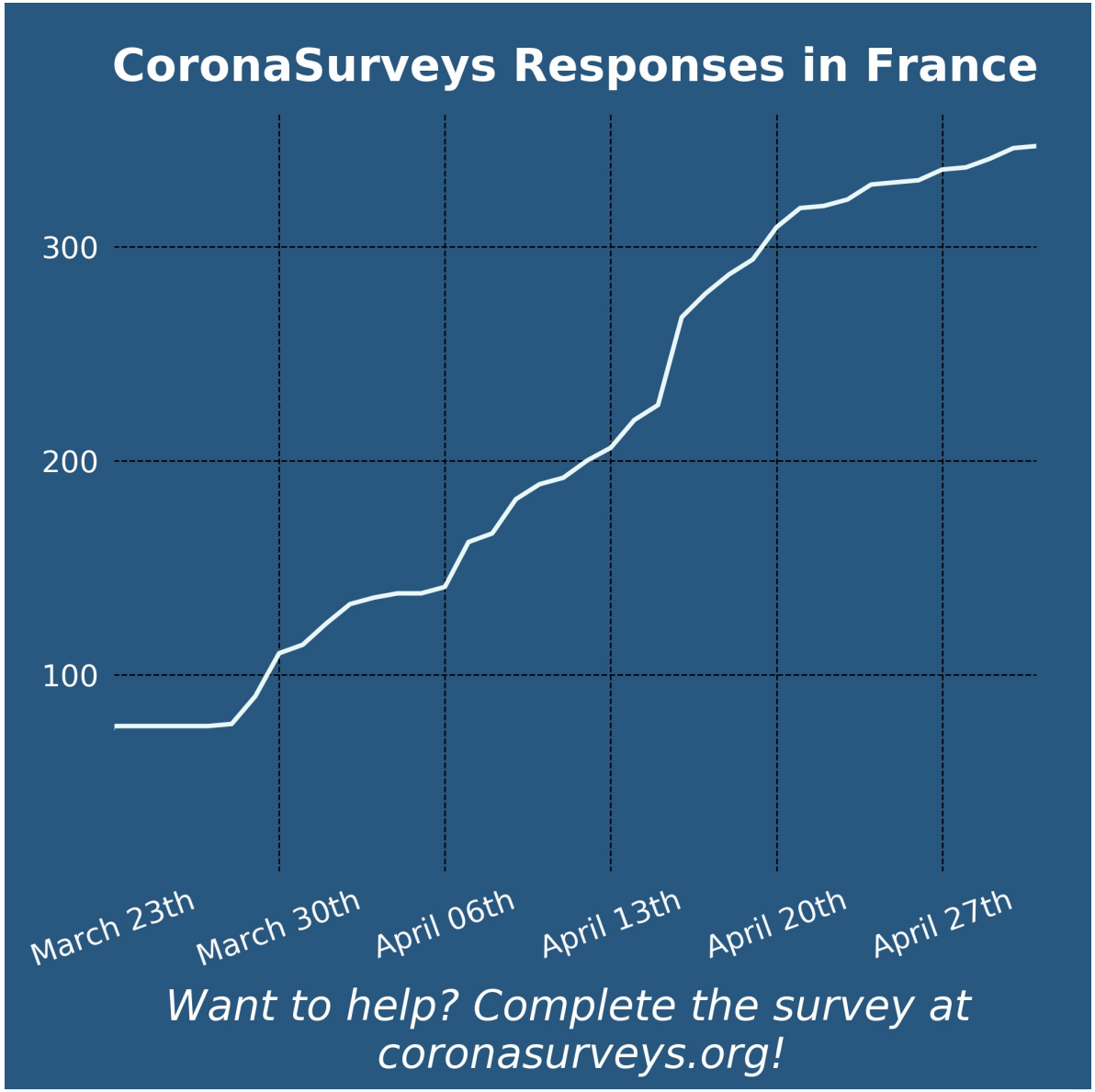
Responses: Spain



Responses: Portugal



Responses: France



Data cleaning

Survey responses, pairs $(cases_i, reach_i)$, are cleaned by identifying and removing outliers:

- $reach_i$ – remove entries above $1.5*$ the interquartile range
- $\frac{cases_i}{reach_i} < 0.3$ – remove entries with very high incidence
- Several datapoints, from batches of 30 or more answers

Estimators

Assume n response pairs $(cases_i, reach_i)$, region of population P
We can produce three estimators:

$$\text{Weighted} \quad E_w = P \cdot \frac{\sum_i cases_i}{\sum_i reach_i}$$

$$\text{Mean} \quad E_m = P \cdot \frac{\sum_i (cases_i / reach_i)}{n}$$

$$\text{Dunbar} \quad E_d = P \cdot \frac{\sum_i cases_i}{n \cdot 150}$$

Only showing Weighted in the following.

- Still, how can we know the data is meaningful?
- We need other independent estimates

Fatality based estimates

Coarse grained: deaths * 400 estimate

Amy Maxmen. How much is coronavirus spreading under radar? Nature News Explainer, March 13th, 2020. https://www.nature.com/articles/d41586-020-00760-8

Current deaths times 400

> goes something like this: Data from China suggest that about three weeks passes between when a person feels sick and dies from COVID-19. And if you assume a case fatality rate of roughly 1%, a back-of-the-envelope calculation suggests that each death represents about 100 cases in the first week. Right now, he adds, you can expect the epidemic to double each week if those cases aren't identified and isolated – bringing the estimate to 400 at the time of death. Because the error bars on each of these variables are large, epidemiologists check their figures against further information.

Coarse grained estimate, only in the initial exponential growth.

Fatality based estimates

Fine grained: cCFR estimate

- Corrected Case Fatality Ratio $\approx$ fatalities over detected cases with known outcomes (about 2 weeks lag)
  - `https://cmmid.github.io/topics/covid19/global_cfr_estimates.html`
- Find a "stable" baseline for the cCFR
  - China, Wuhan, $cCFR = 1.38$. Verity et al. Lancet paper

Fatality based estimates

Fine grained: cCFR estimate

- Corrected Case Fatality Ratio $\approx$ fatalities over detected cases with known outcomes (about 2 weeks lag)
- Find a "stable" baseline for the cCFR
  - China, Wuhan, $cCFR = 1.38$. Verity et al. Lancet Paper

April 17th, Wall Street Journal

Fatality based estimates

Fine grained: cCFR estimate

- Corrected Case Fatality Ratio $\approx$ fatalities over detected cases with known outcomes (about 2 weeks lag)
- Find a "stable" baseline for the cCFR
  - China, Wuhan, $cCFR = 1.38$. Verity et al. Lancet Paper
  - Alternatives: Korean data, Germany Gangelt data
- Calculate a country cCFR from daily ECDC data
- Use proportion to baseline to infer current cases estimate

Caveats: Baseline quality, death reporting policies, cases coverage, calibration of symptoms to death delay, missing date of symptoms.

Fatality based estimates

Fine grained: cCFR estimate

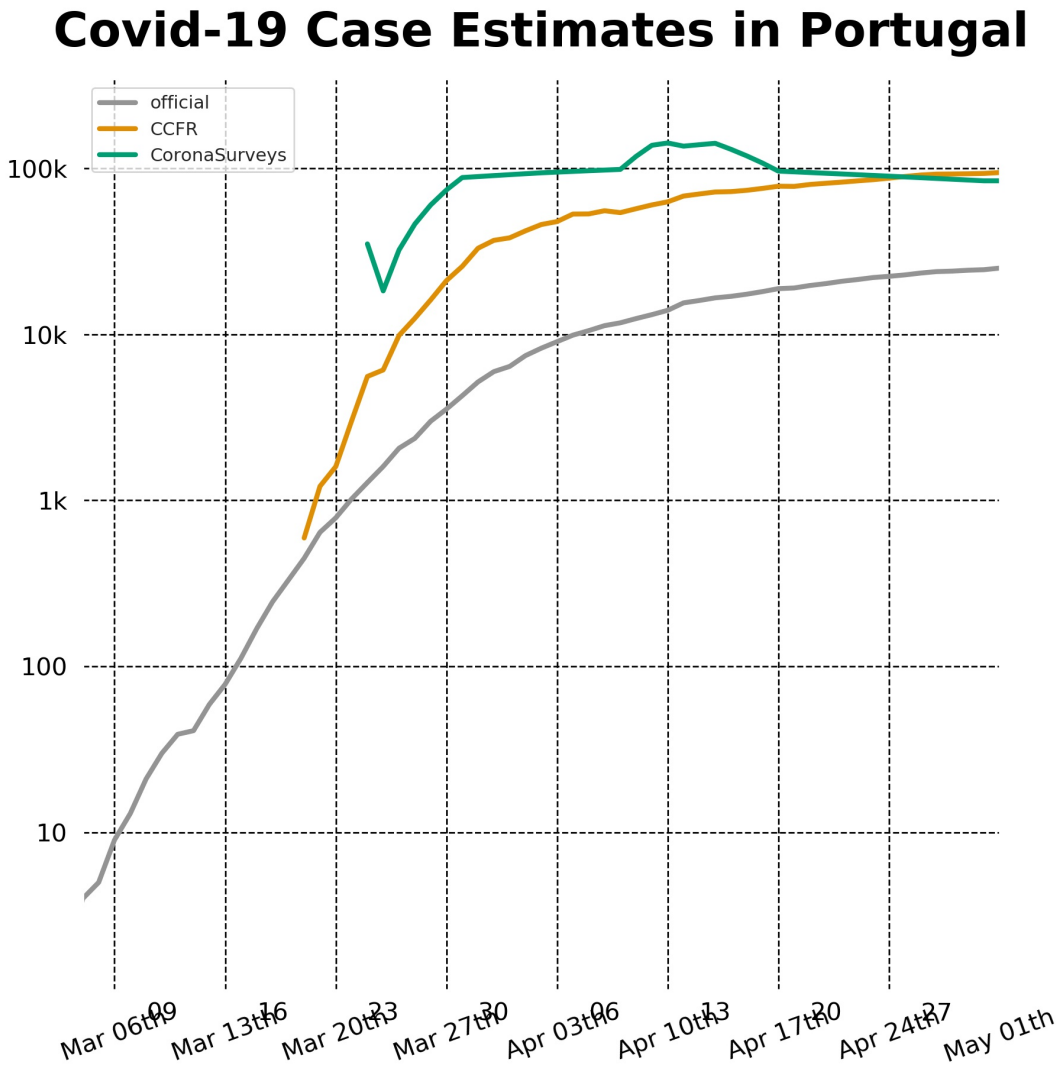
- Corrected Case Fatality Ratio $\approx$ fatalities over detected cases with known outcomes (about 2 weeks lag)
- Find a "stable" baseline for the cCFR
  - China, Wuhan, $cCFR = 1.38$. Verity et al. Lancet Paper
  - Alternatives: Korean data, Germany Gangelt data
- Calculate a country cCFR from daily ECDC data
- Use proportion to baseline to infer current cases estimate

Caveats: Baseline quality, death reporting policies, cases coverage, calibration of symptoms to death delay, missing date of symptoms.

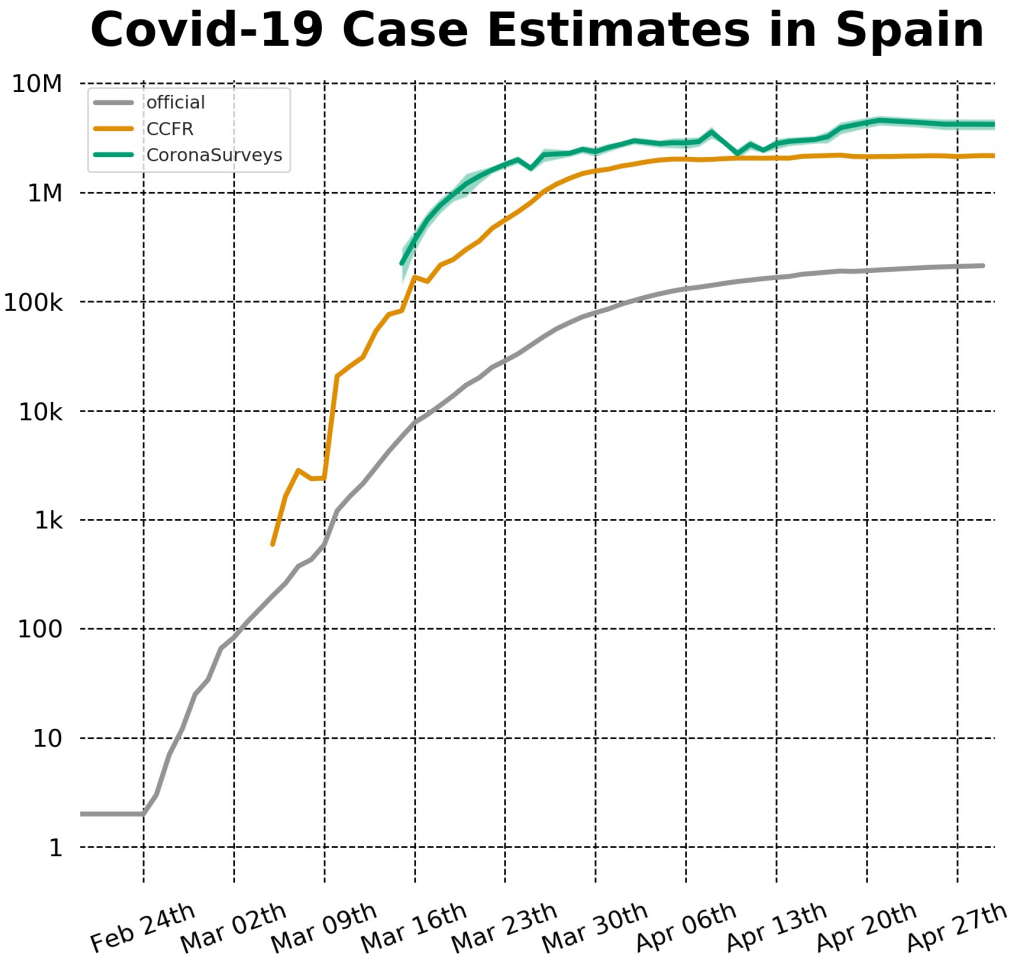
Summary of Estimators

- Official number of cases
- CCFR estimate
- Coronasurveys estimate

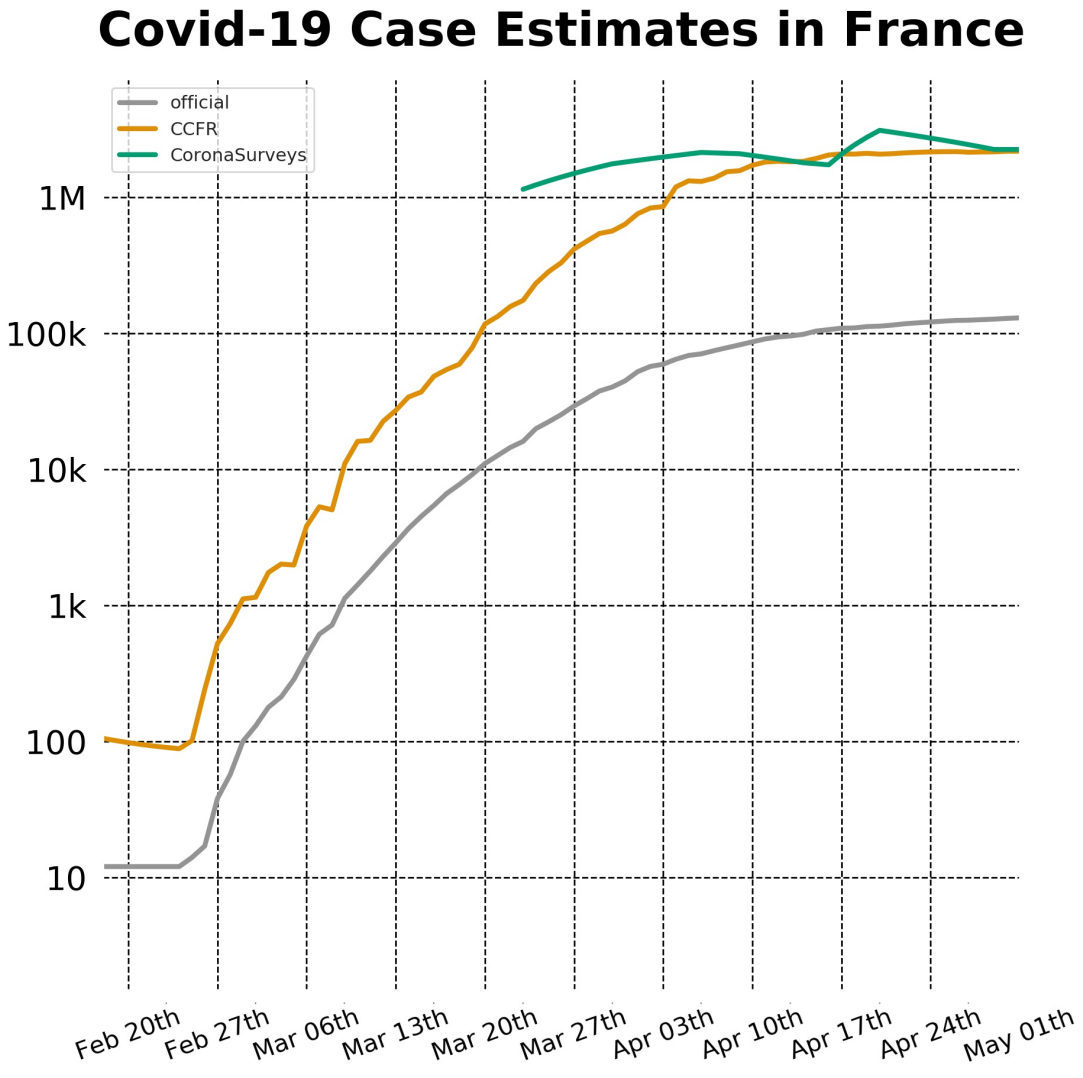
Estimates, Portugal



Estimates, Spain



Estimates, France

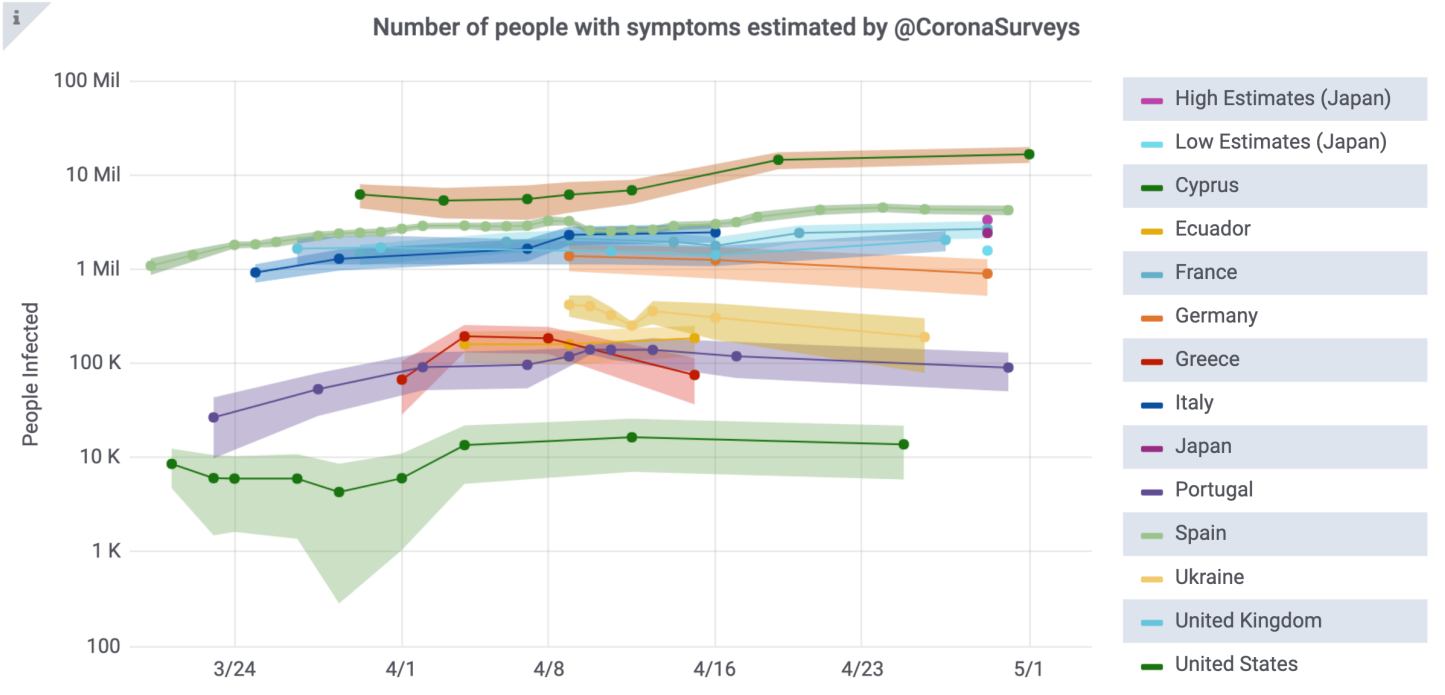


@Coronasurveys, status and potential uses

- Data is being collected since mid March
- We are calibrating as the pandemic and data evolves
- Surveys are open for all globe
- New regional-based estimates
- Survey improvements to reduce bias

- Potential for quick assessment when reliable techniques lack
- Countries with good digital penetration but lacking testing

Portfolio
Countries



Questions?

@CoronaSurveys at http://coronasurveys.org

- https://www.instagram.com/coronasurveys/
- https://tinyurl.com/coronasurveysfrance
- Twitter: @coronasurveys
- https://www.instagram.com/coronasurveys/
- FB: https://www.facebook.com/groups/209076966867175
- Both our raw and processed data is openly available https://github.com/GCGImdea/coronasurveys