# Time Series Analysis
# 1. Prop 99 CA Tax
# 2. Italian Covid-19 Excess Mortality Rates

Chirag Modi, Vanessa Boehm
Uros Seljak
LBNL, UC Berkeley

with Simone Ferraro & George Stein

RPM, LBNL, Berkeley

23rd April 2020

# Links

Non-Parametric Gaussian Process in Fourier Space for estimating counterfactual & Impact
Modi & Seljak 2019, NeurIPS 2019
https://arxiv.org/abs/1910.07178

COVID-19 pre-print (Modi et.al.)
https://www.medrxiv.org/content/10.1101/2020.04.15.20067074v2

Medium post discussing our results-
https://medium.com/bccp-uc-berkeley/how-deadly-is-covid-19-data-science-offers-answers-from-italy-mortality-data-58abedf824cf

Medium post detailing our methodology -
https://medium.com/bccp-uc-berkeley/time-series-analysis-of-italian-mortality-data-ac780e2a1706

COVID-19 data, code, updated draft -
https://github.com/bccp/covid-19-data

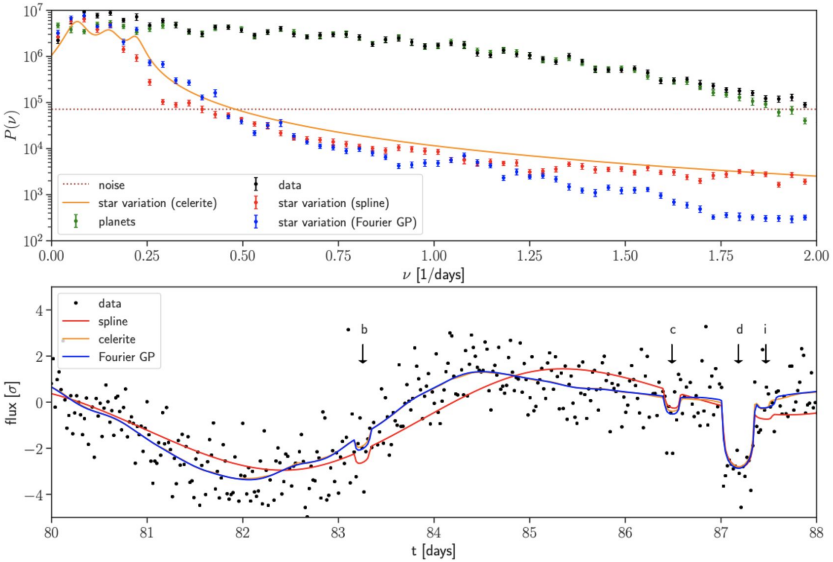
# Why do astronomers care about time series analysis?

A lot of astronomy data are time series:

Example - Kepler time series of a star flux, searching for planet transits in the presence of non-Gaussian noise (outliers!) and star variability (sunspots etc)

Solution: Gaussian Process using Fourier based power spectrum stationary kernel (Robnik & Seljak 2020)

Here we will apply this method to predict counterfactuals

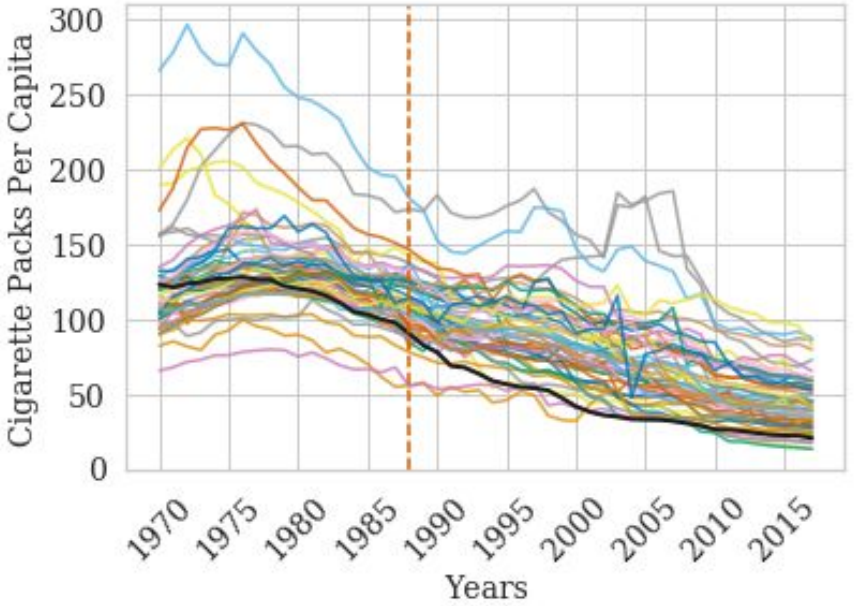


# Part 1: Counterfactual Analysis with Gaussian Process in Fourier Space

# Counterfactual Analysis

*Prop 99 : State of California passed a tax on the sale of cigarette packs in 1988. Did it reduce the sale of tobacco in the state?*

- A set of "control" units as baseline
- A "treated" unit that undergoes a "treatment"
- "counterfactual" - what would the treated unit look like in the absence of a treatment.



# Synthetic Controls Method (Abadie et.al. 2010)

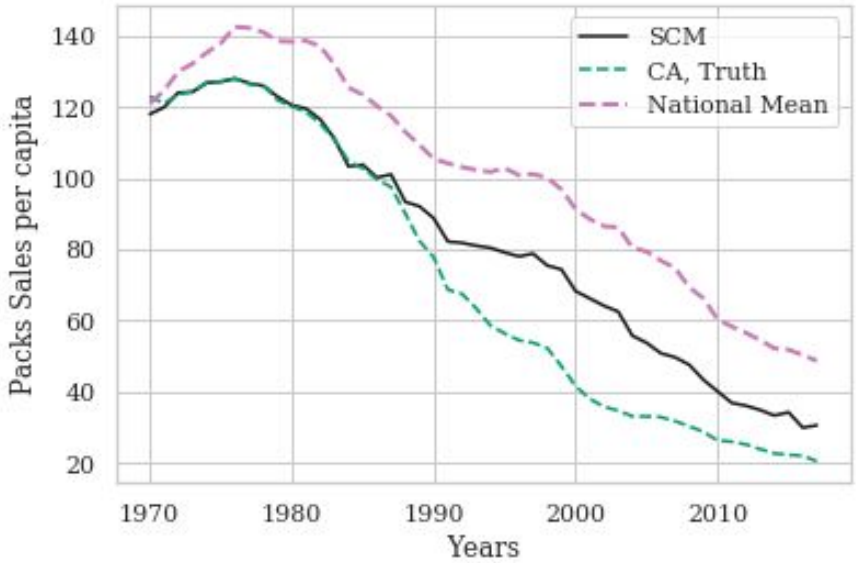
Counterfactual is a weighted linear combination of control units

How to estimate the weights?
Minimize the difference with the pre-intervention observed data

$$W^* = \min_W (W^T \cdot X_0^N - X_1^I)^2 \quad \text{s.t.} \sum_{i=i}^{N} W_i = 1, W_i > 0 \; \forall i$$

Predict

$$Y_1^N = W^* \cdot Y_0^N$$




https://economics.mit.edu/files/11859

# Non-parametric GP in Fourier Space

Modi, Seljak (NeurIPS 2019)
https://arxiv.org/abs/1910.07178

**Basic idea:** learn time-correlations in the data & use long-wavelength modes to predict counterfactual

**Latent variables -** Fourier modes of the time-series of treated unit
**Assumption -** Stationary and Gaussian data (or make it so)

Prior learnt from controls

$$\mathbf{s}_i = \mathbf{u}_i + i\mathbf{v}_i = \mathcal{R}^\dagger \mathbf{d}_i$$

$$\mathcal{P}_i = \mathbf{s}_i \mathbf{s}_i^\dagger, \;\; \mathcal{P}_{i\nu} = u_{i\nu}^2 + v_{i\nu}^2,$$

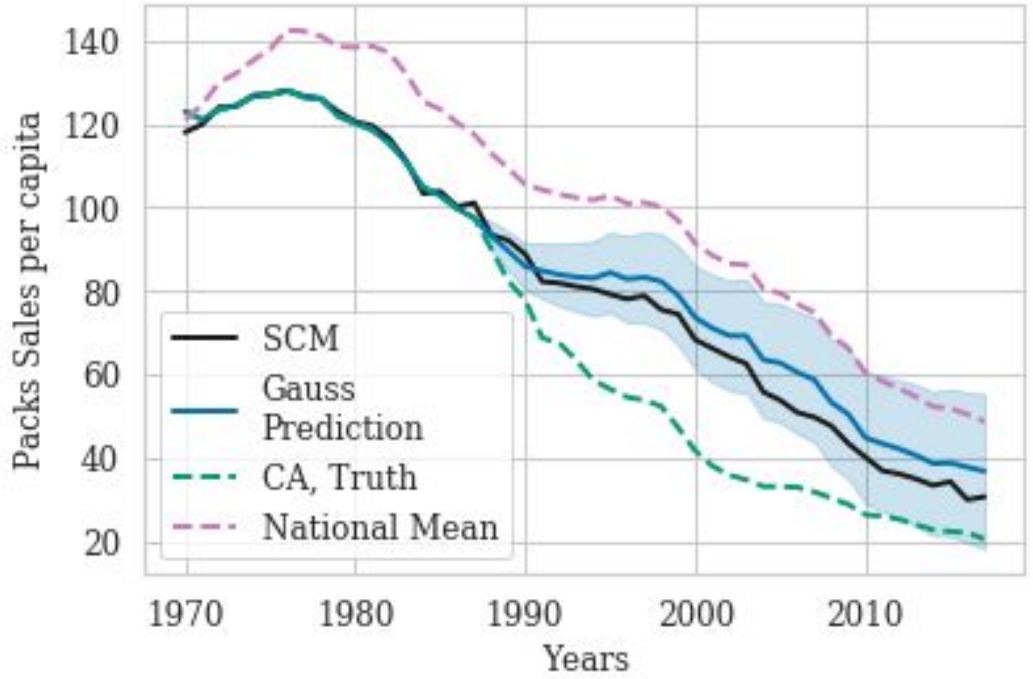
Minimize negative log-posterior to fit Fourier Modes

Covariance on the prediction

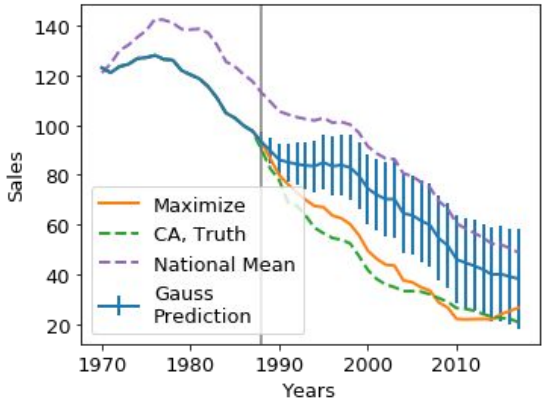
$$\mathcal{L} = \sum_{t=T_i=1}^{T_0} \frac{\left(\mathcal{R} s_{1,t} - d_{1,t}^N\right)^2}{\sigma_t^2} + \frac{\mathbf{s}_1 \mathbf{s}_1^\dagger}{\mathcal{P}_{\rm pr}}.$$

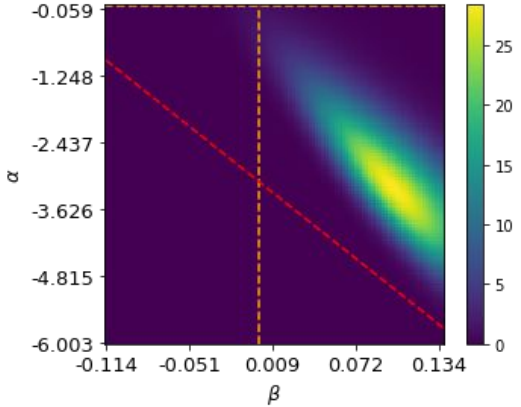
$$C = \mathcal{R}((\mathcal{P}_{\rm pr}{}^{-1} + \mathcal{R}^\dagger \mathcal{N}^{-1} \mathcal{R})^{-1} \mathcal{R}^\dagger.$$

# Counterfactual Prediction



# Hypothesis Testing




Hypothesis A : No impact
Hypothesis B : Significant impact
**Strategy :** Compare likelihood ratios

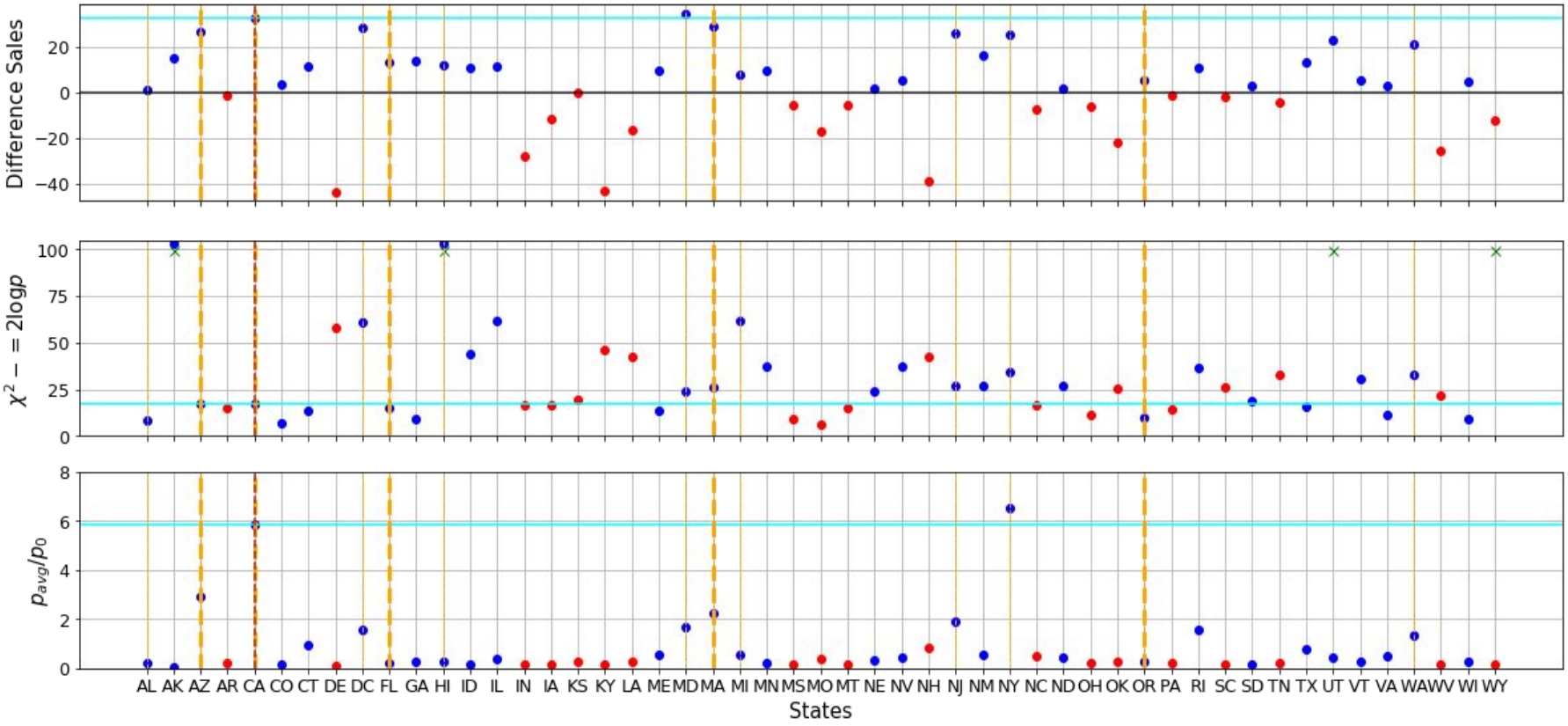
Hypothesis A ⇒ Likelihood of data under predicted counterfactual
Hypothesis B ⇒ Likelihood of data under alternate, parametric counterfactuals to make observations more likely

**a-priori & control driven**

$$m_t(\alpha, \beta) = \hat{Y}_{1,t} + \alpha(T - T_0) + \beta(T - T_0)^2 \quad \forall t \in (T_0, T_f)$$

$$p_B(\mathbf{Y}_1^I) = \int d\alpha d\beta p(\alpha, \beta) p[\mathbf{Y}_1^I | \mathbf{m}(\alpha, \beta)].$$

# Placebo analysis to validate methodology



Modi, Boehm, Ferraro, Stein, Seljak (April 2020)
https://doi.org/10.1101/2020.04.15.20067074

# Part 2:
# Time Series Analysis of Italian Covid-19 Excess Mortality Rates

https://github.com/bccp/covid-19-data

# Time Series of Mortality Data - Motivation

**How deadly is COVID-19?**

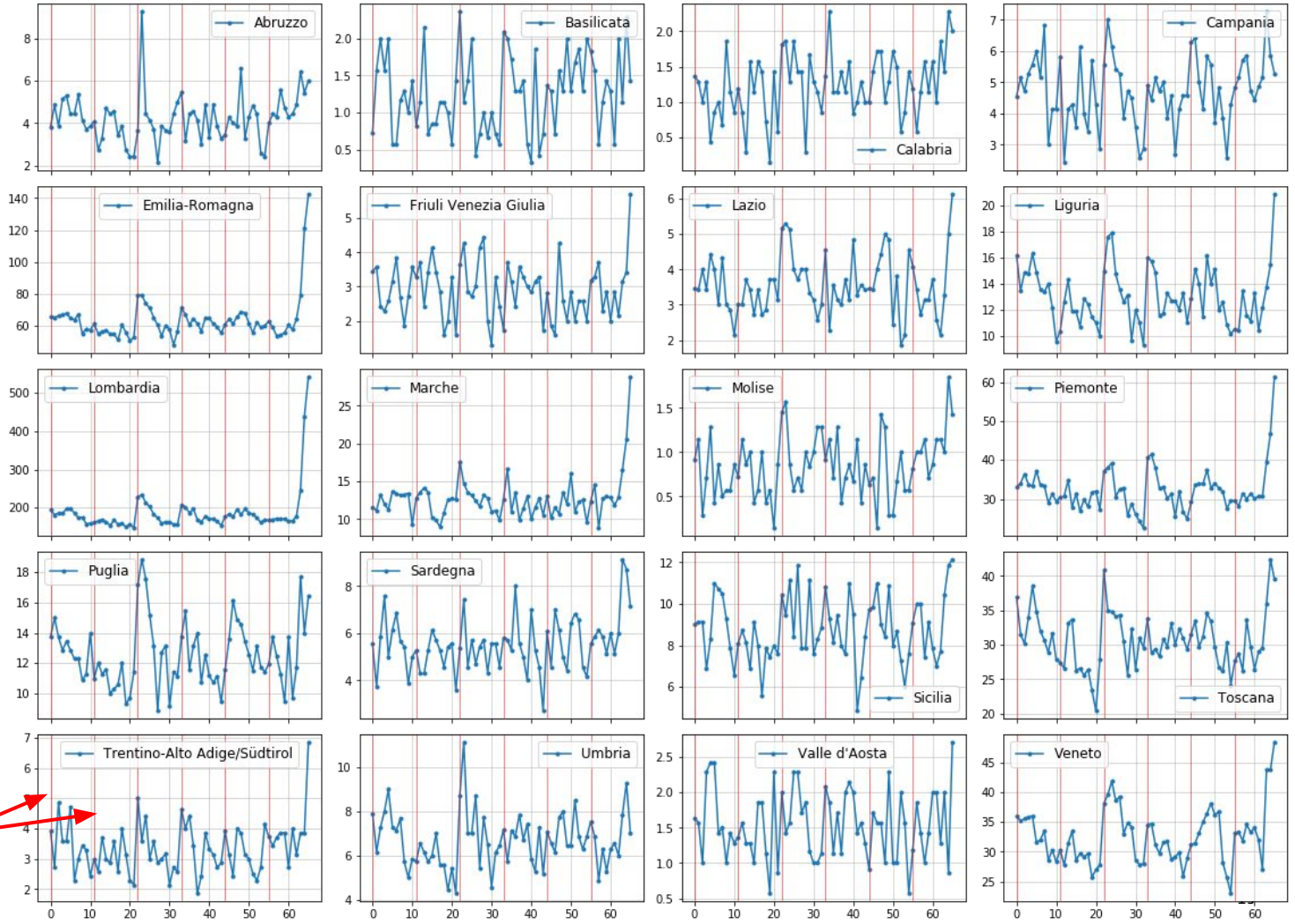
requires knowledge of two numbers:

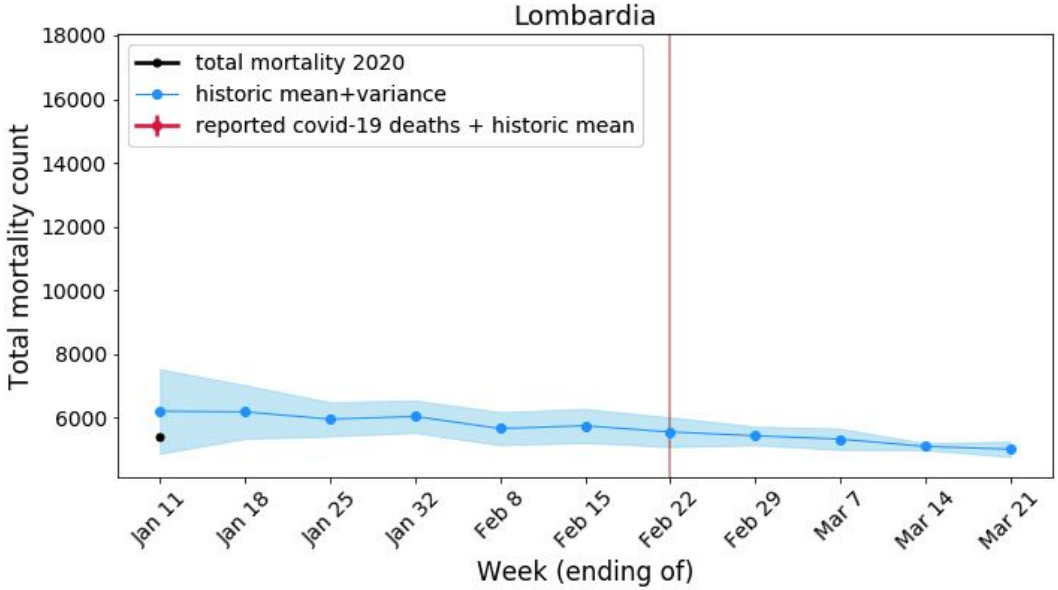
1) total number of deaths
2) total number of infections

a lot of discussion has focused on getting 2) right (e.g. antibody tests)
But are we sure that all deaths have been counted?

*Can we find a dataset that answers 1) independently of testing?*

Data: daily mortality with age information for 1648 towns from all the different regions in Italy from 01/01-04/04 for 2015-2020







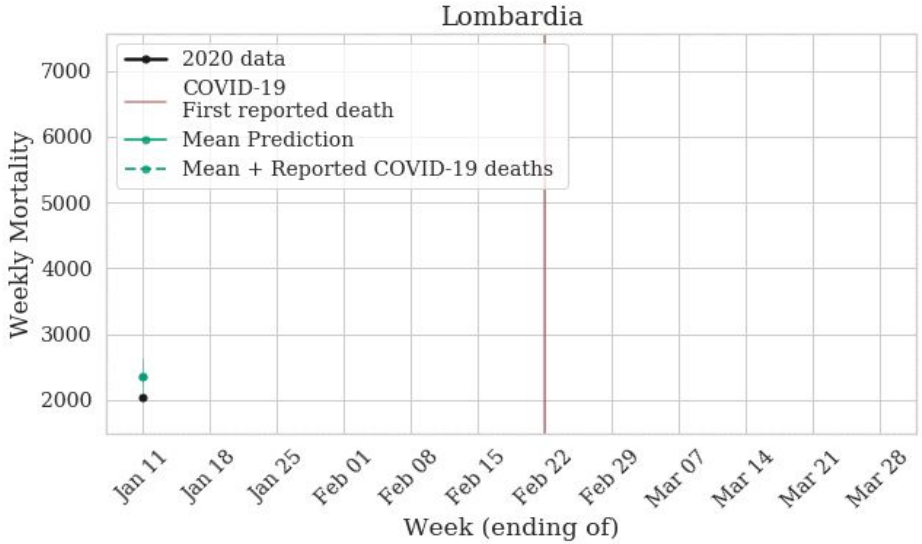
- similar trends observed in other countries
- caught media attention in last few days, e.g.

https://www.nytimes.com/interactive/2020/04/21/world/coronavirus-missing-deaths.html

# Counterfactual Analysis

Mean estimate is
*completely agnostic of any observation.*
↦ suboptimal and possibly biased.

It ignores information
from this year's data and correlations
between weeks.




## Our Analysis:

1. learn a model from historic data
2. condition it on this year's pre-pandemic data
3. make counterfactual (what if there had been no pandemic) prediction for time after onset of pandemic (Feb 22).

# Conditional Mean Gaussian Process

Assumption: data follows a Gaussian distribution

Method: Train a GP and condition on this year's pre-pandemic data

Choice of kernel:

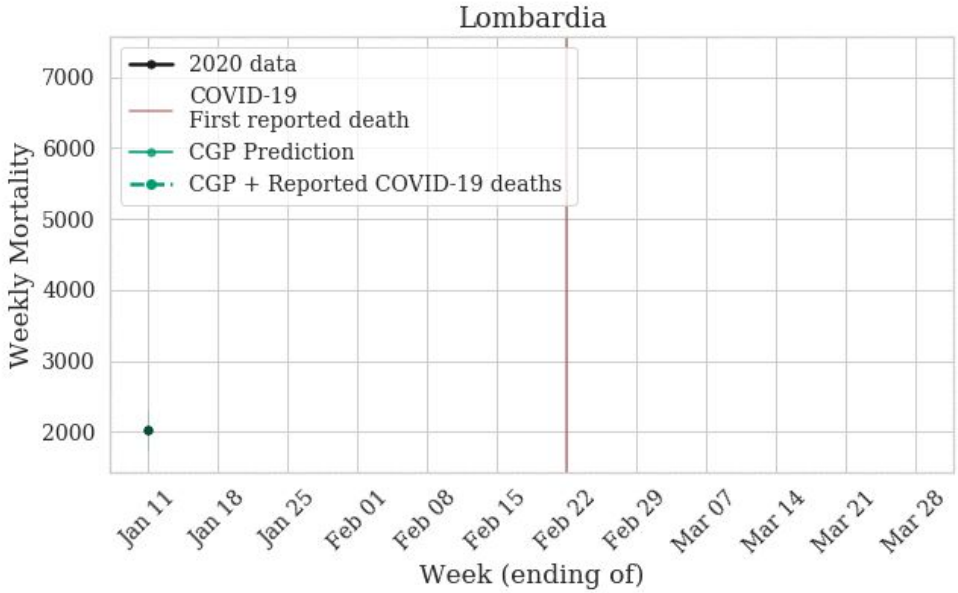
- learn kernel from historical data:
  - 2 component PCA of measured cov (captures 90% of variance in data)
  - add a squared exponential stationary kernel

Prediction is mean:

$$\overline{\boldsymbol{\mu}} = \boldsymbol{\mu}_1 + \Sigma_{12}{\Sigma_{22}}^{-1}(\boldsymbol{a} - \boldsymbol{\mu}_2)$$

Uncertainty from covariance:

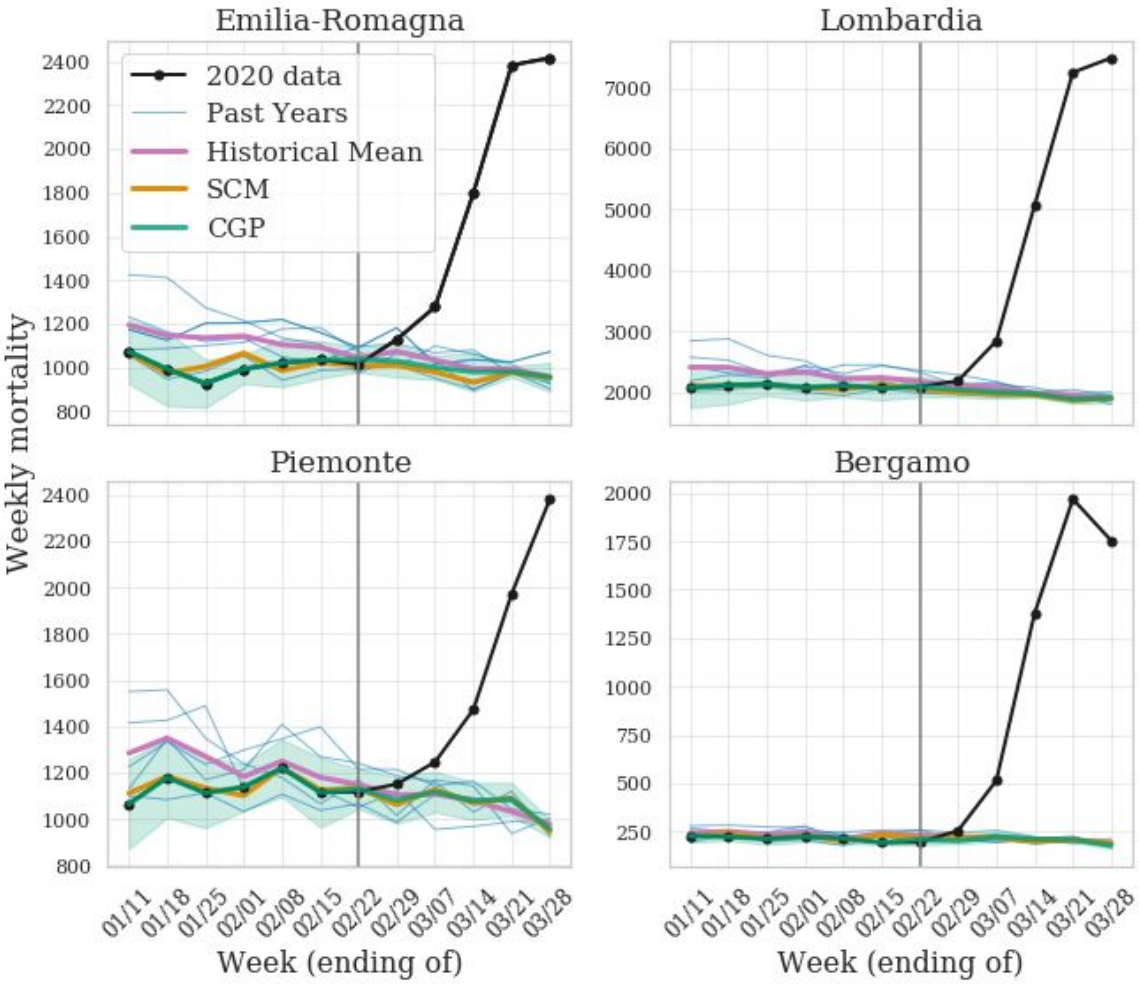
$$\overline{\Sigma} = \Sigma_{11} - \Sigma_{12}{\Sigma_{22}}^{-1}\Sigma_{21}$$



# Validating Counterfactuals on the pre-interventions

Counterfactuals match the observed data better than historical mean

by construction for GP without noise (Poisson noise is negligible in most cases)



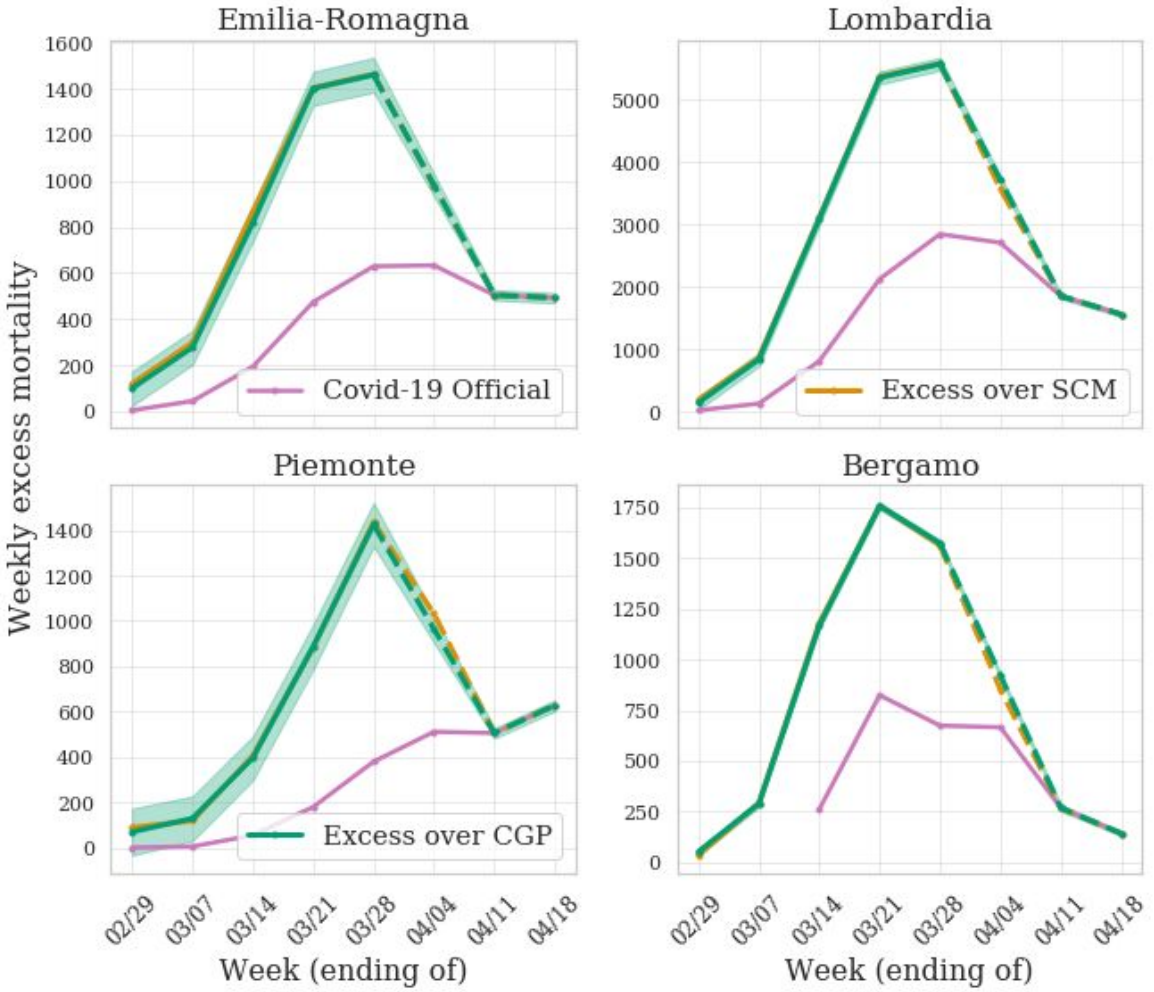
# Weekly Excess Mortality And Extrapolating

Data is *incomplete*:
Lombardia 72%
Bergamo 74%
we scale up assuming at random (35-39% correction).

large excess in mortality over CF
large excess in mortality over official COVID-19 mortality until March 28

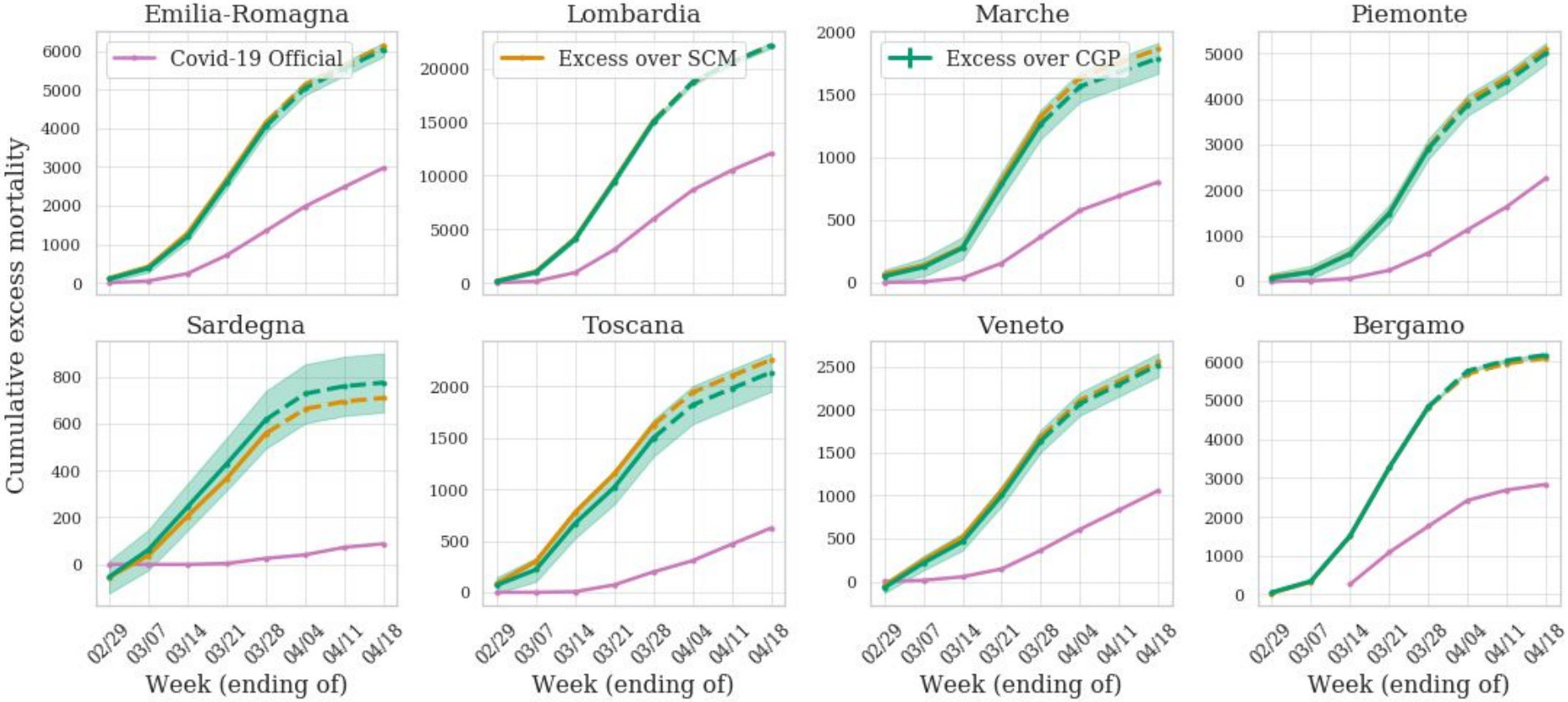
Assumption: excess due to COVID (backed with correlation analysis)

To estimate total mortality we linearly extrapolate to April 11 and then follow COVID-19 deaths



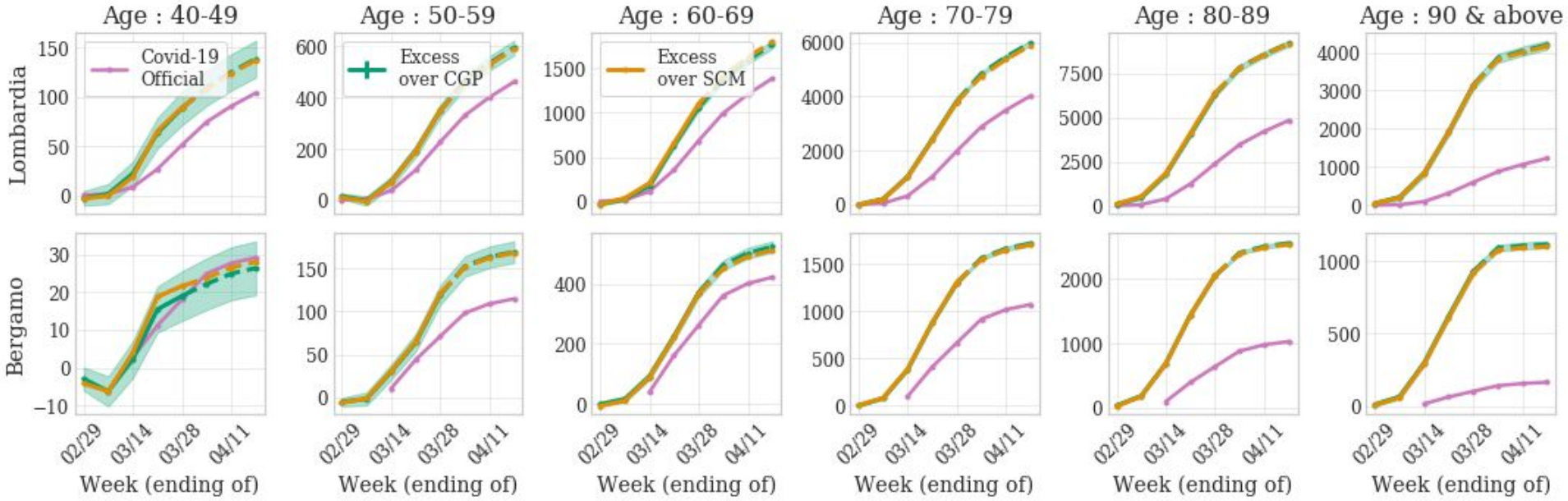
# Cumulative Excess Mortality Count

COVID death toll estimated by this analysis:
52000±2000
factor of 2-4 higher than official numbers



# Cumulative Excess Mortality Count

- Fairly good match with reported COVID-19 deaths for ages below 50 years
- Significant excess that increases with age for ages above 60.
- *Hypothesis:* old people that die at home or in retirement homes; never tested for COVID-19



# Part 3:
# How deadly is COVID-19?

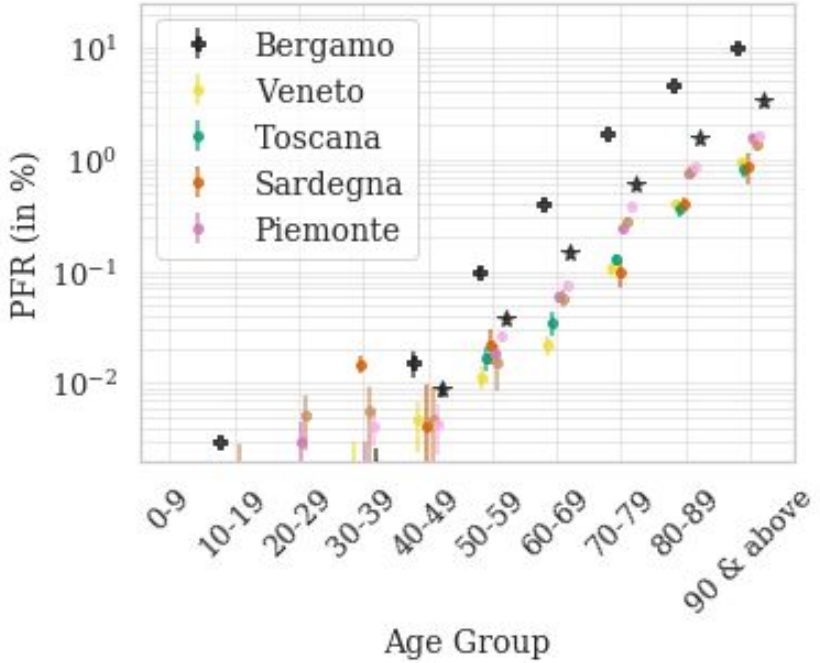
# What can we say about Infection Fatality Ratio (IFR)?

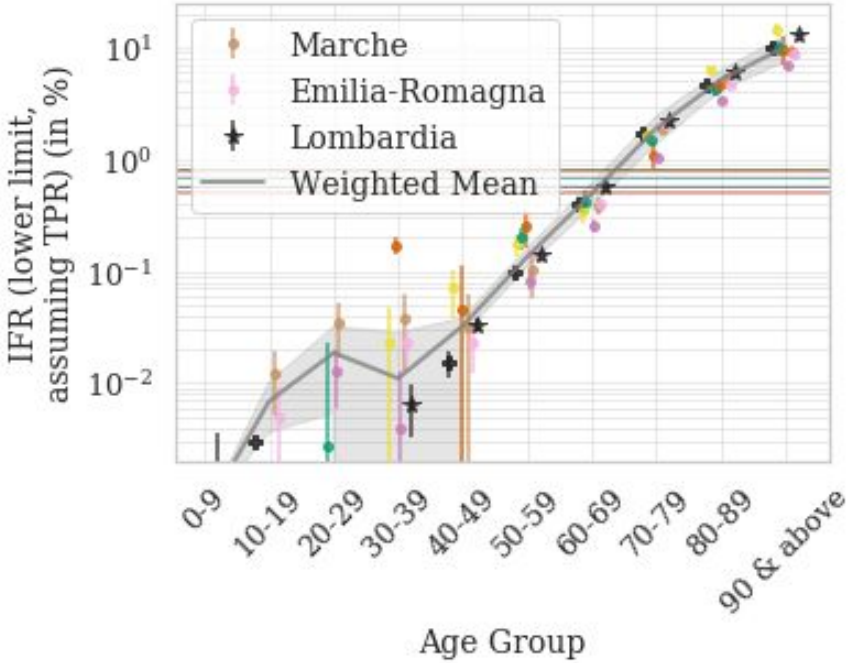
$$PFR = \frac{\#\ deaths\ due\ to\ infection}{population} \qquad IFR = \frac{\#\ deaths\ due\ to\ infection}{population\ x\ IR}$$

- IFR lower bound = PFR with IR = 1, i.e. # Covid-19 deaths / total population
  - 0.56% Bergamo province
- Better IFR lower bound : test positive rate (TPR, fraction of tests that are positive) is an **upper bound** to IR
  - Assumption - sicker people get tested more
  - 0.8% lower bound for Lombardia (from 27% TPR in Lombardia)
- IFR for Lombardia 1% (0.8%-1.6%, 95% c.l.) using Princess Diamond data

# Infection Fatality Ratio (IFR) for age groups

PFR and IFR are a strong function of age: in Bergamo 2% for 70-79, reaching 10% for people above 90 years.



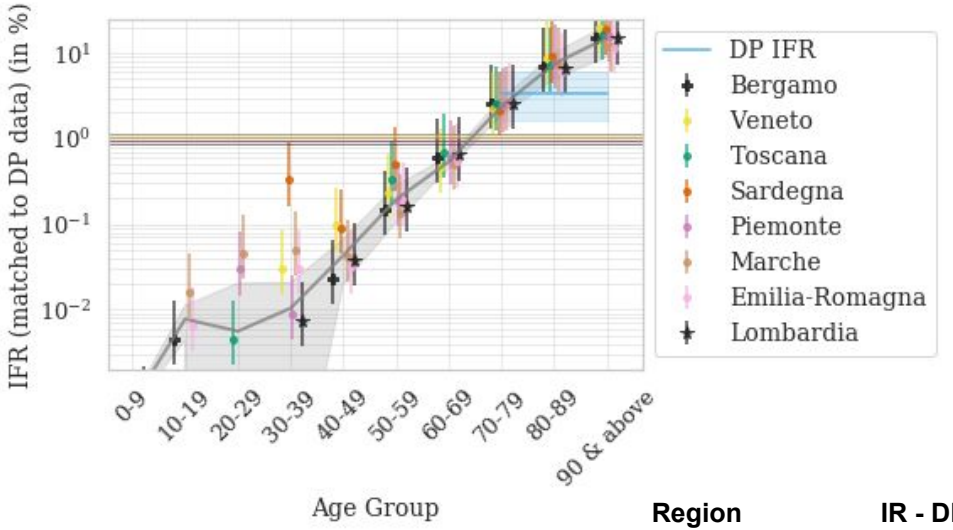



# What can we say about Infection Ratio (IR)?

- **Princess Diamond Cruise -**
  - 11 deaths for age > 70
  - (could be 13 out of 14 total)
  - 330 total infections
  - IFR = 3% (1.5%-5% 95%cl)

- Matching regional age distribution for age>70 & Assume age-independent IF

- Lombardia : PFR ~ 1%
  - Assuming same IFR 70-89: IR ~ 23% (12%-40% 95%cl)
- Bergamo province: 63% (35%-100%): herd immunity?




IR varies a lot between different regions →

| Region | IR - DP |
|---|---|
| Emilia-Romagna | 0.14 (0.07-0.25) |
| Lombardia | 0.23 (0.12-0.41) |
| Marche | 0.11 (0.06-0.20) |
| Piemonte | 0.10 (0.05-0.18) |
| Puglia | 0.03 (0.02-0.05) |
| Sardegna | 0.05 (0.02-0.08) |
| Toscana | 0.05 (0.03-0.09) |
| Veneto | 0.05 (0.02-0.09) |
| Bergamo | 0.67 (0.33-1.19) |

# What can we say about Case Fatality Ratio (CFR)?

- CFR = # fatalities / # positive tests, CFR>IFR
- Varies drastically across the countries, from 0.5% to 10% (Italy), 20% Lombardia
- High IR explains high CFR for Italy
  - Lombardia: 2% of population tested, TPF ~30%, so  0.7% population is positive
  - Estimated IR = 23%, 35x times larger!
  - Tests probing only the tip of the iceberg, population with severest symptoms
- Is 0.3-1.2% CFR Iceland (95% cl) in conflict with our IFR?
  - 10 deaths, 1770 infections. CFR still growing
  - Has younger population, we predict 0.5% IFR lower limit, 0.6% mean IFR.

# What about other regions?

- Model: **IFR mean (lower bound, upper bound) is 1 (0.8, 1.8) of Yearly Mortality Rate**

- New York City
  - YMR = 0.62%, so IFR lower bound = 0.5%
  - PFR = 0.16% (as of today)
  - IR = PFR/IFR = 25% (14%-30% 95%cl), 1.5% tested positive so far, ratio 10-20

- Age distribution of deaths? Why are people dying in NYC younger than in Italy?
  - YMR fraction of deaths in NYC below age of 65 (70) is 26% (34%)
  - Reported NYC COVID-19 fraction of deaths below age of 65 (70) is 23% (32%)
  - YMR fraction of deaths in Lombardy below age of 65 (70) is 10% (15%), same as COVID-19: dying population is older
  - **COVID-19 mortality tracks overall mortality (YMR a good proxy)**
  - a quick estimate what is the chance of dying if you get infected: same as the chance of dying next 12 months

# What about other regions?

- Model: **IFR mean (lower bound, upper bound) is 1 (0.8, 1.8) of Yearly Mortality Rate**

- Santa Clara: 0.5% IFR lower bound, 0.6% mean, 1.1% upper bound.
  - Strongly inconsistent with BenDavid et. al. “Stanford seropositivity study” claiming IFR = 0.12-0.2%
  - Issues with their statistical analysis: observed  IR 1.5%, post-stratification gives them 2.5-4.2% IR, specificity of the seropositive test 0.5-1.5% and so lower errors should be consistent with 0% IR

- Los Angeles county: Assume 0.6% IFR.
  - Observed deaths = 663/8M population, still growing and likely underestimated
  - Assume ~1000 deaths: this gives 2% IR
  - Recent seropositivity study by some of the same people finds 2.8-5.6%, consistent with 2% after specificity correction

# Summary

- Italy official COVID-19 mortality underestimated, counterfactual analysis possibly more reliable
- Infection Fatality Rates 0.5%-1% (contradicts recent claims of 0.1%)
- High Infection Rates in worst hit regions of Lombardy and Bergamo
- This explains the large differences between CFR (20% in Lombardia) and IFR (1%): many more infected than tested
- Steeper IFR age dependence than previously assumed
- A simple way to compute IFR is to use Yearly Fatality Rate at a given location (works also for age dependence)
- COVID-19 tracks overall mortality, kills the weakest members of population

# Thank you