COS-D407. Scientific Modeling and Model Validation

Lecturer: Christina Bohk-Ewald

Week 3

University of Helsinki, Finland
26.10.2020–09.12.2020

Third week's class:

Scientific modeling and model validation in practice

- Q & A: recap of material of previous session
- Present your findings of previous lab session
- Demographic scaling model for estimating total numbers of COVID-19 infections
- Validity of demographic scaling model's infection estimates

Third week's class in the lab: Apply demographic scaling model and critically think about validity of its results.

- Apply demographic scaling model: estimate total numbers of COVID-19 infections for Finland.
  Extra: You could do this again for another country of your choice.
- Critically think about the key assumptions of this model and their implications for the results.

$\rightarrow$ Present and discuss your findings in class at the beginning of the next session on Monday.

Brief Q&A: recap material of previous session:

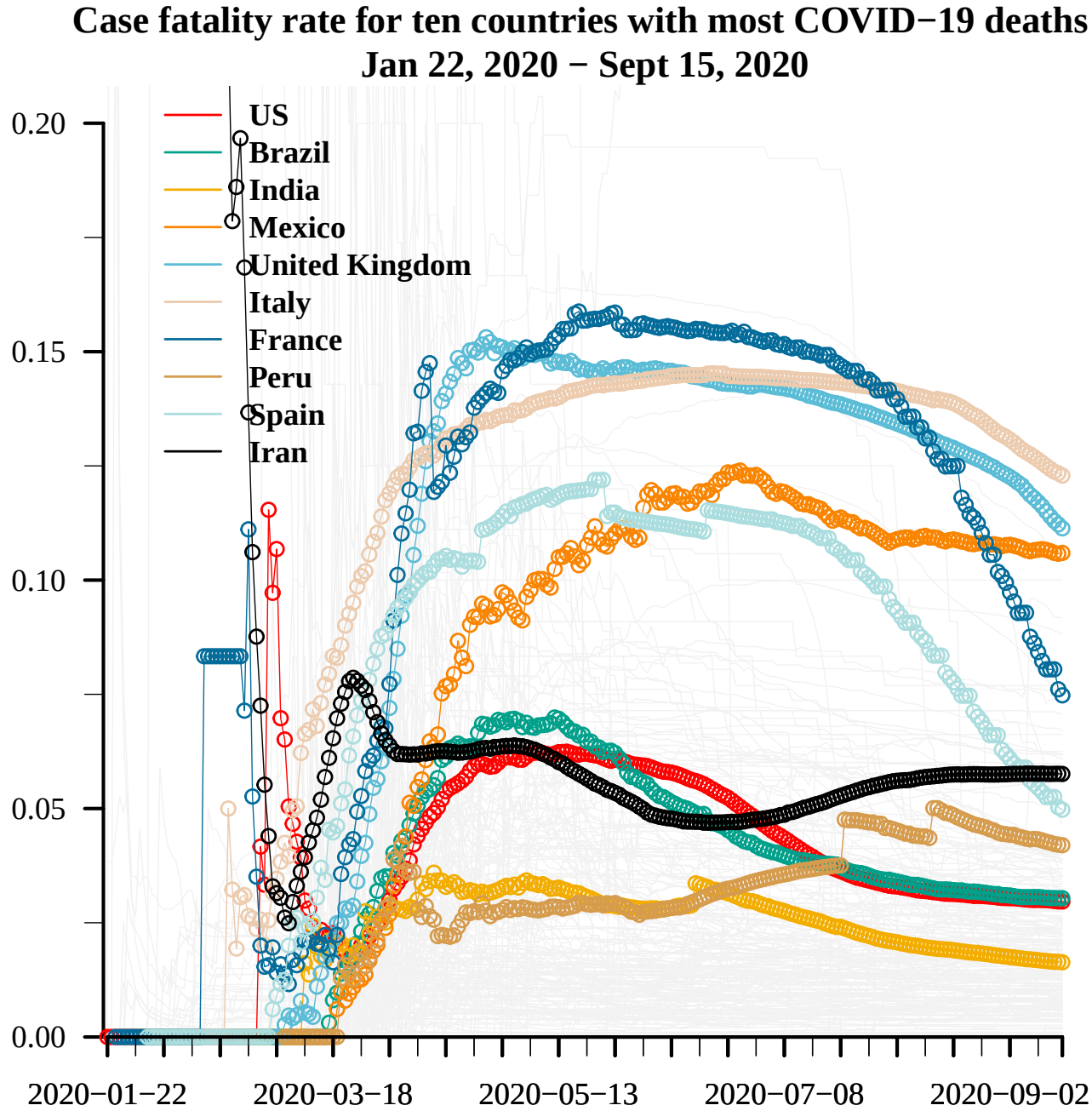
- What are the main steps of the scientific method?
- Why can the scientific method be at least as important as the scientific finding?
- Why is it so important that an explanation for a phenomenon is testable or falsifiable?
- Why is it important to re-test a model with high predictive power for a phenomenon?
- What are common pitfalls during scientific work?

$\rightarrow$ Open questions?

Present your findings of previous lab session:

- What are recent trends in the case fatality rate (CFR) of COVID-19 in the ten countries with most COVID-19 deaths as of last week?
- What are possible explanations for cross-country differences in CFR?
- Have you done this exercise with another scientific finding of your choice?

$\rightarrow$ Open issues?



Data source: JHU CSSE: https://data.humdata.org/dataset/novel-coronavirus-2019-ncov-cases

Cross-country variation in CFR may be due to

- Real differences in mortality attributable to COVID-19
- Age structure of population and, consequently, of deaths and cases attributable to COVID-19
- Stage of progress of COVID-19 outbreak in each country
- Classification of COVID-19 deaths
- Testing practices wrt to test coverage and test specificity
- Capacity and occupancy rate of health-care systems for intensive care
- Effectiveness of control measures to curb severe COVID-19 infections
- ...

$\Rightarrow$ Confirmed cases and reported deaths may be biased

Confirmed cases may underestimate the number of infections

- Cases with mild symptoms or asymptomatic cases might go undetected
- Test coverage
  - Test kits may not be available in large numbers
  - Focus on sub-populations, e.g., cases with proven contact to other COVID-19 cases or hospitalized cases
- False negatives
  - People get tested after the first week of infection, when it is likely that SARS-CoV-2 cannot be detected in pharynx anymore (PCR)
  - Test for antibodies could be be carried out before a body has had a chance to produce them
- ...

Reported deaths may be biased

- Reporting delays may amount to several days.
- Inconsistent practices for classifying COVID-19 deaths within and between countries. For example, only deceased individuals who (1) were hospitalizd or (2) died from COVID-19 as primary and / or secondary cause of death may be counted.
- Test coverage and test specificity may be insufficient. For example, not all deaths are tested for COVID-19. Persons dying at home or in other institutions may not be counted.

A demographic scaling model

for estimating

the total number of COVID-19 infections

Christina Bohk-Ewald, Christian Dudel, and Mikko Myrskylä

Accepted for publication in the International Journal of Epidemiology

Our key question...

How many people have been infected with COVID-19?

...is important but barely answered yet

- Existing seroprevalence studies for COVID-19 have largely relied on samples that are not representative for the total population, and population representative studies are only slowly becoming available.

- Existing approaches to estimate the spread of COVID-19 rely on complex statistical methods that typically have high data demands.

That is why...

We develop and implement the demographic scaling model
for estimating COVID-19 infections
with minimal data requirements,
so that it is broadly applicable
in contexts with both rich and poor data.

The demographic scaling model

- $I_x = P_x \cdot \lambda_x$ (1)

  We want to estimate the number of infections I, which are a fraction λ of the total population size P in each age group x. P is known, I and λ are both unknown.

- Knowing that the infection fatality rate $IFR_x = \frac{D_x}{I_x}$,
  we modify equation (1):

  $D_x = IFR_x \cdot P_x \cdot \lambda_x$ (2)

  and estimate the unknown infection prevalence λ
  with the known number of deaths D, scaled infection fatality rates IFR, and population counts P in each age group x

- $\lambda_x = \frac{D_x}{IFR_x \cdot P_x}$ (3)

The demographic scaling model, ctnd

- We finally estimate the total number of infections as the sum of the population counts P multiplied with the infection prevalence λ over all ages x:

$$I = \sum_x P_x \cdot \lambda_x \quad (4)$$

- Inserting the definition of λ of equation 3 also yields:

$$I = \sum_x \frac{D_x}{IFR_x} \quad (5)$$

$\rightarrow$ The key challenge is to arrive at credible estimates of IFR_x and D_x

How to get credible estimates of IFR_x

The basic model is a mix of statistical modeling and epidemiology.
In order to arrive at credible estimates for infection fatality rates,
we now add a touch of demography :-)

How to get credible estimates of IFR_x

Given that infection fatality rates (IFRs) are rarely available for any country, the question is:

How can we make use of IFR estimates for, e.g., China when we want to estimate the COVID-19 infections for, e.g., Italy?

...considering that countries differ in their vulnerability to COVID-19 due to substantial differences in age structure, health conditions, and medical services.

How to get credible estimates of IFR_x

- If we take Chinese IFRs, China is our *reference country* (RC).
- And if we estimate COVID-19 infections for Italy, Italy is our *country of interest* (COI).
- To account for cross-country differences in the age structure, health conditions, and medical services between the RC and COI, we not only borrow but also scale IFRs from the RC onto the COI.
- This scaling is based on remaining lifetime e_x:

  $$IFR_{e_x}^{COI} = IFR_{e_x}^{RC}$$

  which is a parameter of a life table, and life tables are readily available for many countries.

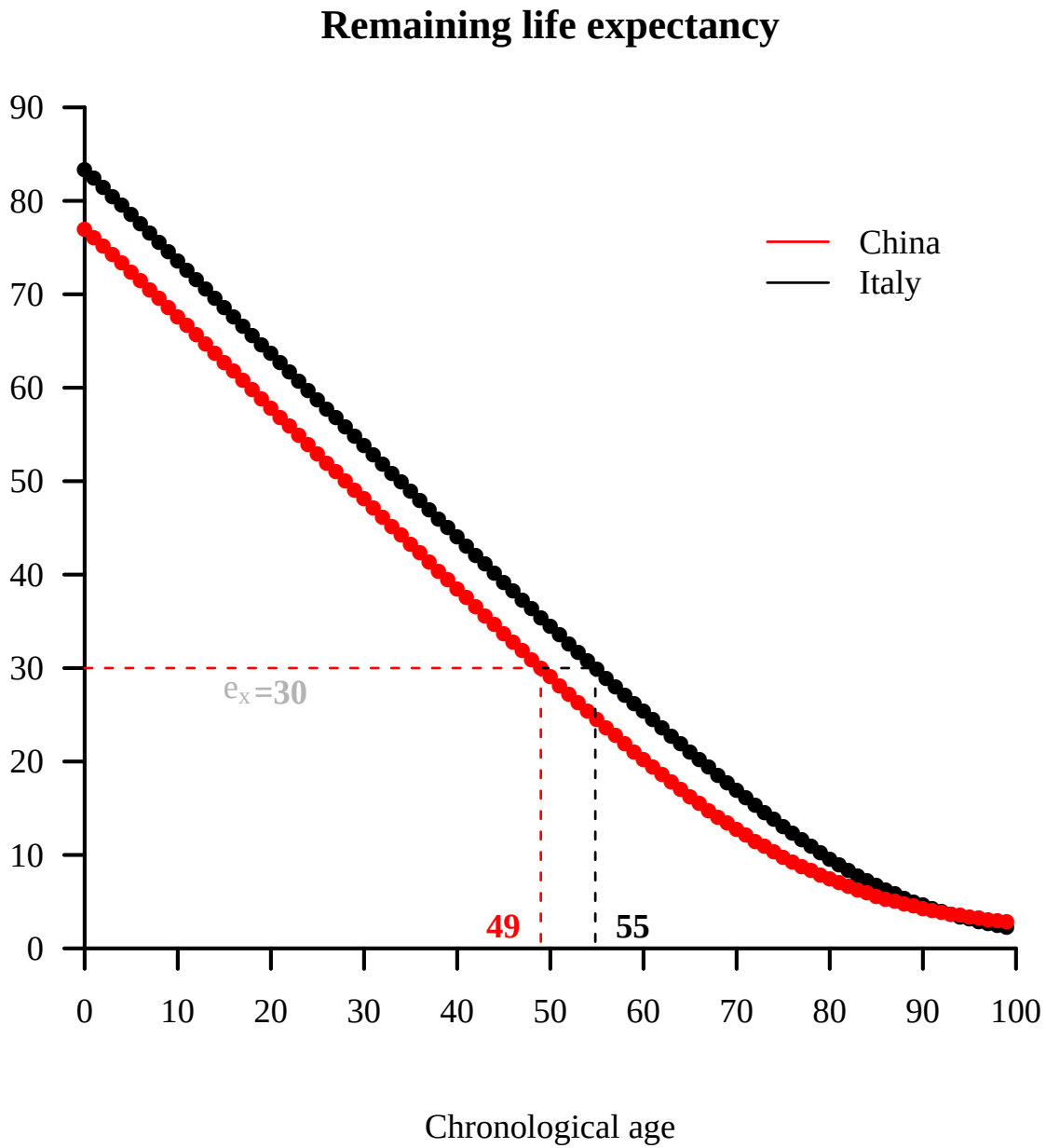
How to get credible estimates of IFR_x, ctnd

- $IFR_{e_x}^{COI} = IFR_{e_x}^{RC}$

  basically says that we assign the same infection fatality rate (IFR) to people in the RC and COI who have, on average, the same number of life years left e_x.

  For example, if 49-year-olds in a RC have, on average, the same number of life years left as 55-year-olds in a COI, we assign the infection fatality rate of the 49-year-olds in the RC to the 55-year-olds in the COI.

How to get credible estimates of IFR_x, ctnd



- e_x is remaining lifetime in years
- x is chronological age in years
- $IFR^{China}_{e_x=30} = IFR^{Italy}_{e_x=30}$

$\Downarrow$

- $IFR^{China}_{x=49} = IFR^{Italy}_{x=55}$

How to get credible estimates of D_x

- Total death counts are available for many countries on a daily basis from Johns Hopkins University CSSE.
- We disaggregate total deaths into age groups using a global average pattern over age (Dudel et al. (2020)).

How does this look like in practice?

We apply the demographic scaling model
for estimating COVID-19 infections
for the ten countries
that have reported most COVID-19-related deaths
as of September 21, 2020.

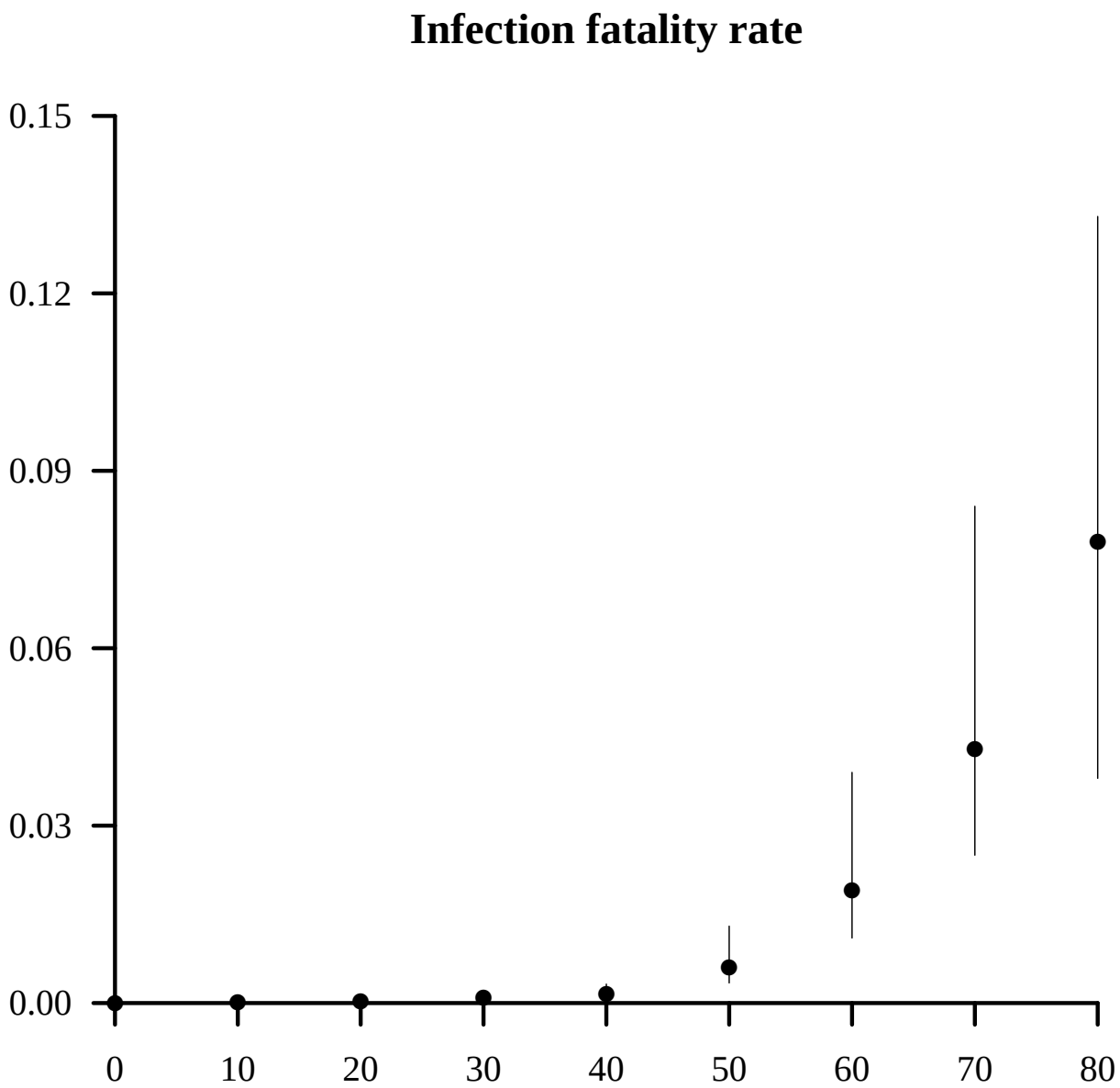
How does this look like in practice?

We start with the input data.

How does this look like in practice?

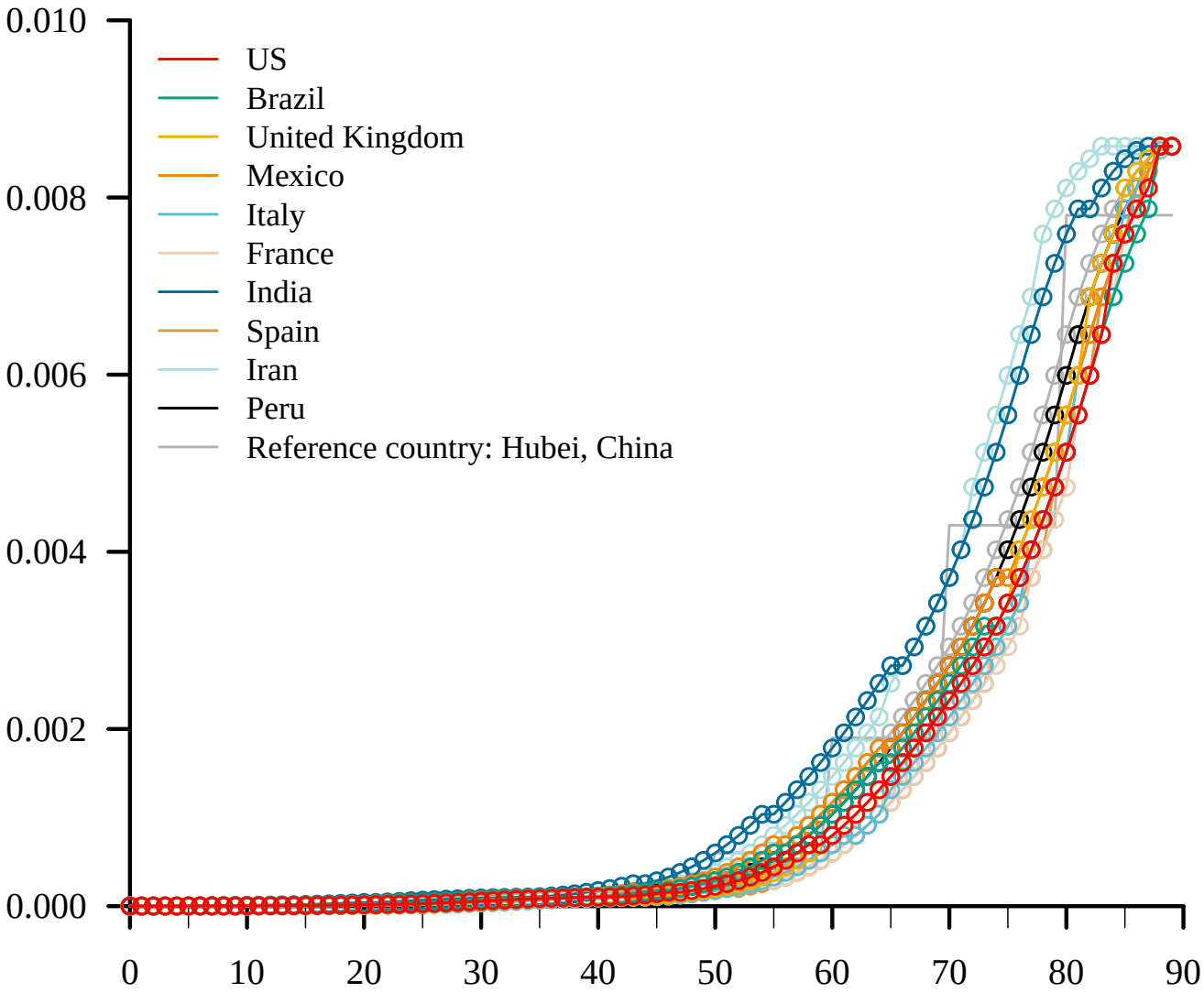
As input we take:

- IFRs of Hubei, China, as reported in Verity et al. (2020), as reference
- Deaths attributable to COVID-19 on a daily basis povided by JHU CSSE (2020)
- Population counts and life tables from UNWPP (2019)



Data source: Verity et al. (2020)

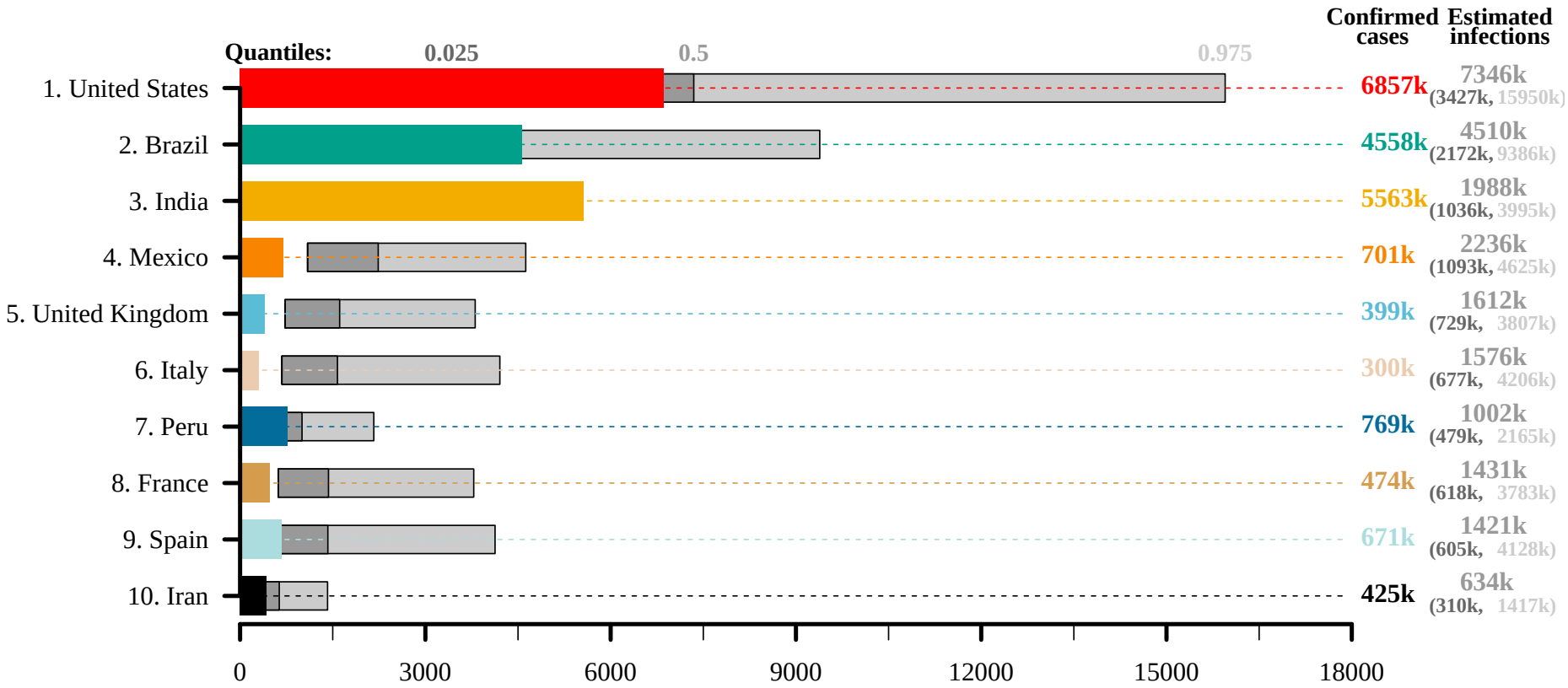
Scaled infection fatality rates based on remaining life expectancy
Reference country: Hubei, China



How does this look like in practice?

We now continue with the estimated COVID-19 infections
for the ten countries
that have reported most COVID-19-related deaths
as of September 21, 2020.

Results: total number of COVID-19 infections

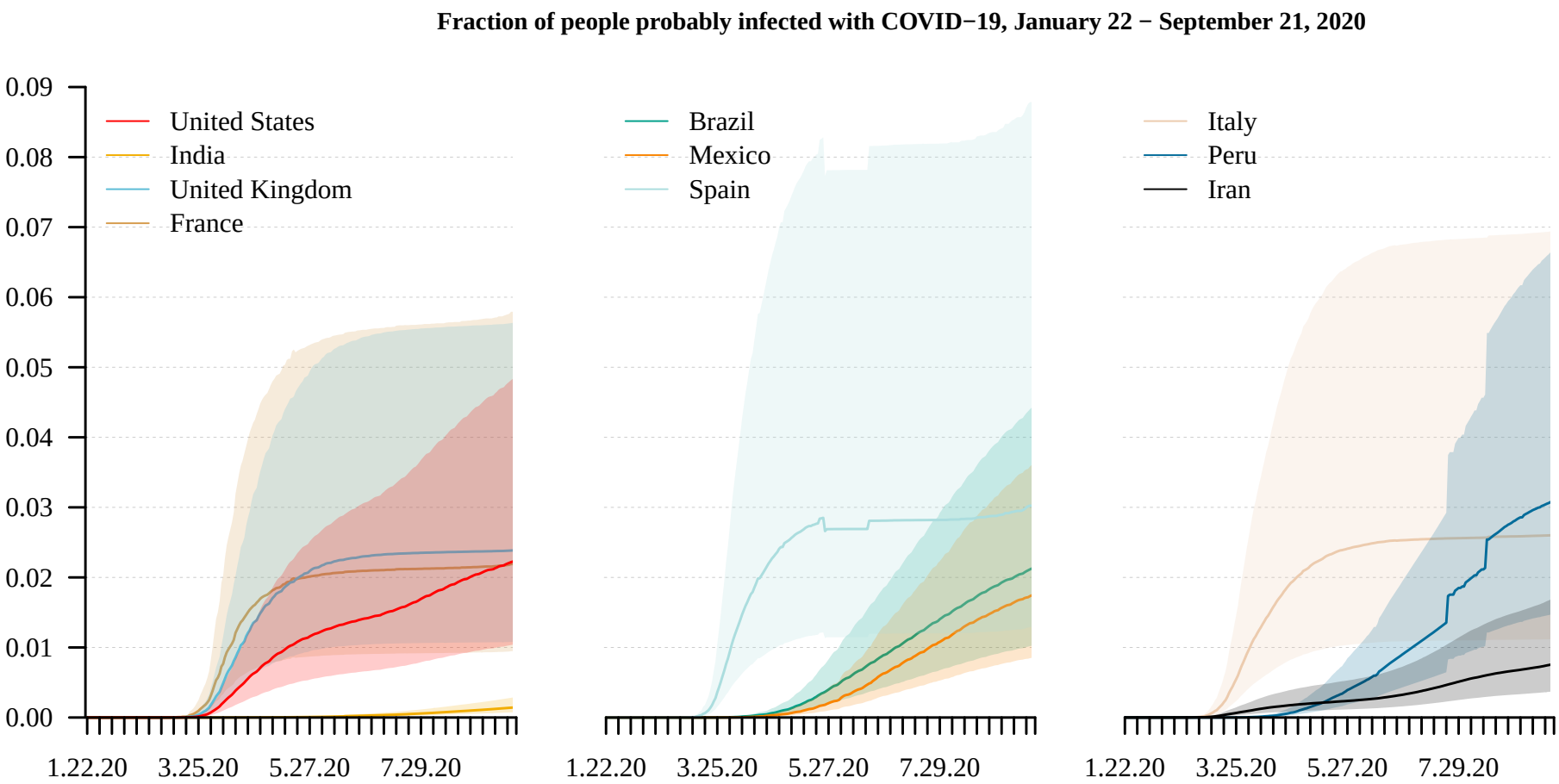


Results: total number of COVID-19 infections

Across the 10 countries with most COVID-19 deaths as of September 21, 2020:

- The total number of infected individuals is more than twice as large as the number of confirmed cases.
- Uncertainty of findings is high. 95% prediction interval: 1 to 6 times as many infections as confirmed cases.
- Infection estimates are uncertain and vary across countries.

Results: COVID-19 prevalence



Results: COVID-19 prevalence

Across the 10 countries with most COVID-19 deaths as of September 21, 2020

- COVID-19 prevalence increases with time.
- Variation across countries is high.
- Central estimate is on average 2%. It ranges between 3.1% in Spain and Peru and 0.2% in India.
- Uncertainty of findings is large. Upper bound includes values as high as 9% for Spain and 7% for Italy.

How accurate are the COVID-19 infection estimates
of
the demographic scaling model?

Key assumptions of demographic scaling model

1. COVID-19-related death counts are fairly accurately recorded.

2. Infection fatality rates from reference country are fairly accurately recorded and
become applicable in a country of interest through proper scaling based on remaining life expectancy.

Key assumptions of demographic scaling model

The two key assumptions may only partially hold at the moment.

However, as soon as better input data will become available,
the demographic scaling model can account for them,
and its COVID-19 infection estimates
are likely to become more accurate.

Key messages

To wrap things up...

Key messages

- The demographic scaling model:
  - is broadly applicable in contexts with both rich and poor data.
  - facilitates the timely monitoring of the spread of the COVID-19 pandemic.
  - allows to estimate the total number and prevalence of COVID-19 infections.
- The infection estimates for the sample of 10 countries indicate that the coronavirus pandemic is more widespread than the numbers of confirmed cases suggest.

What you have learned today about the scientific method in general

- Describe main methodological steps of the demographic scaling model.
- Describe COVID-19 infection estimates for the ten countries with most COVID-19 deaths as of September 21, 2020.
- Describe the temporal development of the COVID-19 infection estimates for the ten countries with most COVID-19 deaths since January 22, 2020.
- Describe the two key assumptions of the demographic scaling model.

Course learning materials

Course learning materials on GitHub:

`https://github.com/christina-bohk-ewald/2020-COS-D407-scientific-modeling-and-model-validation`

Recommended learning material for today's class

- **Bohk-Ewald et al. (to appear)**
  A demographic scaling model for estimating the total number of COVID-19 infections. International Journal of Epidemiology.
  Preprint available on medRxiv:
  `https://doi.org/10.1101/2020.04.23.20077719`
- **Richard P Feynman (1974)**
  Cargo Cult Science. Some remarks on science, pseudoscience, and learning how to not fool yourself.
  Caltech's 1974 commencement address.
  `http://calteches.library.caltech.edu/51/2/CargoCult.htm`
- **Carl Sagan (1997)**
  The Demon-Haunted World: Science as a Candle in the Dark.
  Ballantine Books.

Thank you for your attention!

christina.bohk-ewald@helsinki.fi

Third week's class in the lab: Apply demographic scaling model and critically think about validity of its results.

- Apply demographic scaling model: estimate total numbers of COVID-19 infections for Finland.
  Extra: You could do this again for another country of your choice.
- Critically think about the key assumptions of this model and their implications for the results.

$\rightarrow$ Present and discuss your findings in class at the beginning of the next session on Monday.