COS-D407. Scientific Modeling and Model Validation

Lecturer: Christina Bohk-Ewald

Week 4

University of Helsinki, Finland
01.11.2021–15.12.2021

Fourth 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
- Model validation continued. Case without true realizations.
  - Introduction to the *demographic scaling model* for estimating total numbers of COVID-19 infections.
  - How to validate the infection estimates of the *demographic scaling model* when the true number of infections is unknown?

Fourth 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.

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

Brief Q&A: recap material of previous session:

- How has Finnish mortality developed in the last 150 years?
- What is the general procedure of ex-post forecast validation?
- What kind of forecast error measures do you know of? In what cases are they particularely useful?
- What have you learned from validating the Lee-Carter mortality forecasts for Finland and Italy?

→ Open questions?

Present your findings of previous lab session:

- What are the most important findings for you of the last lab session?
- Do you have some findings / insights from any of the additional exercises?
  - Results for another country?
  - Results for men?
  - Results based on other base periods and / or forecast horizons?
  - ...

Present your findings of previous lab session:

Ricarda has also provided another pdf file for you that you can find on GitHub:

- File name: “example-solutions-for-additional-hands-on-exercises.pdf”.
- It provides one (of many) possible ways to approach the additional exercises of week 3 in R.
- Please note that it does not provide interpretations of results. That is up to you.

Model validation

In the context of demographic forecasting, comparing the model output to its *true realizations* via ex post errors and cross validation is useful to, e.g.,

1. Decide if a method of interest is suitable to forecast mortality in a particular country. Example of Duerst and Bohk-Ewald (2020). Course week 3.
2. Select the most suitable method to forecast parameter of interest from a large basket of possible methods. Example in the context of cohort fertility forecasting. If you are interested, please have a look into this paper:

Bohk-Ewald et al. (2018)
Forecast accuracy hardly improves with method complexity when completing cohort fertility. PNAS 115(37), 9187–9192.
DOI: `https://doi.org/10.1073/pnas.1722364115`

Model validation, continued

In this course, you will be introduced to two basic ways for evaluating a model and the validity of its output.

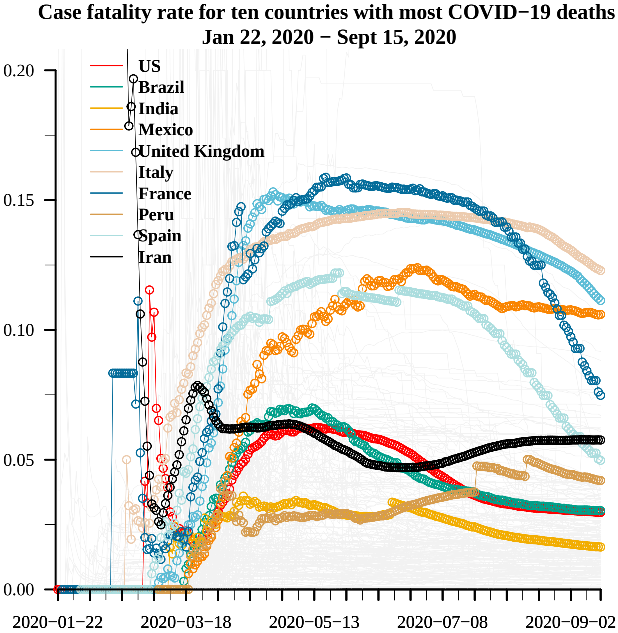
1. Comparing the model output to its *true realizations* using, e.g., ex post errors and cross validation. Week 3.
2. Analyzing the *process* that a model uses to generate its output and, if possible, analyzing the *sensitivity of the results* with respect to, e.g., using input data from different sources. This can also be done if true realizations are unknown. **Weeks 4 & 5.**

$$\text{Input parameters} \Rightarrow \text{Model} \Rightarrow \text{Outcome variables}$$

Model validation, continued

Case without true realizations.

- Introduction to the *demographic scaling model* for estimating total numbers of COVID-19 infections.
- How to validate the infection estimates of the *demographic scaling model* when the true number of infections is unknown?



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

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:

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

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

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

→ 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.

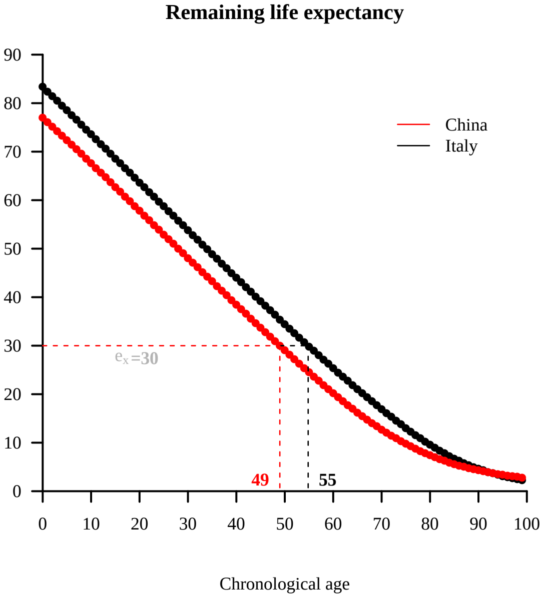
How to get credible estimates of IFR_x, ctnd

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

  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

We take death counts attributable to COVID-19 by age from the COVerAGE-DB (Riffe et al. 2020).

Original method described in paper:

- 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 using data of Dudel et al. (2020).

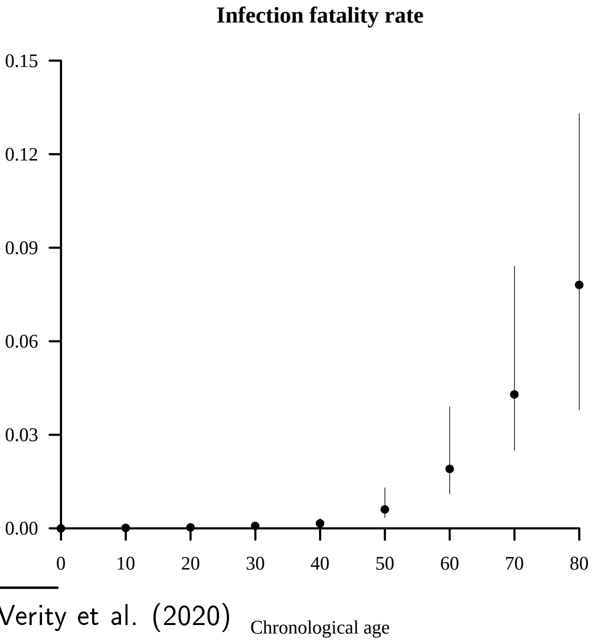
How does this look like in practice?

We apply the demographic scaling model
for estimating COVID-19 infections
in Finland from March 2020 to April 2021.

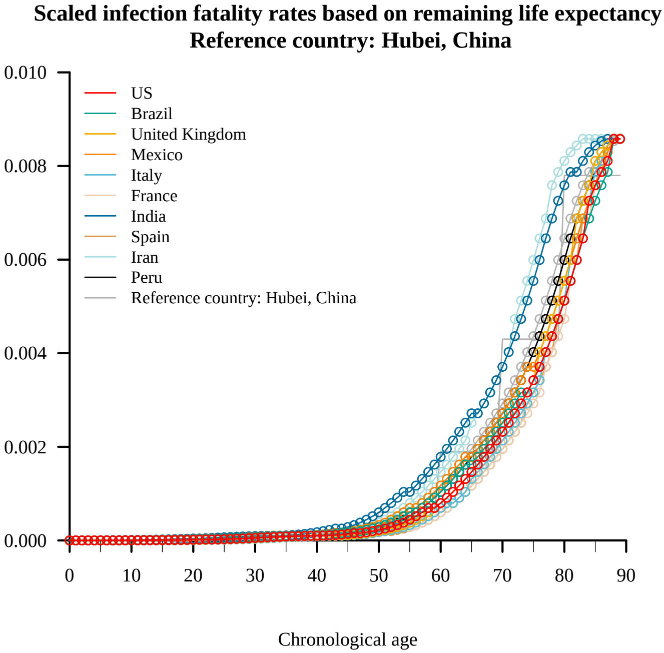
How does this look like in practice?

As input we take:

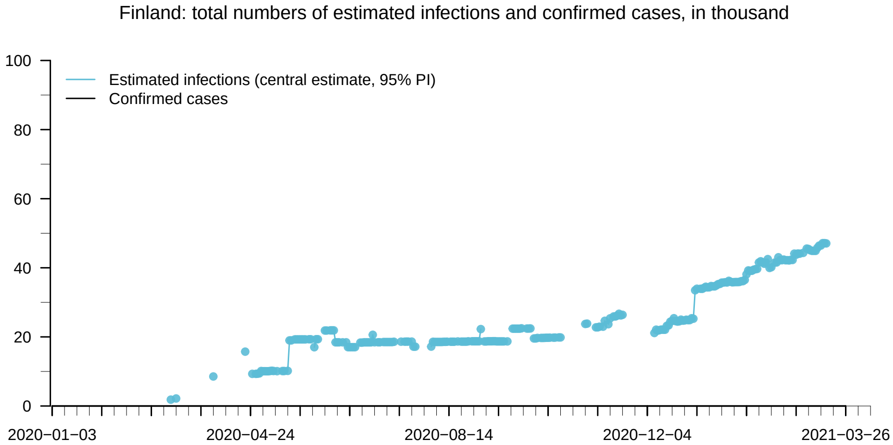
- Reference IFRs of Hubei, China, as reported in Verity et al. (2020), that we scale to Finland.
- Deaths by age attributable to COVID-19 on a daily basis, as reported to the COVerAGE-DB (version 14 April 2021; Riffe et al. 2020).
- Population counts and life tables from UNWPP (2019).



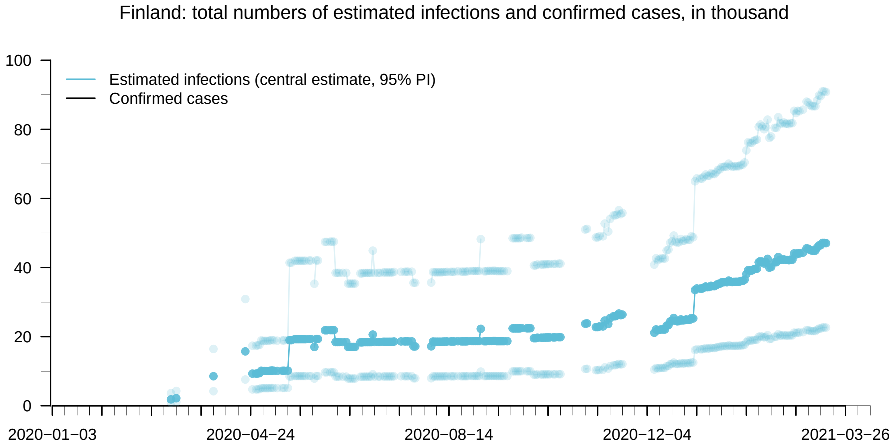
Data source: Verity et al. (2020)


Scaled infection fatality rates based on remaining life expectancy
Reference country: Hubei, China
0.010
0.008
0.006
0.004
0.002
0.000
0
10
20
30
40
50
60
70
80
90
Chronological age
US
Brazil
United Kingdom
Mexico
Italy
France
India
Spain
Iran
Peru
Reference country: Hubei, China

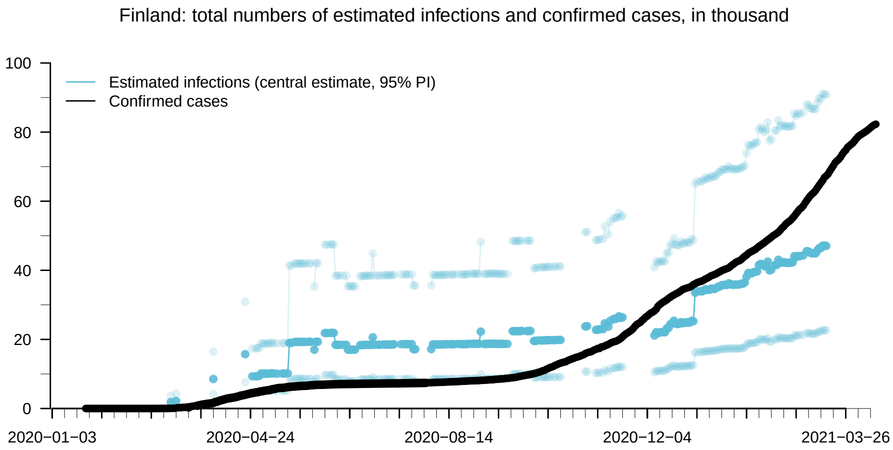
Results: total number of COVID-19 infections in Finland



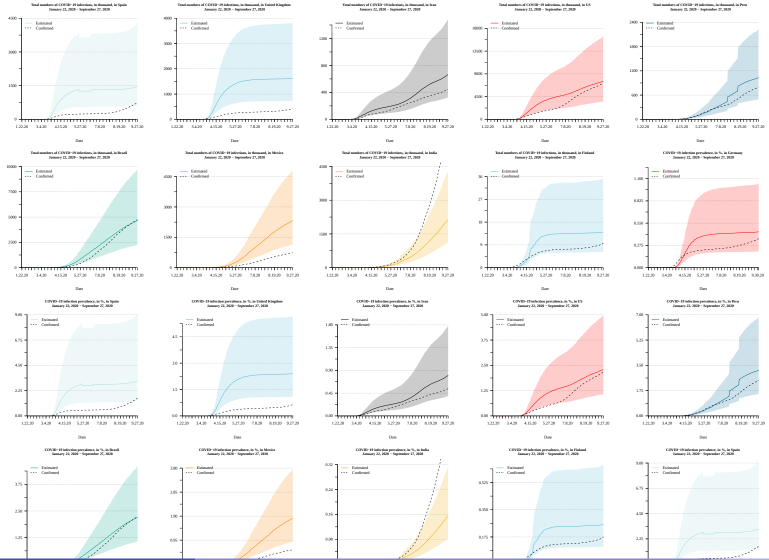
Results: total number of COVID-19 infections in Finland



Results: total number of COVID-19 infections in Finland



Total number of COVID-19 infections (and infection prevalence) can be estimated for many countries



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.

What you have learned today

- Describe main methodological steps of the demographic scaling model.
- Describe the level and temporal development of the COVID-19 infection estimates in Finland 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/2021-COS-D407-scientific-modeling-and-model-validation

Recommended learning material for today's class

- **Bohk-Ewald et al. (2020)**
  A demographic scaling model for estimating the total number of COVID-19 infections. International Journal of Epidemiology 49(6), 1963–1971. DOI: `https://doi.org/10.1093/ije/dyaa198`

Thank you for your attention!

`christina.bohk-ewald@helsinki.fi`

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

- Apply the demographic scaling model to estimate total numbers of COVID-19 infections in Finland since January 2020.
  Extra: You could do this again for another country of your choice.
- How would you evaluate the process of the demographic scaling model? How plausible are the infection estimates, also considering, e.g., the quality of the input data and the key assumptions of the demographic scaling model?

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