# The importance of uncertainty in contact patterns for quantifying risk

## Introduction

Age is a major risk factor for severe disease due to COVID-19 infection, so accounting for the contact patterns of different age groups is particularly important for quantifying the risk faced by each. Age-structured compartmental models, such as the `SEIRDAge` model in the `comomodels` package, use contact matrices to characterise these age-specific patterns.

`comomodels` includes estimates of contact matrices for 152 countries (Prem et al. (2021); with details available in the data documentation). For each country, a contact matrix is provided for four different locations where people may mix with others: at home, at school, at work, and elsewhere. For a contact matrix, $C$, each element, $C_{i,j}$, indicates the expected number of contacts someone from age group $i$ has per day with people from age group $j$, which is given by:

$$C_{i,j} = \frac{\text{total \# contacts between } i \text{ and } j}{\text{size of group } i}. \tag{1}$$

Because $C_{j,i}$ has the size of group $j$ as its denominator, typically $C_{i,j} \neq C_{j,i}$ due to demographic patterns meaning contact matrices are not typically symmetric. Since the age-demographics of a population affect its contact matrices, a contact matrix estimated for a given country should not be repurposed for another without due care (Arregui et al. 2018).

Contact matrices are constructed from survey data, where participants in the study record their contacts throughout the day and include information on the age and location of the contact.

In this tutorial, we show how contact matrices can be used within `comomodels` to simulate infection dynamics using the `SEIRDAge` model. We then show how the considerable uncertainty inherent in a particular, and commonly used, set of contact matrix estimates generates commensurate uncertainty in key model outputs for a model parameterised for the United Kingdom (UK).

## Contact patterns for the UK

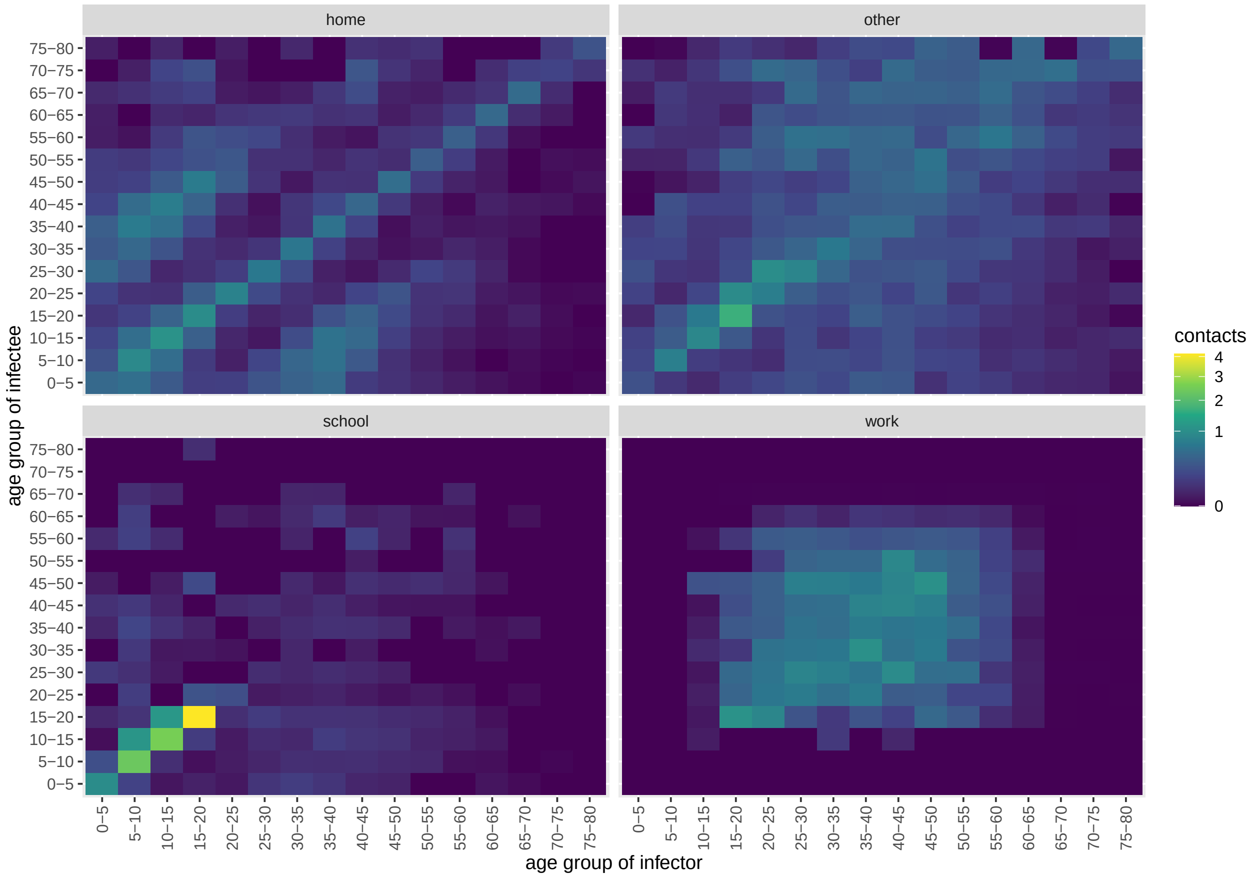
We first visualise the set of contact matrices for the UK. We load all contact matrices for each location (home, school, work and other) and the population demographics of each country which are included in `comomodels`.

```r
library(comomodels)
library(tidyverse)
library(ggplot2)
library(socialmixr)

contact_home <- comomodels::contact_home
contact_work <- comomodels::contact_work
contact_school <- comomodels::contact_school
contact_other <- comomodels::contact_other
population <- comomodels::population
```

We then select a specific country for which we want to visualize the contact matrices, and plot all four contact matrices for this country. In the contact matrices provided by Prem et al. (2021), the oldest age group is for

75-80 year olds; in what follows, we assume that the contact patterns are the same for individuals aged 80+. This is likely a strong assumption, since it neglects the change of circumstances that may occur for many in this age group. But the change in circumstances may act either to increase or to decrease contact rates, so assuming contact patterns remain the same is a reasonable null hypothesis.

```r
# reformat matrices for plotting
ages <- seq(0, 120, 5)
age_names <- vector(length = 16)
for(i in seq_along(age_names)) {
  age_names[i] <- paste0(ages[i], "-", ages[i + 1])
}

format_matrix <- function(contact_matrix, age_names) {
  colnames(contact_matrix) <- age_names
  contact_matrix$age_infectee <- age_names
  contact_matrix %>%
    pivot_longer(all_of(age_names)) %>%
    rename(age_infector=name) %>%
    mutate(age_infector=fct_relevel(age_infector, age_names)) %>%
    mutate(age_infectee=fct_relevel(age_infectee, age_names))
}

c_home <- format_matrix(contact_home$"United Kingdom", age_names) %>%
  mutate(type="home")
c_work <- format_matrix(contact_work$"United Kingdom", age_names) %>%
  mutate(type="work")
c_school <- format_matrix(contact_school$"United Kingdom", age_names) %>%
  mutate(type="school")
c_other <- format_matrix(contact_other$"United Kingdom", age_names) %>%
  mutate(type="other")

c_all <- c_home %>%
  bind_rows(c_work) %>%
  bind_rows(c_school) %>%
  bind_rows(c_other)

# plot all
c_all %>%
  ggplot(aes(x=age_infector, y=age_infectee, fill=value)) + geom_tile() +
  facet_wrap(~type) +
  theme(axis.text.x = element_text(angle=90, vjust=0.5, hjust=1)) +
  scale_fill_viridis_c("contacts",
                       trans="sqrt") +
  xlab("age group of infector") +
  ylab("age group of infectee")
```



These matrices illustrate rich contact patterns for the UK, which are markedly different between locations. At school, unsurprisingly, students mix with many others of similar ages. At home, there is considerable intergenerational mixing. At work, there is more uniform mixing – by vast majority between people of working age. These suggest a common transmission pattern, in which schoolchildren, who have the most daily contacts, infect one another. They then pass infection onto their parents at home, who then pass their infection onto work colleagues. A cluster-controlled trial in the US where "intervention schools" were offered live attenuated flu vaccine with comparable control schools offered no vaccine had significantly fewer influenza-like symptoms in follow up (King Jr et al. 2006). This hints that offering vaccines to schoolchildren may be effective at reducing transmission of COVID-19 (Walter et al. 2022), although other strategies including regular COVID-19 testing of students have been advocated (Asanati, Voden, and Majeed 2021).

The availability of contact matrices for different locations within the package facilitates simulation of age-structured transmission models across geographies. The current set of age-structured models in `comomodels` does not explicitly include location-specific infections. Thus, we obtain a single contact matrix for the age-structured models by summing the four location-specific contact matrices:

$$C_{i,j} = C^{\text{home}}_{i,j} + C^{\text{school}}_{i,j} + C^{\text{work}}_{i,j} + C^{\text{other}}_{i,j}.$$

It is possible, however, to investigate how changes to contacts occurring at a particular location affect outputs: by perturbing a particular contact matrix in the above sum, then running the model using the new overall contact matrix.

## Uncertainty in the contact matrix

When performing forward simulations of the age-structured SEIRD model or when fitting these models to data, the contact matrix is typically provided as a fixed input. But using point estimates for the contact matrix neglects the considerable uncertainty inherent in them.

The purpose of the remainder of this notebook is to investigate the sensitivity of the outputs of the age-structured SEIRD model to the uncertainty in contact matrix estimates. To do so, we use bootstrapped samples of the contact matrix to represent this uncertainty. The bootstrap algorithm works by selecting a random sample (with replacement) of the survey respondents, which it then uses to construct a contact matrix. Across many such samples, the set of contact matrices provides a measure of uncertainty in the number of daily contacts across different age groups.

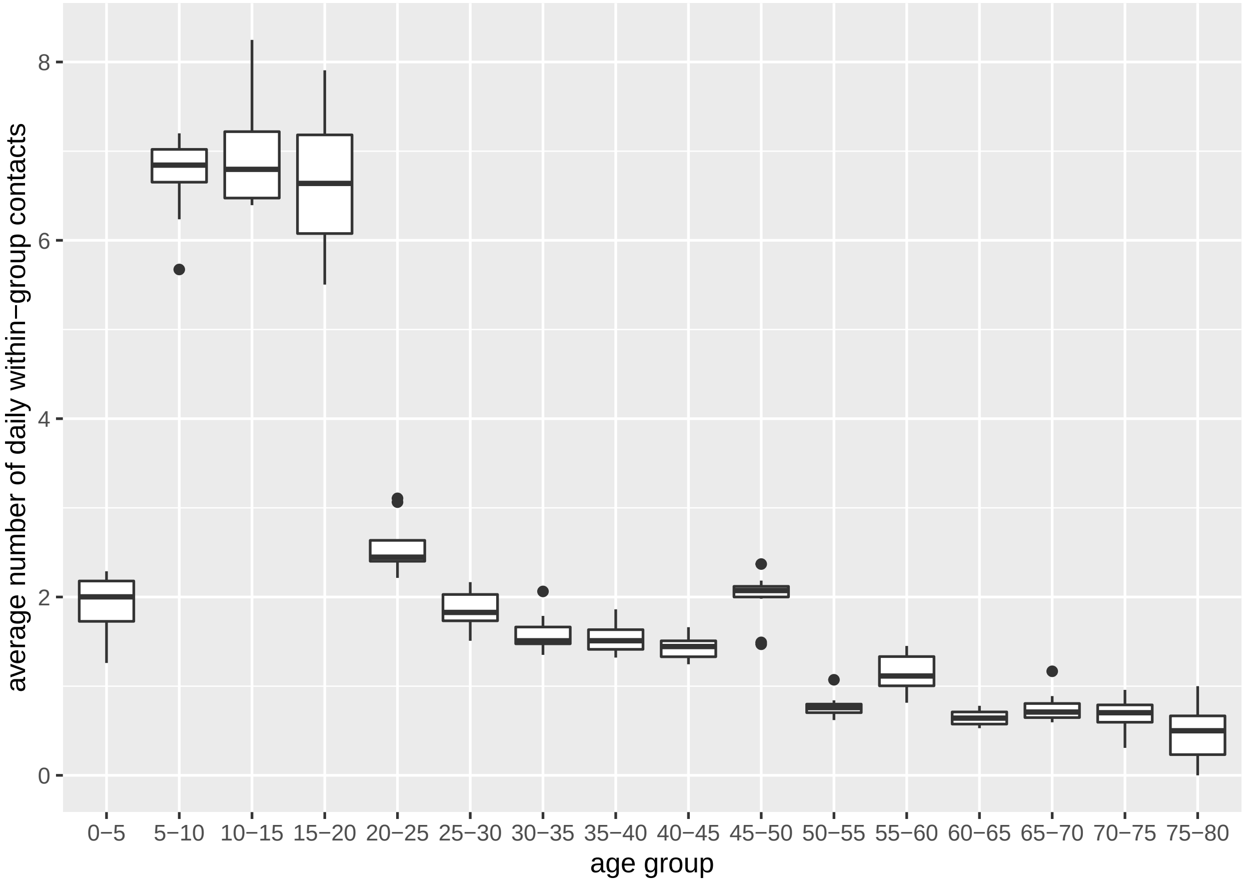
There are a range of factors which the bootstrapped approach to uncertainty quantification does not consider, however. The algorithm assumes that the original survey from which the contact matrices are calculated is representative of the underlying population, which may not be true. For example, if contact data are collected primarily from an urban area in a country whose population is mostly rural, the resulting contact matrices would likely be unrepresentative of the country as a whole. (Indeed, if the two contact matrices differ so markedly, it may be preferable to model the urban and rural populations separately, as in the `SEIRD_RU` model.) The bootstrap algorithm does not allow for such biases in quantifying uncertainty, so likely understates true population-level uncertainty.

We obtain bootstrapped samples of the contact matrix using the `socialmixr` library which accesses the POLYMOD data (Funk 2020; Mossong et al. 2017). We base our analysis on the UK and generate 200 contact matrix draws to represent its uncertainty.

```
# Define age groups and names
ages <- seq(0, 120, 5)
age_names <- vector(length = 16)
for(i in seq_along(age_names)) {
  age_names[i] <- paste0(ages[i], "-", ages[i + 1])
}

# Get population data: merge ages 75+
pops <- population[population$country == "United Kingdom", ]$pop
pop_fraction <- pops / sum(pops)
pop_fraction[16] <- sum(pop_fraction[16:21])
pop_fraction <- pop_fraction[1:16]
n_ages <- 16

# Load the contact matrix data from POLYMOD and get bootstrap samples
n_bootstrap <- 10
data(polymod)
polymod_data <- contact_matrix(polymod,
                               n=n_bootstrap,
                               countries="United Kingdom",
                               age.limits=ages)

# Get the first element of the list, which contains the matrices
# In general, bootstrap sampling fails for the 80+ age group,
# so we assume contact patterns remain same as for 75-80 y.o.s
matrices <- polymod_data["matrices"][[1]]
```

First, we inspect the range of values in the sampled contact matrices. In the plot below, we show the variation in within-age-group daily contacts: i.e. we plot the samples of the diagonal elements of the contact matrix.

```
contacts_same_age <- c()
ages_list <- c()
for (i in 1:n_bootstrap){
  diags <- diag(matrices[[i]][[1]])[1:16]
  contacts_same_age <- append(contacts_same_age,
                              diags)
  ages_list <- append(ages_list, age_names)
}

data <- data.frame(ages_list, contacts_same_age)
data$ages_list <- factor(data$ages_list, levels=age_names[1:16], ordered=TRUE)

ggplot(data, aes(x=ages_list, y=contacts_same_age)) +
  geom_boxplot() +
  xlab("age group") +
  ylab("average number of daily within-group contacts")
```



The plot shows that ages 5–20 (i.e. mostly school children) have the greatest variation in contacts. Most likely, this is because this age group has the most contacts.

## The influence of contact matrix uncertainty on epidemic dynamics

To explore the sensitivity of model output to the entries in the contact matrix, we run the age-structured SEIRD model once for each bootstrap sample of the contact matrix. We use fixed values for the other parameters of the model. Here, we assume the transmission dynamics are representative of ancestral SARS-CoV-2 and obtain representative values from the `covid_transmission_parameters()` function in `comomodels`. For further details on these chosen parameter values, we refer the interested reader to this function's documentation. Initially, we assume that 0.1% of the population has been exposed to infection, with the remainder of the population susceptible.

```
# Age structured parameters
mu <- covid_transmission_parameters(is_age_structured=TRUE)[4]$mu$mu[1:8]
```

```
mu_age_vals <- rep(mu, each=2)

gamma <- covid_transmission_parameters(is_age_structured=TRUE)[3]$gamma$gamma[1:8]
gamma_age_vals <- rep(gamma, each=2)

# Set the non-age structured parameters
parameters <- covid_transmission_parameters()
kappa <- parameters$kappa
gamma <- parameters$gamma
mu <- parameters$mu
R0_target <- parameters$R0
beta <- (mu + gamma) * R0_target
```

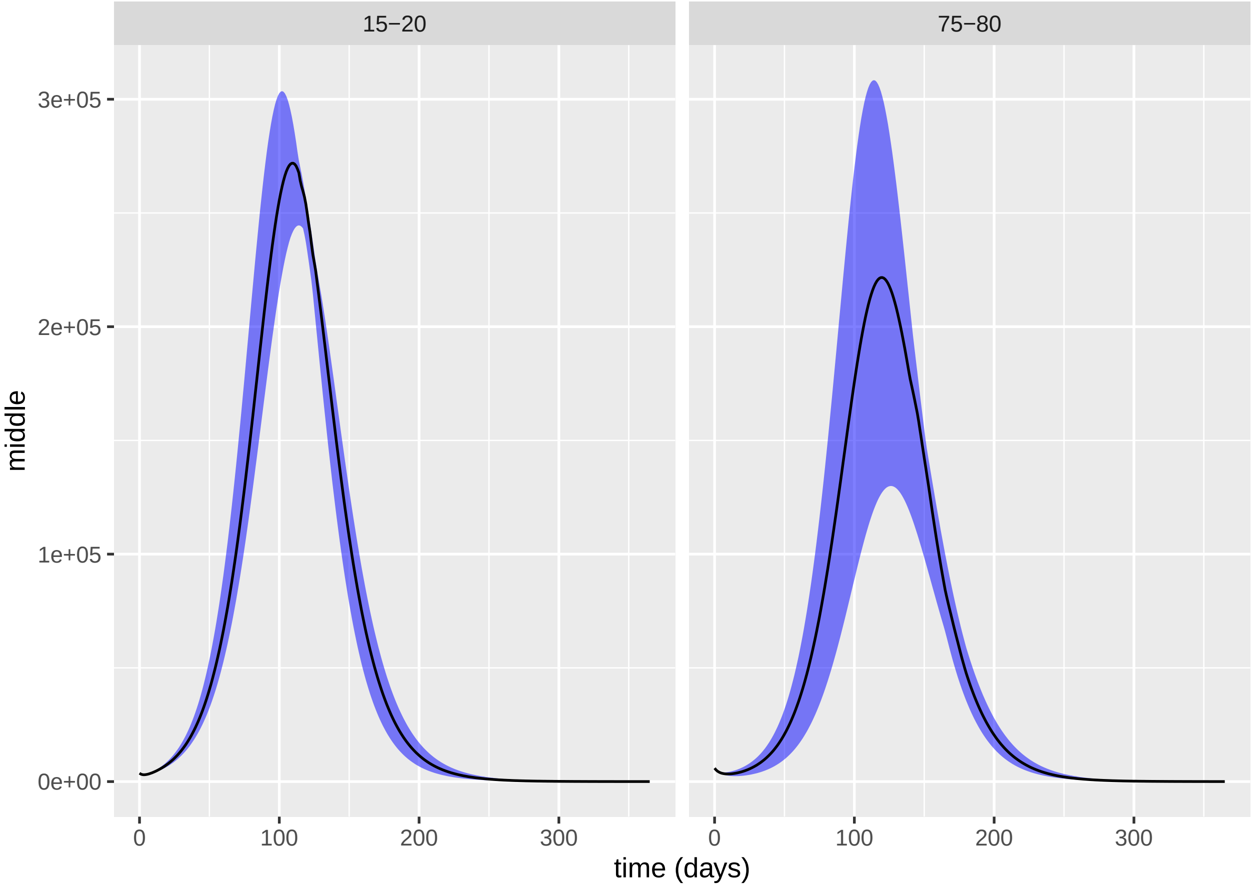
With the parameter values set, we now simulate the `SEIRDAge` model for one year for each of the bootstrap samples of the contact matrix.

```
times <- seq(0, 365, by=1)
for (i in 1:n_bootstrap){
  matrix=matrices[[i]][[1]]

  # Remove the column and row names so the model will accept it
  colnames(matrix) <- NULL
  rownames(matrix) <- NULL

  # Keep the data for ages 0-80, in 5 year increments
  matrix <- matrix[1:16, 1:16]

  model <- comomodels::SEIRDAge(n_age_categories=n_ages,
                   contact_matrix=matrix,
                   age_ranges=as.list(age_names))

  # Set the other parameters of the model
  transmission_parameters(model) <- list(b=beta,
                                         k=kappa,
                                         g=gamma_age_vals,
                                         mu=mu_age_vals)

  initial_conditions(model) <- list(S0=pop_fraction*0.999,
                                    E0=rep(0, n_ages),
                                    I0=pop_fraction*0.001,
                                    R0=rep(0, n_ages),
                                    D0=rep(0, n_ages))

  # Run model
  res <- run(model, time=times)

  # Get states from results
  res <- res$states

  # Save the data for the I and R compartments
  x <- res %>%
    filter(compartment %in% c("I", "R", "D")) %>%
    mutate(iteration=i)
  if (i == 1)
```

```
    all_results <- x
  else
    all_results <- rbind(all_results, x)
}
```

We then use the simulations to generate the central 90% quantiles in the infectious population size for two age groups: 15-20 year olds and 75-80 year olds. We plot these quantiles below.

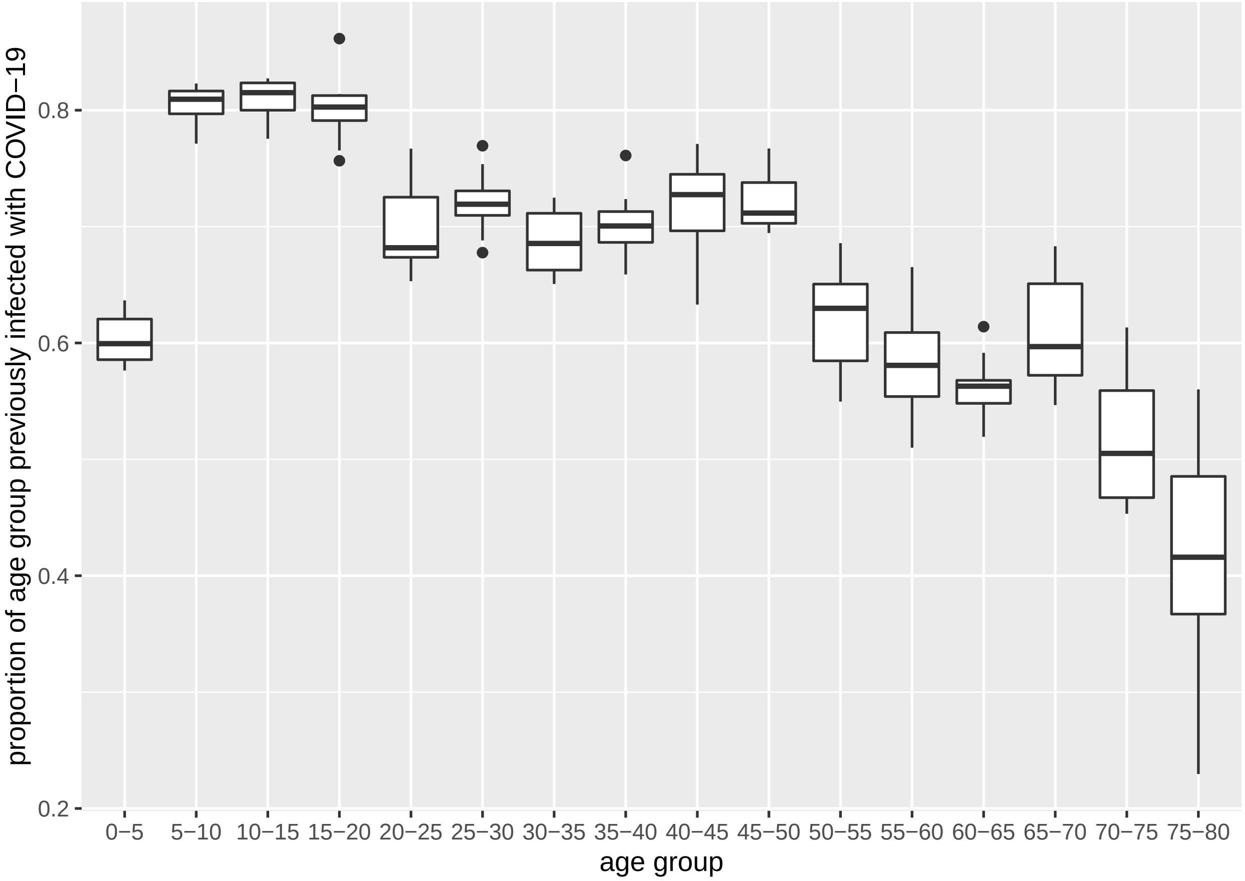
```
# Filter and find quantiles
I_df <- all_results %>%
  filter(age_range %in% c("15-20", "75-80"),
         compartment=="I") %>%
  mutate(value=value * sum(pops)) %>%
  group_by(time, age_range) %>%
  summarise(lower=quantile(value, 0.05),
            middle=quantile(value, 0.5),
            upper=quantile(value, 0.95))

# Plot
ggplot(I_df, aes(x = time)) +
  geom_ribbon(aes(ymin = lower, ymax = upper),
              fill = "blue", alpha = 0.5) +
  geom_line(aes(y = middle)) +
  facet_wrap(~age_range) +
  xlab("time (days)")
```



In both age groups, the results show that there is considerable uncertainty in the peak infectious counts, but with considerably higher variation for the older age group – a result that is particularly worrying given the greater risk of severe disease facaed by older individuals (Verity et al. 2020).

Next, we examine the proportion of individuals infected with COVID-19 at the end of the year, which we plot below. Here, we consider only those individuals who have survived infection.

```
# Filter data
df <- all_results %>%
  filter(compartment == "R") %>%
  filter(time == 365)

# Plot
ggplot(df,
       aes(x=age_range, y=value/pop_fraction)) +
  geom_boxplot() +
  xlab("age group") +
  ylab("proportion of age group previously infected with COVID-19")
```

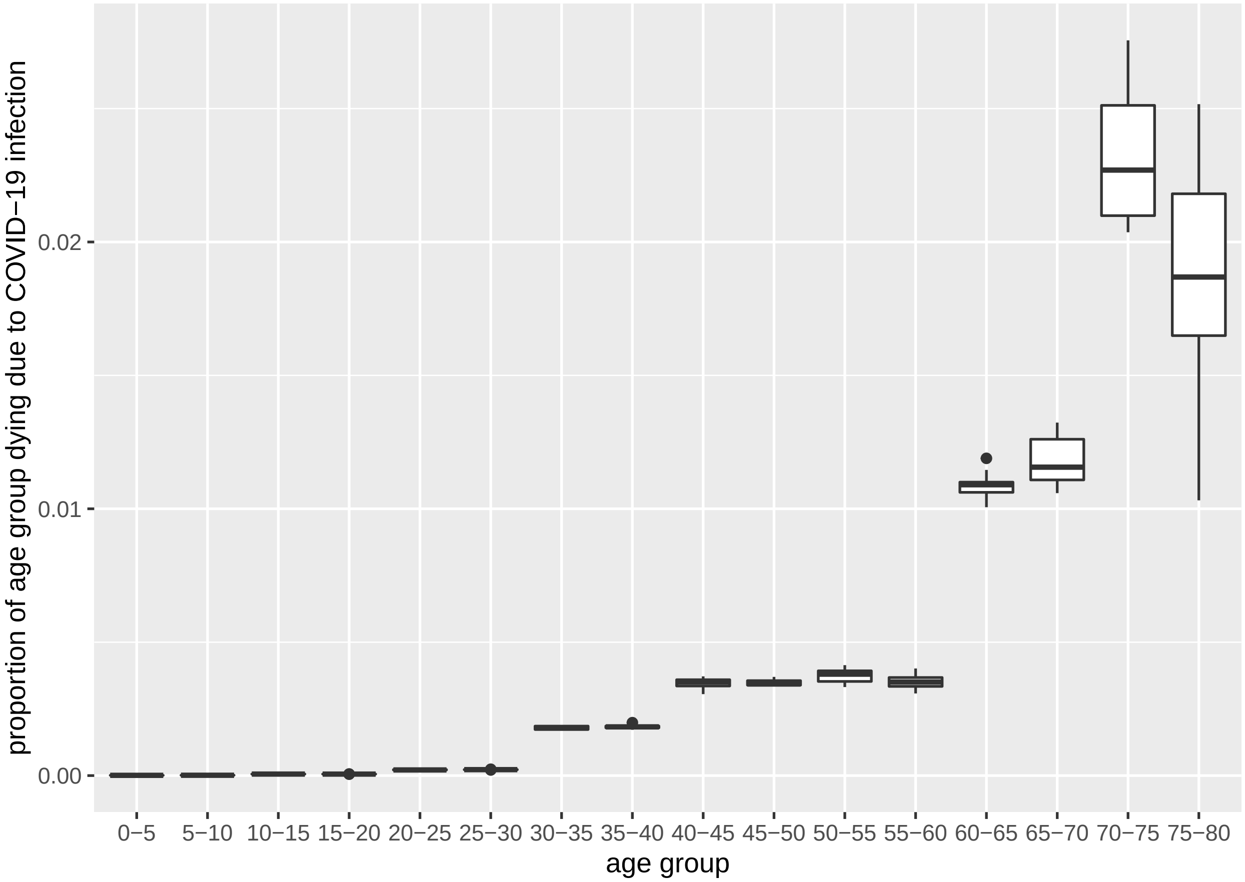


This plot shows that those aged 5–20 are those most likely to have been infected by COVID-19: mainly because these individuals have the highest number of contacts and, because of this, are important drivers of infections within the population.

This plot also illustrates the pronounced uncertainty in the proportion of infecteds in the oldest age group.

Finally, we study the effect of uncertainty in the contact matrix on the number of individuals that die of the infection.

```
# Filter dataset
df <- all_results %>%
  filter(compartment == "D") %>%
  filter(time == 365)

# Plot
ggplot(df, aes(x=age_range, y=value / pop_fraction)) +
  geom_boxplot() +
  xlab("age group") +
  ylab("proportion of age group dying due to COVID-19 infection")
```



Although younger people are more likely to be infected due to their higher number of contacts, deaths occur mainly in the elderly. The bootstrapped samples of the contact matrix generate a wide range of deaths, particularly in the oldest age groups.

## Conclusion

Who we interact with and how often we interact changes as we age. COVID-19 is predominantly spread from close contact with infected individuals and has a strong age-profile of severe disease. Because of this, mathematical models of disease transmission are often acutely sensitive to estimates of age-specific contact patterns.

## References

Arregui, Sergio, Alberto Aleta, Joaquin Sanz, and Yamir Moreno. 2018. "Projecting Social Contact Matrices to Different Demographic Structures." *PLOS Computational Biology* 14 (12): e1006638.

Asanati, Kaveh, Louise Voden, and Azeem Majeed. 2021. "Healthier Schools During the COVID-19 Pandemic: Ventilation, Testing and Vaccination." *Journal of the Royal Society of Medicine* 114 (4): 160–63.

Funk, Sebastian. 2020. *Socialmixr: Social Mixing Matrices for Infectious Disease Modelling.* https://CRAN.R-project.org/package=socialmixr.

King Jr, James C, Jeffrey J Stoddard, Manjusha J Gaglani, Kristine A Moore, Laurence Magder, Elizabeth McClure, Judith D Rubin, Janet A Englund, and Kathleen Neuzil. 2006. "Effectiveness of School-Based Influenza Vaccination." *New England Journal of Medicine* 355 (24): 2523–32.

Mossong, Joel, Niel Hens, Mark Jit, Philippe Beutels, Kari Auranen, Rafael Mikolajczyk, Marco Massari, et al. 2017. "POLYMOD Social Contact Data." *Zenodo.*

Prem, Kiesha, Kevin van Zandvoort, Petra Klepac, Rosalind M Eggo, Nicholas G Davies, Centre for the Mathematical Modelling of Infectious Diseases COVID-19 Working Group, Alex R Cook, and Mark Jit. 2021. "Projecting Contact Matrices in 177 Geographical Regions: An Update and Comparison with Empirical Data for the COVID-19 Era." *PLOS Computational Biology* 17 (7): e1009098.

Verity, Robert, Lucy C Okell, Ilaria Dorigatti, Peter Winskill, Charles Whittaker, Natsuko Imai, Gina Cuomo-Dannenburg, et al. 2020. "Estimates of the Severity of Coronavirus Disease 2019: A Model-Based Analysis." *The Lancet Infectious Diseases* 20 (6): 669–77.

Walter, Emmanuel B, Kawsar R Talaat, Charu Sabharwal, Alejandra Gurtman, Stephen Lockhart, Grant C Paulsen, Elizabeth D Barnett, et al. 2022. "Evaluation of the BNT162b2 Covid-19 Vaccine in Children 5 to 11 Years of Age." *New England Journal of Medicine* 386 (1): 35–46.