Individual-level Modeling of COVID-19 Epidemic Risk

Application of hierarchical Maximum Likelihood Estimation to multi-level data collected in real-time during the pandemic

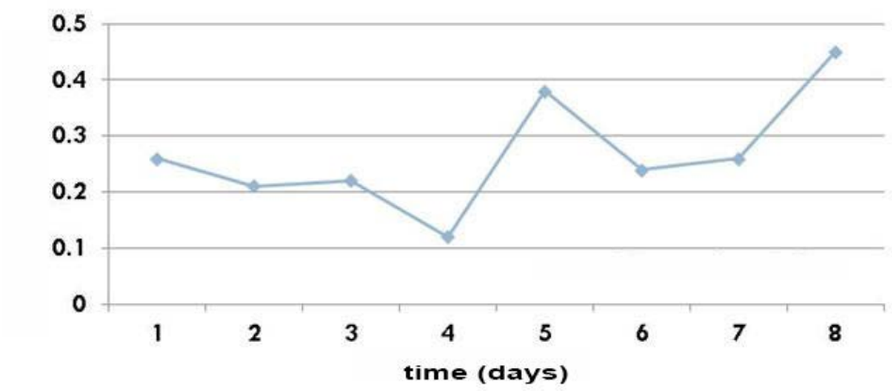
Andres Colubri - Broad Institute

Predicting risk of infection during the COVID-19 pandemic

The goal with this work is to construct a model to predict risk of infection for individuals in a given time window using demographic (e.g.: age, sex), medical (e.g.: pre-existing conditions), symptoms (e.g: fever, cough), and behavioral (e.g.: exposure events)

$$P(\text{infection in } [t\text{-}d, t] \mid D_{[t\text{-}d,\ t]})$$

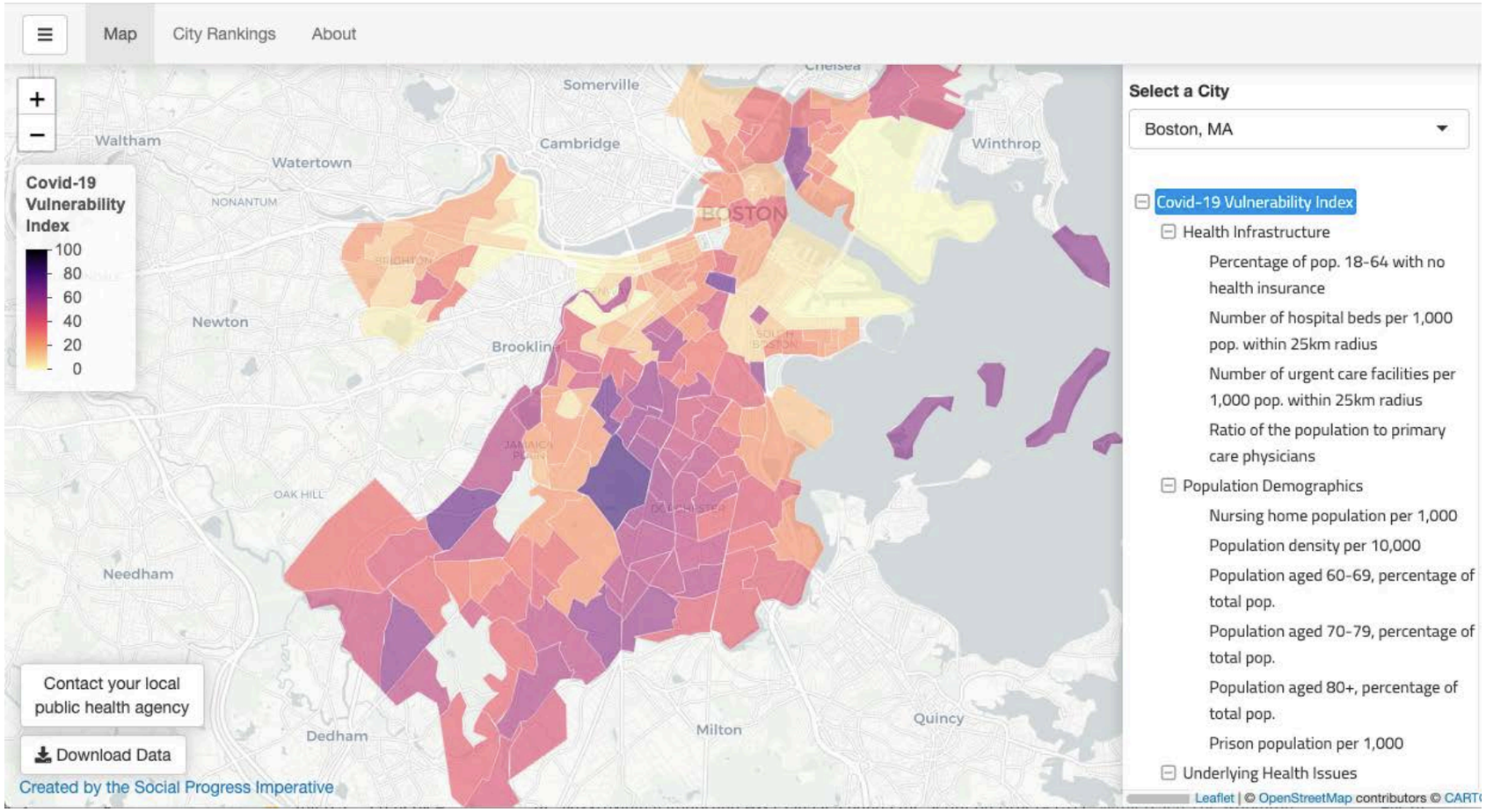
[t -d, t] represents a suitable window of d days up to time t (for instance, by taking d = incubation period). This can give us an individualized time profile:



D stands for some of the data that is available and changes over time (e.g.: daily symptoms). The idea behind using a general conditional form is to enable updating the risk with new information through Bayes Theorem.

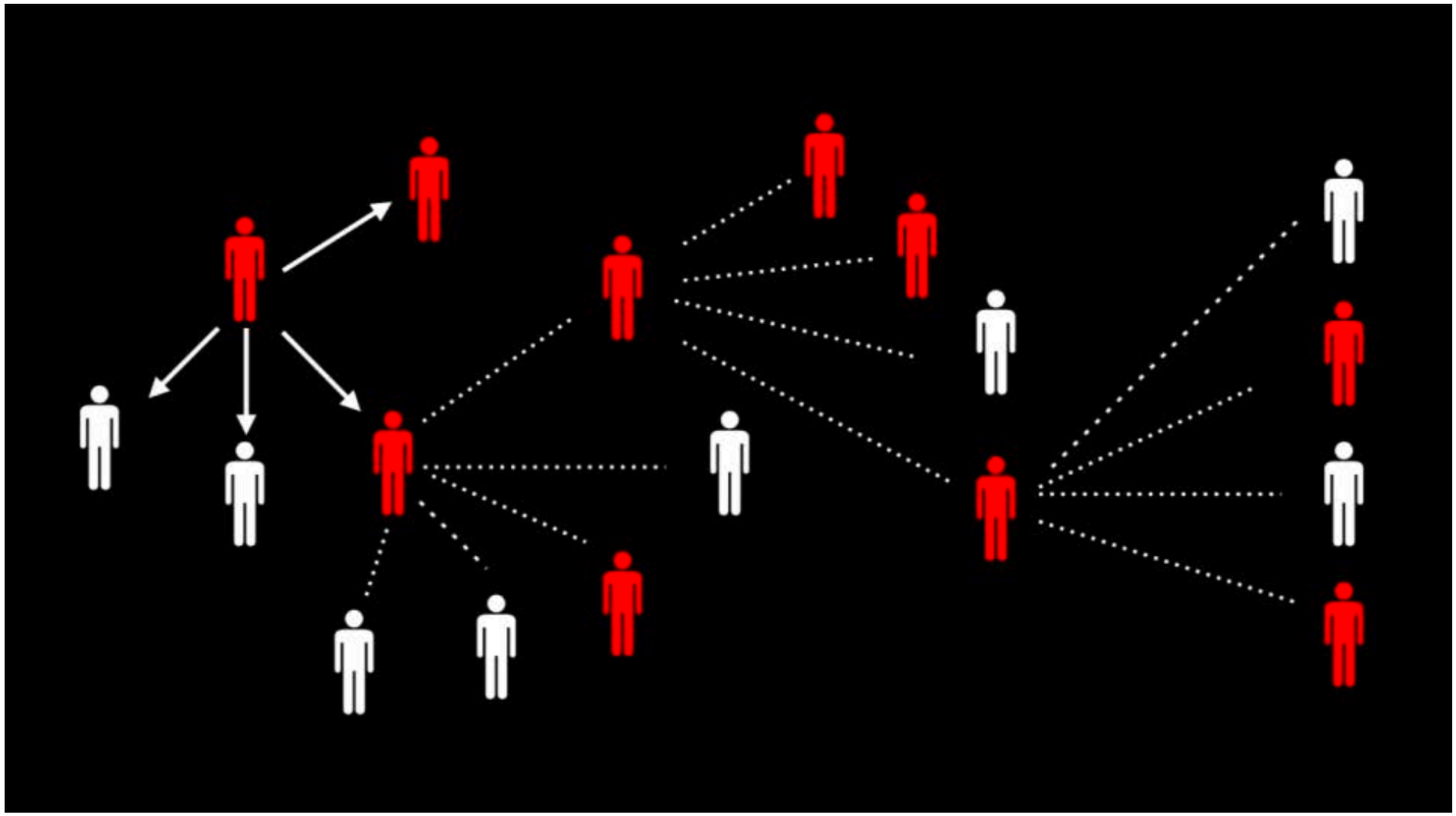
Applications of risk predictions

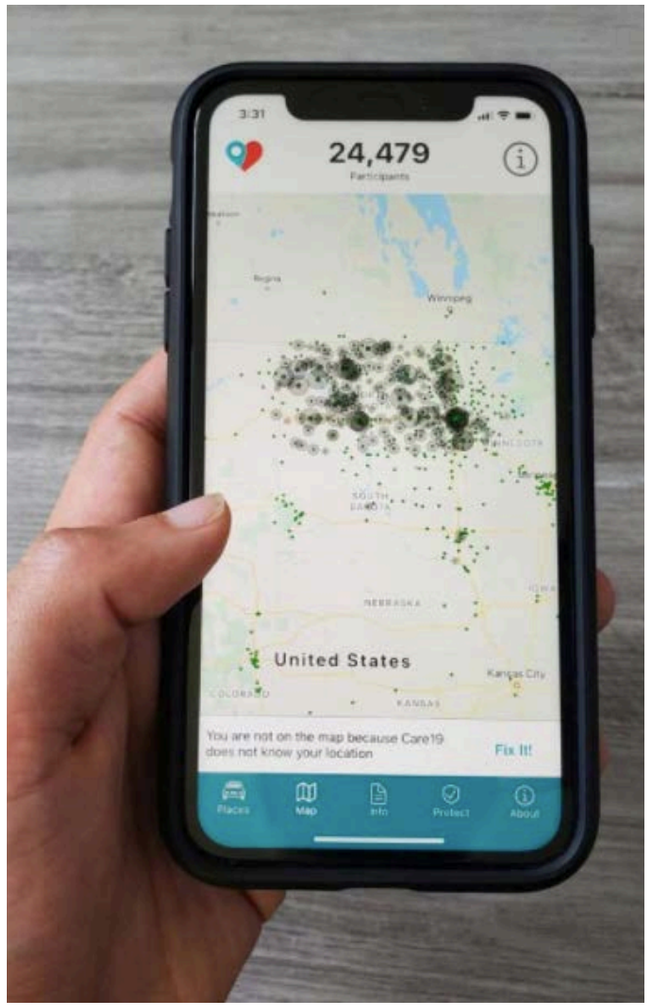
- Risk mapping and identification of hot-spots
- Informing decision-making by healthcare workers
- Risk parameters can be used in other epidemic models



Real-time data collection with mobile apps

The data required by the model can be collected via a mobile app for symptom tracking and contact tracing. The risk prediction would be updated in “real-time” as new data is gathered over time





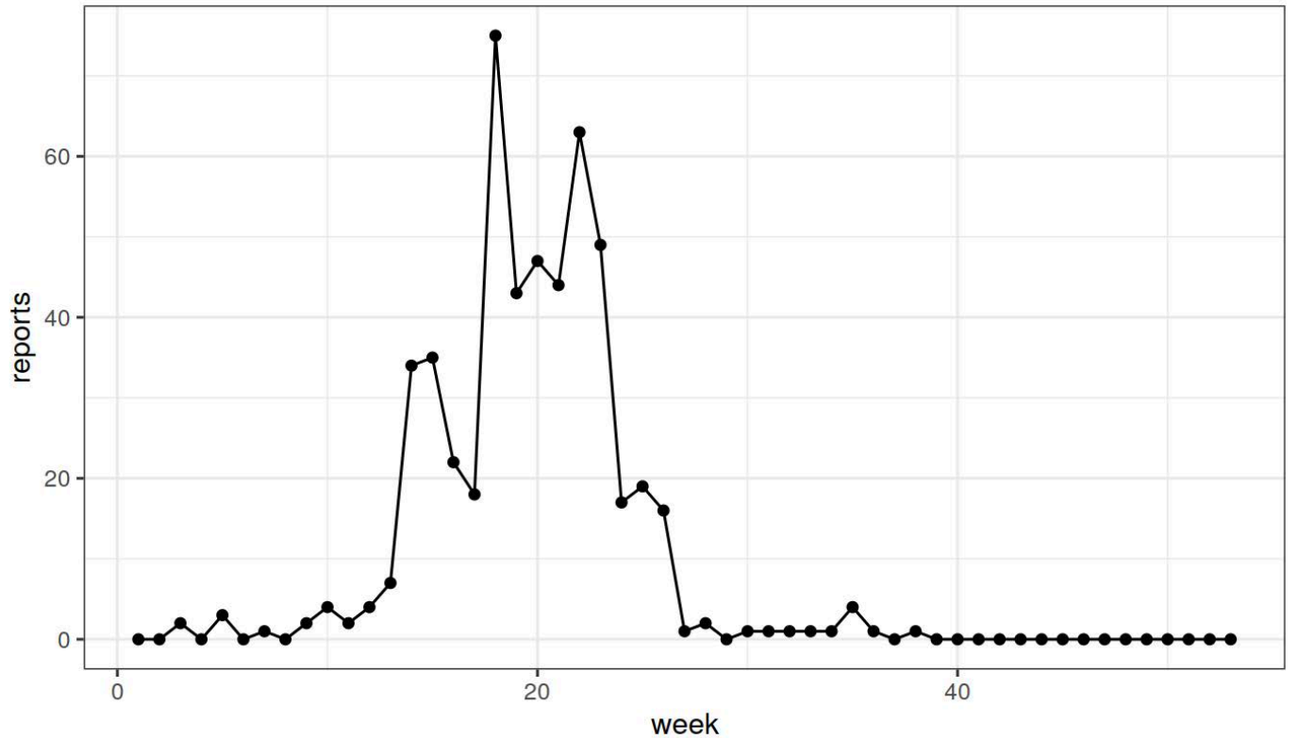
Privacy considerations

- Personal data remains on users' devices where is used to compute the risk prediction, which is then securely uploaded for aggregation or displayed to the user.

- The risk model itself will be trained using data only from consenting individuals and then distributed among all the users of the app

Multi-level epidemiological data

Population level



Individual level

Individual-level Models

In ILMs, we have an expression for the probability of infection of an individual in a given time interval [t, t+1) given the set of exposure contacts C(i, t):

$$P(i,t) = 1 - \exp\Big[\Big(- \Omega_S(i) \sum_{j \in C(i,t)} \Omega_T(j)\kappa(i,j,t)\Big) - \epsilon(i,t)\Big]$$

Where the susceptibility of individual i and infectivity of j are represented as a linear combination of covariates:

$$\Omega_S(i) = a_0 + a_1 X_1(i) + \ldots + a_N X_N(i)$$

$$\Omega_T(j) = b_0 + b_1 Y_1(j) + \ldots + b_M Y_M(j)$$

$\kappa(i,j,t) =$ Infectious kernel, depending on distance, time of contact between i and j

$\epsilon(i,t) =$ Sparks term, accounting for infections not caused by the contacts between individuals

Simplifications to the model

As a first step, we can start with a simple individual-level model with only two covariates for the susceptibility and infectivity functions:

$$\Omega_S(i) = a_0 + a_1 X_1(i)$$

$$\Omega_T(j) = b_0 + b_1 Y_1(j)$$

For example, X_1(i) could represent a binary “immune status” (0=normal, 1=low immunity due to age or pre-existing condition), and Y_1(j) the symptomatic status of infectious individual I (0=asymptomatic, 1= symptomatic)

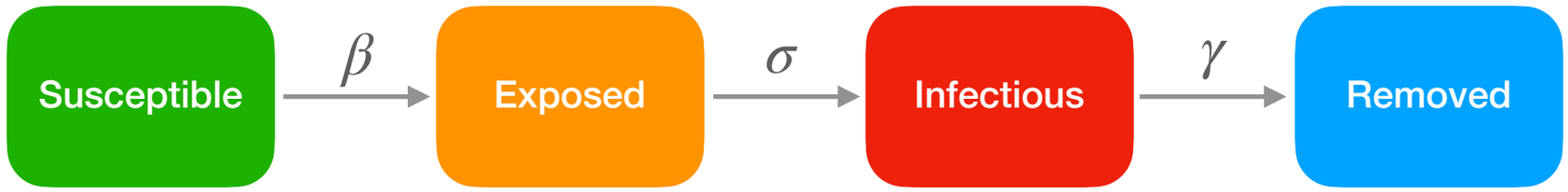
Two more simplifications:

$\kappa(i, j, t) =$ 1 if (i, j) were in contact, 0 otherwise

$$\epsilon(i, t) = 0$$

Population-level Models

SEIR compartmental model



$$\frac{dS}{dt} = -\frac{\beta SI}{N}$$

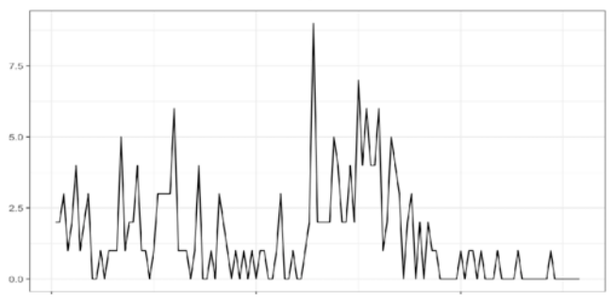
$$\frac{dE}{dt} = \frac{\beta SI}{N} - \sigma E$$

$$\frac{dI}{dt} = \sigma E - \gamma I$$

$$\frac{dR}{dt} = \gamma I$$

$$N = S + E + I + R$$

Parameter estimation using population-level data



β
γ
σ

Parameter estimates

Connecting population and individual parameters

Given a sample $I_S(t)$ of infectious individuals from the population at time t and their contacts $\{C(j,t), j\ I_S(t)\}$, we can estimate the rate of infection in the SEIR model from the individual probabilities of infection p(i, j, t):

$$\hat{\beta} = \frac{1}{|J(t)|} \sum_{j \in J(t)} \sum_{i \in C(j,t)} p(i,j,t)$$

$$\beta = cp$$

c=contact rate

p=probability of transmission

$$p(i,j,t) = 1 - \exp\Big[- \big(a_0 + a_1 X_1(i)\big)\big(b_0 + b_1 Y_1(j)\big)\Big]$$

Considering the covariate values $\{X_n(i)\}$ and $\{Y_m(j)\}$ as part of the observed data, we can think of the rate of infection as a function of the parameters a_0, a_1, b_0, b_1

Partially-observed Markov process

- Maximum likelihood estimation of the parameters with can be done using the partially observed data:

  y_t = new case counts over time

  X_t, Y_t = covariates for infected individuals and their contacts over time

- The Markov property of epidemic dynamics enable the use of iterated filtering form MLE in the context of Partially-Markov processes (POMP).

- The POMP package can be used for this end: https://kingaa.github.io/pomp

Specification of the POMP

The dynamics of the Markov process begin the SEIR model can be specified as:

$$
\begin{aligned}
S(t+1) &= S(t) - B(t) \\
E(t+1) &= E(t) + B(t) - C(t) \\
I(t+1) &= I(t) + C(t) - D(t) \\
R(t+1) &= R(t) + D(t) \\
S(t) + E(t) + I(t) + R(t) &= N
\end{aligned}
$$

$$
\begin{aligned}
B(t) &\sim Binomial(S(t), 1 - \exp(-\beta(t) I(t)/N)) \\
C(t) &\sim Binomial(E(t), 1 - \exp(-\sigma)) \\
D(t) &\sim Binomial(I(t), 1 - \exp(-\gamma))
\end{aligned}
$$

Together with the observation model:

$$y_t | \, C(t) \sim Binomial(C(t), \rho)$$

and the expression for the rate of infection as function of the individual-level parameters:

$$\hat{\beta} = \frac{1}{|\, J(t) \,|} \sum_{j \in J(t)} \sum_{i \in C(j,t)} 1 - \exp\Big[- \big(a_0 + a_1 X_1(i)\big)\big(b_0 + b_1 Y_1(j)\big)\Big]$$

Implementation details: R script and C snippets

```r
obs_data %>%
  pomp(t0 = time_start,
       time = time_unit,
       rprocess = euler(sir_step, delta.t=time_step),
       rinit = sir_init,
       rmeasure = rmeas,
       dmeasure = dmeas,
       globals = extra,
       cdir = code_folder,
       cfile = file_name,
       accumvars=c("C"),
       statenames=c("S", "E", "I", "R", "C"),
       partrans=parameter_trans(
         log=log_trans_params,
         logit=logit_trans_params),
       paramnames = c(free_param_names, fixed_param_names),
       #compile=FALSE,
       verbose = TRUE
  ) -> model
```

```r
sir_init <- Csnippet("
  double m = pop/(S_0 + E_0 + I_0 + R_0);

  S = nearbyint(m*S_0);
  E = nearbyint(m*E_0);
  I = nearbyint(m*I_0);
  R = nearbyint(m*R_0);

  C = 0;
")
```

```r
double beta;
double foi;
double rate[3], trans[3];

beta = calc_beta(t, a0, a1, b0, b1);

// expected force of infection
foi = beta * I/pop;

rate[0] = foi;      // stochastic force of infection
rate[1] = sigma;    // rate of ending of latent stage
rate[2] = gamma;    // recovery

// transitions between classes
reulermultinom(1, S, &rate[0], dt, &trans[0]);
reulermultinom(1, E, &rate[1], dt, &trans[1]);
reulermultinom(1, I, &rate[2], dt, &trans[2]);

S += -trans[0];
E += trans[0] - trans[1];
I += trans[1] - trans[2];
R = pop - S - E - I;

// Assigning the right number to the accumulation variable that's used
// in the observation model is absolutely critical!!!!
C += trans[1];
```

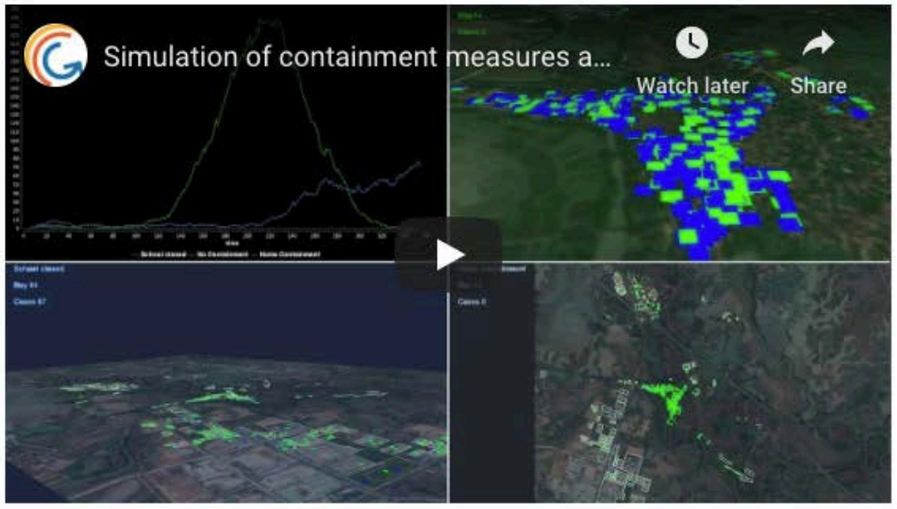
```r
double calc_beta(double td, double a0, double a1, double b0, double b1) {
  static int *indices = NULL;
  static double *contacts = NULL;
  static int max_t = 0;
  static int num_v = 0;

  if (indices == NULL) {
    FILE *file;

    file = fopen(\"MAIN_FOLDER/indices\", \"r\");

    int idx;
    while (fscanf(file, \"%d\", &idx) > 0) max_t++;
    rewind(file);

    indices = (int *)malloc(sizeof(int)*max_t);
    int i = 0;
    while (fscanf(file, \"%d\", &idx) > 0) {
      indices[i] = idx;
      i++;
    }
    fclose(file);

    file = fopen(\"MAIN_FOLDER/contacts\", \"r\");
    float val;
    while (fscanf(file, \"%f\", &val) > 0) num_v++;
    rewind(file);

    contacts = (double *)malloc(sizeof(double)*num_v);
    i = 0;
    while (fscanf(file, \"%f\", &val) > 0) {
      contacts[i] = val;
      i++;
    }
    fclose(file);

    //Rprintf(\"%d %d\\n\", max_t, num_v);
  }

  double beta = 0;

  int t = (int) td;
  if (max_t <= t) t = max_t - 1;
  int idx = indices[t];
  int ninf = 0;
  while (-1 < contacts[idx]) {
    int ncont = (int) contacts[idx++];
    double y = contacts[idx++];
    for (int i = 0; i < ncont; i++) {
      double x = contacts[idx++];
      double p = (a0 + a1 * x) * (b0 + b1 * y);
      beta += 1 - exp(-p);
    }
    ninf++;
  }

  if (0 < ninf) {
    beta /= ninf;
  }

  //Rprintf(\"%lg = %lg\\n\", td, beta);

  return beta;
}
```

Generating synthetic data with ABMs

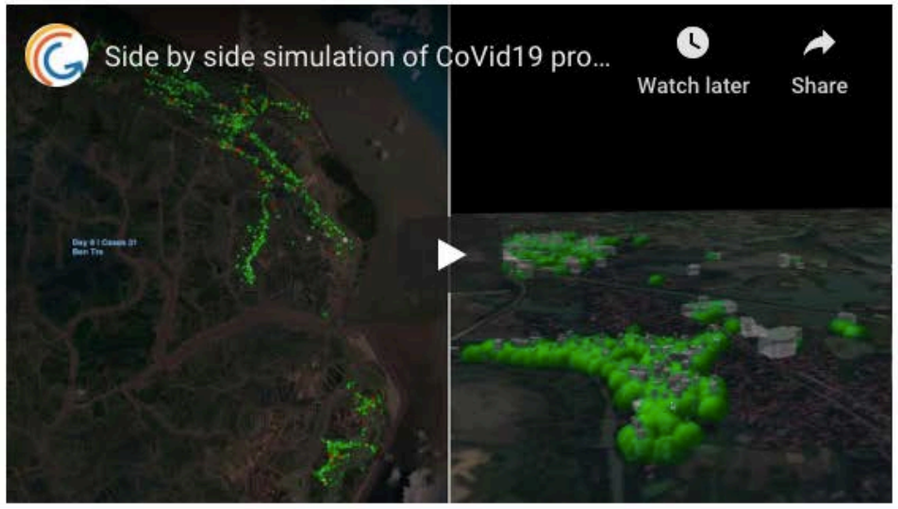
Before we have access to real individual-level data, we can generate a synthetic dataset using agent-based models.

The GAMA platform could be a good option to run ABMs simulations for COVID-19: https://gama-platform.github.io/covid19

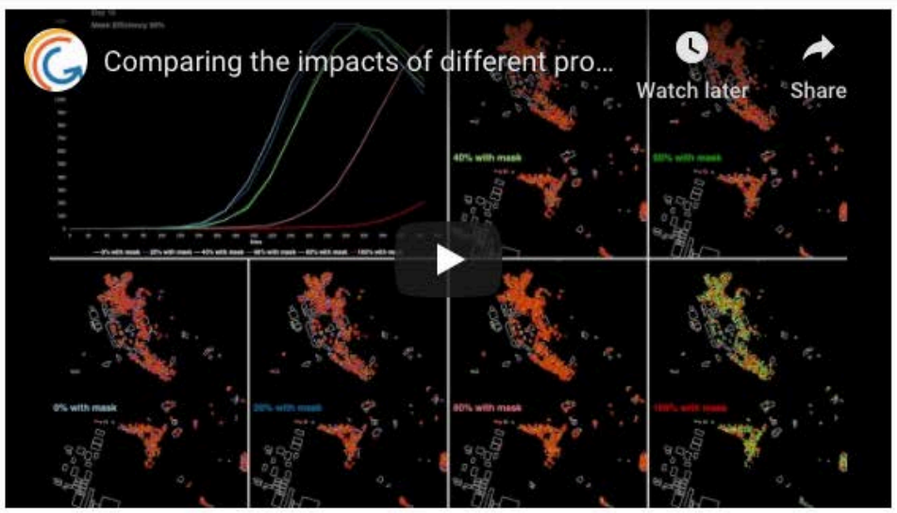
Simulation of containment measures against CoVid19 using GAMA



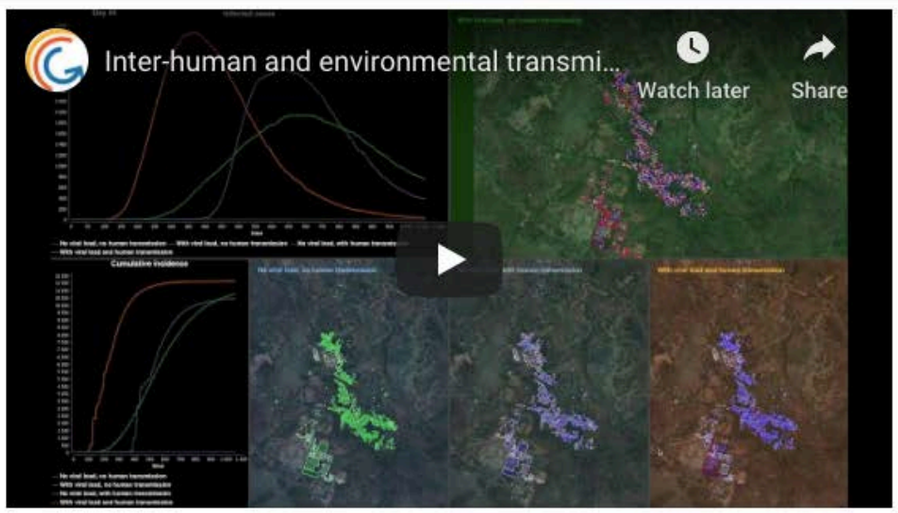
Side by side simulation of CoVid19 propagation in two case studies



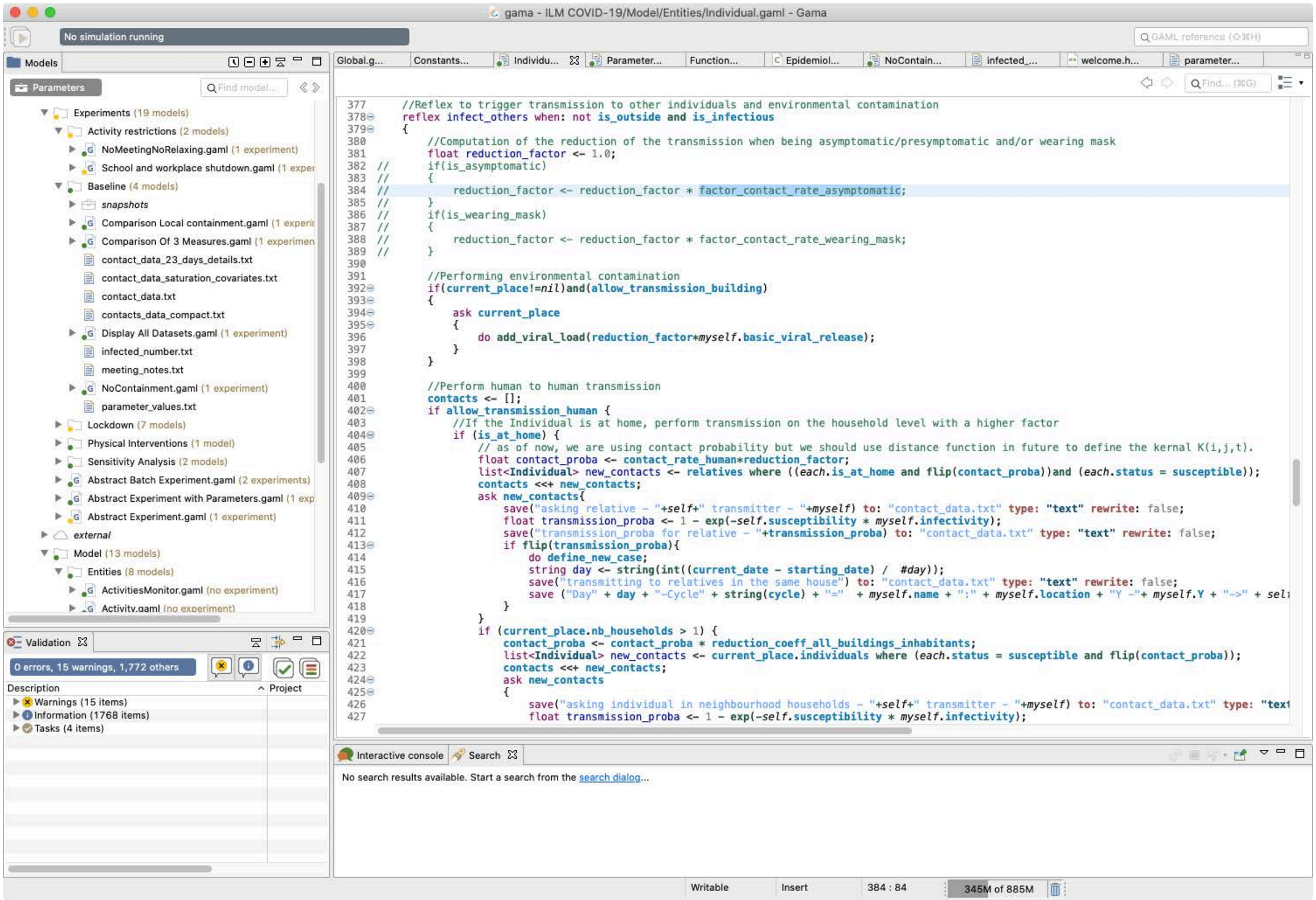
Comparing the impacts of different proportions of people wearing masks



Inter-human and environmental transmissions

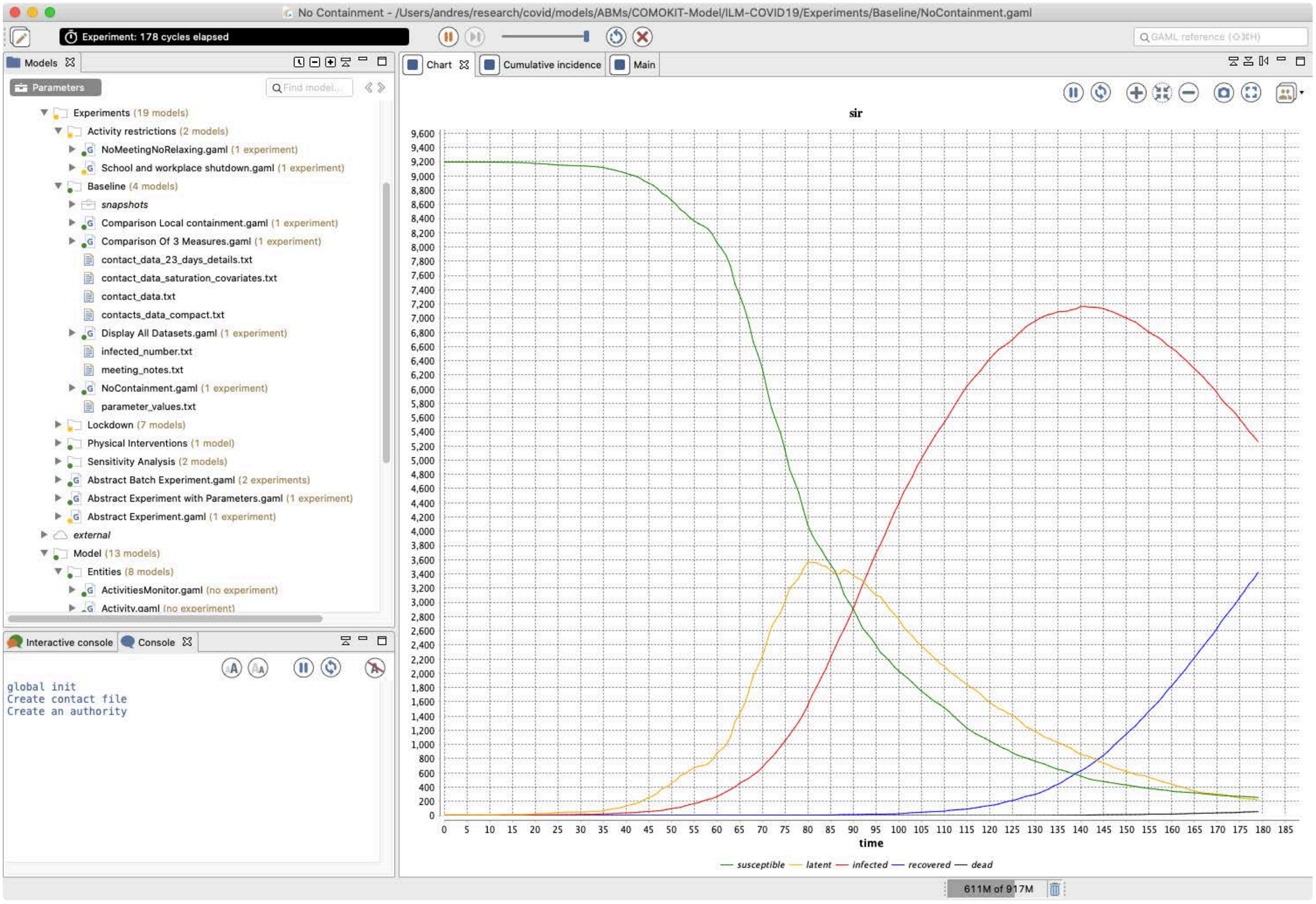


Programming detail epidemic models with GAML




```
377     //Reflex to trigger transmission to other individuals and environmental contamination
378     reflex infect_others when: not is_outside and is_infectious
379     {
380         //Computation of the reduction of the transmission when being asymptomatic/presymptomatic and/or wearing mask
381         float reduction_factor <- 1.0;
382 //      if(is_asymptomatic)
383 //      {
384 //          reduction_factor <- reduction_factor * factor_contact_rate_asymptomatic;
385 //      }
386 //      if(is_wearing_mask)
387 //      {
388 //          reduction_factor <- reduction_factor * factor_contact_rate_wearing_mask;
389 //      }
390
391         //Performing environmental contamination
392         if(current_place!=nil)and(allow_transmission_building)
393         {
394             ask current_place
395             {
396                 do add_viral_load(reduction_factor*myself.basic_viral_release);
397             }
398         }
399
400         //Perform human to human transmission
401         contacts <- [];
402         if allow_transmission_human {
403             //If the Individual is at home, perform transmission on the household level with a higher factor
404             if (is_at_home) {
405                 // as of now, we are using contact probability but we should use distance function in future to define the kernal K(i,j,t).
406                 float contact_proba <- contact_rate_human*reduction_factor;
407                 list<Individual> new_contacts <- relatives where ((each.is_at_home and flip(contact_proba))and (each.status = susceptible));
408                 contacts <<+ new_contacts;
409                 ask new_contacts{
410                     save("asking relative - "+self+" transmitter - "+myself) to: "contact_data.txt" type: "text" rewrite: false;
411                     float transmission_proba <- 1 - exp(-self.susceptibility * myself.infectivity);
412                     save("transmission_proba for relative - "+transmission_proba) to: "contact_data.txt" type: "text" rewrite: false;
413                     if flip(transmission_proba){
414                         do define_new_case;
415                         string day <- string(int((current_date - starting_date) /  #day));
416                         save("transmitting to relatives in the same house") to: "contact_data.txt" type: "text" rewrite: false;
417                         save ("Day" + day + "-Cycle" + string(cycle) + "="  + myself.name + ":" + myself.location + "Y -"+ myself.Y + "->" + sel
418                     }
419                 }
420                 if (current_place.nb_households > 1) {
421                     contact_proba <- contact_proba * reduction_coeff_all_buildings_inhabitants;
422                     list<Individual> new_contacts <- current_place.individuals where (each.status = susceptible and flip(contact_proba));
423                     contacts <<+ new_contacts;
424                     ask new_contacts
425                     {
426                         save("asking individual in neighbourhood households - "+self+" transmitter - "+myself) to: "contact_data.txt" type: "text
427                         float transmission_proba <- 1 - exp(-self.susceptibility * myself.infectivity);
```

Programming detail epidemic models with GAML

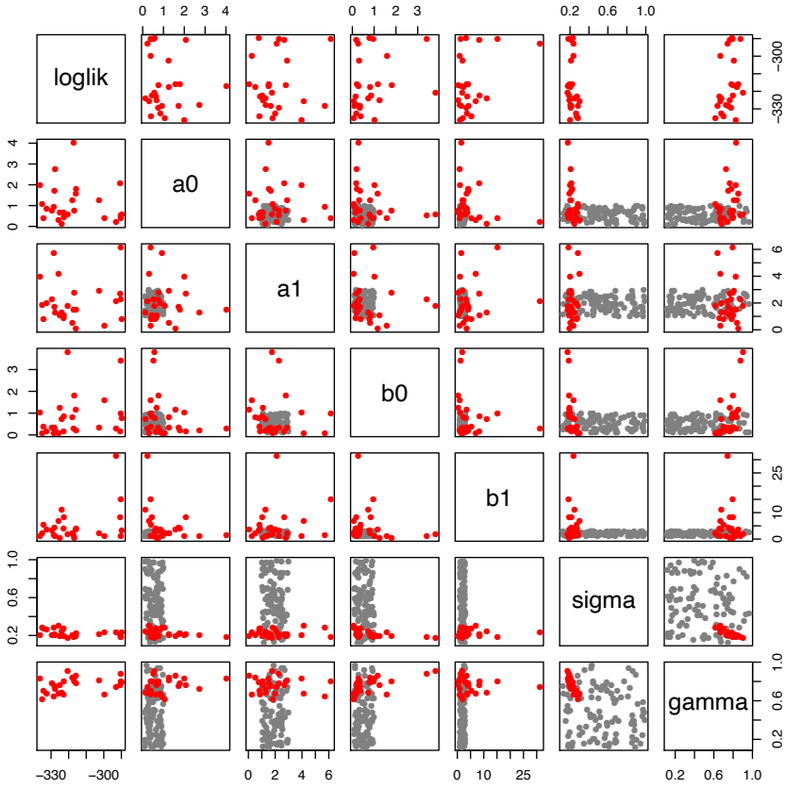


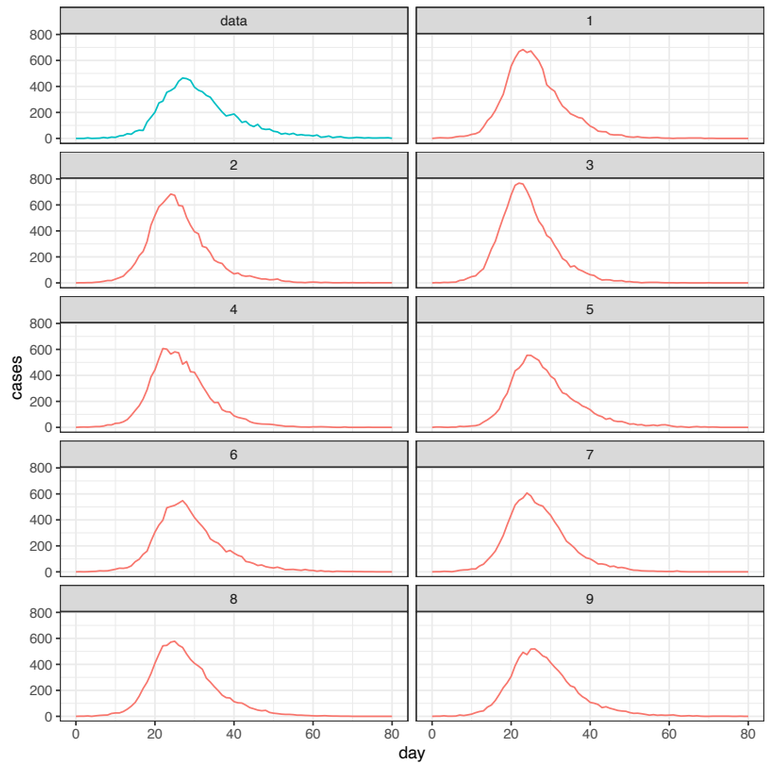
Synthetic individual data

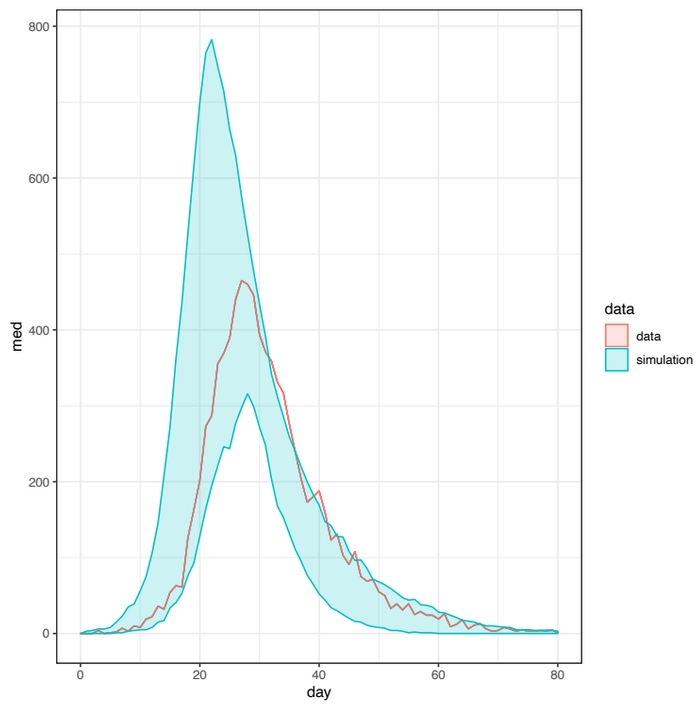
	A	B	C	D	E	F	G	H	I	J	K
1	day	cycle	contact_name	contact_cova	individual_name	individual_cc	latent_time	infectious_time	serial_interval	time_before_hospitalisation	
2	6	25	Individual5312	0	Individual7610	0	4.163419309	17.14059689	8.827247166	0	
3	6	25	Individual848	0	Individual4510	1	3.979964513	29.06000767	4.22052944	0	
4	6	26	Individual5417	1	Individual7610	0	4.163419309	17.14059689	8.827247166	0	
5	6	26	Individual6757	1	Individual6164	0	0.114839473	11.58134946	-1.500446467	0	
6	6	26	Individual7621	1	Individual3227	1	4.366601502	39.84823038	4.511224984	0	
7	6	26	Individual546	1	Individual7738	1	1.313355833	16.56253455	0.340662604	14.9070164	
8	6	27	Individual3885	1	Individual7610	0	4.163419309	17.14059689	8.827247166	0	
9	7	28	Individual7091	1	Individual570	1	3.281420229	10.14013293	0.207872264	7.324098788	
10	7	28	Individual6514	1	Individual7610	0	4.163419309	17.14059689	8.827247166	0	
11	7	29	Individual1002	0	Individual570	1	3.281420229	10.14013293	0.207872264	7.074098788	
12	7	29	Individual1212	0	Individual2728	1	3.766591594	43.58907001	8.686280905	0	
13	7	30	Individual7584	1	Individual7610	0	4.163419309	17.14059689	8.827247166	0	
14	7	31	Individual5571	1	Individual2293	1	3.856000729	27.67758048	4.877699113	0	
15	7	31	Individual2990	0	Individual6128	1	3.012488875	24.12826707	7.70325373	0	
16	7	31	Individual4681	0	Individual7738	1	1.313355833	16.56253455	0.340662604	13.6570164	
17	7	31	Individual7693	0	Individual7738	1	1.313355833	16.56253455	0.340662604	13.6570164	
18	7	31	Individual422	1	Individual1879	1	4.472772245	14.88633611	4.219805624	0	
19	7	31	Individual6605	1	Individual1879	1	4.472772245	14.88633611	4.219805624	0	
20	8	32	Individual113	1	Individual1879	1	4.472772245	14.88633611	4.219805624	0	
21	8	32	Individual5963	1	Individual1879	1	4.472772245	14.88633611	4.219805624	0	
22	8	32	Individual3631	1	Individual1879	1	4.472772245	14.88633611	4.219805624	0	
23	8	33	Individual4744	1	Individual7610	0	4.163419309	17.14059689	8.827247166	0	
24	8	33	Individual1666	0	Individual7610	0	4.163419309	17.14059689	8.827247166	0	
25	8	33	Individual4522	0	Individual7610	0	4.163419309	17.14059689	8.827247166	0	
26	8	33	Individual2536	0	Individual1879	1	4.472772245	14.88633611	4.219805624	0	
27	8	33	Individual7465	1	Individual1879	1	4.472772245	14.88633611	4.219805624	0	
28	8	33	Individual1865	1	Individual1879	1	4.472772245	14.88633611	4.219805624	0	
29	8	33	Individual281	1	Individual6128	1	3.012488875	24.12826707	7.70325373	0	
30	8	33	Individual2437	0	Individual2293	1	3.856000729	27.67758048	4.877699113	0	
31	8	34	Individual1843	1	Individual2728	1	3.766591594	43.58907001	8.686280905	0	
32	8	34	Individual4917	1	Individual2293	1	3.856000729	27.67758048	4.877699113	0	
33	8	34	Individual7926	0	Individual2293	1	3.856000729	27.67758048	4.877699113	0	
34	8	34	Individual2939	0	Individual2293	1	3.856000729	27.67758048	4.877699113	0	
35	8	34	Individual7006	0	Individual4510	1	3.979964513	29.06000767	4.22052944	0	
36	8	34	Individual8013	0	Individual7610	0	4.163419309	17.14059689	8.827247166	0	
37	8	34	Individual8365	1	Individual4974	1	7.821872442	21.02575436	0.055554802	0	
38	8	34	Individual5723	1	Individual7738	1	1.313355833	16.56253455	0.340662604	12.9070164	
39	8	35	Individual762	0	Individual9036	0	3.291059991	19.20771991	6.152979157	0	
40	8	35	Individual1931	1	Individual9036	0	3.291059991	19.20771991	6.152979157	0	

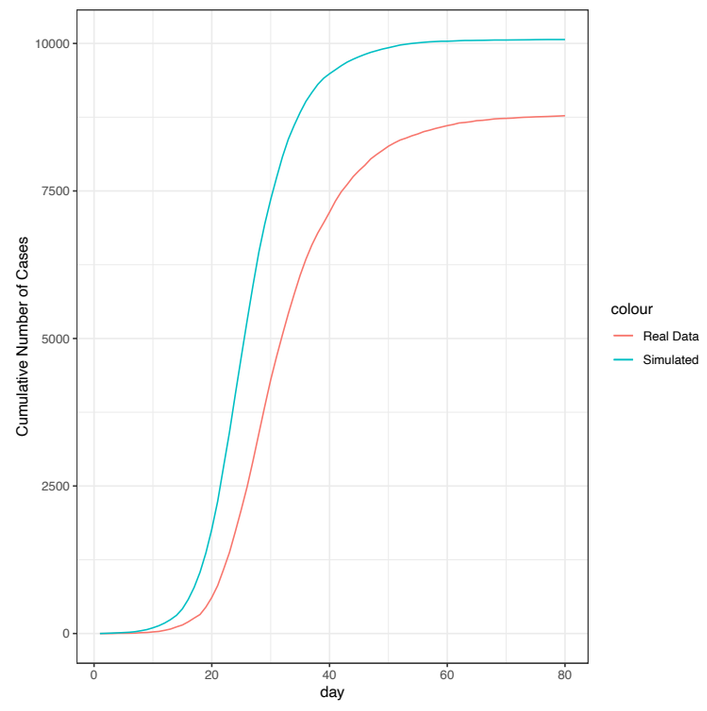
Some preliminary results

parameter	value
a0	0.2
a1	2.0
b0	0.2
b1	2.0









Better estimates after shorter MLE runs…?

Short MLE run

parameter	point estimate
a0	0.50
a1	2.32
b0	1.32
b1	2.3
sigma	0.212
gamma	0.85

Long MLE run

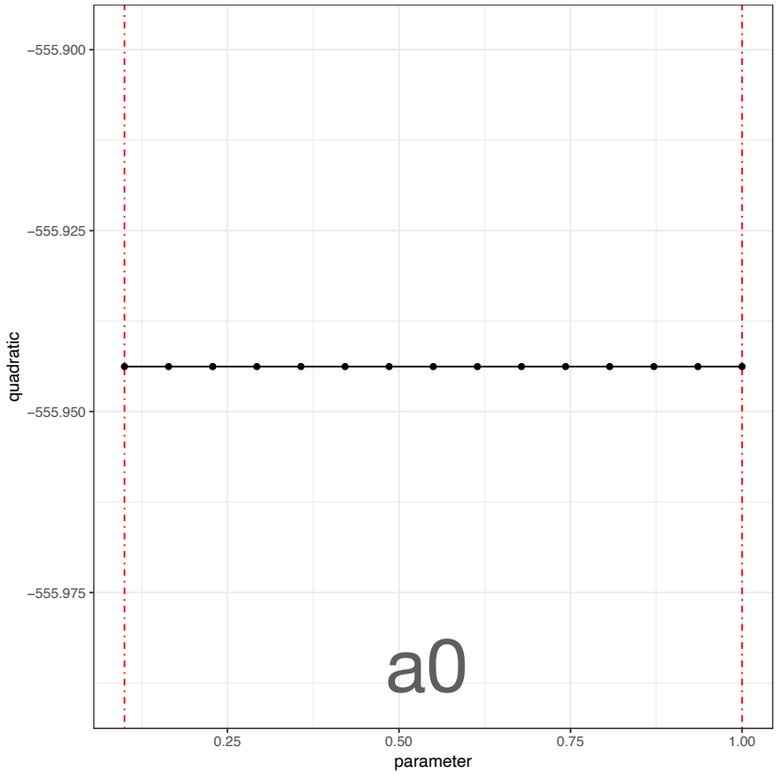
parameter	point estimate
a0	0.59
a1	0.8
b0	0.79
b1	3.32
sigma	0.23
gamma	0.79

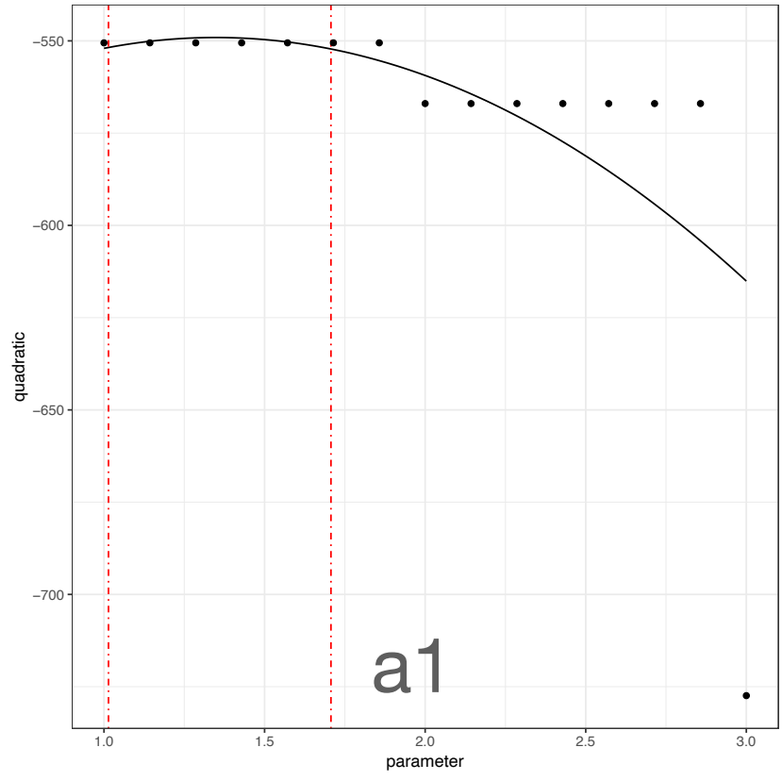
From the synthetic data:

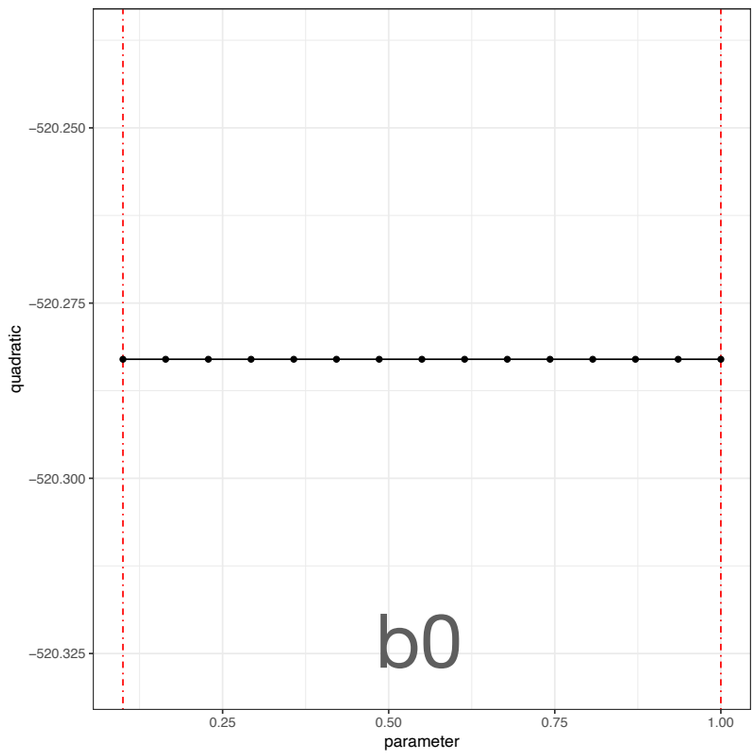
mean latent time is 3.8 days ⟶ sigma=0.26

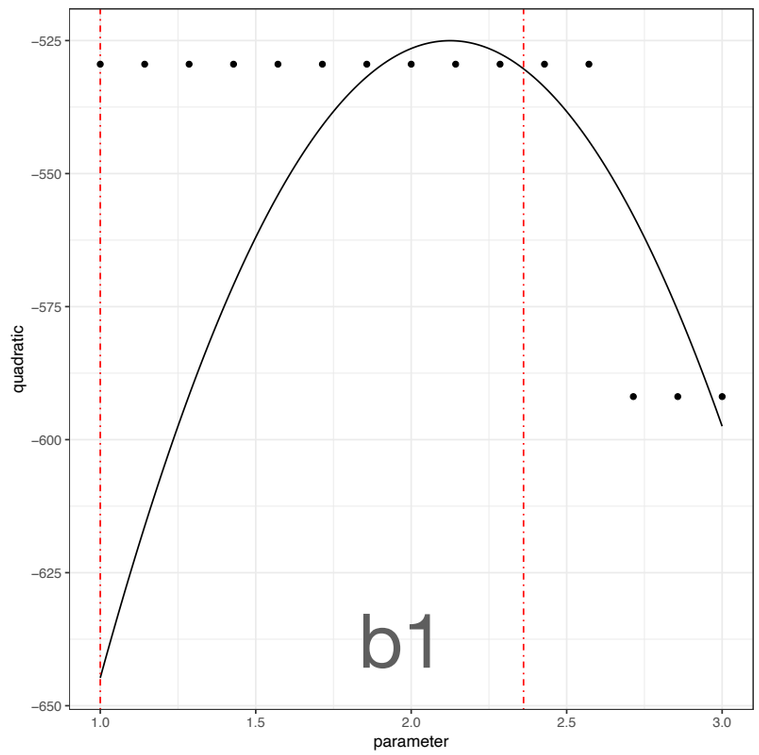
mean infectious time is 1.7 days ⟶ gamma=0.6

Likelihood surface is too flat?









Problem with the parameter estimation

$$\hat{\beta} = \frac{1}{|J(t)|} \sum_{j \in J(t)} \sum_{i \in C(j,t)} 1 - \exp\Big[- \big(a_0 + a_1 X_1(i)\big)\big(b_0 + b_1 Y_1(j)\big)\Big]$$

Sample size decreases as susceptibles become depleted?

Beta is underestimated, parameters are inflated

Symmetry of the dependency on the parameters?

Parameter values can be swapped, likelihood is flat

Re-parametrization of the probability function

$$\hat{\beta} = \frac{1}{|J(t)|} \sum_{j \in J(t)} \sum_{i \in C(j,t)} 1 - \exp\Big[- \Big(c_{00} + c_{01} Y_1(j) + c_{10} X_1(i) + c_{11} X_1(i) Y_1(j) \Big) \Big]$$

We estimate the a00, a01, a10, a00 parameters instead, and obtain the ratios of the original a's and b's by applying the following relationships:

$$\begin{aligned} a_0 b_0 &= c_{00} \\ a_0 b_1 &= c_{01} \\ a_1 b_0 &= c_{10} \\ a_1 b_1 &= c_{11} \end{aligned}$$

$$\frac{a_0}{a_1} = \frac{c_{01}}{c_{11}}$$

$$\frac{a_0}{a_1} = \frac{c_{00}}{c_{10}}$$

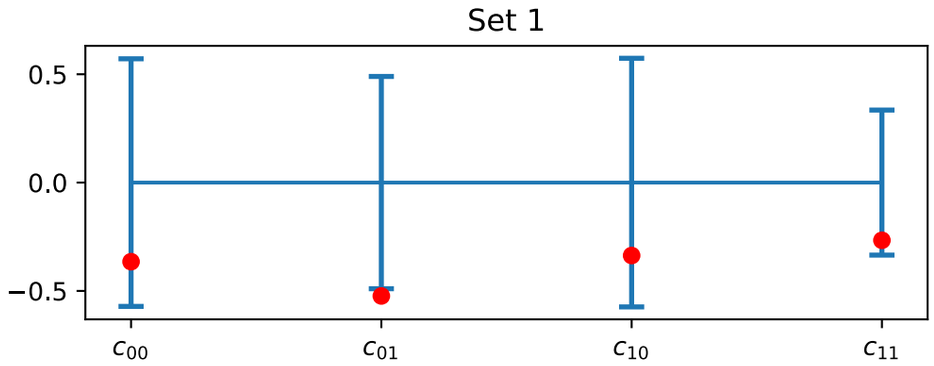
$$\frac{b_0}{b_1} = \frac{c_{00}}{c_{01}}$$

$$\frac{b_0}{b_1} = \frac{c_{10}}{c_{11}}$$

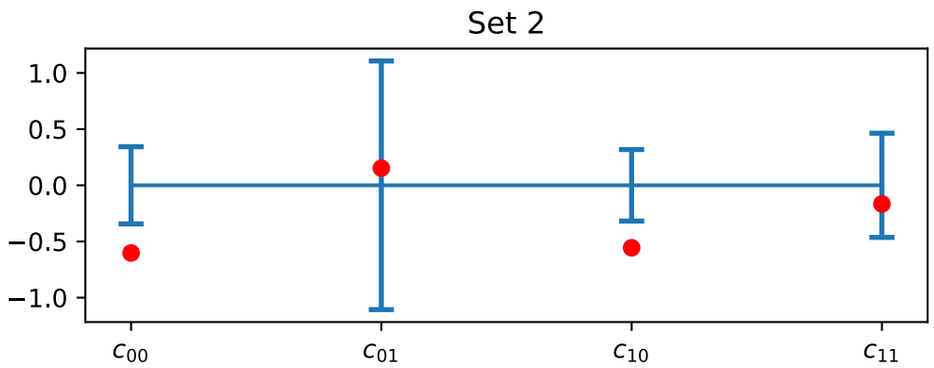
Also, assumed latent and infectious times to be known

MLEs for the three parameter sets

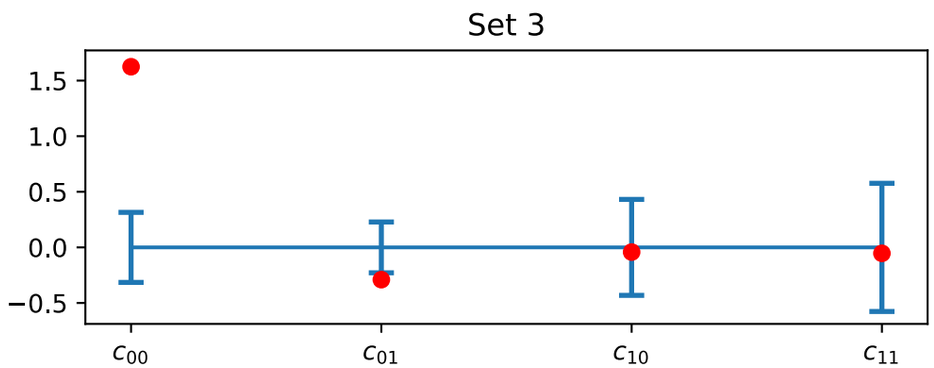
Parameter	True value	MLE mean	MLE sdev
c00	0.04	0.063	0.036
c01	0.40	0.839	0.411
c10	0.40	0.603	0.346
c11	4.00	5.452	1.825



Parameter	True value	MLE mean	MLE sdev
c00	0.25	0.629	0.216
c01	0.40	0.347	0.384
c10	0.75	1.690	0.538
c11	1.20	1.438	0.667



Parameter	True value	MLE mean	MLE sdev
c00	1.0	0.381	0.120
c01	1.0	1.409	0.322
c10	2.0	2.088	0.901
c11	2.0	2.113	1.218



Estimation of parameter ratios for three parameter sets

1: True values

parameter	value
a0	0.2
a1	2.0
b0	0.2
b1	2.0

$$\frac{a_0}{a_1} = 0.1 \qquad 0.5\frac{\overline{MLE}(c_{01})}{\overline{MLE}(c_{11})} + 0.5\frac{\overline{MLE}(c_{00})}{\overline{MLE}(c_{10})} = 0.13$$

$$\frac{b_0}{b_1} = 0.1 \qquad 0.5\frac{\overline{MLE}(c_{10})}{\overline{MLE}(c_{11})} + 0.5\frac{\overline{MLE}(c_{00})}{\overline{MLE}(c_{01})} = 0.09$$

2: True values

parameter	value
a0	0.5
a1	1.5
b0	0.5
b1	0.8

$$\frac{a_0}{a_1} = 0.33 \qquad 0.5\frac{\overline{MLE}(c_{01})}{\overline{MLE}(c_{11})} + 0.5\frac{\overline{MLE}(c_{00})}{\overline{MLE}(c_{10})} = 0.31$$

$$\frac{b_0}{b_1} = 0.62 \qquad 0.5\frac{\overline{MLE}(c_{10})}{\overline{MLE}(c_{11})} + 0.5\frac{\overline{MLE}(c_{00})}{\overline{MLE}(c_{01})} = 1.5$$

3: True values

parameter	value
a0	1.0
a1	2.0
b0	1.0
b1	1.0

$$\frac{a_0}{a_1} = 0.5 \qquad 0.5\frac{\overline{MLE}(c_{01})}{\overline{MLE}(c_{11})} + 0.5\frac{\overline{MLE}(c_{00})}{\overline{MLE}(c_{10})} = 0.42$$

$$\frac{b_0}{b_1} = 1.0 \qquad 0.5\frac{\overline{MLE}(c_{10})}{\overline{MLE}(c_{11})} + 0.5\frac{\overline{MLE}(c_{00})}{\overline{MLE}(c_{01})} = 0.63$$

Individual risk of infection

Once the parameters of the model have been determined through MLE using POMP, in particular the individual-level parameters (a_0, a_1, b_0, b_1 or the alternative parametrization c_{00}, c_{01}, c_{10}, c_{11}) we can compute individual probabilities of infection. And with those, return to the original quantity we want to calculate:

$$R = P(I_{[t-d,t]} \mid D)$$

Applying Bayes' Theorem P(A|B)=P(B|A)P(A)/P(B) give us the following expression:

$$P(I_{[t-d,t]} \mid D) = \frac{P(D \mid I_{[t-d,t]})P(I_{[t-d,t]})}{P(D)}$$

$P(I_{[t-d,t]}) =$ Marginal probability of infection in time [t-d, t]

$P(D) =$ Marginal probability of new data D

$P(D \mid I_{[t-d,t]}) =$ Conditional probability of observing D given infection in [t-d, t]

An expression for the marginal probability of infection

$P(I_{[t-d,t]}) =$ Marginal probability of infection in time [t-d, t]

The event “infection in [t-d, t]” can be expressed as the following union of disjoint events:

Infection occurs in time t-d	$P(t-d)$
Infection does not occur in time t-d and does in t-d+1	$\left[1-P(t-d)\right]P(t-d+1)$
…	…
Infection does not occur until time t	$\prod_{l=0}^{d-1}\left[1-P(t-d+l)\right]P(t)$

So the total probability is the sum of each term:

$$P(I_{[t-d,t]}) = \sum_{n=0}^{d}\prod_{l=0}^{n-1}\left[1-P(t-d+l)\right]P(t-d+n)$$

An expression for the marginal probability of infection

We have an expression for the marginal individual probabilities of infection at time t:

$$P(i,t) = 1 - \exp\Big[\Big(- \Omega_S(i) \sum_{j \in C(i,t)} \Omega_T(j)\kappa(i,j,t)\Big) - \epsilon(i,t)\Big]$$

To calculate this, we need:

- The covariates from individual i (demographics, etc)
- The list of infectious contacts C(i, t), and the covariates for each
- The environmental covariates (optional)

In our simplified model, we have:

$\Omega_S(i) = a_0 + a_1 X_1(i)$ X, Y = immune and symptomatic status

$\Omega_T(j) = b_0 + b_1 Y_1(j)$ a's, b's obtained through MLE using case data

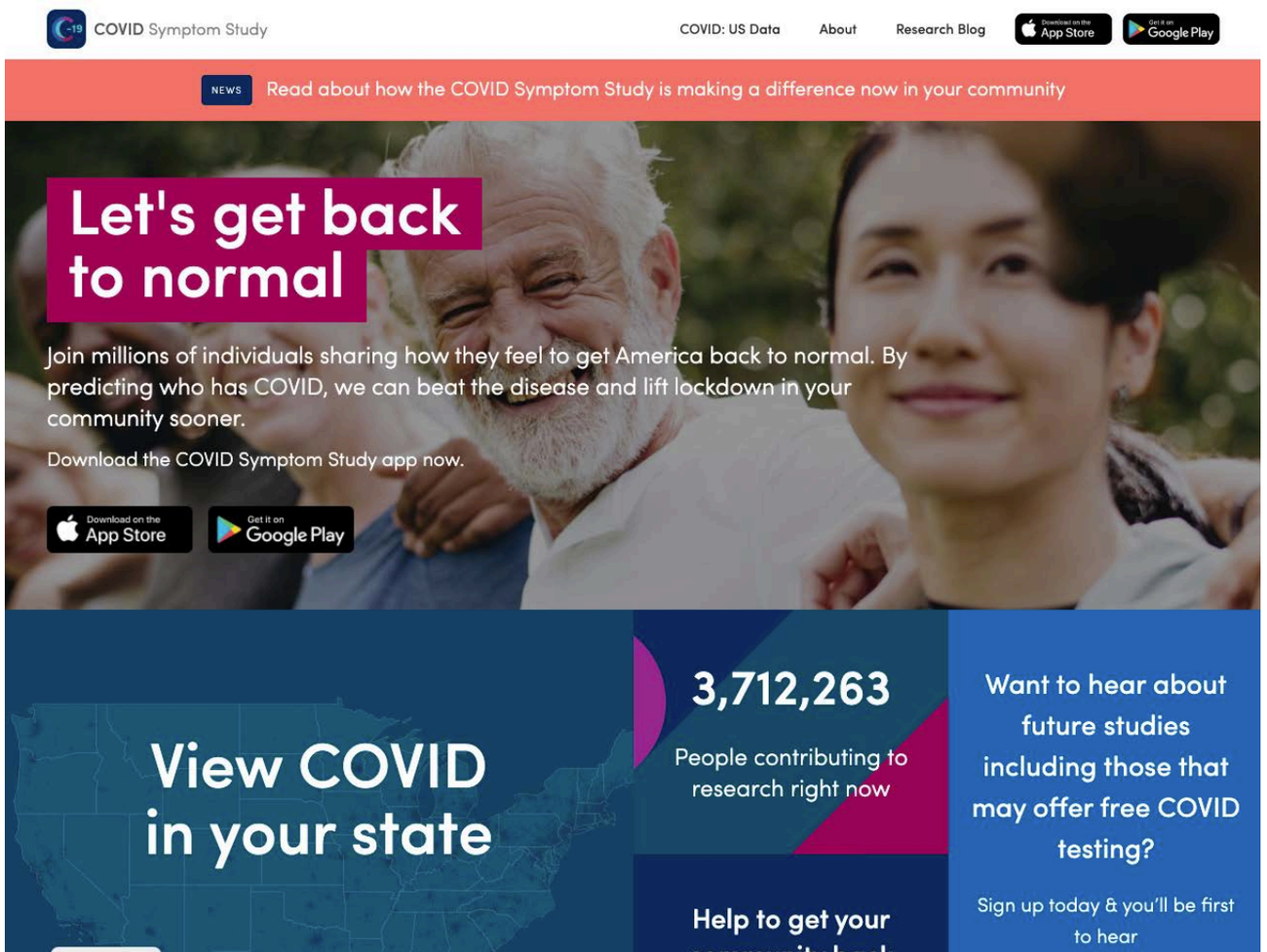
Conditioning data factor

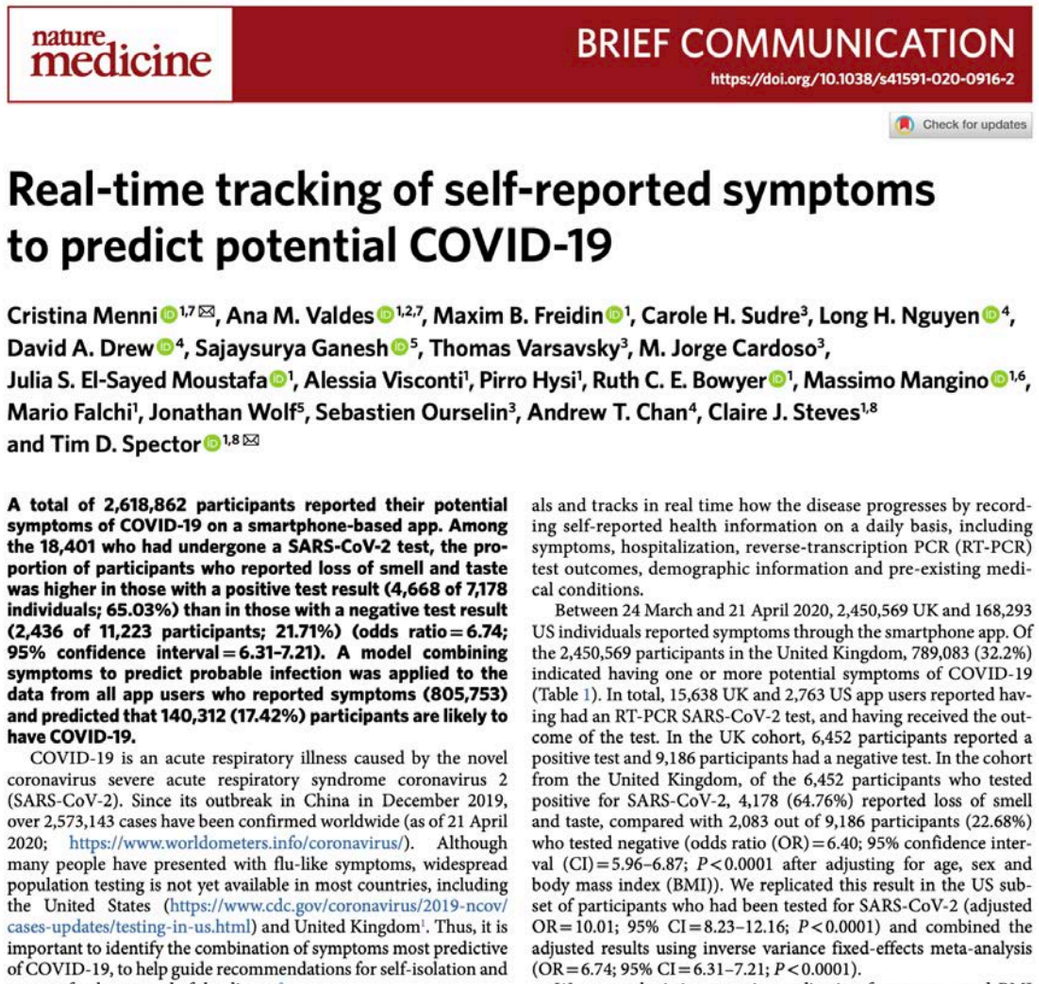
$$r_s = \frac{P(D \mid I_{[t-d,t]})}{P(D)} = \frac{P(S \mid I)}{P(S)}$$

The data D could represent self-reported symptoms, D=S, given d~14 days, $I_{[t-d,\ t]}$ is simply becoming infected at some point

There is a model that calculates P(I|S):

https://covid.joinzoe.com/us





nature medicine

BRIEF COMMUNICATION

https://doi.org/10.1038/s41591-020-0916-2

Check for updates

Real-time tracking of self-reported symptoms to predict potential COVID-19

Cristina Menni[1,7]✉, Ana M. Valdes[1,2,7], Maxim B. Freidin[1], Carole H. Sudre[3], Long H. Nguyen[4], David A. Drew[4], Sajaysurya Ganesh[5], Thomas Varsavsky[3], M. Jorge Cardoso[3], Julia S. El-Sayed Moustafa[1], Alessia Visconti[1], Pirro Hysi[1], Ruth C. E. Bowyer[1], Massimo Mangino[1,6], Mario Falchi[1], Jonathan Wolf[5], Sebastien Ourselin[3], Andrew T. Chan[4], Claire J. Steves[1,8] and Tim D. Spector[1,8]✉

A total of 2,618,862 participants reported their potential symptoms of COVID-19 on a smartphone-based app. Among the 18,401 who had undergone a SARS-CoV-2 test, the proportion of participants who reported loss of smell and taste was higher in those with a positive test result (4,668 of 7,178 individuals; 65.03%) than in those with a negative test result (2,436 of 11,223 participants; 21.71%) (odds ratio = 6.74; 95% confidence interval = 6.31–7.21). A model combining symptoms to predict probable infection was applied to the data from all app users who reported symptoms (805,753) and predicted that 140,312 (17.42%) participants are likely to have COVID-19.

COVID-19 is an acute respiratory illness caused by the novel coronavirus severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). Since its outbreak in China in December 2019, over 2,573,143 cases have been confirmed worldwide (as of 21 April 2020; https://www.worldometers.info/coronavirus/). Although many people have presented with flu-like symptoms, widespread population testing is not yet available in most countries, including the United States (https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/testing-in-us.html) and United Kingdom[1]. Thus, it is important to identify the combination of symptoms most predictive of COVID-19, to help guide recommendations for self-isolation and prevent further spread of the disease[2].

Between 24 March and 21 April 2020, 2,450,569 UK and 168,293 US individuals reported symptoms through the smartphone app. Of the 2,450,569 participants in the United Kingdom, 789,083 (32.2%) indicated having one or more potential symptoms of COVID-19 (Table 1). In total, 15,638 UK and 2,763 US app users reported having had an RT-PCR SARS-CoV-2 test, and having received the outcome of the test. In the UK cohort, 6,452 participants reported a positive test and 9,186 participants had a negative test. In the cohort from the United Kingdom, of the 6,452 participants who tested positive for SARS-CoV-2, 4,178 (64.76%) reported loss of smell and taste, compared with 2,083 out of 9,186 participants (22.68%) who tested negative (odds ratio (OR) = 6.40; 95% confidence interval (CI) = 5.96–6.87; $P < 0.0001$ after adjusting for age, sex and body mass index (BMI)). We replicated this result in the US subset of participants who had been tested for SARS-CoV-2 (adjusted OR = 10.01; 95% CI = 8.23–12.16; $P < 0.0001$) and combined the adjusted results using inverse variance fixed-effects meta-analysis (OR = 6.74; 95% CI = 6.31–7.21; $P < 0.0001$).

We re-ran logistic regressions adjusting for age, sex and BMI

Infection predictor using symptoms alone

Menni et al. constructed the logistic regression predictor for I given S:

$$P(I|S) = 1.32 - 0.01 \times \text{age} + 0.44 \times \text{sex} + 1.75 \times \text{smell and taste loss} + 0.31 \times \text{cough} + 0.49 \times \text{fatigue} + 0.39 \times \text{skipped meals}$$

$$\frac{P(S|I)}{P(S)} = \frac{P(I|S)}{P(I)} = r_s$$

$P(I) =$ overall prelavelence of infection

So we end up with the following formula for the risk of infection at time t for individual i:

$$R(i,t) = \frac{P(I|S_i)}{P(I)} \sum_{n=0}^{d} \prod_{l=0}^{n-1} \left[1 - P(i, t-d+l)\right] P(i, t-d+n)$$

Calculating the risk score in the agent-based simulations

A fist sanity check was to calculate R during ABMs, using the a's and b's coefficients estimated by MLE, and calculating the average risk for susceptible and infected individuals:

Set 1 (a_0=0.3, a_1=2.3, b_0=0.18, b_1=2.0) { R_s= 0.24±0.25 ; R_I= 0.52±0.34

Set 2 (a_0=0.4, a_1=1.3, b_0=0.65, b_1=0.43) { R_s= 0.28±0.24 ; R_I= 0.45±0.32

Set 3 (a_0=1.3, a_1=3.1, b_0=0.7, b_1=1.1) { R_s= 0.55±0.31 ; R_I= 0.63±0.31

The parameter set 3, and 2 to a lesser extent, results in higher probabilities of infection overall, so there is a smaller separation in the risk between infected and non-infected individuals.

Simulating risk-based quarantine

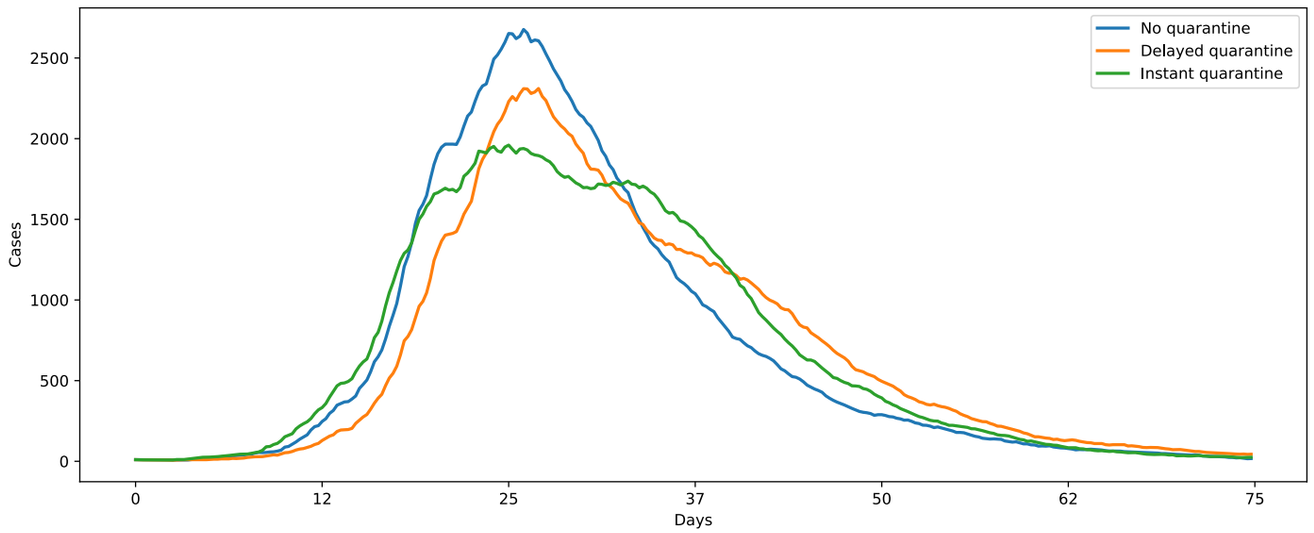
- The individual risk in the ABMs simulations was used to determine quarantining of a susceptible if the risk is over a certain threshold.
- Quarantine would last 14 days
- We considered two scenarios:

Scenario 1: there is a delay of 4 days between the risk calculation and its use to determine quarantine. The idea was to simulate the fact that infected contacts are not determined instantaneously, but with a delay due to the time it takes for the symptoms to appear

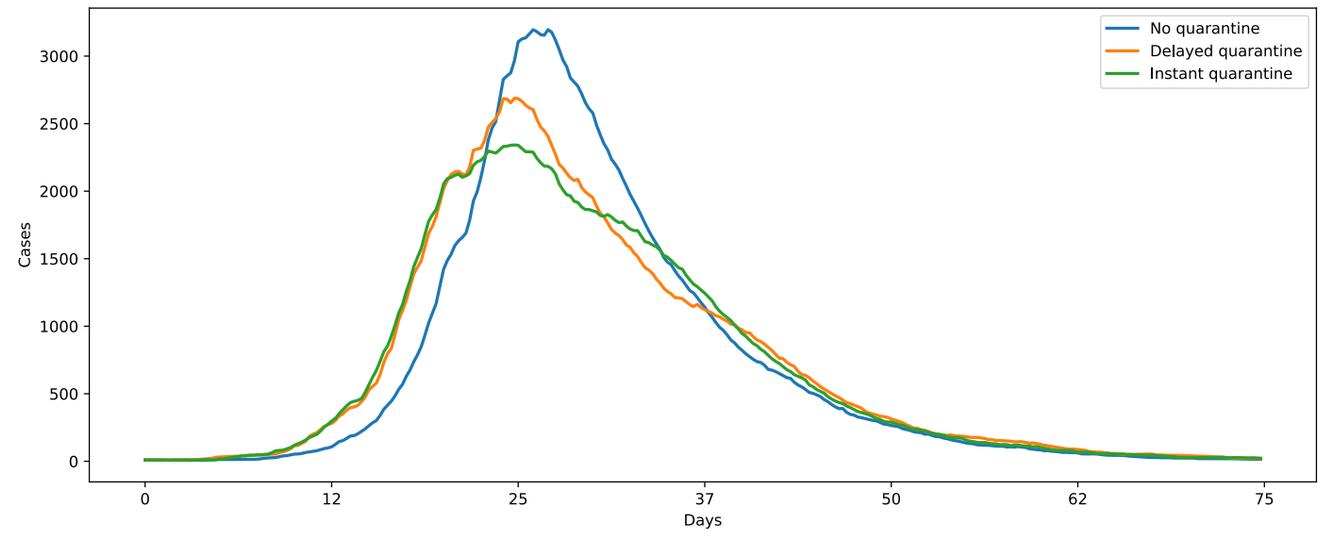
Scenario 2: The risk is updated instantly with the information of the contacts, this represents an unrealistic situation where infectious status is known upon interaction - This gives an upper bound for the performance of the method though

Simulating risk-based quarantine - Results

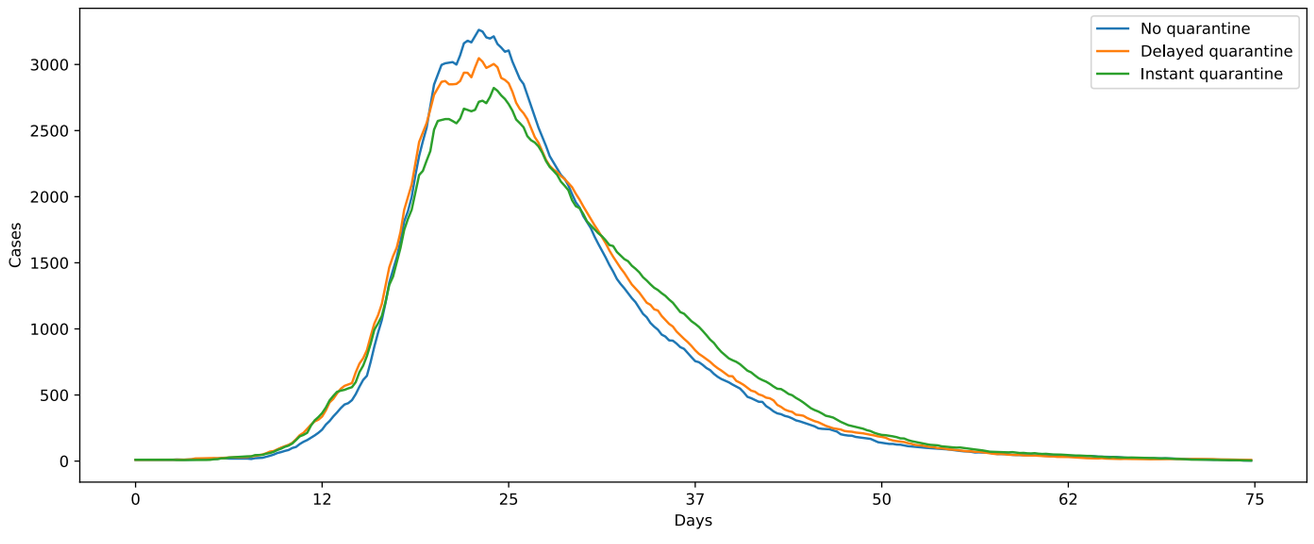
Parameter set 1



Parameter set 2



Parameter set 3



Conclusions

- We constructed an statistical inference framework that allows to obtain individual-level epidemic parameters by applying MLE to population-level case data

- We tested this framework on simulated data using an agent-based model to generate epidemic data resolved at the individual level

- We defined an individual-level epidemic risk model that depends on data such as demographics, medical condition, self-reported symptoms and contact tracing information. Additional data could be added as well

- The initial simulation experiments are promising and suggest that is possible to:

    1. Obtain good estimates for the individual-level parameters by applying MLE on the population level data

    2. Interventions based on the individual-level risks, such as quarantine, could help in lowering the peak of the epidemic, i.e.: “flattening the curve”

Limitations

- The individual-level models are very simple, including only two covariates
- Results were obtained from simulated experiments, this method need to be validated on real data
- Risk calculation need knowledge of confirmed cases to determine the exposure events, other approaches (Zdeborova, Bengio) are based on the idea of estimating the probabilities of each state (susceptible, infected, recovered) of all individuals using message-passing algorithms