Lecture 22.
Friday 11th December, 2020.
Compiled: Thursday 31st December, 2020.

0.0.1 R_0 from the cluster size

We will now provide another tool to estimate R_0: we will try to infer it from the **cluster size**. This is useful specially when *infections are not highly transmissible*, that is to say when we do **not** observe an **exponential growth** of cases, but only sporadic infections.

Let us discuss which are the possible infections with $R_0 < 1$, in particular what are their main features. These are for example *zoonotic infections*, such as $MERS$[1] that was transmitted through dromedaries and camels and whose mortality rate was high, or infections that are *close* to *eradication threshold*, such as measles. It is important to analyze such diseases and to understand how far we are from the epidemic threshold, since the higher the number of infected individuals, the higher the risk of a mutation in the pathogen or a decrease in vaccine uptake.

We want to understand how to deal with this problem: a first approach is to *assume* that the *few cases* we observe are **linked** to a **unique cluster**.

We define as a **cluster** all cases generated by the so called *index case*. A cluster is an entity like the one in Fig. 1, whose most important quantities are:

- **Index case**: infection caused by an external source.
- **Offsprings**: cases infected by the index case.

In the case where we deal with many clusters, the distribution of their size s depends on R_0 according to $P(s|R_0)$.

Figure 1: Estimation of R_0 using a cluster approach when observing a small amount of cases.

Some examples of $P(s|R_0)$ probability density might be, depending on the assumptions we make on the heterogeneity of the network:

- All infectious individuals behave equally and generate on average R_0 transmissions. The number of *offsprings* is k and are distributed as a Poisson:

$$k \sim \text{Pois}(k|R_0) \implies P(s|R_0) = \frac{(sR_0)^{s-1}e^{-sR_0}}{s!}$$

- Introducing a *continuous-time* SIR dynamics, since R_0 is below 1 stochastic effects are important. The size of the cluster therefore follows a *Markovian birth and death process* of the kind:

$$P(s|R_0) = \frac{(2s-2)!}{s!(s-1)!}\frac{R_0^{s-1}}{(R_0+1)^{2s-1}}$$

[1] Middle East Respiratory Sindrome (MERS) coronavirus, 2013. Breban, et al. The Lancet 2013

However, it might happen, specially when cases are few, that many of them are not reported: data we have is almost for sure biased in this sense. Hence we need to account also for **under-reporting**. Each case may go unobserved with probability p_{miss}: as a consequence if a cluster has real size s, we may observe a cluster of size $o \leqslant s$.

The **probability of observing a cluster** of size $o \geqslant 1$, given R_0, is Binomial wrt missed cases:

$$P(o|R_0, p_{miss}, o \geqslant 1) = \frac{\sum_{s \geqslant 0} P(s|R_0)\binom{s}{o} p_{miss}^{s-o}(1 - p_{miss}^{o})}{1 - o(o = 0|R_0, p_{miss})}$$

where $P(s|R_0)$, can be any of the aforementioned probabilities. And the **likelihood**:

$$\mathcal{L}(o_i|R_0, p_{miss}) = \left(\sum_i o_i\right)! \prod_i \frac{1}{o_i!} P(o|R_0, p_{miss}, o \geqslant 1)$$

where, as an example, the first factor sums over all the possible sizes of the cluster that we may have observed, and all the possible permutations ("!").

Exploiting the tools we have just introduced epidemiologists were able to provide an estimate for R_0. It is interest to notice that at early spread, in February this was $R_0 \simeq 2$. More likely the R_0 for the Eastern strain of Coronavirus was computed. Indeed a couple of months later scientists were able to estimate R_0 for the European strain, and it resulted to be $R_0 \simeq 3$. This can explain why in UK the Chinese strain was overtaken by the European one, despite it appeared earlier and number of infected people was way larger.

0.1 Incubation period estimation

One other relevant quantity we need to estimate as soon as possible is the **incubation period**. It is really important since, under some assumptions, we can relate it to the *generation times* distribution from which to compute R_0. In order to do it, high quality data is needed. However, if there is none, we can still provide an estimation by using similar diseases one, even though obviously this analysis might lead to wrong outcomes.

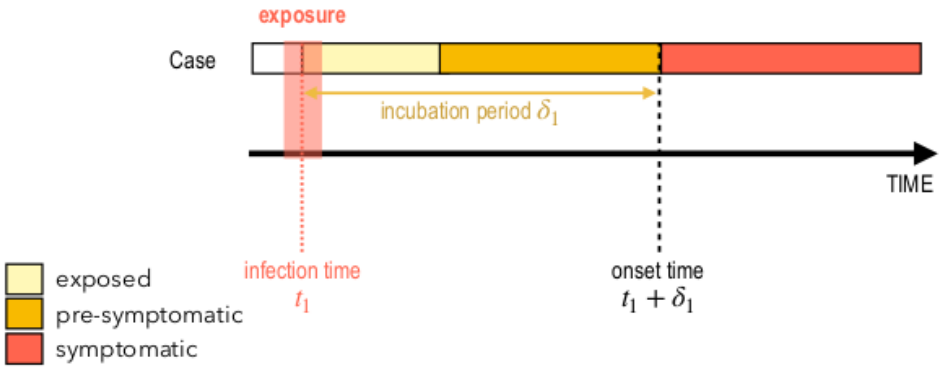

Figure 2: Incubation period analysis.

The **incubation time** is defined as the time elapsed between the **onset** of symptoms and the *infection time*, namely when the **exposure** occurred (Fig. 2). The former is easy to collect, for instance when a person has been visited by a doctor and start showing any symptoms. In addition, incubation time is also a measure of the delay in the response of restriction policies in infection curve. However, exposure time is very difficult to retrieve, nonetheless is very important: pre-symptomatic

phase relevance is even higher for COVID-19, since pre-symptomatic individuals turn out to be really infectious. Some ways deal with this problem are through **contact tracing**, **case investigation**. Once a case is confirmed, contacts are investigated. They are contacted, isolated and go through clinical and virological assessment. This allows us to collect infector-infected pairs and more likely when contact occurred.

One other approach one may want to follow to estimate the *incubation period* is to proceed with the **analysis of travelling cases**. With referral to Fig. 3 and COVID-19 outbreak: Backer et al.[2] analyzed 88 cases detection starting from January, 20th through January, 28th. Their travel history (to and) from Wuhan was known, as well as their symptom onset date. During this early stage of the epidemic, it is most likely that travellers were infected in Wuhan. Consequently, their time spent in Wuhan can be assumed to be the duration of exposure to infection without any contact tracing procedure. As said, we know *for sure* the date of the onset of symptoms, but we need to infer when infection occurred. One should note that the shorter the stay in a risky area, the more precision we have in inferring the duration of infection period. After providing these estimations, data was fitted and compared using 2-parameter continuous distributions supported on a semi-infinite intervals, such as Gamma, Weibull and Log-normal distribution. Later, they proceeded to maximize the Likelihood and, by the means of a strictly positive flat prior for the two parameters, since there was no guess about their value, they tried to infer them. The most credible estimation for the **incubation period** was:

$$\text{incubation period } \sim 6.4 \text{ days} \qquad \text{C.I. } 95\%[2.1, 11.1] \text{ days} \tag{1}$$

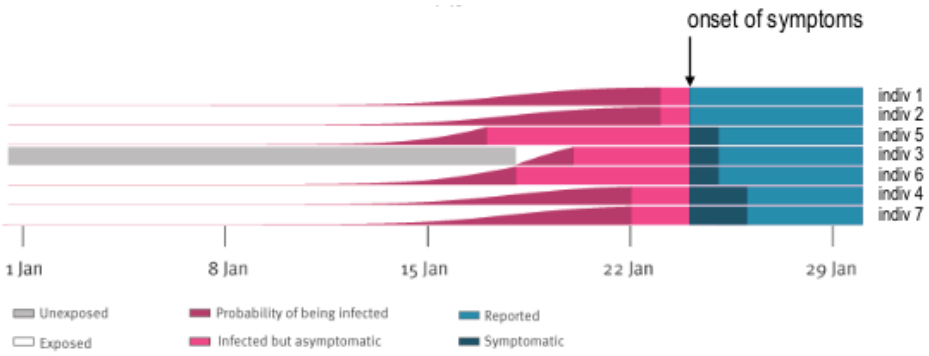

Figure 3: Pioneer analysis of travelling cases for COVID-19 outbreak.

0.2 Generation time and serial interval estimation

Let us recall the definition of the **generation time**: it is the time elapsed between the moment when the infector was infected and the moment when he infects someone as one can see in Fig. 4). It is indeed really difficult to obtain, nonetheless it can be computed thanks to the **serial interval** T_s that are easier to compute. It is the time *elapsed* between the *onset of symptoms* for two individuals, where we assume one to be the infector and the other one the infected. Therefore, the serial interval is a random variable that is linked to both generation times distribution (τ) and the incubation periods distribution (δ) in the following way:

$$\tau_S = \overset{\text{onset time case 2}}{(t_2 + \delta_2)} - \overset{\text{onset time case 1}}{(t_1 + \delta_1)} = \underset{\text{serial interval}}{(t_2 - t_1)} + (\delta_2 - \delta_1) = \tau + (\delta_2 - \delta_1) \tag{2}$$

[2]Backer, Klinkenberg, Wallinga. Incubation period of 2019 novel coronavirus (2019-nCoV) infections among travellers from Wuhan, China, Euro Surveill. 2020;25(5).

where we sample $\tau \sim w(\tau|\theta_2)$, $\delta \sim g(\delta|\theta_1)$, $\tau_s \sim f(\tau_s|\theta_1, \theta_2)$. This indeed is true *on average*, and can be done since we assume that the average generation time and the average infectious duration are equal. Therefore, **serial interval** is often used as a **proxy** for the **generation time**, but with a warning! This argument is valid since $\langle\delta_1\rangle = \langle\delta_2\rangle$, therefore $\langle(\delta_2 - \delta_1)\rangle = 0$, but we should recall that $\tau_S \neq \tau$: *on average* they are the same ($\langle\tau_s\rangle = \langle\tau\rangle$), but their *variance* is different ($\sigma_{\tau_S} > \sigma_\tau$) [3]!

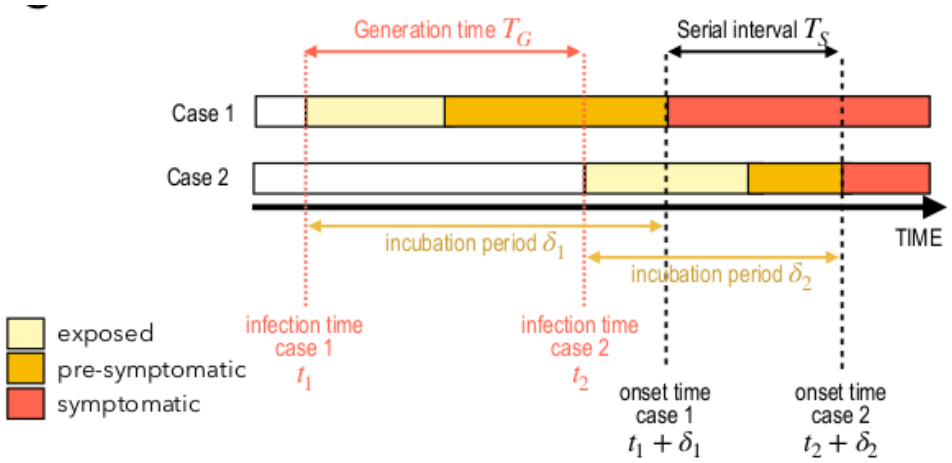

Figure 4: Generation time and serial interval relation.

Moreover, the variance of the generation times distribution f is greater than the one of w of infectious duration: indeed we can observe infections even caused by pre-symptomatic individuals: **pre-symptomatic phase** is important for transmission. In addition, variance of f is also greater than the variance of G and this might lead to an underestimation of R_0 since it holds that, relating the latter to the serial interval distribution, $R_0 \simeq e^{GT_G - (1/2)G^2\sigma^2}$.

As an example, let us consider the case of COVID-19 data and proceed with the estimation of the generation times distributions. Researchers[4] reported 91 confirmed cases in Singapore and 135 in Tianjin, and relied their paper on previous estimates of incubation period, whose distribution is:

$$\delta \sim g(\delta|\theta_1) = \text{Gamma}(\delta|\theta_1) \tag{3}$$

Data we are talking about consisted in information about infector and infected pairs. We know that **generation times** is $\tau_S = \tau + (\delta_2 - \delta_1)$, therefore distributes according to the convolution between the two distributions h and w:

$$f(\tau_S|\theta_1, \theta_2) = \int_{-\infty}^{+\infty} w(\tau - x|\theta_2)h(x|\theta_1)dx = \mathcal{L}(\{\tau_{s,i}\}|, \theta_2\theta_1) \tag{4}$$

where $x = \delta_2 - \delta_1$, $x \sim h(x|\theta_1)$ and $\{\tau_{s,i}\}$ is the set of generation intervals observed.

Once we have proceeded to the estimation of θ_2 we can numerically compute:

$$P(t_2 < t_1 + \delta_1) = P(\tau_S < \delta_1) \tag{5}$$

that gives us the **proportion of cases generated by pre-symptomatic transmission**. The final results are the following for the COVID-19:

$$\text{generation time } = 5.20 \text{ days} \qquad \text{C.I. } 95\% \ [3.78 - 6.78] \text{ days} \tag{6}$$

[3] Indeed, $Var(X + Y) = Var(X) + Var(Y)$.

[4] Ganyani, Kremer, Chen, Torneri, Faes, Wallinga, Hens. Estimating the generation interval for coronavirus disease (COVID-19) based on symptom onset data, Euro Surveill. 2020;25(17):pii=2000257

and the *proportion of pre-symptomatic transmission* being around 50%, that is a high number. This implies that infected people tend to infect more if they are not reported, since this number is not negligible. A natural consequence is that, when we spot a case, it will have already infected almost half of the people it would infect during the whole infectious period.

The opposite case actually tells us why Ebola did not spread: despite a really high mortality rate ($60 - 70\%$), the infectiousness was not constant in time and was proportional to the time individuals stay infected as long as with the severity of the symptoms. Those individuals could be successfully isolated in time as soon as any symptom shows up, thus avoiding the spread. In this case, **isolation** for infected individuals is a successful solution to stop the disease.

0.3 Infection severity

One of the most important quantities we may want to compute is the **infection severity**. Indeed, we need to know the full spectrum of symptoms including the ones of **pauci-symptomatics** and **asyntomatics**. It gives us a key to interpret data and observations, moreover allows us to estimate the **actual number of infections** and its **cumulated**: this might be helpful to understand how many people are susceptible to the disease at the moment. Another reason to explain why it is important is if, once recovered, we acquire some sort of immunity, even temporary, actual susceptible people become less in number. Moreover, it is important to compute the **infection fatality ratio** and to provide **projections on hospital needs**: we will know how many people would need ICU or hospitalization, based on the fraction of people that develop mild or more serious illness. These fractions are normalized wrt *total* number of infected people, including asymptomatic, so one should easily understand how much hard is to compute these ratios.

On the other hand this is a really difficult task to estimate these proportions, but this goal can be pursued through **contact tracing studies**, routine testing, or cohorts[5]. Indeed these are actually very difficult and expensive procedures, specially in terms of resources, and may require even formation for people which were assigned by these task. Moreover some natural experiments could be done. That is the case of the *Diamond princess*, a cruise ship in Japan where some people were tested positive. Therefore, others were quarantined on the boat for some time: eventual infections were registered, reported their eventual symptoms and finally led to the hospital. However, this brought someone to raise some ethical objections over this scientific non conventional approach. Another experiment was due to repatration flights, where people were tested in airports while returning home in addition to the reporting of their eventual symptoms.

In conclusion, the *true proportion* of asymptomatics of COVID-19 is still not certain ($\sim 20\% - 50\%$) and there is a strong dependence on age.

0.4 Reproductive ratio R_t

Let us compute the most relevant quantity of the epidemiological process, namely the **Reproductive ratio** R_t. This is different from the **basic reproductive ratio** R_0 and we recall that the latter is average number of cases an infectious individual infects in a *fully susceptible population* during the course of his/her infectious period. Moreover, this is *computed at the initial stage of an outbreak*. However, this is not

[5]We choose a sample of people/volunteers among a population, and proceed to follow their evolution in time in terms of infected, recovered, times distributions and so on.

realistic, being the number of susceptible dependent on time, for instance due to acquired immunity.

Instead, the **Reproductive ratio** R_t is the **average number of cases** an infectious individual infects **at a given time** t during the course of his/her infectious period. This is the *natural extension* of R_0 to the later outbreak stage. It is an **indicator** of how the transmission potential of the epidemic evolves in time.

The **dependence in time** has to be introduced to take into account that as the outbreak unfolds the population is not fully susceptible anymore. We retrieve the simple version of the Kermack and McKendrick model, where the new infections:

$$I(t+\delta t) = I(t) e^{\left[\mu \int_t^{t+\delta t} \left(R_0 \frac{S(t')}{N(t')} - 1\right) dt'\right]} \tag{7}$$

That in turn can be locally approximated to:

$$I(t+\delta t) \simeq I(t) e^{\delta t \mu (R_t - 1)} \tag{8}$$

where $R_t = R_0 \frac{S(t)}{N(t)}$.

In reality things are more complex: R_t does not change only due to the depletion of susceptible (immunity building): it might change as effect of interventions, behavioural change of population ($\langle k_t \rangle$ term):

$$R_0 = \frac{\beta \langle k \rangle}{\mu} \longrightarrow R_t = \frac{\beta \langle k_t \rangle}{\mu} \frac{S(t)}{N(t)} \tag{9}$$

Now, we drop the assumption of the simple SIR and put ourselves in a more general framework. We will follow two different paths and interpretations to compute R_t.

Cori method

The first one is the method used almost in all Western countries, and was developed by **Cori** et al.[6]. The main pro is that it captures immediate changes in number of contacts thanks to lockdown or other restrictions, and this makes it really useful. It starts from a generic generation time distribution, namely $w(\tau)$, and it is based on the Lotka Euler equation:

$$I(t) = \int_0^\infty I(t-\tau) \beta(\tau) d\tau \tag{10}$$

where $\beta(\tau) = w(\tau) R_0$. One should note that the number of infected people at time t depends on the number of infected at time $t - \tau$ and on the model parameters at time τ!

Let us generalize the last expression. In order to do so, we make the assumption that the reproductive ratio varies in time, being the infectiousness time dependent as well $\beta(\tau, t) = w(\tau) R_t$. Therefore introducing R_t:

$$I(t) = \int_0^\infty I(t-\tau) w(\tau) R_t d\tau \tag{11}$$

note as it was introduced the dependence over the absolute time t through the variable R_t.

Reverting the last equation and discretizing the time we obtain:

$$R_t = \frac{I_t}{\sum_{s=1}^{t} I_{t-s} w_s} \tag{12}$$

[6]Cori, Ferguson, Fraser, Cauchemez, A New Framework and Software to Estimate Time-Varying Reproduction Numbers During Epidemics, American Journal of Epidemiology, 178, 2013.

where the denominator indicates the total infectiousness of individuals that are infectious at the time t. According to this **interpretation**, R_t *is the average number of secondary cases that each infected individual would infect if the* **conditions remained as they were at time t**. This method is called "**real time method**", since it links the actual situation to the past one through the generation times distribution finally providing an estimation for R_t. Given the total number of newly infected people today we can assume that background situation will not change in the close future, finally trying to predict the future number of infections that is function of R_t. The R_t analysis in France is illustrated in Fig. 5.

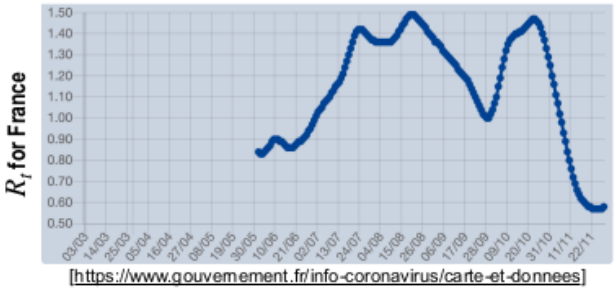

Figure 5: R_t analysis for France throughout 2020. Changes are mainly due to restriction and variation of $\langle k_t \rangle$ and most likely not due to immunity building.

Wallinga method

The second interpretation we introduce is the one made by **Wallinga** et al.[7]. It is used to infer *who infected whom* from available information. When we have an incidence curve, the **only information** regarding a case is the **date** in which a case was recorded. Hence, the relative **probability** p_{ij} that case i is infected by case j, given the time elapsed $t_i - t_j$ depends on the **generation interval** and assuming a case is registered the date in which it was infected is:

$$p_{ij} = \frac{w(t_i - t_j)}{\sum_{i \neq k} w(t_i - t_k)} \tag{13}$$

The denominator denotes all the case we have, and acts as a sort of normalization term. Sometimes, however, this is not realistic since there might be some delays in reporting the cases. We now introduce the **cohort reproduction number**:

$$R_j = \sum_i p_{ij} \tag{14}$$

it counts the *average number of secondary transmissions caused by a cohort* that was infected at time step t. It is the infection potential of a cohort (might be even a single individual) at time t. We recall that a cohort is *not* a cluster, but is a group of cases that we follow from now on in the future, as a sample individuals. This method takes into account naturally the variability in the transmission potential of all individuals, since we are not making any assumption for it. We are trying to quantify **number of transmission generated by cases at time** t.

The **price to pay** for using this method is that it can be used **only retrospectively**: we are trying to compute when the secondary cases, generated by the infected at time t, have already been infected.

[7]Wallinga, Teunis, Different Epidemic Curves for Severe Acute Respiratory Syndrome Reveal Similar Impacts of Control Measures, American Journal of Epidemiology, 160, 6, 2004.

Method comparison

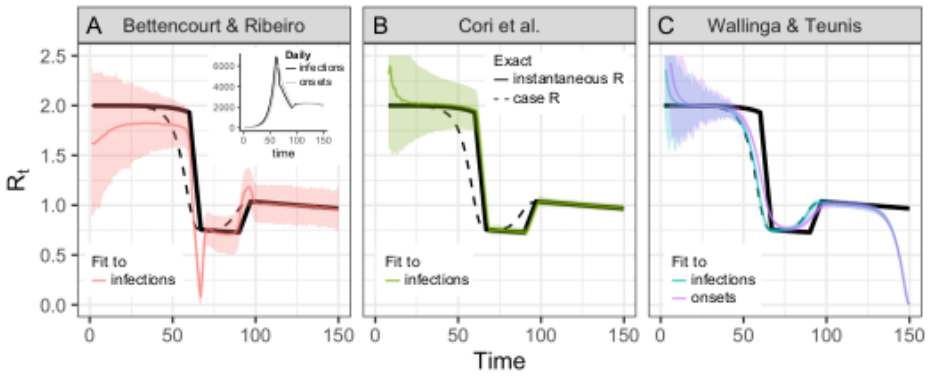

Figure 6: Difference between Bettencourt, Wallinga and Cori approaches for computing R_t. We see that Wallinga's has a cut-off. This behavior was expected since this approach analyzes future cases that are manually set to 0, nonetheless this kind of estimation is wrong. Cori's instead is more realistic.

We now want to make a little comparison between the two above described models (Fig. 6).

Cori describes the number of new infections at time t and link them to *past infections*. Hence, it **look backwards in time**. On the other hand, **Wallinga** relates the number of today's infections at time t to the *future cases* they will generate. Hence, it **looks forward in time**. The last model however cannot be used in real time analysis, but it sticks more to the R_t definition.

Let us use these approaches with a more practical task. Let us compute the actual life span of some individuals that was born in 2013. Using Wallinga's retrospective approach, we would need to wait until all individuals die out, and then proceed to estimation. Conversely, using Cori's approach, we could estimate the same quantity by assuming that death rates in the future will be similar to present ones. It is a more physical approach: indeed it takes into account actual conditions.

In reality, we do not have as data the time of infection, and generally we **report only delayed effects** (symptoms) of some events that had occurred some time in the past (Fig. 7). So, every report contains an **intrinsic delay** within itself, and even a delay of a single day can be risky for certain type of diseases. In this framework, hospital data is currently the best one since it relies on an uniformity of testing. In addition, outpatient testing data [8] is useful, but its availability depends on many variables such as "population testing" policies for a given country. In conclusion, if we want to compute R_t we either build a **compartmental model** and, sticking to it we estimate parameters using *maximum likelihood*, or alternatively, we build a **statistical model** and try to infer infection times and all related quantities by the means of *deconvolution*. Another last possibility is to build a model accounting for **latent states**, where the observables become *hospital admissions date* and the latent state is the one that immediately follows the infection. However, since we do not observe the time of infection, we can still use maximum likelihood to infer R_t.

[8] People that get tested under medical prescription or needs to travel and therefore have to be tested.

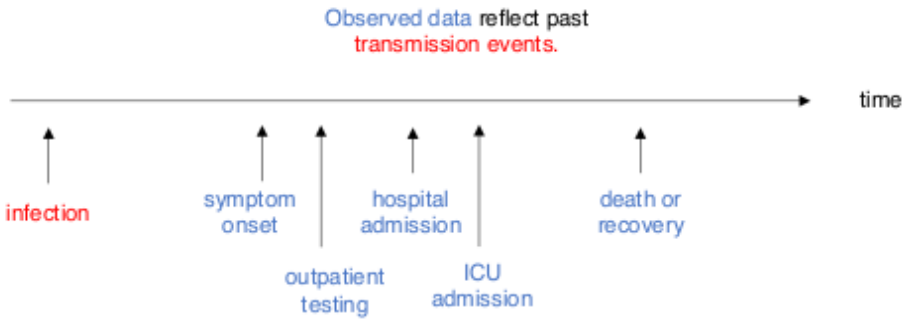

Figure 7: Generation time and serial interval relation.