The Mathematics and Statistics of Infectious Disease Outbreaks MT3002 S20
Department of Mathematics
Stockholm University

Per Helders, project 2

The multiple potential exposures problem when estimating incubation periods

Abstract

The multiple potential exposures problem is the problem concerning what role the data of infected individuals, who report having had more than one potential infectious contact during the approximate incubation time, should have when estimating the incubation time more precisely.

This report will investigate two problems:

1. Of which kind and size are the consequences of excluding individuals with multiple potential infectious exposures from the data when it comes to estimating the mean and the distribution of lengths of incubation period?
2. What other methods, apart from excluding these individuals, can be used in order to address the multiple exposures problem?

Introduction

The treatment of these two problems will here give rise to two separate theoretical models. The first (Model 1) will show the size of and the reason for the bias that occurs when we are excluding the individuals mentioned above from the data, given different parameters input. The second (Model 2) will provide a proposal for handling cases where individuals have reported multiple potential infectious exposures. Together with that model we will shortly discuss some other proposals that have been used or discussed.

This report will focus on the estimation of the mean length of incubation period and it´s distribution, as these measures is crucial for estimating the generation time and the basic reproduction number, which are two concepts of importance when deciding how to handle an outbreak of an infectious disease. Hence, it will not provide further details concerning the specific effects on those two concepts caused by our (hopefully more precise) knowledge of the length of the incubation time.

This report will be accompained by R code. This code covers both Model 1 and Model 2, with different comparisons to other methods and also some simulations (see **Simulations** and **Comments on the code**).

Background and purpose

A central feature when judging consequences and preventives regarding infectious desease outbreaks as COVID-19 is to have knowledge about the incubation time, i.e. the time between an individual´s exposure for infection and the outbreak of symptoms. But since the occasion of exposure is not an observation, but rather an estimation, so is the incubation time. Additionally, different diseases have not only different incubation times, but also different distribution of incubation periods. For instance, COVID-19 is said to have an approximate incubation time of between 2 and 15 days with a mean around 5,5 days, which is a rather large span. An important tool for estimating the incubation time is back-tracing. An individual, testing positive for (e.g.) COVID-19, is interviewed about earlier contacts. Sometimes it can be quite obvious when he or she got infected, other times there are zero (not discussed here) or more than one possible occasion reported when infection possibly can have occured, and we don´t know on which of these occasions the infection actually was transmitted. How do we handle these cases statistically?

One way that has been used, for instance at the ebola outbreak in West Africa 2014-2015 (discussed in Britton/Scalia Tomba 2019, page 8), is to simply exclude these cases from the data. This means fewer data point, which is a disadvantage, but the rest of the data seems unambiguous which would benefit a reliable result of the statistical analysis. However, it can be shown (and we will show this) that doing this creates a bias, which does certainly *not* benefit the reliability of the analysis. One purpose of this report is to actually show this theoretically, and also to provide a code-based tool, or model, for judging the size of this error, given certain chosen data. There are other ways to go than excluding multiple cases from the data. Thus, the second purpose of this report is to discuss some other ways to handle this problem, and also present a second model to minimize bias and to actually use the data provided by these cases.

Incubation time

In this report the expressions ”incubation time” and ”incubation period” are treated as synonyms.

It should be made clear, that in order to provide a tool for estimating the incubation time accurately (Model 2), we need the incubation time and it´s distribution (in some form) as one of the input parameters. Obviously, this looks problematic. However, we can always describe the different sizes of error obtained by excluding multiple potential infection exposures cases, given different (hypothetical) initial incubation times. Additionally, even an initially approximate or even inaccurate estimation of the distribution of the lengths of the incubation period will show to be useful, because for new data, estimates of the distribution of lengths of incubation periods for data including multiple exposure reports, can still be made more precise. Iteration of this procedure will further refine the estimations (see also attached code **Model2.R**).

Input data

The data we need are reports from infected individuals concerning the time (or times, if multiply exposed) they suppo-

sedly have gotten infected. From this data the mean incubation periods and their distribution can be deduced.

We shall assume that different occasions for potential infectious contacts are spread uniformly random and independent of whether an individual is susceptible or latent. Note that these data are obtained by interviews, and dependent of how an infected individual judge the situations where he/she has experienced what is thought of as potential infectious contacts.

When looking at typical data of incubation times for COVID-19 in form of a graph (for example graph 1), we can see that this distribution looks somewhat like continuous probability density functions as gamma or log-normal distributions. But Model 1 and Model 2 will not be based on these kinds of generalizations, as here the purpose is limited to the multiple exposures problem. Thus, Model 1 and Model 2 will use only basic data as described above. The probability functions that will be used are not expressed as continuous functions, but as discrete, with the relation between domain and range expressed by these sets themselves directly (see def: $p(t)$, Model 1, page 4).

Regarding cases reporting a third or more potential infectious contact, this will in Model 1 not be treated at all, as an individual would be disqualified from the data already with two potential infectious contacts. For Model 2 though, this situations will be taken into consideration.

Data can, for instance, be obtained from tables and/or diagrams showing number of cases as a function of the length of incubation period. Input of probability of getting infected or infection rate etc will not be made. What is relevant here is not that risk, but the number of reports of multiple potential contacts, and (in Model 2) their multiplicity and their location on the timeline, and thus their impact on estimations of risk and rate.

Model 1 will judge the bias of mean and distribution of incubation time caused by excluding multiple cases. Model 2 will produce mean and distribution where multiple cases are not excluded, but dealt with.

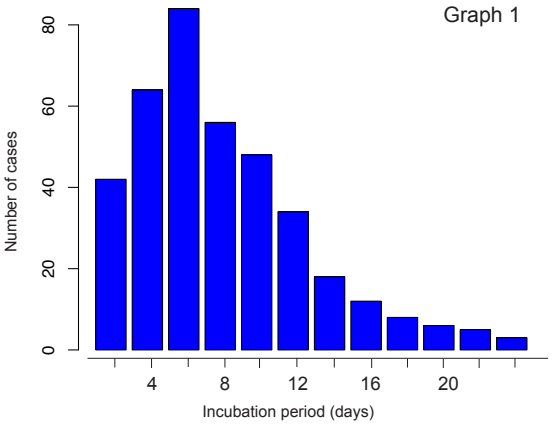
Description of bias when excluding multiple exposures from data (Model 1)

The reason for why excluding multiple cases leads to a bias is that the probability for having a second potential exposure is bigger the longer incubation time a certain individual has, which then leads to the fact that excluding multiple cases is to exclude cases with longer incubation times, which will lead to a bias of the mean towards underestimation of, not only incubation times, but also generation times and serial intervals, as well as underestimation of R_0 (the basic reproduction number).

Two basic quantities are crucial for estimating the bias:

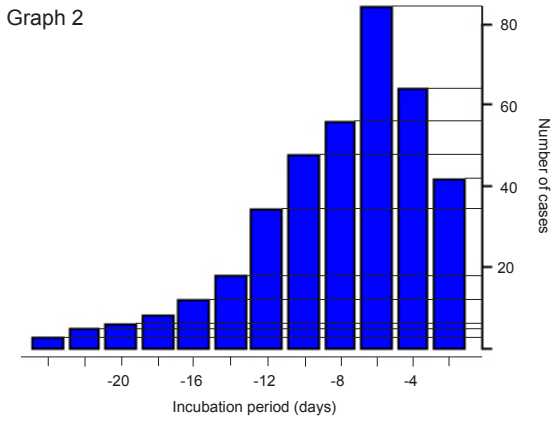
1. The length of the incubation period.
2. The number of individuals that is associated with the different lengths of incubation periods (i.e. the distribution).

An ordinary diagram showing the number of cases as a function of the length of the incubation period can be used to obtain these quantities (see graph 1). Think of Model 1 as



this: Consider a material of only single exposures. Distribute randomly a number of second cases, and then exclude these cases. Compare the means before and after exclusion.

We start with graph 1 as an example. This is a hypothetical diagram, but designed to be plausible for COVID-19 (see **Comments on the data**, page 8). We are going to reorganize this information in two ways. First, we want to look at these data backwards in time, so we will flip the whole graph 180° horizontally (see below). We define time 0 as the time when symptoms break out, and we are looking backwards. Secondly, we observe that the two quantities can be composed to one by multiplying them. The new quantity can, as looked upon from a geometrical point of view, be regarded an area (or rather several areas, one for each period of incubation showed in graph 2), but with the addition of also showing the number of cases. This is illustrated in graph 2 by the lines going from the top of each bar to the y-axis. The bars together with the lines and the two axis create rectang-



les, one for each period of incubation time. The fractions consisting of the area of each rectangle divided by the area of all rectangles together will directly give to the probabilities for a second infectious exposure of an individual to be associated with a certain incubation period. Thus, the product of the number of cases and the length of incubation time gives the probability. To show this even clearer, we will display these rectangles side by side, not one onto the other as in graph 2. This will give us a graph (with a y-axis that is split up) like the one displayed on next page (graph 3). The table in graph 3 shows the data behind the graphs 1-3.

The coloured areas together can be seen as the total probability for next infectious contact to hit an individual that is already infected (which all are, thus = 1). If a "second" infectious exposure will hit in the red area, then excluding this case will have the consequence that the mean will be biased towards smaller values, and the opposite will happen if the next infectious exposure ends up in the blue area. All upcoming "second" infectious exposures will be distributed randomly over the graph. We can see already by overlooking the graph that the red area is bigger than the blue, hence this shows that a bias towards smaller mean will occur if we exclude cases of additional infectious exposures. Some calculations give that the probability p for an additional infectious exposure to hit the red area is equal to the red area divided by the total area: p = 0,58 (calculations: see code).

Conclusion: In this example, with the data presented, the probability that excluding an individual having a second infectious exposure will result in a smaller mean is p = 0,58.

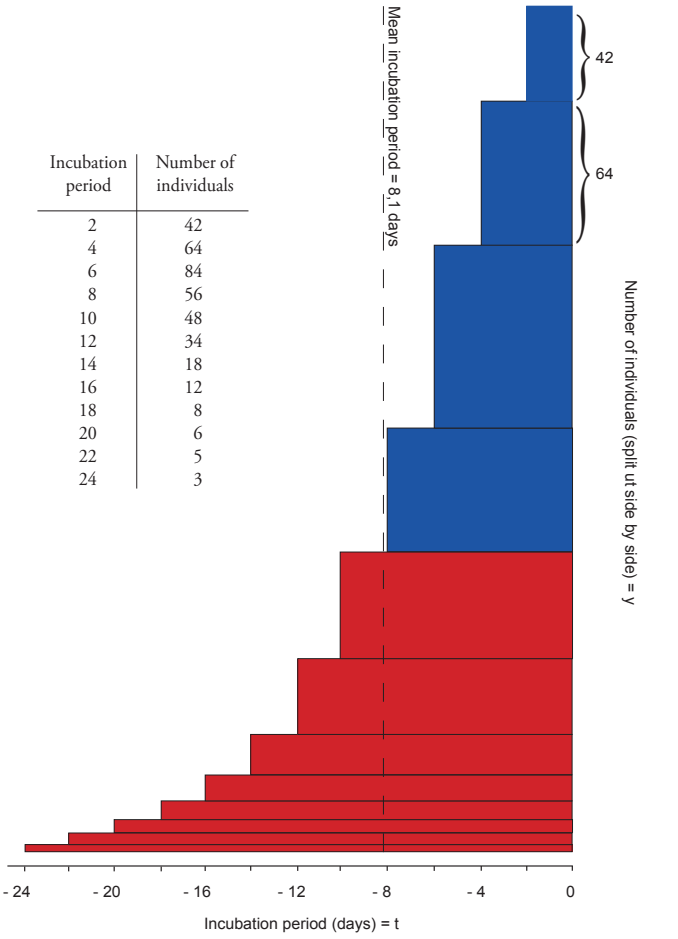
Note: The *area* corresponing to a certain incubation time corresponds to the probability that next second infectious contact occasion will occur within that incubation time (if the total area is set to 1).

We want to ask if this something general, i.e. is it always true that we will have such a bias regardless of which data we put in? The answer is yes.

Probability graph of incubation periods
(reorganisation of graph 2)

Graph 3

Incubation period	Number of individuals
2	42
4	64
6	84
8	56
10	48
12	34
14	18
16	12
18	8
20	6
22	5
24	3



Proof:
Consider any non-uniform distribution of a probability mass function. There will always be a mean value E_k of the set of values in the domain governed by the range. Consider two subsets of these values; those bigger than E_k (with a mean value E_r) and those smaller than E_k (with a mean value E_b).

Then we have $E_r > E_k > E_b$. The area as shown in graph 3 is $A_k = t_i \cdot y_i$ for any t in the domain and any y in the range.
Any mean value in the domain is

$$E_k = \frac{t_k \cdot y_k + t_{k+1} \cdot y_{k+1} + \ldots + t_i \cdot y_i}{y_k + y_{k+1} + \ldots + y_i} \quad \text{thus} \quad E_r > E_k > E_b \implies$$

$$\frac{t_r \cdot y_r + t_{r+1} \cdot y_{r+1} + \ldots + t_m \cdot y_m}{y_r + y_{r+1} + \ldots + y_m} > \frac{t_b \cdot y_b + t_{b+1} \cdot y_{b+1} + \ldots + t_n \cdot y_n}{y_b + y_{b+1} + \ldots + y_n} \implies$$

$$\implies m + n = i \implies t_r \cdot y_r > t_b \cdot y_b \implies A_r > A_b.$$

Then we have $E_r > E_k > E_b \implies A_r > A_b$. □

Next step is to estimate how big this bias will be. We cannot just use p = 0,58 from our example, because the bias just proved will occur also in the red and blue groups respectively. To minimize this bias we can intuitively realize that it is a matter of grading on the x-axis in graph 2. Actually, we can apply the proof above for any interval on the x-axis, always partition it in two (by the mean of the interval), of which one produces a bigger mean value in the domain, and the other a smaller. There is a limit of this process, reasonably the minimum quantity in the domain, which would most often be 1 day if we deal with incubation time.

In our example we have grading 2 days as interval. We will use this to calculate the size of the bias as an example, and then do a generalization. Assume that in our data (see table graph 3), we have a total of 100 multiple cases (which is nearly 25% of those initially infected). Some calculations will then give the new mean ≈ 7,1. Compare this with the original 8,1 (see graph 3, and for calculations see the R code).

Result: The new mean is 7,1 days, which is ≈ 12,4% shorter incubation time as before excluding the multiple cases.
Conclusion: This, together with the proof above, shows that the finer gradation we have (in terms of incubation time), the more of the bias occuring when excluding multiple cases becomes visible.

Probability of having multiple cases

The bias will of course be affected by changes of the fraction of cases being multiple. The smaller fraction of cases that is multiple, the smaller the bias will be.

In the example we used a hypothetical number occu-

rences of multiple cases. As mentioned 100 cases (in our example) is nearly 25% of the population. Is this reasonable?

In fact, it is difficult to estimate this number. It will always be a result from interviewing (subjective) infected individuals. In the case of Covid-19, there are numerous symptoms that are shared with other diseases. With some other diseases, it might well be easier to determine whether it is probable that a certain contact might have been infectious.

The fact that we have a number of symptoms associated with Covid-19 makes the uncertainty concerning what may have been a infectious contact bigger. This will lead to a bigger bias of the kind investigated here, as we may get a big fraction of individiuals who have experienced contact that they regard as possibly infectious. Also general contact rate, social density etc. has great importance for the occurence of potential multiple exposures.

Generalization model 1

We will now generalize these calculations. Let us think of a data material that contains only single exposures. This material will most likely be biased, but that is not crucial for this purpose. Let N be the whole population. Establish the mean incubation time $E(N)$ for this material. It is given by

$$E(N) = \frac{t_1 \cdot y_1 + t_2 \cdot y_2 + \ldots + t_i \cdot y_i}{y_1 + y_2 + \ldots + y_i}$$

where t_1 is the first value in the domain (length of incubation time as positive integer), y_1 is the number of individuals associated with that length of incubation time and i is the number of incubation time length intervals (on the x-axis).

Let´s randomly distribute a number n of secondary cases, which will all have the same probability to hit any given point on the coloured surface in graph 3 (equivalent to that any point will have the same probability to get hit). We consider only second cases, not third or more, because it is not relevant. These are treated as second cases and not contributing to n. Thus, with n we mean number of cases with more than one potential infectious contact (= those that will be excluded).

What we now want to do is to compare $E(N)$ with $E(N - n)$, where $(N - n)$ is the population after excluding the multiple cases.

We introduce some definitions and abbreviations to save space:

$A_1 = t_1 y_1$ (Definition of the first of the areas in graph 3)

$T_A = t_1 y_1 + t_2 y_2 + \ldots + t_i y_i$ (Definition of the total area in graph 3)

$p_1 = \frac{A_1}{T_A}$ (The probability for a second potential contact to hit the incubation time bin = t_1)

$z_1 = y_1 - p_1 y_1$ (Definition of the new value for the number of individuals associated with length of incubation time = t_1, i.e. after excluding multiple cases)

$P_A :=$ The probability for a new second case to cause bias of $E(N)$ towards smaller values (= underestimation).

Now we are ready for the general definition of model 1:
A. The new mean after excluding multiple cases.
B. The general probability for exclusion of multiple cases to cause underestimation.
C. The size of bias of the mean length of incubation time.
D. The new distribution of the lengths of incubation time.

A. $E(N - n) = \frac{t_1 \cdot z_1 + t_2 \cdot z_2 + \ldots + t_i \cdot z_i}{z_1 + z_2 + \ldots + z_i}$

B. $P_A = \frac{(A_1 + A_2 + \ldots + A_i) \mid t > E(N)}{T_A}$

C. The numerical size of bias of the mean length of incubation time = $E(N) - E(N - n)$

D. The distribution of the lengths of incubation periods is given by the new probability mass function (PMF) $p(t)$:

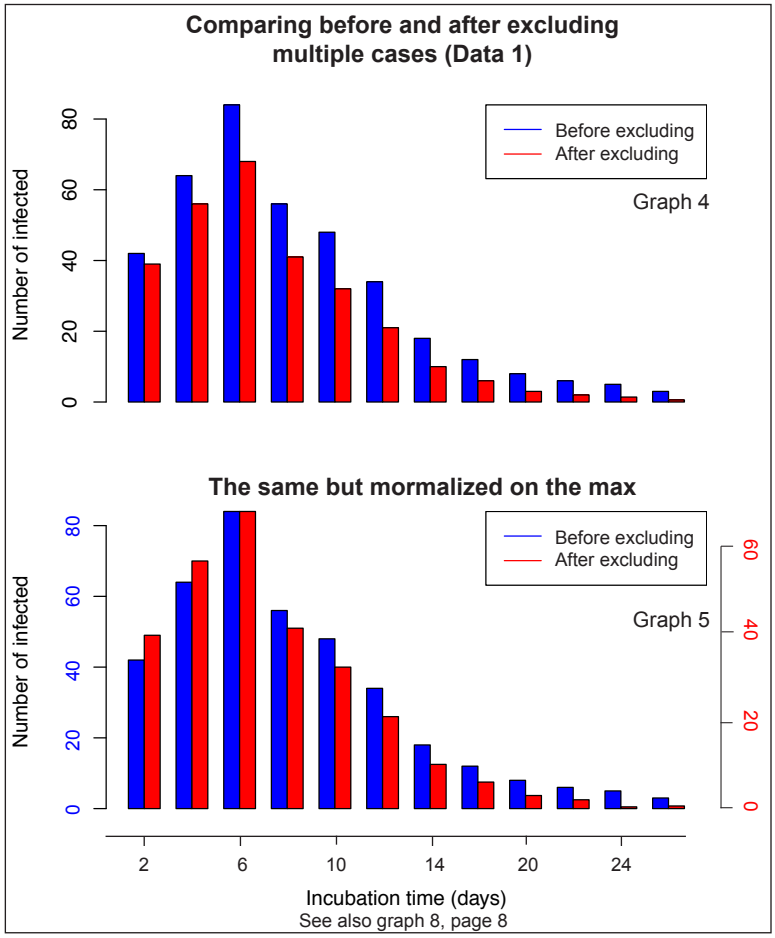
$p(t) := \{t_1, t_2, t_3, \ldots, t_i\} \mapsto \{z_1, z_2, z_3, \ldots, z_i\}$

We can look visually at the difference by applying the data from our example and compare graph 1 with the new graph. Graph 1 has blue bars and the new graph has red bars (see next page).

In graph 4 we can see the difference in terms of excluded data points (red bars). In graph 5 the data are the same, but the new values normalized to the maximum of the blue bars. This shows clearly the offset of the main, with longer red bars for the smaller values, and shorter bars for the larger values. Note that the scale is different for the red bars compared to the blue, so this is used only to show the offset (there are still more cases with incubation time of 2 days after the exclusion, as it was before. See also graph 8).

How to treat the multiple exposure problem? (Model 2)

In Britton/Scalia Tomba (2019) some alternative proposals are given and discussed for the treatment of this problem. These proposals are not put forward as useful methods in their paper, rather they are discarded, but here we will call them Method 1-4 and use them (only) as comparison objects.



These methods are:

1. To only consider the earliest exposure when dealing with individuals with multiple exposures.
2. To only consider the most recent exposure for these cases.
3. To consider all potential infectious contacts as independent cases, i.e. one individual with (say) two exposures is treated as two individuals.
4. To consider all potential exposures as having the same probability.

We will discuss these alternatives shortly, but also add yet another alternative, which we will call **Model 2:**

5. For all multiply exposured individuals in the data, replace the potential incubation times for all multiple exposures at each such individual by a single value *v*. Then treat these individuals as single exposures occured at $t_k = v_k$.

In Britton/Scalia Tomba (2019) we are given the following model for the likelyhood L of infection transmission at potential time points $e_1, \dots, e_k$:

$$L(e_1, \dots, e_k, s) = \left[e^{-\int_0^s \lambda(u)du} \prod \lambda(e_i) \right] \times \left[\sum_{i=1}^{k} p(1-p)^{i-1} g(s-e_i) \right] \quad \text{(Eq 1)}$$

where at time t, the rate of infection exposure is $\lambda(t)$ and the probability of infection at exposure equals p (the same p for all contacts with all infectives) and $g(t)$ is the density distribution of the incubation period (s = onset of symptoms).

The discussion in Britton/Scalia Tomba concerning the alternatives 1-4 above shows:

1. This alternative will lead to an overestimation, as is quite obvious. Without any doubt, a number of multiple exposures will *not* have their earliest time of potential exposures as the actual one.

2. This alternative is the opposite to 1, and therefore biased the other way because of the fact that surely not all multiple exposures will have thier latest one as the actual one. So this will lead to underestimation of the incubation time.

3. Britton/Scalia Tomba points out that alternative 3, compared to (Eq 1), gives rise to an underestimation of the incubation time (along with the serial interval and the R_0) caused by the fact that in (Eq 1) lower weight is given to the shorter incubation times. Bias is also caused by a biased (too big) number of individuals.

4. This alternative also leads to an overestimation of values in less likely intervals.

The model given by Britton/Scalia Tomba gives us the distribution of the likelyhood for an individual to get infected over the potential occasions. We will, in model 2 (see below), handle this slightly different.

The alternative to exclude multiple cases is also discussed in Britton/Scalia Tomba, with the same conclusion as here, i.e. that individuals with longer incubation periods will be more likely to have multiple exposures, and thus their exclusion leads to underestimation of the mean of the incubation time.

Note that when we were looking at Model 1, the situation could be described as on the coloured surface, each point had the same probability to be hit by a second potential exposure, but as areas are not of the same size, these sizes govern the probability for any multiple exposure to be of a certain length of incubation time. Here in Model 2 we are looking at another probability. The probability for an individual to get a second potentially infectious exposure is not the same thing as the probability that one of, say, two potential infectious exposures of one individual is more likely to be the true than the other. This second probability is governed only by the distribution of the incubation period lengths, i.e. the probability mass function, PMF, or $g(t)$ (probability density function, i.e. a contiunous function) in Britton/Scalia Tomba (see under 5. below).

It is mentioned by Britton/Scalia Tomba that there are several more detailed models to use for estimating the incubation time, but they require more detailed information from the data. As mentioned above, concerning for example COVID-19 it is difficult to know to what extent back-traced data are reliable. In alternative 5 we will not pay attention to the probability for an individual to *get* an infectious

contact. Our purpose is only to find a method to handle multiple cases such that they cause as small bias as possible.

5. As just mentioned, alternative 5 will not say anything about the probability to catch an infection, but only about how to optimize v (thus we do not use $\lambda(t)$ or p found in (Eq 1), only $g(t)$ but as a discrete function, which we call $p(t)$). The data that will be used may contain a high degree of variation regarding the number of individuals who claim to have multiple exposures. And as we have seen, the fraction of multiple exposures in the material affects the bias when excluding multiple cases (or applying any of the methods 1-4). But if we can find a good value for v, then the fraction of multiple exposures becomes less important for the result. So, how can we find the value v?

One tempting way is to let v be the mean of the multiple potential incubation times. We will call this Method M. It seems reasonable to think that this will give a better result that any of the methods 1-4 (or 1-3 if we interpret 4 as precisely this way to find v), and better than excluding the multiple cases. We can see that Method M is exactly the same as Model 2 using a uniform distribution for the PMF. It will never be the true value (as this will remain unknown) but the overestimations and the underestimations would cancel each other out, as it is plausible to think that, if we have for instance a group of individuals who claims to have two potential exposures at t_1 and t_2, then for half of them the mean (or any value v located between t_1 and t_2) will be an overestimation and the other half will be underestimated. However, this leads to overestimating of incubation time and thus underestimating of R_0. As multiple cases are randomly and evenly spread over the data, longer incubation times would be favored (for similar reasons as why multiple cases in general is more likely to be overrepresented in longer incubation times, see Model 1. See also code **Model2.R**).

We would come closer to the truth by taking into consideration the probability for t_1 and t_2 respectively to be the true value. An example: An individual have symtoms outbreak at time 0, and is diagnosed with COVID-19. Back-tracing leads to a report where the individual had one potential infectious exposure 15 days ago, and a second 6 days ago. It is obvious that if we take the mean of these two, it will be less adequate than weighing the probability for a 6 days incubation time as heavier than a 15 days incubation time, given the probability mass function of the distribution of lenghts of incubation periods. Here is the crucial point where we will have to use the data we are searching. But even if we do an approximate estimation of the distribution of the length of the incubation periods, this estimation would shows to be better that saying that t_1 and t_2 are equally likely. This should be part of the considerations behind the equation provided by Britton/Scalia Tomba (1), but in the present report formulated in a simpler way with fewer parameters and thus with less complexity (it is not invovling $\lambda(t)$ and p), but still more accurate that the methods 1-4 and the method of excluding multiple exposures.

To find the value v we start by looking at the probability mass function of the distribution of lenghts of incubation periods (PMF) = $p(t)$. This is to be regarded as a discrete function and formulated as a set, in our example case consisting of 21 probabilities $\{p_1, p_2, \ldots, p_{21}\}$ as a function $p(t)$ of the set of 21 lengths of incubation period in days {2, 3, 4, ... , 22}.

We exclude the possibilities for incubation time to be shorter that 2 days or longer that 22 days. In fact, we can restrict ourselves to use the incubation time interval [2,14] or so (for the COVID-19 case). The risk by doing that can be assessed, but for the moment we just assure not to underestimate R_0 by underestimating the incubation time and generation time. Our data that we used so far, graphs 1-3, does not provide the grading $t_{k+1} - t_k = 1$, but we will use other data in the R code, with a maximum incubation period of 14 days, with the grading $t_{k+1} - t_k = 1$ (see e.g. **Data2.xlsx**). It is likely that a data material (for COVID-19) consisting of the results of interviews, would be detailed down to single days of incubation time. Of course this can be different for other diseases. Then we define v as follows (**Model 2**):

$$v = \frac{\sum_{k=1}^{i} (p(t_k)\, t_k)}{\sum_{k=1}^{i} p(t_k)} \qquad (2)$$

where $(t_1, t_2, \ldots, t_i)$ is the timeline of the multiple potential incubation times, and v is the new single incubation time (in days) substituting the multiple ones. Here the time t is given as a positive integer (not negative as in graph 2 and 3). This way we can weigh v on the probabilities given, and thus obtain values closer to the reality than if we had used only the mean incubation time of the multiple potential exposures. However, v will not be equal to the true time-point of infection, but it will be closer than if t_1 and t_2 were treated as equally likely, and deviations would typically cancel out so that the probability for a bias $\rightarrow 0$ (for large populations). We assume that there are no significant changes of R_0 or transmission rate during the incubation time.

By running the code **Model2.R** we can see (e.g.) that the conclusions regarding Methods 1-4 holds. Regarding

Summary of mean and standard deviation for substitution value v. Simulation with 10000 cases, each with 5 potential exposures (see page 8, column 1, Simulations)

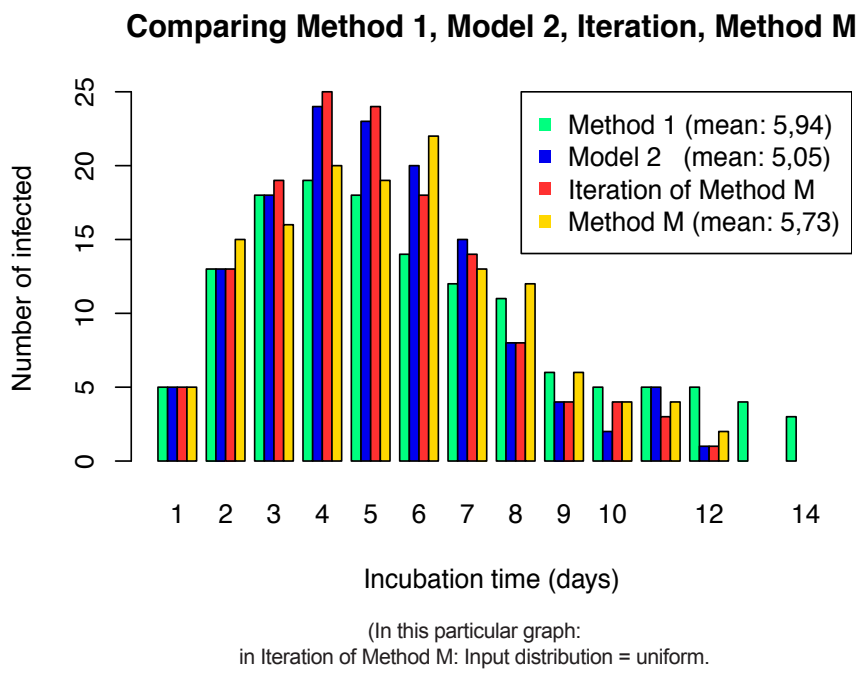
	Mean	SD
Substitution value v for longest and shortest incubation time only	4.79	1.32
Substitution value v for the three incubation times when longest and shortest are excluded	6.40	2.14
Substitution value v for all five potential exposures incubation times	5.53	1.39

multiple (more than two) potential exposures we can also compare what happens if we have, say, five potential multiple exposures per individual and apply Model 2 only on the two values that consist of the longest and the shortest incubation, or if we apply Model 2 on the three values that lies between the longest and the shortest incubation times, and finally also when we apply Model 2 on all given values. This is done in the code **Simultip.R** and the results are shown in the table page 6 and in graph 7 (see also under Simulations below). Note that the values in the table on page 6 only consider cases with, in that case, five potential exposures (e.g. no single exposures). Thus these are not final values for any complete data, as exclusion of single exposures will cause overestimation for the same reason as exclusion of multiple exposures causes the opposite.

Conclusions model 2

We can use (2) as a model in order to estimate the incubation time given a material with multiple potential exposures reported by back-tracing. When v is settled for all individuals with more than one potential exposure in the data set, we can derive mean and distribution and thus create a new graph showing the adjusted probability mass function.

We have used an approximate PMF to estimate v, but if we repeat the process (by using the code) with the new PMF as the tool to have a yet newer v, we will see that that we get a mean very close to the first one. **Conclusion: The mean obtained when using Model 2 is to a much bigger extent based on the information contained in the data than on possible deviations following from the PMF input. Even with a inaccurate input PMF we will (by iteration if needed) by Model 2 get a good estimation of v already with two iterations. This can be shown by running the attached code. Note how red and blue bars are following each other (graph 6, see also R code: Model2.R).**



(In this particular graph:
in Iteration of Method M: Input distribution = uniform.
2 iterations gave a mean approx. 5,06)

Graph 6

General conclusions

As we have seen (and also mentioned in the introduction), both model 1 (theoretical estimation of bias cause by excluding cases with multiple potential infectious contacts) and model 2 (estimating a single substitute for the multiplicity in these cases) are dependent on some idea of the distribution of the lengths of the incubation time periods. Surely, this seemed problematic, but to what extent?

Looking at model 1, we can draw some conclusions already with a uniform distribution. A uniform distribution would mean that all the bars in graph 1 would have the same height. This would lead to a bigger underestimation than in our example, but we note that we can, anyway, see that there *is* a bias. The same would be seen if a normal distribution had been used. **In fact, as the proof on page 3 is valid for any distribution, we can conclude that there will always be a bias when excluding multiple potential cases.**

Concerning the size of this bias, model 1 is strictly theoretical, comparing what happens if a given situation (distribution) gets some "second" cases. Of course, it would not happen that way, as these events has already occured. In fact, we do not typically know if the distribution we use have excluded multiple cases or not. It seems reasonable to argue that this information would be possible to get with the data, but efforts made for the present report to get this kind of information from Folkhälsomyndigheten resulted in the answer that, for the majority of the COVID-19 cases, they do not even have any information about day for outbreak of symptoms, which implies that reliable data about potential multiple exposures would be even more unlikely.

Model 2 seems to have the same problem: the weighing of v presupposes an estimated distribution. However, by running the code we can see that this problem is minor. In fact, a handy way (with fully accurate results) is to apply for example Method 1 on the input data in order to obtain a input PMF for Model 2. With already one iteration the result is accurate (this can be examined in the code **Model 2.R**). The reason for this phenomenon is that **the input data contains such large fraction of the essential information** that even a uniform distribution of the input PMF would do (after iterations), and even more so the distribution obtained with the help of Method 1.

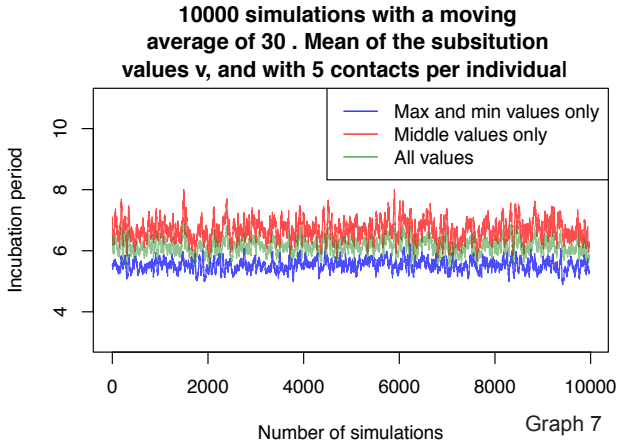
The conclusions also leed to the insight that back-tracing is important. Without back-tracing we cannot hope for reliable estimations of central concepts as R_0 and generation time. So one conclusion may be that efforts should be done to, already before an outbreak occurs, secure that the resources and the knowledge to quickly make back-tracing are present.

Simulations

Attached to this report are some simultions of which one is more extensive (**Simultip.R**). In this simulation incubations periods are spread randomly over a population of optional size. The number of exposures is also optional. The purpose of the simulation is to investigate differences concerning *v* (see model 2) for different segments of the incubation periods. What *v* will be obtained if we apply Model 2 on

1. only the shortest and longest of the given incubation periods?
2. the rest of the values except the shortest and longest?
3. all values?

Intuitively we can think that the three mean values for the simulations over *v* would have the result that 2 is longer than 1 (because the bias described in Model 1 will occur within the interval of the two extremes). Alternative 3 would be in the middle, and would be the most accurate. Alternative 2 would also have the biggest variance because that *v* is obtained from data with the smallest span (on average). This is confirmed by the simulation, see graph 7 and the attached code **Simultip.R**. **Note:** the results of the simulation (and graph 7) cannot be directly compared to results of Model 2, as the simulation only deals with multiple potential exposures of which all are, say, five contacts (no single contacts). The results from one example of simulation are shown in the table on page 6 and in the graph below (in this graph the variance is reduced by a moving average, but all alternatives can be observed in **Simultip.R**).



Graph 7

Some simulation can also be done in **Model1.R** to see different scenarios for different input of number/fraction of second potential exposures (Model 1).

Comments on the data

Some data sets are attached as xlsx-documents. Data1 is a simple distribution of incubation periods with no multiple potential exposures and Data2 is a larger material with some dual multiple exposures that is treated in the code for Model 2. A third data set (Data3) is more complex, with up to 6 exposures.

The data sets are hypothetical, but can be regarded as likely for COVID-19 (except Data3). The fact that they are not authentical has no impact on the purpose for this report and the calculations or conclusions performed.

Data1 and the Data in graph 1-3 and the example on page 3 in the report concerning Model 1 is obtained (and somewhat modified) from Yang, Dao, Zhao et al (2020).

Comments on the code

All comments can be found in the code itself. The code for Model 2 is in two parts; **Model2.R** is for input data with only single and double potential exposures. **Model2.1.R** is for input data with 1-6 potential exposures (as e.g. Data3). **Simultip.R** handles optional number of simulations and optional number of exposures (from 3 - 6).

References and attachments

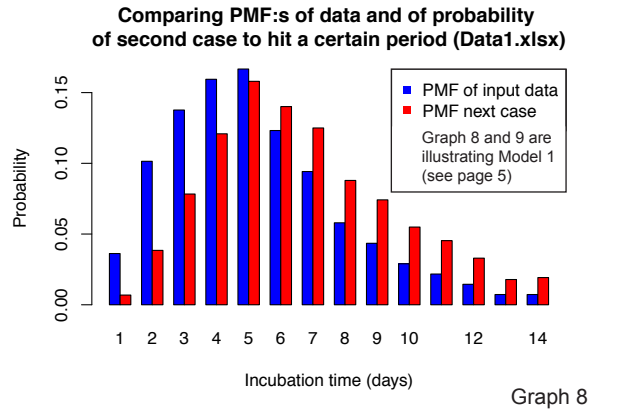
References

Estimation in emerging epidemics: Biases and remedies (T Britton and G Scalia Tomba 2019, Interface, royalsocietypublishing.org)

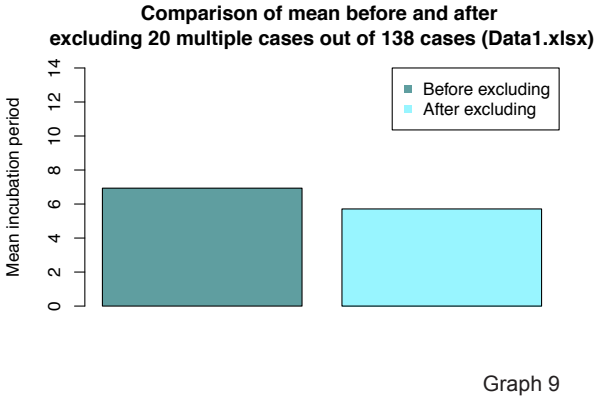
Estimation of incubation period and serial interval of COVID-19: analysis of 178 cases and 131 transmission chains in Hubei province, China (L Yang, J Dao, J Zhao et al, Cambridge Core, Cambridge University Press 2020, DOI: https://doi.org/10.1017/S0950268820001338)

Attachments

Calculations: see relevant R code.
R code: Model1.R, Model 2.R, Model2.1.R, Simultip.R
Data: Data1.xlsx, Data2.xlsx, Data3.xlsx



Graph 8



Graph 9