**Adapting our infection date estimation process to cater for explicit ‘modern’ infection date estimation in terms of ‘date of detectable infection’ (DDI) intervals which effectively provide non-naïve priors for the recency-assay-based infection date estimate (IDE)**

*29 April 2023*

Shelley’s crucial question on the call raises two critical questions: (see below for definitions and other details)

1. Is our use of CEPHIA (or other primary) data sets for the estimation of P\_R(t) curves correctly adapted to use the language of DDI?
2. How do we handle the explicit uniform priors for the estimate of the DDI, rather than just treating the recency test date both as time 0 and the LPDDI?

On issue 1: We need to confirm that the data points being read in which report recency assay results per individual, treat the time variable as “days since the individual’s estimated DDI” – correctly handled behind the scenes before the calibration data is imported into the P\_R(t) estimation process. This doesn’t affect the primary code which is handling the estimation process – but it needs to be checked.

The remainder of this document is about adapting the existing code so that it handles explicit ‘infection date estimates’ in the sense that have been developed by various papers led Shelley, Eduard (attached – Joseph, you might want to spend a few minutes at some point refreshing on this).

We introduced the notion of DDI – date of detectable infection. It is the (usually not known explicitly) date on which infection would first be detectable IF the person in question were being tested by a super-sensitive assay which picks up a viral load of one copy per ml. The point being that we ran into lots of objections when attempting to talk about ‘date of actual infectious exposure’ as being just a little too far into the not directly observable aspects of the process.

So it’s important to understand the use of some form of

‘earliest plausible date of detectable infection’ (EPDDI)

And

‘latest plausible data of detectable infection’ (LPDDI)

Where ‘earliest plausible’ refers to estimation windows – and it NOT the same use of ‘earliest’ as implied by the meaning of that word hidden in DDI. This is a terribly important point. DDI always refers to the first/earliest date on which infection would have been detectable. The use of the word ‘earliest’ or ‘first’ in

‘earliest plausible date of detectable infection’

Refers to the earliest date of an estimation window, in which DDI is estimated to lie. Hence there is nothing weird or incorrect about the full meaning of

‘latest plausible data of detectable infection’

Which is actually answering the question:

‘Given available data, what is the **latest** possible date, within our constructed estimation window, on which we believe the infection would **first** have been detectable by a specified super sensitive assay.

It’s super important to be clear that DDI has nothing to do with when infection was in fact detected, or at which scheduled visit it should have been detectable, or anything else with an operational meaning. DDI is just a day in the history of a person, post infectious exposure, when the infection could in principle be detected by a specific kind of viral load assay. It is understood that when we estimate DDI, we do so based on data available at the time, and the estimates are uncertain. It is also understood that DDI is not a precisely known time after infectious exposure – and we accept that from case to case, the time between infectious exposure and DDI will vary – perhaps substantially – but we keep in our back pocket that there is some usual delay between infectious exposure and DDI. I’m not sure what the latest and best thinking is about what this delay might be, on average.

So much for the terminology; the punchline being that:

* EPDDI is basically a few days before the last negative test – the exact number of days being determined by the ‘diagnostic lag/delay’ of the **most sensitive** part of whatever diagnostic algorithm was used on that last negative specimen.
* LPDDI is basically a few days before the first positive test – the exact number of days being determined by the ‘diagnostic lag/delay’ of the **least sensitive positive test** within the diagnostic algorithm was used on that first positive specimen.

So now we need to be able to deal with a data set of this form:

|  |  |  |  |
| --- | --- | --- | --- |
| S\_id | assay val | LPDDI | EPDDI |
| 1 | 0.23 | 10 | 400 |
| 2 | 1.89 | 20 | 300 |
| 3 | 0.94 | 30 | 200 |
| 4 | 1.44 | 20 | 300 |

The code developed so far basically presumes that all the values for LPDDI are ZERO.

My goal is to MAP the existing code onto the time windows implied by the full data – rather than change anything about the core logic of the existing code. I think this can be done straightforwardly enough, as long as a few steps are done in the right order and in the right way.

To begin, we can ignore the LPDDI column and just use the ‘assay values’ (in this case Geenius index) to generate the relevant ‘un-normalised likelihoods’

Now, without discarding the original values of LPDDI and EPDDI (Both of which need to be whole numbers, because we will casually use them as either numbers or array index ranges/labels/offsets), we can take the vector of un-normalised likelihood function values (evaluated at one day steps) and simply cut off the first LPDDI of them – and make a note that the new list of values now has a length **EPDDI minus LPDDI**.

In other words, in terms of the language of the existing code**: T = EPDDI - LPDDI.**

Now we can use the existing code to:

* Normalize the posterior over the T **(= EPDDI minus LPDDI)** remaining days in the estimation window
* Generate a cumulative posterior density
* Find a mode
* Find various percentiles
* Find the F\_T and F\_P based estimation windows which are the heart and soul of the proposed estimation procedure

The key point to bear in mind is that we have somewhat sneakily shifted our origin. **Therefore:** once all this is done – **we need to ADD the value LPDDI back on to ALL the calculated dates** – including the end dates.

To make sure that we can track this process – I suggest that the code produce at least intermediate results where we can track the process per individual – so we can see the value of T, etc and we can confirm that for assay values which push the estimation window strongly towards either the most recent or the earliest piece of the estimation window, we end up with windows which are terminated by the given values of LPDDI and EPDDI.

It's very important to generate the un-normalised likelihood BEFORE we do any of this date shifting – because the logic of the likelihood function needs to be based on the original meaning of the one-day-steps relative to the actual date of taking the sample for the genius test.

Note that this whole procedure now also works even if the recency (Geenius) test date is not exactly the day of first positive test. For example, someone comes in, tests positive, and then is called back a week later for a recency test. This would be fine, as long as the LPDDI and EPDDI are calculated correctly ‘relative to the recency test date’.