# Incubation and latency time estimation for SARS-CoV-2

Arntzen, V.H.

# Summary

The incubation (infection to symptom onset) and latency time (infection to start-of-infectiousness) are crucial quantities to inform control measures at the beginning of an infectious disease outbreak. An example is the duration of quarantine for potentially infected individuals, a control measure that was frequently imposed after SARS-CoV-2 emerged in 2020. This thesis is inspired by the Vietnamese context at the time. With limited intensive care capacity and a long-stretched border with China, Vietnam initially strived to prevent any transmission of SARS-CoV-2. Government-allocated quarantine facilities in which individuals were regularly tested for presence of the infection provided a unique data set.

What is known about the time of infection is usually limited to the window of exposure, i.e. the interval from the first to the last possible moment of infection, such that the start point of incubation time observations is *interval censored*. Common practice is to assume that (i) the risk of infection within the exposure window is constant and (ii) incubation time follows a *gamma*, *lognormal* or *Weibull* distribution (think of three different baking forms for cake dough). However, during the beginning of an outbreak, the daily number of new cases grows exponentially and coronaviruses are known to have a long, right-tailed incubation time distribution, leaving (i) and (ii) unrealistic. In Chapter 2, we investigated this issue by generating toy data sets. We observed that a model that allows for more flexibility in the shape of the distribution provides a better fit to the right-hand tail. For the estimation of the latent period of SARS-CoV-2 with data from Vietnam we assumed an increasing risk of infection within the exposure window (Chapter 4).

Contract tracing aims to notify cases soon after infection to prevent further transmission. Exposure information is retrieved retrospectively, through interviews with notified cases. Estimates of incubation time typically use such contact tracing data. Differential recall may occur: at the time of the interview, infected individuals may recall recent exposure more precisely than less recent exposure. To mitigate the impact of violating assumption (i), analysis is often restricted to observations with well-defined exposure. However, in the presence of differential recall this restriction leads to underestimation of the incubation time.

In Chapter 3, we discuss this issue and another phenomenon that has been overlooked in early estimates of the incubation time distribution of SARS-CoV-2.

Quarantine length for SARS-CoV-2 was typically informed by estimates of the incubation time, even though two out of five infected individuals will never develop symptoms. Latency time would be a more logical quantity for informing this decision, but its estimates are sparse for two reasons: the required data is rarely available and estimation is complicated further because start-of-infectiousness cannot be exactly observed. Utilizing RNA shedding as a proxy for infectiousness, the start-of-infectiousness likely took place between the last negative and first positive test for SARS-CoV-2. Therefore, observations of latency time consist of two windows such that standard methods cannot be used. We developed an `R` software package (`doublIn`) suitable for this type of observations. Using unique data from Vietnam and realistic model assumptions, we estimated the latency time of the SARS-CoV-2 Delta variant for unvaccinated and non-immune individuals in Vietnam in 2021 (Chapter 4).

The incubation and latency time are examples of time-to-event data, the type of data that the survival analysis field within statistics studies. Another example is the age at which a woman develops breast cancer. Understanding the risk associated with genetic variants like BRCA1/2 is crucial as it informs decisions about precautionary measures such as preventive surgical removal of the breast. Estimates are typically based on data from high-risk families with multiple affected members, requiring a tailored approach to provide estimates applicable to individuals from low-risk families as well. In Chapter 5, we generalized the state-of-the-art approach. Our software is openly available for researchers working on similar estimation problems (`R` package `wcox`).