1

Introduction

Contents

1.1 Incubation time and latency time distribution

Incubation and latency time of an infectious disease are crucial quantities to understand and manage infectious disease outbreaks. The incubation time is the time interval from infection to symptom onset, and the latency time refers to the period between infection and the start of infectiousness. The corresponding distributions are typically estimated at the beginning of an outbreak with a novel pathogen. Table 1.1 gives some examples of different ways in which incubation time is used [Nishiura, 2007].

Table 1.1: Common uses of the incubation period distribution of infectious diseases [Nishiura, 2007]. The author distinguishes the major field of use and the various functionalities.

Major field of use	Explanation and example
Clinical practice	Rough estimates of the time of exposure of bedside cases (e.g., to determine the causes and/or sources of infection)
	Development of a treatment strategy that extends the incubation period (e.g., antiretroviral therapy for HIV/AIDS)
	Early projection of disease prognosis when the incubation period is clearly associated with clinical severity (e.g., diseases caused by exotoxin)
	Clinical investigations of the impact of infecting dose on the clinical appearance of a disease (i.e., the dose-response mechanism)
Public health practice	Determination of the length of quarantine required for a potentially exposed individual (e.g., limiting the movement of those exposed to SARS within a household)
Epidemiologic study	Determination of the eradicability of a disease (e.g., determination of the effectiveness of isolation measures)
	Estimation of the time of exposure during a point source outbreak (e.g., in identification of the source of infection during large-scale food poisoning)
	Determination of the end of a point source outbreak (i.e., statistical tests that determine if case onset is over)
	Reconstruction of epidemic curves and short-term predictions of slowly progressing diseases (e.g., backcalculation of HIV/AIDS and prion diseases)
	Estimation of the transmission potential and infectiousness relative to disease-age (e.g., estimation of the relative infectiousness of smallpox)
Ecological study	Determination of the adaptation strategy of a parasite (e.g., evolution of vivax malaria owing to seasonal selection pressure)

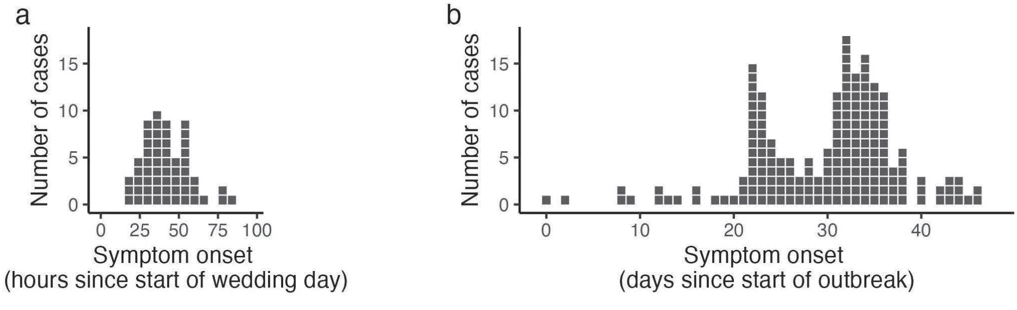
Estimates of the incubation time and latency time are not necessarily the same because when an infected individual starts to be infectious may not coincide with symptom onset. For SARS-CoV-2, presymptomatic transmission was observed [Tindale et al., 2020].

The time elapsed between the start of infectiousness and symptom onset has been important. Namely, a long time lag between start of infectiousness and symptom onset indicates that isolation of symptomatics is most likely ineffective [Nishiura, 2007], i.e. when latency time is considerably shorter than incubation time.

While all individuals infected with SARS-CoV-2 experience a latency period, the same cannot be said for the incubation time. Among SARS-CoV-2 infected individuals, 40.5% did not develop symptoms until recovery [Ma et al., 2021]. Individuals with asymptomatic infection are typically more challenging to notify. The latency time is one of the factors determining the efforts required to control the spread of an infectious disease [Demers et al., 2023]. Despite its relevance, estimates of the latency time are rare and therefore, public health measures are typically informed by incubation time instead.

The incubation time differs considerably by type of infection, from few hours for toxic food poisoning to sometimes a few decades for tuberculosis, AIDS and variant Creutzfeldt-Jakob disease [Nishiura, 2007]. The incubation time for a specific infectious disease varies as well and may, amongst others, depend on age, transmission route, received pathogen dose, vaccination status and natural immunity [Held et al., 2019].

Figure 1.1: Epidemic curve related to two infectious disease outbreaks: (**a**) point source outbreak of salmonellosis among 57 cases visiting a wedding reception in Malahide, Ireland (1996) [ECDC, 2018]; (**b**) propagated (person-to-person transmission) outbreak of measles in Hagelloch, Germany (1861) in 187 cases [Groendyke et al., 2010]. Each square represents one case.



In some infectious disease outbreaks all infected individuals are thought to be infected at the same calendar time by a shared source and the infection is not transmitted further. These so called point source outbreaks provide the most straightforward scenario for estimating the incubation time. Figure 1.1a is the *epidemic curve*, a specific type of histogram, corresponding to an outbreak of salmonellosis related to a wedding reception in Ireland [ECDC, 2018]. Each box represents an infected individual. The notified cases (y-axis) are organized by the six-hour time window in which each respective case exhibits the first symptom(s) (x-axis). Contaminated turkey served at the wedding was identified as the most likely vehicle of infection. Since the infection probably occurred at the wedding reception, the distribution of individual symptom onset days can be directly observed, as it approximately equals the incubation time of salmonellosis. The median incubation time of salmonellosis is known to be 45 hours [Eikmeier et al., 2018].

Many outbreaks propagate as direct transmission takes place between individuals or indirect transmissions mitigated by vectors like mosquitos. The epidemic curve curve of such a propagated outbreak is different from a point-source related curve. Figure 1.1b visualizes the daily number of individuals with measles [Groendyke et al., 2010], a highly infectious disease transmitted from person to person. The epidemic curve no longer resembles the incubation time distribution as individuals acquire the infection on different calendar dates. Often, the moment of infection is not precisely observed but it is, at best, known to occur within a specific time window. The information needed to estimate the incubation time distribution typically includes the exposure window along with the symptom onset day. Additional assumptions are needed to estimate the distribution of incubation time.

1.2 Statistical assumptions: a short history

The first estimate of the incubation time for influenza dates back to 1919 [McKendrick, 1925; Nishiura, 2007]. McKendrick, who gained recognition primarily for his infectious disease transmission models, estimated the incubation time on data from 92 maritime ships that left different harbours in Australia. He used the counts of individuals with symptom onset each day t since departure, i.e. $I(t) = 64, 17, 5, 2$ cases on the 1st, 2nd, 3rd and 4th day ($t = 1, 2, 3, 4$), respectively. Assuming that infection took place on shore and no transmission took place on board, he used the idea that individuals that developed symptoms on the second day since departure ($t = 2$) were exposed *at least* two days before when he estimated the daily probability Z_r of the incubation time r days after exposure as

$$Z_r = p(1-p)^{r-1}. \tag{1.1}$$

where $r \geq t$. Until today, the uncertainty of the infection moment that McKendrick acknowledged remains a major challenge in incubation time estimation. Nowadays, it is common to assume that infection is equally likely to occur on all exposure days (in the example: 1, 2, 3 or 4 days before departure). We will revisit the validity of the latter assumption later in the thesis.

The first attempt to model the incubation time distribution for infectious diseases using a continuous parametric distribution was attributed to John Miner in 1916 [Miner, 1922; Nishiura, 2007]. Miner suggested employing the right skewed Pearson I distribution while Philip E. Sartwell later proposed the lognormal distribution as an alternative [Sartwell, 1950; Nishiura, 2007]. The rationale for choosing a lognormal distribution was that pathogens were thought to grow exponentially within a host. While there is little evidence to support this reasoning for all infectious diseases, the lognormal distribution remains part of the commonly assumed triplet of right-tailed distributions today, alongside the gamma and Weibull distribution.

Coronaviruses are known to have a relatively long tailed incubation time distribution. The WHO has expressed concern about the validity of the commonly assumed parametric distributions, as they may not adequately capture the tail behaviour of the incubation time distribution of corona viruses [WHO, 2003]. The mismatch is particularly problematic since the percentiles are of particular interest; for instance, the 95^{th} percentile is typically used to choose the minimum duration of quarantine for potential cases.

In Chapter 2, we investigate the impact of using parametric distributions through a simulation study and assess the performance of a more flexible alternative. In Chapter 4, we assume another flexible distribution for the SARS-CoV-2 latency time distribution that includes the gamma and Weibull distributions as a special case.

Two common statistical concepts for observations of time-to-event complicate estimation of incubation and latency time. A time-to-event or survival time is the time interval between an initial event and the occurrence of an event of interest such as death, disease-progression, relapse, et cetera.

Often, the start- or endpoint of such an interval of interest cannot be observed precisely, which is referred to as *censoring* (Section 1.4). Moreover, in observational data it is common that certain individuals are observed whereas others go unnoticed, which may lead to *truncation* (Section 1.5). Survival analysis is the statistical discipline devoted to studying time-to-event data, which can be incubation time, latency time, the age of breast cancer diagnosis et cetera.

Several concepts from survival analysis are relevant to the infectious disease context. Estimation of the time from infection to a certain event, such as initial multiplication of gametocytes, a stage of malaria parasites, in the human body, is complicated when some individuals recover from the infection before the endpoint occurred [Andolina et al., 2023; Ramjith et al., 2022]. Clearance of the infection is referred to as a competing risk. There are parallels between survival analysis models and those for spread of infectious disease as well, in specific with the stochastic SIR model [Putter et al., 2024] that models how individuals migrate through the susceptible, infectious and recovery stages. However, the data available early onwards in an infectious disease outbreak is typically fuzzy, stressing the need of tailored approaches for the infectious disease context in specific.

In the applications of this thesis, the estimates rely on observational data. This type of data contrasts clinical trial data in which individuals are assigned specific treatments at known time points and are monitored during follow-up. Before the statistical concepts relevant to our estimation problem are discussed in more detail, the spatiotemporal context and the corresponding data that inspired this thesis are introduced.

1.3 Contact tracing data: SARS-CoV-2 in Vietnam

Acknowledging its limited intensive care capacity and the long-stretched, 1297 km long border with China, the policy of Vietnam was characterised by stringent and early policy measures such as complete border closure. The country initially strived to prevent any introduction and local transmission of SARS-CoV-2. The main pillars of the elimination policy were extensive contact tracing of infected individuals and quarantining of potential infecteds. The quarantine policy for each potential case depended on the closeness to an infected individual, which is referred to as the ’F-system’ and is unique to the pandemic response of Vietnam [Hardy et al., 2020]. For direct contacts of an infected individual, quarantine typically took place in designated quarantine facilities. Further details are provided in Chapter 4.

During contact tracing, notified cases were typically asked to recall their potential risk exposures and if so, when they first exhibited symptoms. In the designated quarantine facilities in Vietnam, swabs were taken regularly to test individuals for SARS-CoV-2. This context provides a unique data set that allowed to estimate the latency time for the SARS-CoV-2 Delta variant, which to the best of the author's knowledge is the first estimate based on data from outside of China.

1.4 What is observed and what is not: censoring

Whereas symptom onset is typically observed up to a day precise, the knowledge of the moment of infection and the start of infectiousness is generally limited to the time interval during which these start- and endpoints occurred. Typically, RNA shedding is used as a proxy for infectiousness. Common practice is to assume that the start of infectiousness occurs between the last negative and first positive test for SARS-CoV-2, such that instead of the exact start of infectiousness, a time window containing the endpoint of latency time is observed. Hence, an observation of incubation time consists of an exposure window and the symptom onset day, while an observation of latency time consists of an exposure window and a start-of-shedding window. Observations of incubation and latency time are *single interval censored* (time origin) and *doubly interval censored*, respectively.

Standard methodology is available when the endpoint is interval censored rather than the time origin. It is common to assume a constant risk of infection within the exposure window. As discussed in more detail in Chapter 2, this assumption is convenient because the likelihood can be rewritten with a reversed time axis and yields an interval censored endpoint. Therefore the incubation time can be estimated using available software. Unfortunately, the validity of this assumption is doubtful in the context of an evolving outbreak. At the beginning of an outbreak of a novel pathogen we are confronted with an exponential growth of new infections and this implies that the constant risk assumption is unrealistic. We performed simulation studies to investigate the impact of the constant risk assumption on the estimates of the percentiles of the SARS-CoV-2 incubation time distribution. Real data from the beginning of the pandemic are used as illustration (Chapter 2).

Another bias that may occur is related to the imperfection of our memory. Recalling when risk exposure took place becomes more challenging when it occurred a long time ago. We refer to this phenomenon as ’differential recall’. Due to uncertainty in recall, exposure windows of less recent exposure may become relatively wide, increasing the risk of bias due to violation of the constant risk assumption. To limit the latter bias, the analysis is often restricted to observations with a narrow exposure window. Even though the term recall bias is frequently mentioned in papers estimating incubation time, to the best of the author’s knowledge, differential recall of exposure, where recent exposures are memorised more precise than exposures longer ago, has never been explicitly studied in this context. In Chapter 3 we show that in the presence of differential recall selecting observations with narrow exposure windows leads to an additional bias.

While censoring concerns incomplete information and can be seen as a specific type of missing data, another statistical challenge in incubation and latency time estimation consists of observations that remain unobserved. In the following section we will elaborate on this.

1.5 Who is included and who is not: biases related to ascertainment

Random sampling is at the core of unbiased estimation. Every individual in the population of interest should have the same probability of being included in the sample. The latter can be challenging, especially when data is collected retrospectively. This thesis discusses three examples of observational data in which certain individuals from a study population were included with a higher probability than others. We briefly introduce the three concepts and refer to the respective chapters for further details.

The earliest estimates of the SARS-CoV-2 incubation time distribution were based on data from individuals who left Wuhan around the Lunar New Year and developed symptoms on or after their travel day [Backer et al., 2020]. Individuals who developed symptoms before travelling, i.e. those with short incubation times, were less likely to be included in the analyzed data. What in survival analysis is referred to as *late entry* or *left truncation* was not addressed in the estimates for SARS-CoV-2 in literature. We examined the impact on the estimates in a simulation study (Chapter 3).

Observations are *right truncated* when those with a relatively long time-to-event are less likely to be included in the data set. This phenomenon has been described for observations of the SARS-CoV-2 latency time from China [Xin, Li, Wu, Li, Lau, Qin, Wang, Cowling, Tsang and Li, 2021]. Right truncation occurred in the data from Vietnam, as individuals were only included in the data when tested positive for SARS-CoV-2 before the end of quarantine or the last day of sampling which in our data was marked by the start of a decline in case incidence due to reporting delay: contact tracing system became overwhelmed with large numbers of cases that could no longer be investigated as thoroughly as before. Right truncation is addressed in our analysis in Chapter 4, a consideration that is not immediately apparent for doubly interval censored observations.

Facing non-random samples is not unique to the infectious disease context and may also occur in other settings. For example, in breast cancer research, individuals are likely to attend a clinic for genetic risk when several family members developed breast cancer. When breast cancer is not frequent in the family, the presence of a genetic component is less likely to be observed. To examine the high risk due to specific genetic variants, researchers include data available from genetic clinics which typically concerns high-risk families with multiple affected individuals. This data is a non-random sample of the population, leading to biased estimates of the increased risk associated with a genetic variant. By means of a tailored weighting method, we restore the data composition such that the results can be extrapolated to the population of interest (Chapter 5). The robustness of our method is investigated by simulations and a two real data applications are provided.

1.6 This thesis

This thesis is a collection of four papers concerning different themes in survival analysis and (infectious disease) epidemiology. Table 1.2 presents an overview of the estimation problems we addressed in each chapter and the field of application. In Chapter 6, we place our work in a broader context by discussing future directions.

Table 1.2: Overview of the estimation problems discussed in this thesis.

Application	**Cause**	**Effect**	**Remedy**
Estimation of incubation and latency time	Assuming a constant risk of infection is not realistic	Overestimation in exponential growth phase (**Chapter 2**)	Assuming that the risk of infection within the exposure window increases congruently with the infection incidence during the exponential growth phase (**Chapter 4**)
	Assuming a gamma, lognormal and/or Weibull distribution for the time-to-event; subsequent choice based on AIC or LOO IC	Biased estimates of the tail percentiles (**Chapter 2**); potential misfit between true and chosen distribution (**Chapter 2**)	A flexible modelling choice, such as Penalized Gaussian Mixture (**Chapter 2**); fitting a generalized gamma distribution that includes gamma, lognormal and Weibull as special cases (**Chapter 4**)
	Differential recall	Underestimation when observations with narrow exposure windows are selected (**Chapter 3**)	Analyze all observations, including also wider exposure windows (**Chapter 4**)
	Delayed entry (left truncation)	Overestimation (**Chapter 3**)	Not straightforward as the late entry time is not observed exactly due to the interval censored infection time
	Right truncation	Underestimation (**Chapter 4**)	Addressed in the analysis; available as a ready-to-use `R` package (`doublIn`) (**Chapter 4**)
The genetic risk of breast, ovarian or prostate cancer	Family-based sampling	Ascertainment bias: underestimation of the risk associated with a genetic variant or polygenic risk score (PRS, **Chapter 5**)	A weighting approach that generalizes the state-of-the-art method; available as a ready-to-use `R` package (`wcox`) (**Chapter 5**)