7

Modelling misreported data

Leonardo S. Bastos

Fundação Oswaldo Cruz, Brazil

Luiz M. Carvalho

Fundação Getulio Vargas, Brazil

Marcelo F.C. Gomes

Fundação Oswaldo Cruz, Brazil

CONTENTS

During a pan/epidemic, data on cases and deaths must be opportune. That is, data collection and availability on official databases must be provided in a timely manner. To make data-driven actions, the faster the information is obtained, the better. However, the quality of information is inversely related to the speed the information is gathered. In this chapter, we describe some data quality issues like reporting delays and disease misclassification in the context of the COVID-19 pandemic and severe acute respiratory illness (SARI) surveillance in Brazil. Moreover, for each of these issues, a statistical model-based solution is presented to incorporate/propagate the uncertainty related

DOI: 10.1201/9781003148883-7

to the lack of data quality into the inference process, making the analysis more adequate. Although the examples are focused on those two specific contexts, they can be easily translated to other diseases with structured notification systems, being relevant for public health surveillance in general.

7.1 Issues with the reporting of epidemiological data

In any passive surveillance[1] system, be it related to outpatient cases such as influenza-like illness (ILI) and arboviruses surveillance, or to hospitalised cases such as severe acute respiratory illness (SARI) surveillance, there is an intrinsic time delay between the event of infection and its notification in the corresponding database. This is due to the fact that notified cases are only identified once symptomatic individuals are attended to at a health care unit. This delay will depend on characteristics of the disease itself, the typical incubation period which will define how long it takes for infected individuals to develop symptoms, as well as cultural and structural characteristics of the exposed population, which affects how long symptomatic individuals usually wait before seeking medical attention. In situations where the notification is not made automatically in the official database, there will also be a potential delay between the notification date (the date in which the case is identified at the healthcare unit) and the digitisation date (the date in which the notification sheet is typed into the digital database). It is clear then that notification delays have both a structural component that can be minimised by infrastructure investments, such as the hiring of dedicated staff for filling notification forms and inserting this information into the digital databases, migration from paper forms to electronic notification, good quality internet access at health care units, and so on (Lana et al., 2020); and an intrinsic component that is related to the patient itself which can be mitigated by ease-of-access to health care facilities and information campaigns to motivate early medical attention. To illustrate this process, we will describe the timeline of a hypothetical COVID-19 case captured by local surveillance.

Let's assume the following example: a person has just found out that he was recently in contact with someone infected by SARS-CoV-2. We shall assume this encounter occurred at time t_0. At time t_1, the patient presented the first symptoms: shortness of breath, fever, anosmia (no sense of smell) and ageusia (loss of taste). Then these symptoms get worst at time t_2 and he, the patient, sought medical assistance where he were tested and hospitalised. The test result was negative for SARS-CoV-2, but given their symptoms the doctor was quite convinced that it was a COVID-19 case. So she, the doctor, asked for another test. At time t_3, she received the test result and, as she expected, it

[1] Passive surveillance: that in which cases are reported as patients seek medical care.

was positive for SARS-CoV-2, that is, indeed a COVID-19 case. The attending health care professional should enter the case into a surveillance system, but when they received the result they were too busy to do it. Eventually they reported it at time t_4. The patient's condition got worse and he was transferred to the intensive care unit ICU at time t_5. Unfortunately, at time t_6, the patient passed away, which was then reported to the surveillance system at time t_7. The timeline is qualitatively represented in Figure 7.1.

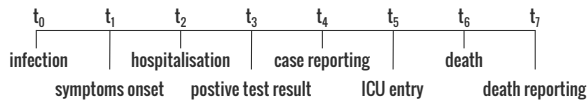

FIGURE 7.1: Qualitative timeline of events related to a hypothetical hospitalised COVID-19 case which required ICU and evolved to death.

The previous example illustrates a COVID-19 case from infection to death. Notice that it could be any infectious disease confirmed by a laboratory test. Recording the times $t_0, t_1, t_2, \ldots$ are essential for the surveillance team to learn about the dynamic of each epidemic. Surveillance systems gather information about several patients like the previous example in a structured data set. It is compulsory for diseases to be reported immediately as soon as a case is identified, for instance yellow fever, which is a mosquito-borne disease transmitted to humans by infected mosquitoes. There are some diseases for which it is compulsory hospitalised cases and/or deaths to be reported, like dengue fever, which is another mosquito-borne disease transmitted to humans very frequently in Latin America and Southeast Asia. And there are some diseases for which reporting is not compulsory, but for which there is a sentinel surveillance system that monitors cases, such as influenza-like illness in most countries in Latin America, the USA and in Europe. In this chapter, we will use COVID-19 as an example and assume that reporting of hospitalised cases and deaths is mandatory.

The time t_0 representing when the person was infected is usually unknown. A person may be able to describe a list of situations where they were exposed but it is nearly impossible to correctly say when the infection occurred. On the other hand, the time of onset symptoms, t_1, is more likely to be remembered by the patient and it is usually a time recorded in a surveillance system. Still there is some memory bias, since the patient may not be certain about when the first symptoms started (t_1). The time when the person seeks medical care is very clear, t_2, which would be the day when the person is evaluated by a health professional. This is usually called the notification date, and is also called the occurrence date. Once a case of a particular disease, COVID-19 in this case or at least a suspected case, is identified, it should be reported to the surveillance team. This notification may occur at the same time the person seeks assistance, but it may also be reported later on at time t_4, called the digitisation date in the case of digital databases. The delays given by the

differences $t_4 - t_1$ and $t_4 - t_2$ are very important. The first delay represents the time until the surveillance team is aware of an infectious disease case. The longer it takes, the more people may be infected from this person. Hence, if the surveillance team is aware of this delay, then it may be able to warn the health units about an outbreak of an infectious disease occurring in a particular area and act in order to mitigate the disease spreading. The second delay, $t_4 - t_2$, is the time between a case being identified until it is reported in a surveillance system, reflecting the capacity of the health unit to report cases. For instance, suppose a health unit is in an isolated area without a computer. In this case, a notification is filed in a paper-based form and eventually it will be recorded in a surveillance system.

Naturally, the longer the notification/digitisation delay, the more inadequate the current case count is for alert systems and situation analysis based on daily or weekly incidence. Not only that, since current incidence will be underestimated, it also impacts the adequacy of this data as input for mathematical and statistical forecasting models. This is where the modelling of reporting delays comes into play, providing *nowcasting* estimates. That is, estimates for what happened up-to-now, in contrast to forecasting which estimates what will come next. Based on the distribution of these delays, the surveillance team may be able to predict the actual number of cases that occurred but have not been notified yet, plan public response in light of the current situation, as well as devise strategies in order to reduce these delays when they become overwhelming. Nowcasting models are also described as a backfill problem in the literature. The description of this process and application examples are described in Section 7.2.

Diseases like COVID-19 also require a biological test to confirm the infection, so on top of reporting delay, there is an extra time associated with testing. It could be a rapid test or a test run in a laboratory, and after knowing the result, the patient record in the surveillance system must be updated. So the surveillance team have to be careful while analysing counts of a disease that can only be confirmed after a biological test. The counts are affected by delay and even under notification since the information, namely here the test result, for some patients might not be updated in the surveillance system.

By definition, public health surveillance strategies based on syndromic definitions will potentially capture cases from multiple pathogens, not only cases related to a single one. In the case of SARI, several respiratory viruses are associated with those manifestations such as Influenza A and B, respiratory syncytial virus (RSV), adenovirus, the coronavirus family, including SARS-CoV-2, and many others. Therefore, SARI cases can be classified as a suspect case of infection by any of those viruses. In this scenario, sample collection for laboratory confirmation of the associated pathogen is fundamental for assessing incidence and prevalence of specific diseases, such as COVID-19. On the other hand, surveillance systems based on confirmed cases alone will not mix multiple pathogens in the same database, but case notification will only happen if and when positive laboratory results are obtained.

In the example given here, at time t_2 the patient was tested and the result of this test was reported to their doctor at time t_3, defined as the laboratory delay. And it is a structural delay since it depends on collecting, transporting and analysing a sample from the patient added to the time until the doctor and the patient get back the result. It also depends on the test itself, since there are both rapid tests and tests that need a proper laboratory infrastructure. Also the surveillance team is informed about the laboratory results of a patient with yet another possible delay. It could be automatically informed as soon as the result is known, but it could also be informed only after the patient and their medical doctor know the result, at which time the health professional should update the patient record in the surveillance system.

Notice that at any moment the sample may be damaged leading to a misclassified result; for example, it could be poorly handled by a technician without experience, or during the transport the sample might not be properly stored at the appropriate temperature, etc. Hence a person may get a false negative result due to any of these reasons and their combination. Even if the sample is not damaged, most laboratory tests are not perfect. They have non-null probabilities for false results. The time between exposure, symptoms onset, and sample collection also affect the ability to identify the associated pathogen by any given test, even if proper care regarding sample collection and handling were in place. And different tests have different optimal time windows (La Marca et al., 2020). The probability of a positive result when the patient is really infected is called sensitivity, whereas the probability of a negative result given the patient is not infected is called specificity. In Section 7.3, we present some statistical models to incorporate external information regarding imperfect tests in order to correct infection incidence and prevalence.

7.2 Modelling reporting delays

As illustrated in the previous section and also in Chapter 2, reporting delay is a well-known issue in infectious disease surveillance. The timeliness of case reporting is key for situation analysis based on surveillance data (Centers for Disease Control, 1988; World Health Organization, 2006; Lana et al., 2020). Although efforts should be made to have case reporting be as timely as possible, in practice it is always expected to have at least some level of delay, especially so for passive surveillance. In order to estimate current or recent cases taking into account the reporting delays described in the previous section, it is fundamental that the notification database store all the relevant dates related to the desired count. For suspected cases or general syndromic cases, the minimum information necessary are the symptoms onset and notification date. For simplicity, in this section we will assume that digitisation and notification dates are the same. When this is not the case, digitisation date is

of the utmost relevance, since this is the actual date at which each case will be available in the database for situation analysis. Nonetheless, the presence of notification date is important for public health authorities in assessing to what extent the observed digitisation delay is due to administrative/structural issues, by evaluating the delay between notification and digitisation date. Since our focus here is to describe nowcasting techniques aimed at case counts for situation analysis and/or to be used as input for forecasting models, the only thing that matters is how long it takes for the information to be available in the database, not the intermediate steps themselves.

For confirmed cases, along with symptoms onset date, one would need access to not only the date of the test result but, more importantly, the date on which the result was inserted into the database for each notified case; call it the test result digitisation date. In the same way, for deaths one would need the date of death and the death digitisation date. For simplicity, we will discuss only the nowcast of suspected cases but, as long as those dates are available, the methods can easily be translated to nowcasting of confirmed cases or deaths.

To illustrate the limited information available by the end of each epidemiological week in a real setting, Figure 7.2 shows the time series of SARI data from Brazil during the 2020 season, consolidated at different epidemiological weeks: 15, 25, 35, 45, and 53, the last epidemiological week of 2020. It is clear that the incompleteness of recent weeks due to notification delay not only affects the magnitude, but also the current trend.

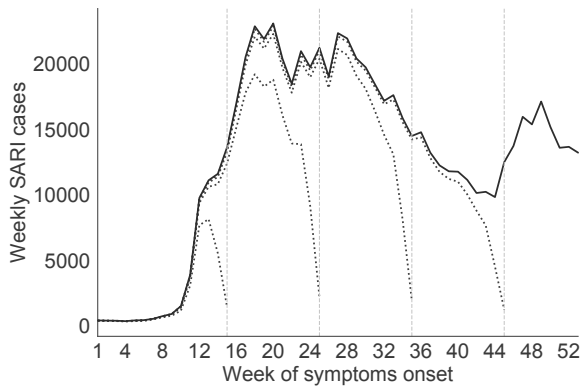

FIGURE 7.2: Weekly cases of severe acute respiratory illness (SARI) notified in Brazil during the 2020 season, by symptoms' onset week. Comparison between the weekly cases digitised up to the last epidemiological week, 53 (solid black line), and what had been digitised up to the end of epidemiological weeks 15, 25, 35, and 45 (dotted grey lines). Vertical lines indicate each of those weeks for reference.

7.2.1 Weekly cases nowcast

In this chapter we will assume that the quantity of interest is the weekly number of new cases, y_t. Nonetheless, the methodology can transparently be applied to any scale of interest, be it daily, weekly, or monthly. The process of estimating current or recent cases is defined as a nowcasting exercise, a nomenclature based on the term forecasting. While forecast is the exercise of estimating what is expected for the coming weeks, nowcasting tries to evaluate what has already happened but has not been reported yet.

Since cases can be inserted into the database retrospectively, at each following week $t+1, t+2, \ldots, t+n$, cases corresponding to week t can still be registered. Let us define $y_{t,d}$ the number of new cases that occurred at week t, based on symptoms onset, but inserted into the database d weeks after week t, based on digitising date. Therefore, the total number of notified cases from week t can be written as

$$y_t = \sum_{d=0}^{d_m} y_{t,d}, \tag{7.1}$$

where d_m is the maximum delay between symptom onset and digitisation. This can be defined by the surveillance team beforehand, or extracted from the data. As it will be clear in the discussion ahead, the higher d_m is, the more computationally expensive the nowcasting will become. Exploratory analysis of historical data can be used to define a cut-off $d_m > 0$ for which $y_{t,d_m} \approx 0$ or $y_{t,d_m} \ll y_t$ for all t as well.

The problem is that, for the most recent complete week T, the database will only have $y_{T,0}$. For the previous week, $T-1$, it will only have $y_{T-1,0}$ and $y_{T-1,1}$, and so on until week $T - d_m$, for which it will have the complete information $y_{T-d_m,0}$, $y_{T-d_m,1}$, $\ldots$, y_{T-d_m,d_m}. Therefore, the process of nowcasting can be described as the process of estimating the not-yet-available data $y_{T,1}, y_{T,2}, \ldots, y_{T,d_m}, y_{T-1,2}, \ldots, y_{T-1,d_m}, \ldots, y_{T-d_m+1,d_m}$. By representing the information in a matrix where lines are the weeks increasing downwards, and columns are the delay increasing from left to right, the missing cells would form a nice triangle, known as the runoff triangle (Mack, 1993):

$$\begin{array}{l} \\ T-d_m \\ T-d_m+1 \\ T-d_m+2 \\ \vdots \\ T-1 \\ T \end{array} \begin{array}{c} \begin{array}{ccccc} 0 & 1 & \cdots & d_m-1 & d_m \end{array} \\ \begin{pmatrix} y_{T-d_m,0} & y_{T-d_m,1} & \cdots & y_{T-d_m,d_m-1} & y_{T-d_m,d_m} \\ y_{T-d_m+1,0} & y_{T-d_m+1,1} & \cdots & y_{T-d_m+1,d_m-1} & \text{NA} \\ y_{T-d_m+2,0} & y_{T-d_m+2,1} & \cdots & \text{NA} & \text{NA} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ y_{T-1,0} & y_{T-1,1} & \cdots & \text{NA} & \text{NA} \\ y_{T,0} & \text{NA} & \cdots & \text{NA} & \text{NA} \end{pmatrix} \end{array} \tag{7.2}$$

There are several ways to tackle this problem, and most approaches can be broadly grouped in two approaches: ones that are hierarchical in nature,

modelling $y_{t,d}$ conditional on y_t, where the latter is given by a Poisson or Negative Binomial distribution and the former is a multinomial process with probability vector of size d_m (Noufaily et al., 2016; Höhle and an der Heiden, 2014); and others that focus on the distribution of the random variables $y_{t,d}$ themselves (Bastos et al., 2019; Barbosa and Struchiner, 2002), also known as the chain-ladder technique from actuarial sciences (Mack, 1993; Renshaw and Verrall, 1998).

In this chapter we will focus on the method described in Bastos et al. (2019), a Bayesian approach to estimate $y_{t,d}$ which was successfully applied to SARI and Dengue surveillance in Brazil, generating a publicly available weekly nowcast reported by InfoGripe (http://info.gripe.fiocruz.br) and InfoDengue (http://info.dengue.mat.br) (Codeço et al., 2018) systems. The implementation of this method for SARI surveillance within InfoGripe was able to provide an early warning for the impact of COVID-19 cases in Brazil even before laboratory confirmation of SARS-CoV-2 predominance among SARI cases (Bastos et al., 2020).

We assume $y_{t,d}$ follows a Negative Binomial distribution with mean $\mu_{t,d}$ and scale parameter ϕ

$$y_{t,d} \sim \text{NegBin}(\mu_{t,d}, \phi), \quad \mu_{t,d} > 0,\ \phi > 0. \tag{7.3}$$

The adopted parameterisation provides $\text{Var}[y_{t,d}] = \mu_{t,d}\,(1 + \mu_{t,d}/\phi)$. The logarithm of the mean of $y_{t,d}$ can be decomposed in a way that takes into account random effects associated with a mean temporal evolution α_t, the delay effect β_d, interaction effects between the two $\gamma_{t,d}$ to accommodate structural changes affecting digitisation timeliness, as well as other relevant temporal covariates $\mathbf{X_{t,d}}$ such as weather data, for example, with its corresponding parameters vector $\vec{\delta}$. Therefore, it can be described as

$$\log(\mu_{t,d}) = \mu + \alpha_t + \beta_d + \gamma_{t,d} + \mathbf{X'}_{t,d}\vec{\delta}, \tag{7.4}$$

where μ is the logarithm of the overall mean number of weekly cases.

The temporal and delay effects can be modelled as a simple first-order random walk

$$\alpha_t \sim N(\alpha_{t-1}, \sigma_\alpha^2), \quad t = 2, 3, \ldots, T, \tag{7.5}$$

$$\beta_d \sim N(\beta_{d-1}, \sigma_\beta^2), \quad d = 1, 2, \ldots, d_m. \tag{7.6}$$

The interaction between time and delay is also modelled as a first-order random walk

$$\gamma_{t,d} \sim N(\gamma_{t-1,d}, \sigma_\gamma^2). \tag{7.7}$$

Since timeliness can be affected by several factors over time, it is important to incorporate this interaction effect. For example, during peaks of hospital capacities, it is possible to have an increase in the notification delay if there

is no dedicated staff for filling out notification sheets, since patient care is prioritised by medical staff. Conversely, if local authorities hire dedicated staff in anticipation or as a response to an outbreak, delay can be reduced. Note that other factors can be easily integrated into the model described in Eq 7.4, including spatial effects (Bastos et al., 2019).

Given the described model, the posterior distribution of the parameters, $\boldsymbol{\Theta} = (\mu, \{\alpha_t\}, \{\beta_d\}, \{\gamma_{t,d}\}, \sigma_\alpha^2, \sigma_\beta^2, \sigma_\gamma^2, \phi)$ given the set of notified data $\mathbf{y} = \{y_{t,d}\}$ can be expressed by

$$p(\boldsymbol{\Theta}|\mathbf{y}) \propto \pi(\boldsymbol{\Theta}) \prod_{t=1}^{T} \prod_{d=0}^{d_m} p(y_{t,d}|\boldsymbol{\Theta}), \tag{7.8}$$

with $p(y_{t,d}|\boldsymbol{\Theta})$ the Negative Binomial defined in Equation 7.3, and $\pi(\boldsymbol{\Theta})$ the joint prior distribution given by the random effects distributions and the prior distributions of their corresponding scale parameters.

Samples from the posterior distribution $p(\boldsymbol{\Theta}|\mathbf{y})$ can be obtained by integrated nested Laplace approximation (Rue et al., 2009; Martins et al., 2013) (INLA), for which there is a publicly available package for `R` at https://www.r-inla.org. Although this process could also be implemented using classical Monte Carlo Markov Chain (MCMC) models (Gamerman and Lopes, 2006), INLA provides a significantly accelerated performance, which is a must for its usefulness in situation rooms during ongoing surveillance, especially when dealing with analysis from multiple locations at once such as those from state or national health secretariats. This characteristic was paramount for the feasibility of weekly nowcast of arboviruses and SARI notifications at the municipal, state and national levels implemented by InfoDengue (Codeço et al., 2018) (http://info.dengue.mat.br) and InfoGripe (http://info.gripe.fiocruz.br) in Brazil, as illustrated in Section 7.2.2.

With the samples from the posterior for the parameters obtained from Eq 7.8, for example by means of `inla.posterior.sample()` function, it is a simple question of implementing a Monte Carlo process to generate a sample of the missing values $\{y_{t,d}\}$ from the matrix illustrated in Equation 7.2. From those, we obtain the estimated marginals $\{\hat{y}_t |\ t = T-d_m+1, T-d_m+2, \ldots, T\}$, which is the actual nowcast of weekly cases. For each generated sample of the posterior distribution of the parameters, we have an estimated trajectory for the weekly cases. Therefore, this process allows for point estimates and credible intervals for the weekly cases. This can then be used to propagate the uncertainty to predictive or forecasting models discussed in other chapters in this book.

Open-source codes for implementing this model can be found at two main repositories maintained by the authors of this chapter: https://github.com/Opportunity-Estimator-EpiSurveillance/leos.opportunity.estimator and https://github.com/lsbastos/Delay.

7.2.2 Illustration: SARI notifications in Brazil provided by InfoGripe

The InfoGripe platform (http://info.gripe.fiocruz.br) was developed as a joint effort by Brazilian researchers from Fiocruz[2] and FGV[3] with the Influenza Technical Group of the Health Surveillance Secretariat (*GT-Influenza, Secretaria de Vigilância em Saúde*) from the Ministry of Health. It is used as an analytical tool for the national surveillance of SARI in Brazil (Ministério da Saúde et al., 2018), providing situation assessment by means of alert levels based on current incidence, typical seasonal profile by state, epidemiological data such as age and sex stratification, weekly cases by respiratory virus, and so on.

The current SARI surveillance scope in Brazil is based on notification from every health care unit with hospital beds (Ministério da Saúde et al., 2019), following a syndromic case definition inspired by the World Health Organization guidelines for SARI surveillance (World Health Organization, 2013). It started in 2009 as a response to the 2009 H1N1pdm09 Influenza pandemic and, in 2012, laboratory testing protocol was extended to include other respiratory viruses of interest (Ministério da Saúde, SINAN, 2012). In Figure 7.3 we show the point estimates for the weekly cases, $\hat{y}_t$ (dashed lines), provided by InfoGripe using the method described in the previous section. We compared it to the consolidated time series, y_t (solid lines), and the weekly counts for each week available by the end of the corresponding week, $y_{t,0}$ (dotted lines), for the aggregated data for Brazil as well as state counts for selected states.

In the face of the COVID-19 pandemic in 2020, it was only natural to use this surveillance system to monitor the corresponding hospitalisations and to incorporate SARS-CoV-2 testing for differential diagnosis (Ministério da Saúde and Secretaria de Vigilância em Saúde, 2020). Even before SARS-CoV-2 testing for notified SARI cases was widely implemented in Brazil, the nowcast model, along with the complementary epidemiological information provided by InfoGripe, were able to provide early warning of the impact of COVID-19 in terms of SARI cases in Brazil (Bastos et al., 2020), illustrating the usefulness of this methodology for action planning and response. Since the modelling approach provides estimated trajectories, it can also be used to assess the trend itself, which can be even more relevant than the actual count estimates. Early detection of a trend indicating an increase in the number of weekly cases, even if the actual number of cases is relatively low, allows for rolling out of mitigation strategies to prevent the collapse of medical resources and high disease burden. With that in mind, the InfoGripe platform was extended to provide short- and long-term trend based on a linear model applied as a rolling window to the last 3 and 6 consecutive weeks, respectively. This information was provided on a weekly basis in the form of situation reports

[2] *Fundação Oswaldo Cruz*, Fiocruz: https://portal.fiocruz.br/
[3] *Fundação Getúlio Vargas*, FGV: https://portal.fgv.br/

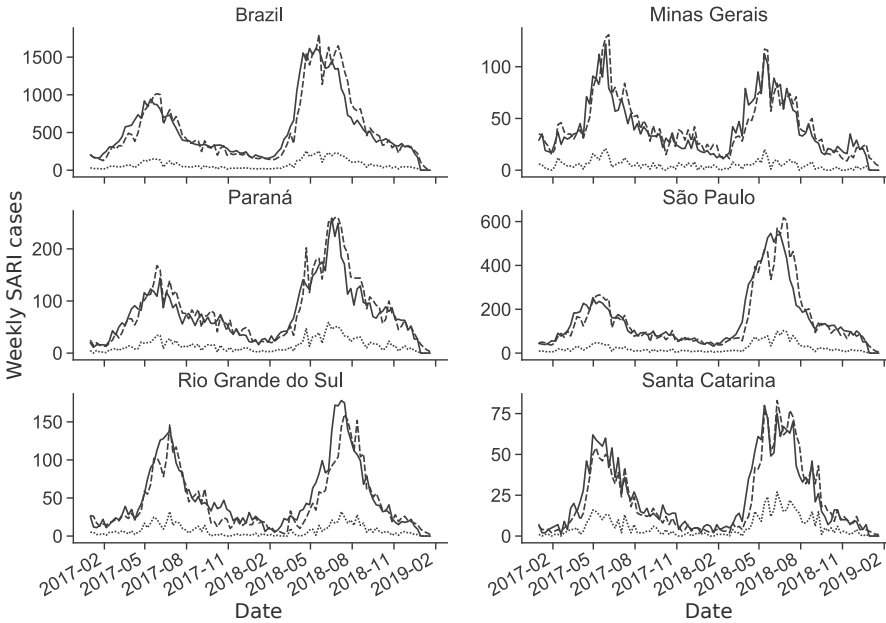

FIGURE 7.3: Weekly cases of severe acute respiratory illness (SARI) notified in Brazil during the 2017 and 2018 seasons, by symptoms' onset week. Comparison between the consolidated weekly cases, y_t (solid line), the point estimated provided by the end of each corresponding week, $\hat{y}_t$ using the described nowcasting model (dashed line), and the number of cases digitised up to the end of each epidemiological week, $y_{t,0}$ (dotted line), for the whole country (top left panel), the states of Minas Gerais (top right panel), Paraná (centre left), São Paulo (centre right), Rio Grande do Sul (bottom left), and Santa Catarina (bottom right). The first two are within the Southeastern region, while the remaining three are in the South region of Brazil.

forwarded to the Health Surveillance Secretariat at the Brazilian Ministry of Health (SVS/MS), and State Health Secretariats subscribed to InfoGripe's mailing list, as well was deposited in a public online repository (`http://bit.ly/mave-infogripe`). Figure 7.4 shows a snapshot taken from the bulletin published by the end of epidemiological week 37, as an illustration of real case application on the field. It provides the nowcast of the weekly incidence in the municipality of Manaus, capital of Amazonas state, along with the likelihood of increasing/decreasing trend in the short and long term up to weeks 36 and 37, and up to weeks 32 to 37, respectively.

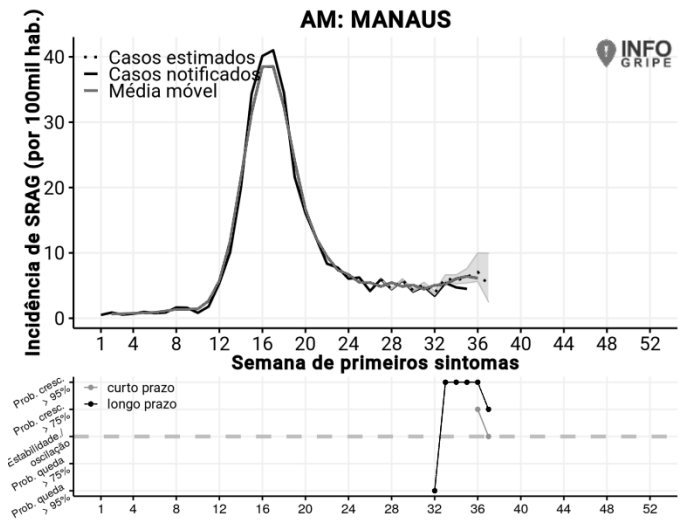

FIGURE 7.4: Snapshot from InfoGripe's Bulletin from epidemiological week 37 of 2020, available at https://gitlab.procc.fiocruz.br/mave/repo/-/blob/master/Boletins%20do%20InfoGripe/boletins_anteriores/Boletim_InfoGripe_SE202037_sem_filtro_febre.pdf. The upper plot shows the weekly incidence per 100k inhabitants of Manaus, capital of the Amazonas state, by symptom onset, with notified cases (*Casos notificados*, solid black line), estimated incidence (*Casos estimados*, dotted black line) with 90%CI, and centred 3-week rolling average of the point estimate (*Média móvel*, solid grey line). The bottom plot presents the trends in terms of likelihood of increase (*Prob. cresc.*) or decrease (*Prob. queda*) for the short term (*curto prazo*, grey line) and long term (*longo prazo*, black line).

7.3 Prevalence estimation from imperfect tests

An important question during an epidemic is how many individuals have been exposed to the disease at time t, i.e., the (true) number of cumulative cases, Y_t^{true}. Let M denote the population size. Assuming permanent and complete immunity, the proportion $\theta_t = Y_t^{\text{true}}/M$, called the **prevalence**, is a key epidemiological quantity because it measures the overall immunity of the population and whether the threshold for collective (herd) immunity has been reached.

However, factors such as reporting delays, underreporting and a large fraction of asymptomatic individuals might make it difficult to ascertain Y_t^{true} accurately. In such a scenario, serological and other diagnostic-based surveys provide a valuable source of additional data that can be used to uncover the underlying pattern of immunity to the disease in a population. The chief idea is to randomly sample individuals from the target population and, using a diagnostic test, measure their levels of antibodies against the causative agent of the disease in question. Assuming for simplicity that all tests are performed at time t, the number x of individuals that test positive out of n provides information about θ_t. Once θ_t has been estimated, we can project the number of exposed individuals $\hat{I}_t = M\hat{\theta}_t$. Unfortunately, diagnostic tests are rarely perfect measurement instruments; one might obtain false positives or false negatives (see Section 7.3.1). One thus needs a statistical model in order to account for this misclassification error.

This section is concerned with providing a statistically principled, model-based treatment of prevalence estimation, having the explicit goal of estimating Y_t with the appropriate level of uncertainty. The theory laid out here is general and encompasses any imperfect test, not just serological diagnostic tests, although the latter provide the main motivation for the techniques discussed.

7.3.1 Preliminaries: Imperfect classifiers

Before we discuss prevalence estimation, it is useful to establish the basic concepts and notation pertaining to (imperfect) binary classifiers. Let $R_i \in \{0, 1\}$ be the random variable which records whether the i-th individual in the sample has a given condition which we are interested in detecting. Also, let $T_i \in \{0, 1\}$ be the random variable which records whether the same individual will produce a positive test ($T_i = 1$). An imperfect test can be evaluated in the presence of a **gold standard**. The gold standard is a measurement device with the highest accepted accuracy and which can be taken to represent the ground truth about R_i. In the context of infectious diseases, a gold standard can be a very precise molecular test that detects the presence of the pathogen's genetic material, polymerase chain reaction (PCR) for instance.

If we measure J individuals with both the test under evaluation and the gold standard, we can tally the numbers of true positives/negatives (TP and TN, respectively) and false positives/negatives (FP and FN) such that TP + FP + TN + FN = J and produce a two-by-two table as the one in Table 7.1.

TABLE 7.1: Hypothetical counts from a (binary) diagnostic test against a gold standard. Here, a value of 1 represents a positive result, and a value of 0 represents a negative result.

		Test under evaluation	
		0	1
Gold Standard	0	TN	FP
	1	FN	TP

The counts from Table 7.1 can in turn be used to estimate key probabilistic quantities such as the sensitivity and specificity of the test. These quantities, their definitions and how to estimate them from a gold standard essay are shown in Table 7.2. As we shall see in the next sections, the sensitivity (δ) and specificity (γ) of a test are the key quantities to consider when estimating the prevalence. The positive and negative predictive values (PPV and NPV) are useful at an individual level because they measure the probability that one has (does not have) the condition being tested conditional on a positive (negative) test outcome. An important aspect of diagnostic tests is that their accuracy for measuring the status of any particular individual depends on the underlying prevalence of the condition being tested. In other words, the PPV and NPV of a test depend on the prevalence θ_t. This means that the inferential value of the results to which any one individual taking the test will strongly depend on the underlying prevalence. If the prevalence is high, the results are accurate at an individual level, but if the disease/condition is rare, the test has little value to inform whether one has the condition in question. This will hopefully become clear in the next section. For a review of the clinical applications of predictive values, sensitivity and specificity, see Parikh et al. (2008).

7.3.2 Prevalence from a single imperfect test

Having established the main accuracy measures of a (binary) test, we can now move on to how to estimate the prevalence θ_t from test data alone, without the gold standard. Estimation of prevalence from imperfect tests is a long-standing and well-studied problem in the Medical Statistics literature (Rogan and Gladen, 1978; Greenland, 1996; Diggle, 2011) and, as we shall see, follows naturally from basic probability manipulations.

TABLE 7.2: Key probabilistic quantities in imperfect classification. We show the probabilistic definition of each quantity, as well as how to estimate it from the counts in Table 7.1.

Quantitity	Definition	Estimate
Sensitivity	$\delta := \mathrm{P}(T = 1 \mid R = 1)$	$\frac{\mathrm{TP}}{\mathrm{TP+FN}}$
Specificity	$\gamma := \mathrm{P}(T = 0 \mid R = 0)$	$\frac{\mathrm{TN}}{\mathrm{TN+FP}}$
Positive predictive value (PPV)	$\zeta := \mathrm{P}(R = 1 \mid T = 1)$	$\frac{\mathrm{TP}}{\mathrm{TP+FP}}$
Negative predictive value (PPV)	$\xi := \mathrm{P}(R = 0 \mid T = 0)$	$\frac{\mathrm{TN}}{\mathrm{TN+FN}}$

For concreteness, we will focus on the case where one surveys a sample of n individuals with a test that detects antibodies against a given pathogen. In this context we thus define R_i to be the random variable which records whether the i-th individual, once exposed, will produce a measurable immune response in the form of antibodies. Recalling the notation in the previous section, we say that the prevalence, $\theta_t := \mathrm{P}(R = 1)$, is the probability that an individual has antibodies and is the main quantity of interest (q.o.i). We can compute $p := \mathrm{P}(T_i = 1)$, sometimes called the **apparent prevalence**:

$$\begin{aligned} p &= \mathrm{P}(T_i = 1 \mid R_i = 1)\,\mathrm{P}(R_i = 1) + \mathrm{P}(T_i = 1 \mid R_i = 0)\,\mathrm{P}(R_i = 0), \\ &= \delta\theta_t + (1 - \theta_t)(1 - \gamma), \end{aligned} \tag{7.9}$$

where δ and γ are the **sensitivity** and **specificity** of the test, respectively. This computation amounts to marginalising over the unobserved quantities in the model and thus can be seen as a Rao-Blackwellisation of the full model, i.e, $p = E[E[T_i|R_i]] = \sum_{r=0}^{1} \mathrm{P}(T_i = 1, R_i = r)$, where the first equality follows from the law of total expectation. The relevant (sufficient) statistic from this experiment is $x = \sum_{i=1}^{n} T_i$. Assuming the tests are conditionally independent given the latent status variables $\boldsymbol{R} = \{R_1, \ldots, R_n\}$, we conclude that x has a binomial distribution with n trials and success probability p. This assumption is reasonable when the sample is random and $n \ll M$. If sampling is not random or if n is large, then the dependency between individuals induced by the contagion process will not be negligible.

Frequentist analysis

A first stab at estimating θ_t can be made by correcting the naïve estimate $\hat{p} = x/n$, along with its usual normal approximation confidence interval, for the sensitivity and specificity of the test. A straightforward manipulation of (7.9) gives rise to the Rogan-Gladen estimator (Rogan and Gladen, 1978):

$$\hat{\theta}_t^{\mathrm{RG}} = \frac{\hat{p} - (1 - \gamma)}{\delta + \gamma - 1}. \tag{7.10}$$

Notice that if $\hat{p}$ is less than the false positive rate (FPR) of the test, $1 - \gamma$, the Rogan-Gladen estimator will yield a meaningless negative estimate. One

might also wish to provide interval estimates for θ_t. For a confidence level $0 < \alpha < 1$, if one constructs a $100 \times \alpha\%$, confidence interval $(\hat{p}_\text{l}, \hat{p}_\text{u})$ for p, an interval $(\hat{\theta}_\text{l}, \hat{\theta}_\text{u})$ can be obtained for θ (Diggle, 2011):

$$\hat{\theta}_\text{l} = \max\left\{0, \frac{\hat{p}_\text{l} - (1-\gamma)}{\delta + \gamma - 1}\right\}, \tag{7.11}$$

$$\hat{\theta}_\text{u} = \min\left\{1, \frac{\hat{p}_\text{u} - (1-\gamma)}{\delta + \gamma - 1}\right\}, \tag{7.12}$$

under the reasonable assumption that $\delta + \gamma > 1$. This approach however ignores uncertainty about γ and δ and thus tends to produce overly-confident estimates (Izbicki et al., 2020). Whilst this can be remedied with bootstrap techniques (see Cai et al. (2020)), a Bayesian approach is able to seamlessly incorporate all the sources of information in order to produce sensible estimates (Flor et al., 2020).

Bayesian analysis

As demonstrated by Greenland (2009) and Gelman and Carpenter (2020) amongst many others, taking a Bayesian approach to the estimation of θ_t naturally allows one to incorporate several sources of uncertainty and produce straightforward estimates of the q.o.i. from the posterior distribution. This section should serve as an up-to-date account of the state-of-the-art. See Branscum et al. (2005) for a review of older Bayesian methods.

As a starting point, consider the model

$$\begin{aligned} \gamma &\sim \text{Beta}(\alpha_\gamma, \beta_\gamma), \\ \delta &\sim \text{Beta}(\alpha_\delta, \beta_\delta), \\ \theta_t &\sim \text{Beta}(\alpha_\theta, \beta_\theta), \\ x &\sim \text{Binomial}(n, p), \end{aligned} \tag{7.13}$$

with p given as in (7.9). This model accommodates uncertainty about test characteristics (sensitivity and specificity) whilst allowing for the incorporation of prior information about θ_t. Elicitation of the joint prior on (δ, γ), π_{DG}, is straightforward if one has information of the form in Table 7.1. This information is usually released by test manufacturers.

For simplicity, assume that $\pi_{DG}(\delta, \gamma) = \pi_D(\delta)\pi_G(\gamma)$ (see Section 7.3.7, however). Further, assume uniform —Beta(1, 1)— distributions on δ, γ. Then the distributions in the model in (7.13) can be seen as posterior distributions with $\alpha_\delta = \text{TP} + 1$, $\beta_\delta = \text{FN} + 1$ and $\alpha_\gamma = \text{TN} + 1$, $\beta_\gamma = \text{FP} + 1$. Sometimes researchers are only able to measure (TP, FN) and (TN, FP) in separate essays, that is, not able to measure sensitivity and specificity jointly. This is a minor complication that changes very little in the elicitation procedure. See Section 2 of Gelman and Carpenter (2020) for an analysis with separate sensitivity and specificity data. There might be situations where the actual counts from a validation experiment are not available. If the mean sensitivity (specificity)

is reported along $100 \times \alpha\%$ uncertainty intervals, hyperparameters can more often than not be approximated by a simple optimisation procedure that attempts to find the hyperparameters that yield mean and quantiles close to the measured values. If one only has access to the mean $E[\delta] =: m_\delta$, say, our advice is to find the Beta distribution with the highest entropy under the constraint that $\alpha_\delta/(\alpha_\delta + \beta_\delta) = m_\delta$—likewise for γ.

The focus on careful elicitation stems from two main reasons: (i) the need to properly propagate uncertainty about test characteristics and (ii) the fact that the likelihood contains no information about δ and γ. These parameters need to be given strong priors in a Bayesian analysis. Hence, the posterior

$$\begin{aligned} \xi(\theta_t, \gamma, \delta \mid x, n) &\propto p^x (1-p)^{n-x} \pi_{DG}(\delta, \gamma) \pi_{\mathrm{P}}(\theta_t), \\ &\propto p^x (1-p)^{n-x} \pi_D(\delta) \pi_G(\gamma) \pi_{\mathrm{P}}(\theta_t), \end{aligned} \tag{7.14}$$

with p given by (7.9) is only interesting in the margin $\xi_T(\theta_t \mid x, n)$ since we expect $\xi_D(\delta \mid x, n)$ and $\xi_G(\gamma \mid x, n)$ to closely resemble the π_D and π_G, respectively. The joint posterior (7.14) is intractable and expectations need to be approximated. This is usually accomplished through the use of Markov chain Monte Carlo (MCMC) methods—see Chapter 5 in this book. Whilst the relatively simple structure of the model lends itself to a Metropolis-within-Gibbs scheme that exploits conjugacy where appropriate, recent treatments of this model and its variations (see Section 7.3.6) have employed a general solution using Hamiltonian Monte Carlo (Gelman and Carpenter, 2020), an approach we favour here also.

As far as inferential summaries go, the posterior mean and median are traditional Bayesian point estimates. The construction of credibility intervals deserves a bit more consideration, however. In the beginning of an epidemic caused by a pathogen to which the population has no previous immunity, prevalence will generally be low. As θ_t is constrained to lie in $(0, 1)$, the usual equal-tailed $100 \times \alpha\%$ credibility interval will not be a good inferential summary as it will very likely include 0, which is not reasonable. Gelman and Carpenter (2020) propose using the shortest posterior interval of Liu et al. (2015), which in the case of a unimodal, univariate distribution such as $\xi_T(\theta_t \mid x, n)$ corresponds to the highest posterior density (HPD) interval. The HPD will usually be tighter and exclude the boundaries whilst still allowing a principled treatment of uncertainty about the quantities of interest.

7.3.3 Re-testing positives

In this section we illustrate how the probability calculus previously discussed can be applied in a slightly more complicated scenario. Since the test is imperfect, one strategy is to do confirmatory tests on the samples which test positive in a triage (two-stage) fashion. For convenience, we will drop the individual-level subscript i in the presentation that follows. Suppose now that we have two tests, $T^{(1)}$ and $T^{(2)}$, but only run $T^{(2)}$ on samples (individuals)

for which $T^{(1)} = 1$. Let θ_t be the prevalence as before, and let γ_1, δ_1, γ_2, δ_2 be the specificity and sensitivity of tests $T^{(1)}$ and $T^{(2)}$ respectively.

Let Z be the outcome of a re-testing positives-only strategy, i.e., $Z = 1$ if tests 1 and 2 are both positives, and $Z = 0$ otherwise. As before, define $w := \mathrm{P}(Z = 1)$. Marginalising over the relevant latent quantities, we have

$$\begin{aligned}
w &= \mathrm{P}(Z = 1, R = 1) + \mathrm{P}(Z = 1, R = 0) \\
&= \mathrm{P}(Z = 1 \mid R = 1)\,\mathrm{P}(R = 1) + \mathrm{P}(Z = 1 \mid R = 0)\,\mathrm{P}(R = 0) \\
&= \mathrm{P}(T^{(2)} = 1, T^{(1)} = 1 \mid R = 1)\,\mathrm{P}(R = 1) \\
&+ \mathrm{P}(T^{(2)} = 1, T^{(1)} = 1 \mid R = 0)\,\mathrm{P}(R = 0) \\
&= \mathrm{P}(T^{(2)} = 1 \mid T^{(1)} = 1, R = 1)\,\mathrm{P}(T^{(1)} = 1 \mid R = 1)\,\mathrm{P}(R = 1) + \\
&+ \mathrm{P}(T^{(2)} = 1 \mid T^{(1)} = 1, R = 0)\,\mathrm{P}(T^{(1)} = 1 \mid R = 0)\,\mathrm{P}(R = 0) \\
&= \delta_2\delta_1\theta_t + (1 - \gamma_2)(1 - \gamma_1)(1 - \theta_t),
\end{aligned}$$

where the last line follows from the assumptions that the two tests are conditionally independent given R. The (Bayesian) analysis of this model proceeds in a similar fashion to what has already been discussed, with informative priors on δ_1, γ_1 and δ_2, γ_2. Let $v = \sum_{i=1}^{n} Z_i$. Then

$$\begin{aligned}
\gamma_1 &\sim \mathrm{Beta}(\alpha_{g1}, \beta_{g1}), \\
\gamma_2 &\sim \mathrm{Beta}(\alpha_{g2}, \beta_{g2}), \\
\delta_1 &\sim \mathrm{Beta}(\alpha_{d1}, \beta_{d1}), \\
\delta_2 &\sim \mathrm{Beta}(\alpha_{d2}, \beta_{d2}), \\
\theta_t &\sim \mathrm{Beta}(\alpha_\theta, \beta_\theta), \\
v &\sim \mathrm{Binomial}(n, w).
\end{aligned} \tag{7.15}$$

This setup leads to more accurate estimates of θ_t, especially if $T^{(2)}$ is more accurate than $T^{(1)}$. Moreover, with widespread testing becoming more common, this is a model variation worth exploring.

7.3.4 Estimating underreporting from prevalence surveys

While θ_t is an interesting quantity to estimate in its own right, another major epidemiological goal is to estimate the number of actual cases that have occurred, which is very unlikely to be correctly captured by official figures. One way to obtain such an estimate is to enact the simple correction $Y_t^{\mathrm{corr}} = M\theta_t$. We might also be interested in estimating the fraction p_{d} of cases which are detected by epidemiological surveillance (i.e., reported). To this end, the model

in (7.13) can be extended with the simple approximate model:

$$\begin{aligned}
\gamma &\sim \text{Beta}(\alpha_\gamma, \beta_\gamma), \\
\delta &\sim \text{Beta}(\alpha_\delta, \beta_\delta), \\
\theta_t &\sim \text{Beta}(\alpha_\theta, \beta_\theta), \\
x &\sim \text{Binomial}(n, p), \\
p_\text{d} &\sim \text{Beta}(\alpha_\text{d}, \beta_\text{d}), \\
Y_t &\sim \text{Binomial}\left(\lfloor M\theta_t \rfloor, p_\text{d}\right),
\end{aligned} \tag{7.16}$$

where $\lfloor a \rfloor$ is the largest integer less than $a \in \mathbb{R}$, also known as the floor function. The underreporting fraction is thus $1 - p_\text{d}$. This model is able to incorporate the uncertainty about δ, γ propagated through θ_t. Many fruitful extensions of this model are possible. For instance, if one has a collection of triplets (Y_{tj}, n_j, x_j) from J locations (countries, states, counties, etc.), one can add a regression component to the probability of detection such as, for instance,

$$\text{logit}(p_{\text{d}j}) = \beta_0 + \beta_1 X_{j1} + \ldots + \beta_P X_{jP},$$

where the X_j are relevant predictors for each location, such as gross domestic product (GDP), human development index (HDI), number of hospitals *per capita* and other variables thought to be explanatory of a location's ability to detect cases. Epidemiologically speaking, such a model is useful insofar as it allows one to understand the socioeconomic factors influencing local government's capacity for detecting and reporting the disease, which may be useful when deciding where and how to allocate resources. See Chapter 8 for more on regression models.

7.3.5 Illustration: COVID-19 prevalence in Rio de Janeiro

In this section we will illustrate the application of prevalence estimation methods to a real data set from a serological survey conducted in Brazil by the EPICOVID study (Hallal et al., 2020b,a). The study is a nationwide survey of 133 sentinel cities for which a random sample of up to 250 households was selected. Individuals were then interviewed and tested for antibodies (IgG and IgM) against SARS-CoV-2 using a finger prick sample. Data were collected in three temporal sampling windows, henceforth called phases: phase 1 took place between the 14th and 21st of May 2020, phase 2 between the 4th and 7th of June 2020 and phase 3 between the 14th and the 21st of June. The data are publicly available at `http://www.rs.epicovid19brasil.org/banco-de-dados/`. According to the manufacturer of the test used in the EPICOVID study (Wondfo), $\delta = 0.848$ with 95% confidence interval (CI) $(0.814, 0.878)$ and $\gamma = 0.99$ with 95% CI $(0.978, 0.998)$. Using this information, we elicit the approximate hyperparameters $\alpha_\delta = 312$, $\beta_\delta = 49$, $\alpha_\gamma = 234$ and $\beta_\gamma = 1$.

The data, along with prevalence estimates obtained using the methods discussed here, are presented in Table 7.3. We show the naïve estimate along

TABLE 7.3: Prevalence estimates (%) for Rio de Janeiro from the EPICOVID data. We present a naïve estimator, the Rogan-Gladen estimator with correction for (fixed) sensitivity and specificity, along with results from fully Bayesian (FB) estimation (see text for details).

	Phase 1	Phase 2	Phase 3
Data (x/n)	5/243	16/250	22/250
Naïve (CI)	2.1 (1.1, 3.0)	6.4 (4.9, 7.9)	8.8 (7.0, 10.6)
Rogan-Gladen (CI)	1.9 (0.8, 3.0)	6.9 (5.1, 8.7)	9.7 (7.7, 11.8)
FB, mean (BCI)	2.3 (0.3, 4.9)	7.3 (4.1, 11.5)	10.1 (6.2, 14.7)
FB, median (HPD)	2.3 (0.0, 4.5)	7.2 (3.9, 11.2)	10.0 (5.9, 14.4)
FB-detection, median (HPD)	2.0 (0.8, 3.5)	6.1 (3.0, 9.3)	8.9 (5.3, 13.1)

with the Rogan-Gladen corrected estimator. Moreover, we show the results for the two fully Bayesian models discussed earlier. For the simpler model in (7.13), we also show the shortest posterior interval (which in this case coincides with the highest posterior density (HPD) interval) discussed by Gelman and Carpenter (2020) along with the more traditional equal-tailed Bayesian credibility interval (BCI). As expected, the HPD is usually tighter than the BCI, although in this example, that difference is unlikely to be of inferential import.

The results for the model with probability of detection in (7.16) are largely in agreement with those from the simpler model, but lead in general to slightly lower estimates for the prevalence. Recall that this model utilises case data and prior information about underreporting and thus in theory contains more information about the process (see below). In agreement with the recent literature (Gelman and Carpenter, 2020; Izbicki et al., 2020), these results make it very clear why it is important to take uncertainty about test characteristics into account.

We have also used the models discussed here to obtain corrected numbers of cumulative infections, I_t^{corr}. In most cases this amounts to a simple rescaling of the estimate for θ_t, that is, $\hat{I}_t^{\text{corr}} = \hat{\theta}_t M$. For the fully Bayesian models, this takes the form of a transformation of the posterior $\xi_T(\theta_t \mid x, n)$. Moreover, the model presented in Section 7.3.4 uses additional information about the probability of detection, p_{d}. This however necessitates the construction (elicitation) of a prior distribution for p_{d}. To achieve this, we used data from Wu et al. (2020), who suggest that overall, the number of actual infections is between 3 and 20 times the reported figures in the United States. This suggests that p_{d} is between 0.05 and 0.334. Covering these two values with 95% probability under a Beta distribution leads to $\alpha_{\text{d}} = 4$, $\beta_{\text{d}} = 20$ and $E[p_{\text{d}}] = 0.16$. These calculations are justifiable in our view because whilst Brazil —and, in particular Rio de Janeiro—and the USA are very different countries, the large variation found by Wu et al. (2020) is likely to encompass the true underreporting in our location of interest.

TABLE 7.4: Estimates of the actual number of cumulative cases in Rio de Janeiro, Y_t^{corr}. We use the same estimators as those in Table 7.3. Notice that for this quantity we report the usual equal-tailed 95% Bayesian credibility interval (BCI). RG = Rogan-Gladen; FB = Full Bayes, FB-det. = Full Bayes with detection probability.

	Phase 1	Phase 2	Phase 3
Cumulative cases (Y_t)	18.7	36.1	53.3
Naïve (CI)	138.8 (77.4, 200.3)	431.9 (327.4, 536.3)	593.8 (472.9, 714.7)
RG (CI)	128.0 (56.6, 199.5)	468.8 (347.3, 590.2)	657.1 (516.5, 797.7)
FB, mean (BCI)	158.5 (22.6, 331.6)	494.7 (275.5, 774.5)	681.2 (415.8, 991.1)
FB-det., mean (BCI)	142.2 (65.7, 256.4)	421.2 (228.2, 658.2)	607.1 (368.5, 901.0)

We estimate the probability (and 95% BCI) of detection in 0.15 $(0.07, 0.28)$ for the first phase, 0.09 $(0.05, 0.16)$ for the second phase and 0.09 $(0.06, 0.14)$ for the third phase. The drop in detection probability could be explained by a surge in cases, which overwhelms the healthcare system and leads to a smaller fraction of cases being reported.

Using notification data up to April 2020 and estimates of the case-fatality ratio (CFR) from the World Health Organisation (WHO), the reporting rate in Rio de Janeiro was estimated by Do Prado et al. (2020) at 0.072 with 95% confidence interval $(0.071, 0.073)$. Our estimates thus encompass those by Do Prado et al. (2020) but incorporate substantially more uncertainty, befitting of a Bayesian analysis.

Table 7.4 shows our estimates of the actual number of cumulative cases, Y_t^{corr}, along with data on the observed cases obtained from `https://brasil.io/home/`. It is important to note how the projections from the fully Bayesian models propagate the uncertainty about model parameters and lead to substantially wider BCIs. Nevertheless, none of the intervals obtained includes the observed number of cases, showing decisively that there is substantial underreporting. Code and data to reproduce the analyses presented here can be found at `https://github.com/maxbiostat/COVID-19_prevalence_Rio`.

7.3.6 Model extensions

We now move on to discuss extensions to the single test model in (7.13) beyond the retesting presented in (7.13). Extensions fall along two main lines: (i) multiple tests and multiple populations and (ii) multilevel (random effects, hierarchical) formulations of the single test model. We shall discuss these in turn.

The first issue one might be confronted with, especially during an epidemic of a new disease, is the absence of a reliable gold standard. If, however, one has access to two (or more) imperfect tests, $T^{(1)}$ and $T^{(2)}$, one can still estimate θ_t by testing n individuals and analysing the resulting contingency

table under an appropriate model for the data. One can either assume conditional independence or conditional dependence between $T^{(1)}$ and $T^{(2)}$ and perform Bayesian model selection and averaging to obtain estimates of θ_t that take model uncertainty into account. The choice between conditional dependence or independence is largely problem-specific: if $T^{(1)}$ and $T^{(2)}$ measure different biological processes, then conditional independence is a valid assumption (Branscum et al., 2005). See Black and Craig (2002) for a detailed treatment of the problem of estimating prevalence without a gold standard.

Another issue one might be confronted with is having two tests and two populations with conditional dependence between $T^{(1)}$ and $T^{(2)}$ from which one can use the assumption that (δ_1, γ_1) and (δ_2, γ_2) do not change across populations to alleviate identifiability problems and estimate θ_{t1} and θ_{t2} efficiently. An important observation is that all derivations for model variations of type (i) follow the same basic structure of those presented here: basic conditional probability manipulations. See Branscum et al. (2005) for an excellent treatment of more model variations such as the three tests and k populations model.

We now consider model extensions of type (ii) above: multilevel models. As shown by the illustration in Section 7.3.5, serological surveys are usually performed across locations and time points and their study designs are more often than not non-ignorable. Fortunately, in multilevel (also called random effects or hierarchical) models, the analyst has a powerful tool set at their disposal.

If one has samples from multiple cities from multiple states, one can construct a hierarchical model where there is a general country-wide "effect" which is shared across states, state effects which are shared across cities and city-specific effects for the prevalence. Moreover, one can construct a model which takes spatial dependence into account (see Section 8.5). This may be particularly important in the context of novel a pathogen because disease spread in a naïve population is closely linked to mobility patterns, which have an obvious spatial component. It must be observed that the spatial component in this case cannot be accurately described by a simple nearest-spatial-neighbour structure (Coelho et al., 2020) and that external mobility data must be incorporated into the model in order to construct a meaningful spatial dependence structure.

Another main line of inquiry is taking measurements over time, since prevalence is a dynamic quantity, under the assumption of complete and permanent immunity, θ_t is a monotonically non-decreasing function of time. Dynamic and time series models (cf. Section 8.3) can be leveraged to account for temporal dependence between measurements taken from the target population longitudinally. Flexible priors, such as Gaussian processes, can be employed to include monotonicity and other epidemiologically-motivated constraints (Riihimäki and Vehtari, 2010).

Additionally, spatial and temporal structures can be combined in order to account for complex sampling patterns. A main consideration, however, is

computational tractability. The analyst has to balance one hand capturing the sampling structure of the available data with formulating a tractable model for which parameters and other quantities of interest (such as latent variables) are estimable on the other.

A chief application of multilevel models is obtaining projections of θ_t for the whole population from which the sample under analysis was taken. This entails accounting for sample characteristics such as sex, race and socio-economic composition and then projecting what the outcome of interest would be in the general population. In a Bayesian setting, once one has obtained the posterior distribution of model coefficients, one can sample from the posterior predictive distribution using the population characteristics in order to obtain detailed projections of prevalence. This procedure is called multilevel modelling and poststratification (MRP) and has been recently applied to COVID-19 prevalence estimation (Gelman and Carpenter, 2020).

7.3.7 Open problems

Despite (Bayesian) prevalence estimation being a long-standing problem, there are still many open avenues of research. Most approaches, including the one presented in this section, assume independence between a test's sensitivity and its specificity. In reality, this is not a reasonable assumption. Mathematically, we expect δ and γ to be negatively correlated. Thus, explicitly incorporating (prior) dependence between sensitivity and specificity is still an open problem deserving of consideration.

In an epidemic context, temporal dependence is induced by a very specific process: disease transmission. A very important line of investigation is integrating epidemic models of the SIR/SEIR type and prevalence estimation. In addition to explicitly modelling temporal dependence, using epidemic models allows for the estimation of scientifically relevant quantities such as the transmission rate and the effective reproductive number (R_t). More importantly, one can use the fitted models to make epidemiological projections under different scenarios, an almost impossible task with unstructured time-series models. See Larremore et al. (2020) for an application of this framework to COVID-19 in the United States of America.

Finally, we shall address a crucial assumption made throughout this section: that the disease leaves permanent and complete immunity. For COVID-19, for instance, this assumption has been shown to not hold completely. In practice, processes such as antibody waning might interfere with our ability to estimate θ_t precisely if we do not account for the fact that δ and γ are now functions of time. Moreover, there is recent evidence that a certain fraction of asymptomatic individuals never develop detectable levels of antibodies (Milani et al., 2020). It is possible to use the same techniques discussed in this section to derive a model that accounts for a fraction of non-responding individuals. To the best of our knowledge however, this remains an unexplored line of inquiry in the literature.

Acknowledgments

The authors thank the Brazilian National Influenza Surveillance Network (Central Laboratories, National Influenza Centers, state and municipal health secretariats' surveillance teams, and the Influenza Working Group, Department of Immunization and Communicable Diseases of the Health Surveillance Secretariat, Brazilian Ministry of Health) for their partnership.

Bibliography

Barbosa, M. T. S. and Struchiner, C. J. (2002) The estimated magnitude of AIDS in Brazil: A delay correction applied to cases with lost dates. *Cadernos de Saúde Pública*, **18**, 279–285.

Bastos, L. S., Economou, T., Gomes, M. F. C., Villela, D. A. M., Coelho, F. C., Cruz, O. G., Stoner, O., Bailey, T. and Codeço, C. T. (2019) A modelling approach for correcting reporting delays in disease surveillance data. *Statistics in Medicine*, **38**, 4363–4377.

Bastos, L. S., Niquini, R. P., Lana, R. M., Villela, D. A. M., Cruz, O. G., Coelho, F. C., Codeço, C. T., Gomes, M. F. C., Bastos, L. S., Niquini, R. P., Lana, R. M., Villela, D. A. M., Cruz, O. G., Coelho, F. C., Codeço, C. T. and Gomes, M. F. C. (2020) COVID-19 and hospitalizations for SARI in Brazil: A comparison up to the 12th epidemiological week of 2020. *Cadernos de Saúde Pública*, **36**, e00070120.

Black, M. A. and Craig, B. A. (2002) Estimating disease prevalence in the absence of a gold standard. *Statistics in Medicine*, **21**, 2653–2669.

Branscum, A., Gardner, I. and Johnson, W. (2005) Estimation of diagnostic-test sensitivity and specificity through Bayesian modeling. *Preventive Veterinary Medicine*, **68**, 145–163.

Cai, B., Ioannidis, J., Bendavid, E. and Tian, L. (2020) Exact inference for disease prevalence based on a test with unknown specificity and sensitivity. *arXiv preprint arXiv:2011.14423*.

Centers for Disease Control (1988) Guidelines for evaluating surveillance systems. *Morbidity and Mortality Weekly Report*, **37**.

Codeço, C. T., Villela, D. A. M. and Coelho, F. C. (2018) Estimating the effective reproduction number of dengue considering temperature-dependent generation intervals. *Epidemics*, **25**, 101–111.

Coelho, F. C., Lana, R. M., Cruz, O. G., Villela, D. A., Bastos, L. S., Pastore y Piontti, A., Davis, J. T., Vespignani, A., Codeço, C. T. and Gomes, M. F. (2020) Assessing the spread of COVID-19 in Brazil: Mobility, morbidity and social vulnerability. *PloS one*, **15**, e0238214.

Diggle, P. J. (2011) Estimating prevalence using an imperfect test. *Epidemiology Research International*, **2011**.

Do Prado, M. F., de Paula Antunes, B. B., Bastos, L. D. S. L., Peres, I. T., Da Silva, A. D. A. B., Dantas, L. F., Baião, F. A., Maçaira, P., Hamacher, S. and Bozza, F. A. (2020) Analysis of COVID-19 under-reporting in Brazil. *Revista Brasileira de Terapia Intensiva*, **32**, 224.

Flor, M., Weiß, M., Selhorst, T., Müller-Graf, C. and Greiner, M. (2020) Comparison of Bayesian and frequentist methods for prevalence estimation under misclassification. *BMC Public Health*, **20**, 1–10.

Gamerman, D. and Lopes, H. F. (2006) *Markov Chain Monte Carlo Stochastic Simulation for Bayesian Inference*. New York: Chapman and Hall/CRC, 2nd edn.

Gelman, A. and Carpenter, B. (2020) Bayesian analysis of tests with unknown specificity and sensitivity. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **69**, 1269–1283. URLhttps://rss.onlinelibrary.wiley.com/doi/abs/10.1111/rssc.12435.

Greenland, S. (1996) Basic methods for sensitivity analysis of biases. *International Journal of Epidemiology*, **25**, 1107–1116.

— (2009) Bayesian perspectives for epidemiologic research: III. Bias analysis via missing-data methods. *International Journal of Epidemiology*, **38**, 1662–1673.

Hallal, P. C., Barros, F. C., Silveira, M. F., Barros, A. J. D. d., Dellagostin, O. A., Pellanda, L. C., Struchiner, C. J., Burattini, M. N., Hartwig, F. P., Menezes, A. M. B. et al. (2020a) EPICOVID19 protocol: Repeated serological surveys on SARS-CoV-2 antibodies in Brazil. *Ciência & Saúde Coletiva*, **25**, 3573–3578.

Hallal, P. C., Hartwig, F. P., Horta, B. L., Silveira, M. F., Struchiner, C. J., Vidaletti, L. P., Neumann, N. A., Pellanda, L. C., Dellagostin, O. A., Burattini, M. N. et al. (2020b) SARS-CoV-2 antibody prevalence in Brazil: Results from two successive nationwide serological household surveys. *The Lancet Global Health*, **8**, e1390–e1398.

Höhle, M. and an der Heiden, M. (2014) Bayesian nowcasting during the STEC O104:H4 outbreak in Germany, 2011. *Biometrics*, **70**, 993–1002.

Izbicki, R., Diniz, M. A. and Bastos, L. S. (2020) Sensitivity and specificity in prevalence studies: The importance of considering uncertainty. *Clinics*, **75**.

La Marca, A., Capuzzo, M., Paglia, T., Roli, L., Trenti, T. and Nelson, S. M. (2020) Testing for SARS-CoV-2 (COVID-19): A systematic review and clinical guide to molecular and serological in-vitro diagnostic assays. *Reproductive Biomedicine Online*, **41**, 483–499.

Lana, R. M., Coelho, F. C., Gomes, M. F. d. C., Cruz, O. G., Bastos, L. S., Villela, D. A. M. and Codeço, C. T. (2020) The novel coronavirus (SARS-CoV-2) emergency and the role of timely and effective national health surveillance. *Cadernos de Saúde Pública*, **36**, e00019620.

Larremore, D. B., Fosdick, B. K., Bubar, K. M., Zhang, S., Kissler, S. M., Metcalf, C. J. E., Buckee, C. and Grad, Y. (2020) Estimating SARS-CoV-2 seroprevalence and epidemiological parameters with uncertainty from serological surveys. *medRxiv*.

Liu, Y., Gelman, A. and Zheng, T. (2015) Simulation-efficient shortest probability intervals. *Statistics and Computing*, **25**, 809–819.

Mack, T. (1993) Distribution-free calculation of the standard error of chain ladder reserve estimates. *ASTIN Bulletin: The Journal of the IAA*, **23**, 213–225.

Martins, T. G., Simpson, D., Lindgren, F. and Rue, H. (2013) Bayesian computing with INLA: New features. *Computational Statistics & Data Analysis*, **67**, 68–83.

Milani, G. P., Dioni, L., Favero, C., Cantone, L., Macchi, C., Delbue, S., Bonzini, M., Montomoli, E. and Bollati, V. (2020) Serological follow-up of SARS-CoV-2 asymptomatic subjects. *Scientific Reports*, **10**, 1–7.

Ministério da Saúde and Secretaria de Vigilância em Saúde (2020) Guia de Vigilância Epidemiológica Emergência de Saúde Pública de Importância Nacional pela Doença pelo Coronavírus 2019: Vigilância de Síndromes Respiratórias Agudas COVID-19.

Ministério da Saúde, Secretaria de Vigilância em Saúde and Coordenação-Geral de Desenvolvimento da Epidemiologia em Serviços (2019) *Guia de Vigilância Em Saúde*. Brasília: Ministério da Saúde, third edn.

Ministério da Saúde, Secretaria de Vigilância em Saúde and Departamento de Vigilância das Doenças Transmissíveis (2018) Plano de Contingência para Resposta às Emergências de Saúde Pública: Influenza – Preparação para a Sazonalidade e Epidemias.

Ministério da Saúde, SINAN (2012) Ficha de Registro individual destinada para unidades com internação. Síndrome Respiratória Aguda Grave (SRAG) - internada ou óbito por SRAG.

Noufaily, A., Farrington, P., Garthwaite, P., Enki, D. G., Andrews, N. and Charlett, A. (2016) Detection of infectious disease outbreaks from laboratory data with reporting delays. *Journal of the American Statistical Association*, **111**, 488–499.

Parikh, R., Mathai, A., Parikh, S., Sekhar, G. C. and Thomas, R. (2008) Understanding and using sensitivity, specificity and predictive values. *Indian Journal of Ophthalmology*, **56**, 45.

Renshaw, A. E. and Verrall, R. J. (1998) A stochastic model underlying the chain-ladder technique. *British Actuarial Journal*, **4**, 903–923.

Riihimäki, J. and Vehtari, A. (2010) Gaussian processes with monotonicity information. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, 645–652.

Rogan, W. J. and Gladen, B. (1978) Estimating prevalence from the results of a screening test. *American Journal of Epidemiology*, **107**, 71–76.

Rue, H., Martino, S. and Chopin, N. (2009) Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **71**, 319–392.

World Health Organization (2006) World Health Organization. Communicable disease surveillance and response systems: Guide to monitoring and evaluating.

— (2013) *Global Epidemiological Surveillance Standards for Influenza.*

Wu, S. L., Mertens, A. N., Crider, Y. S., Nguyen, A., Pokpongkiat, N. N., Djajadi, S., Seth, A., Hsiang, M. S., Colford, J. M., Reingold, A. et al. (2020) Substantial underestimation of SARS-CoV-2 infection in the United States. *Nature Communications*, **11**, 1–10.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com