Analysis of the COVID-19 pandemic in Bavaria: adjusting for misclassification

Felix Günther[1,2],Andreas Bender[1], Michael Höhle[3], Manfred Wildner[4], Helmut Küchenhoff[1]

[1]Statistical Consulting Unit StaBLab, LMU Munich, Germany
[2]Department of Genetic Epidemiology, University of Regensburg, Germany
[3]Department of Mathematics, Stockholm University, Sweden
[4]Bavarian Health and Food Safety Authority / Pettenkofer School of Public Health, Oberschleißheim, Germany

September 29, 2020

Abstract

We present a method for adjusting the observed epidemic curve of daily new COVID-19 onsets for possible misclassification in infection diagnostics. We discuss different assumptions for specificity and sensitivity of the person-specific COVID-19 diagnostics based on PCR-tests, which are the basis for the daily reported case counts. A specificity of less than one implies occurrence of false positive cases, which becomes particularly relevant with an increased number of tests. The recent increase in cases in Bavaria could therefore be smaller than reported. However, an increase in case counts can still be seen from Mid-July until September 2020. The additional consideration of a sensitivity less than one, i.e., the occurrence of false negative tests, results in an epidemic curve in which the daily case counts are increased by a constant factor, but the structure of the curve does not change considerably.

1 Introduction

Data on the daily number of newly reported COVID-19 cases are used at different regional levels to monitor the course of the epidemic. These figures are an important source of information on the current state of the epidemic and, along with other information, serve as a basis for decisions regarding the implementation or relaxation of political and social measures to control the epidemic dynamics. There are three main problems for the interpretation of these numbers: First, the temporal assignment is problematic because the reporting day of a case does not coincide with the day of disease onset or infection. Such reporting delays can lead to misjudgements with respect to the current state of the epidemic. Second, not all infected persons are recorded, because many infected persons have no or only mild symptoms, because persons refuse testing, or

because the testing capacities are insufficient. Third, there may be errors in the testing procedure, which lead to misclassification bias.

To address the first problem we developed a Bayesian *nowcasting* approach, which adjusts for reporting delay by utilizing individual-specific data on disease onset and reporting dates and yields an estimate of the epidemic curve, i.e., the daily number of new disease onsets (Günther et al., 2020).

In this work, we address the third problem of possible errors (misclassification) in the diagnostic test results. Basically, there are two types of error. On the one hand, a person that is infected can have a negative disease diagnostic (false negative). This relates to the sensitivity of the testing procedure, i.e., the probability that an infected individual has a positive test, $P(\text{Test positive}|\text{individual infected})$. Problems associated with false negatives have been discussed in the literature (e.g., Woloshin et al. (2020); Pepe (2004)). One particular danger of false negative testing in the context of infectious diseases is that infected individuals are not isolated (for this a positive test is needed) or are released from quarantine too early and, hence, are able to transmit the disease. On the population level, a low sensitivity leads to an under-estimation of the number of infected individuals, which adds to the under-estimation due to untested missed cases.

On the other hand, a person that is not infected can have a positive disease diagnostic. This relates to the specificity of the testing procedure, i.e., the probability that an individual that is not infected has a negative test result, $P(\text{Test negative}|\text{individual not infected})$. One minus the specificity is then the probability of a false positive. On the individual level, a possible consequence is a superfluous isolation, which also leads to quarantining of presumed contact persons, further contact tracing and further testing, wasting sparse time and resources of the public health workforce. On the population level, false positive results lead to an over-estimation of the number of infected individuals and, hence, could be the cause of intervention measures stricter than necessary. Since the total number of false positive cases increases with the number of performed tests (Seifried et al., 2020), there was a recent discussion in Bavaria, whether the observed increase of the number of COVID-19 cases in August and September of 2020 was caused by a massive extension of testing rather than the result of a real increase of cases (cf. Echtermann (2020)).

This discussion motivated our work. Our aim is thus to quantify possible effects of misclassification on the estimated epidemic curve in Bavaria under different assumptions for sensitivity and specificity of COVID-19 testing. We use a statistical method (*matrix method*) to adjust the daily reported case numbers for misclassification and combine the adjustment with our nowcasting approach (Günther et al., 2020).

The paper is organized as follows. We discuss available evidence and assumptions about the specificity and sensitivity of the COVID-19 diagnostic test (Section 2) and give an overview of the data used in the analysis (Section 3). In Section 4, we present a statistical method for the misclassification adjustment. Results are presented in Section 5 followed by a discussion in Section 6.

2 Evidence on the performance of COVID-19 diagnostics

We are interested in evaluating the effect of misclassification in COVID-19 diagnostics on the epidemic curve estimated from daily case counts by nowcasting (Günther et al., 2020). The

observation units in the analysis are persons and the characteristic of central interest is their current COVID-19 infection status at the time of testing. We therefore focus on the effect of a person-specific misclassification on the current infection status: a person that is currently infected by COVID-19 might be diagnosed as not infected (false negative) or a person that is currently not infected might be diagnosed as infected (false positive). Routine COVID-19 diagnostics is done based on the detection of unique sequences of virus RNA using PCR-tests on clinical respiratory tract specimens of examined individuals (World Health Organization, 2020). Several different commercially available PCR-tests that target different (one or several) viral genes are used by the laboratories. The *analytic* sensitivity and specificity of the PCR-tests applied to an adequately collected and handled specimen are generally reported to be very high (European Commission, 2020; Robert Koch Institut, 2020). There is, however, a lack of concrete numbers for different tests. In a proficiency test of testing laboratories from April 2020, the authors found an average target-specific specificity between 97.8% and 98.6%, and a sensitivity between 98.9% and 99.7%.

These numbers on the analytic performance in a laboratory setting do, however, not directly relate to the person-specific performance of COVID-19 diagnostics in a clinical or screening setting. With respect to false negative results, it is reported that the performance of PCR-tests in infected persons varies strongly depending on the time point of the test after infection. Kucirka et al. (2020) report a sensitivity of PCR-tests close to zero directly after infection, which increases to 80% on day 8 (i.e., three days after typical symptom onset) and then decreases again. A further source of error for false negative tests, apart from quantitatively insufficient viral RNA in the early pre-symptomatic phase, can be an incorrect pre-analytical collection and/or handling of samples. Since an infected person who tests negative in the pre-symptomatic phase might get tested a second time after the onset of symptoms, it is difficult to quantify the average overall sensitivity of the person-specific COVID-19 diagnostics. It could be anywhere in the range of 50% to 90%. There are results of further studies that support assumptions in this range (e.g., Watson et al. (2020); Padhye (2020)).

With respect to false positive results, it is in accordance with WHO recommendations to perform PCR-tests that target two genes (*dual-target* tests) and that tests with unclear or divergent results should be evaluated by experienced physicians or repeated (Robert Koch Institut, 2020). This should decrease the probability of false positive cases considerably compared to the specificity of a single PCR-test on a single target. False positive test results might also be induced by swaps or cross-contamination between samples of infected and not infected individuals. The frequency of such errors can, however, be expected to be rather low. Based on data on the number of examined individuals and positive cases it is possible to establish a plausible lower bound for the probability of false positive diagnostics by looking at the number of examined individuals and the number of reported cases assuming that all examined individuals are in truth not infected (see Section 5.2).

3 Data

We use data on reported COVID–19 cases from Bavaria from the mandatory notification data based on the German Infection Protection Act (IfSG). The data is provided by the Bavarian Health and Food Safety Authority (LGL) on a daily basis. It originates from the reports of the local health authorities (*Gesundheitsämter*) to the LGL (for details see Günther et al. (2020)).

Furthermore, the LGL also provides daily numbers of performed tests as well as the number

of positive tests from different Bavarian COVID-19 testing facilities (laboratories). Because the testing data represent a different data source than the IfSG case numbers (the testing data are reported directly from the laboratories to the LGL on an aggregated basis), and due to potential delays in the reporting of a positive test to the local and regional health authorities, the reported number of positive tests in the second source do not exactly match the number of daily new cases at the local Bavarian health authorities as given in the first data source. Furthermore, the daily number of positive tests also does not map 1-1 to the number of cases, because some individuals are tested several times (e.g., testing of COVID-19 cases in order to be discharged from the hospital) and the laboratories can also test persons that do not live in Bavaria, while Bavarian citizens may also be tested in laboratories outside Bavaria. In the first case, positive test results would be reported the local health authorities at the corresponding place of residence and the case would not be part of the Bavarian COVID-19 case data. In the second case positive test results may be reported to the local health authorities directly from outside Bavaria without contributing to the aggregate laboratory report.

Figure 1 shows the daily number of performed tests as reported by the laboratories (Figure 1A) as well as the number of reported positive tests from the laboratories and the number of COVID-19 cases reported by the local Bavarian health authorities (Figure 1B). Data is shown from March, 16, until September, 20. There is a strong increase in the number of performed tests starting from July. Looking at the number of positive tests and reported cases it can be seen, that the daily number of positive tests reported by the laboratories is bigger than the number of reported COVID-19 cases in Bavaria for most of the days after April, 15. Altogether, the laboratories reported 3,543,112 performed tests and 74,609 positive results in the considered time period, while there were 65,066 positive COVID-19 cases registered at the local Bavarian health authorities.

Looking at the comparison of reported positive tests per day and the number of newly registered cases in Figure 1B we find a change of their association roughly in Mid-April. During the first outbreak in March, there were more cases reported at the health authorities than positive reported tests by the laboratories. Reasons for this might be the incompleteness of the reporting laboratories, and errors in data handling in the critical situation. In our subsequent analyses, we will focus on the period after May 1st, since the available testing data before this time point appears to be questionable and the period after May 1st is also the most interesting with respect to changes in testing activity and potential effects of misclassification on the epidemic curve.

4 Methods

We denote a binary indicator variable Y_i for the event that a person i is currently infected. $Y_i = 1$ means that a person is infected and $Y_i = 0$ that a person is not infected. The result of the person's COVID-19 diagnostic (e.g., one or multiple sequential PCR-tests) is denoted by Y_i^*. $Y_i^* = 1$ corresponds to a classification as infected and $Y_i^* = 0$ to a negative COVID-19 diagnostic. Furthermore, the sensitivity (i.e., the probability of a true positive examination) is defined by $\text{sens} = P(Y^* = 1|Y = 1)$ and the specificity (probability of true negative examination) is denoted by $\text{spec} = P(Y^* = 0|Y = 0)$. Then the probability of a positive examination is given by

$$P(Y^* = 1) = P(Y = 1) \cdot \text{sens} + P(Y = 0) \cdot (1 - \text{spec}) \tag{1}$$

We denote the (unknown) number of examined individuals on a certain day t by NT_t and the

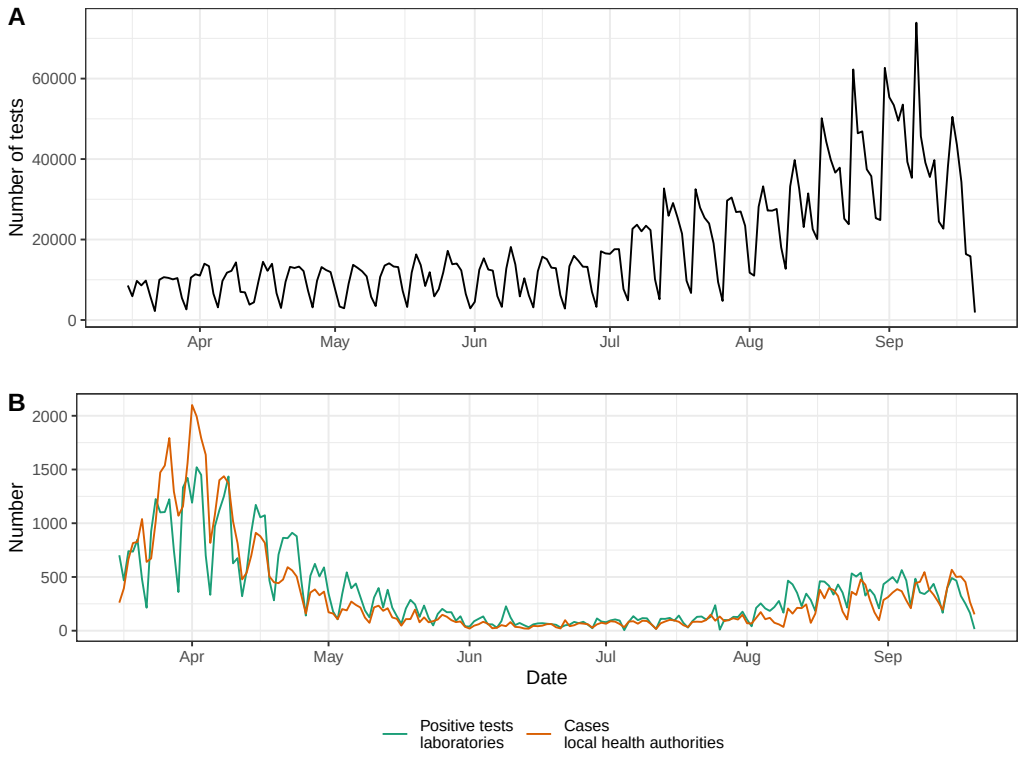

Figure 1: Number of performed COVID-19 PCR-tests per day as reported by the laboratories (Panel A). Number of positive COVID-19 tests per day as reported by Bavarian testing facilities (green) and number of COVID-19 cases as reported by local Bavarian health authorities (*Gesundheitsämtern*, orange, Panel B).

number of positive diagnostic procedures (i.e., reported cases) by N_t^*. Then the expected number of observed cases on day t is

$$E(N_t^*) = N_t \cdot \text{sens} + (NT_t - N_t) \cdot (1 - \text{spec}), \tag{2}$$

where N_t denotes the (unknown) number of infected individuals that are tested on day t. Equation (2) is derived from equation (1) by replacing the probabilities $P(\cdot)$ by the corresponding relative frequencies, $P(Y^* = 1) = N_t^*/NT_t$, $P(Y = 1) = N_t/NT_t$, and $P(Y = 0) = (NT_t - N_t)/NT_t$, and multiplying both sides with the number of conducted tests, NT_t.

Equation (2) shows that the effects of sensitivity and specificity on the observed case counts are different. While a low sensitivity leads to an under-estimation of the number of cases by the factor sens, the effect of specificity is additive and depends on the number tested individuals.

If all information, i.e., the number of reported cases N_t^*, the number of examined individuals NT_t, and the sensitivity and specificity of the diagnostic procedure are known, the number of

true cases can be estimated based on (2) by

$$\hat{N}_t = \frac{N_t^* - NT_t \cdot (1 - \text{spec})}{\text{sens} + \text{spec} - 1} \tag{3}$$

This estimator relates to the well known matrix method, see e.g., Rogan and Gladen (1978).

Since only positive COVID-19 examinations of people residing in Bavaria, N_t^*, are directly reported to the Bavarian health authority, the corresponding number of examined individuals, NT_t, is unknown. However, the overall number of performed tests is separately reported by the laboratories. As described in Section 3, this data can have a difference in temporal allocation and, furthermore, there is an upward bias in the reported (positive) test numbers from the laboratories due to multiple tests on single individuals and tests of individuals living outside of Bavaria. To establish the relationship between the two quantities, we utilize the positive test results and model the number of reported cases at the local health authorities based on the number of positive tests reported by the laboratories from the current and previous days (lagged time series of reported positive tests from the laboratories). We utilize two different models and consider different degrees for the lag-number of positive tests. The first model corresponds to standard linear regression with linear effects of the (lag-)number of positive tests:

$$E(N_t^*) = \sum_{l=0}^{L} \alpha_l \cdot NL_{t-l}, \quad t = \text{May 1st}, \ldots, T. \tag{4}$$

Here, the number of positive tests reported on day t by the laboratories is denoted by NL_t and T corresponds to the most current day, in our case September, 13. We estimate the model based on different values for the maximum lag $L = 0, \ldots, 7$. As a second model, we consider a varying-coefficient model (Hastie and Tibshirani, 1993), in which the linear effect α_l of the (lagged) number of positive tests, NL_{t-l}, varies smoothly over time:

$$E(N_t^*) = \sum_{l=0}^{L} \alpha_l(t) \cdot NL_{t-l}, \quad t = \text{May 1st}, \ldots, T. \tag{5}$$

This model is estimated for different values of $L = 0, \ldots, 7$ as well. From all 14 estimated models, we select the best performing model based on the Bayesian information criterion (BIC).

Assuming a similar fraction of infected individuals in the tests relevant for the Bavarian data and in all tests from the laboratories, we estimate the number of relevant tests NT_t by using the estimated parameters of the selected model (4) or (5):

$$\widehat{NT}_t = \sum_{l=0}^{L} \hat{\alpha}_l(t) \cdot NL_{t-l}, \tag{6}$$

where NTL_t denotes the total number of test reported on day t by the laboratories and $\hat{\alpha}_l(t)$ are the estimated effects of the (lag-)number of positive tests reported by the laboratories. Depending on whether a model of type (4) or (5) is selected, they correspond to time-constant (linear) effects $\alpha_l(t) = \alpha_l$ or (linear) effects that vary smoothly over time.

Plugging in the daily observed numbers of positive COVID-19 cases, N_t^*, and the derived number of relevant examinations per day, $\widehat{NT}_t$ from (6), into (3), we obtain adjusted estimates for the daily number of COVID-19 cases at local Bavarian health authorities, $\hat{N}_t$.

In our nowcasting approach, we estimate the number of cases with disease onset on a specific day based on a complex Bayesian hierarchical model using individual-specific data on the reporting and disease onset date (cf. Günther et al. (2020)). To apply the proposed adjustment for misclassification, we first focus on the scenario of no false negative examinations (sensitivity equals one) and a reduced specificity smaller one. We then calculate the expected number of false positives reported to the health authorities on a certain day based on the difference of

$$\hat{N}_t = \frac{N_t^* - NT_t \cdot (1 - \text{spec})}{\text{spec}} \tag{7}$$

and the reported number of new cases N_t^*. Then, we randomly delete this number of observations from our data set and apply the nowcasting to the reduced data set to estimate the *false positive adjusted* epidemic curve. Based on the results estimated epidemic curve from nowcasting it is possible to estimate the effective time-varying reproduction number $R_e(t)$ as described in Günther et al. (2020).

To take possible false negatives into account, we can rewrite formula (3) by

$$\hat{N}_t = \frac{N_t^* - NT_t \cdot (1 - \text{spec})}{\text{spec}} \cdot \frac{\text{spec}}{\text{sens} + \text{spec} - 1} \tag{8}$$

and plug in different values for the sensitivity, sens < 1. The first term of (8) corresponds to the false positive from (7) and the second part is a constant factor independent of the number of examinations per day.

Since the *false negative* adjustment relies on a constant factor which is independent of the number of tested individuals, the factor $\text{spec}/(\text{sens} + \text{spec} - 1)$ can be directly applied to the result of the false positive adjusted nowcasting procedure, which reduces the computational effort considerably compared to a repeated application of the nowcasting to (upsampled) data. Note, that $\text{spec}/(\text{sens} + \text{spec} - 1) \approx 1/\text{sens}$ for a specificity close to one. The false negative adjustment corresponds therefore roughly to a point-wise up-scaling of the estimated epidemic curve by the reciprocal sensitivity.

Following the discussion in Section 2 we perform calculations based on an assumed specificity of spec $\in \{0.995, 0.997, 0.999\}$ and sensitivity of sens $\in \{0.8, 0.9\}$ for the person-specific COVID-19 examination and compare them to the analysis based on the daily reported case counts, i.e., assuming a sensitivity and specificity of one.

All calculations were done using the statistical programming environment `R` (R Core Team, 2020). To estimate the (varying-coefficient) regression model of the reported number of cases on the reported number of positive tests reported by the laboratories, we used the `mgcv` package (Wood, 2011). The Bayesian hierarchical model for nowcasting was implemented in `rstan` (Stan Development Team, 2020) and estimation of $R_e(t)$ was based on code of the `R0` package (Obadia et al., 2012) for each MCMC sample. Methodological details can be found in Günther et al. (2020) and corresponding code is available at `https://github.com/FelixGuenther/nc_covid19_bavaria`.

5 Results

5.1 Model for the relation between the tests reported from laboratories and the reported cases from health authorities

To derive the relevant number of performed COVID-19 examinations for a specific day, we perform the two-step approach of modeling the reported number of cases per day based on the number of positive COVID-19 tests from the laboratories in a regression model and plug in the total number of COVID-19 tests in the estimated model (see Section 4).

As described in Section 4 we perform a model selection based on the BIC to decide between time-constant and time-varying effects as well as the maximum considered lag-order. With respect to the BIC, the best performing model is the model (5) with time-varying linear effects of the number of positive tests at the laboratories and a maximum lag of $L = 2$.

Figure 2A shows the time-series of both measurements as well as the estimated number of cases based on the selected regression model. Starting from May, 1st, there are 29,408 positive tests reported by the laboratories and 22,408 COVID-19 cases reported at the local Bavarian health authorities. Based on the model, we would expect 22,271 reported COVID-19 cases at the local health authorities over the whole period. The model appears to fit the data with respect to the daily and cumulative numbers of new cases well and has an explained variance of $r^2 = 94.7\%$.

Figure 2B shows the total number of daily PCR-tests reported by the laboratories and the estimated number of COVID-19 examinations that would be reported as new cases to the local Bavarian health authorities on a given day in case of a positive result, as defined in (6).

5.2 Case numbers registered at local health authorities adjusted for false positive cases

Based on the estimated total number of relevant examinations per day and assumptions regarding the specificity of the person-specific diagnostic testing procedure, we adjust the daily number of new cases for false positives. Figure 3 compares the observed number of new cases at the local health authorities (specificity of one, no false positives) to the adjusted numbers assuming a specificity of 0.999, 0.997, and 0.995. Here, we assume a perfect sensitivity of the diagnostic procedure. While the effect of the adjustment is rather low in May, the absolute differences in the observed and adjusted counts get bigger with an increasing number of examined individuals in starting in July. Assuming a specificity of 99.5% yields the same or more expected false positive cases as actually reported by the health authorities in 26 of the 143 considered days (from May, 1st until September, 14; 18.2%; adjusted case numbers are set to zero for those days in Figure 3). Assuming a specificity $\leq$ 99.5% for the person-specific COVID-19 diagnostic appears therefore like a very strong assumption that is not supported by the Bavarian COVID-19 data.

5.3 Adjusted nowcast

Figure 4A shows the adjusted nowcast estimates of the epidemic curve assuming a sensitivity of one and different values of specificity. The figure shows a lower increase of disease onsets when assuming a specificity of 0.997 or 0.995 compared to the unadjusted nowcast. However, the increase in the estimated epidemic curve starting in Mid-July can still be identified from the adjusted curves. Panel B of Figure 4 shows the ratio of the estimated adjusted number

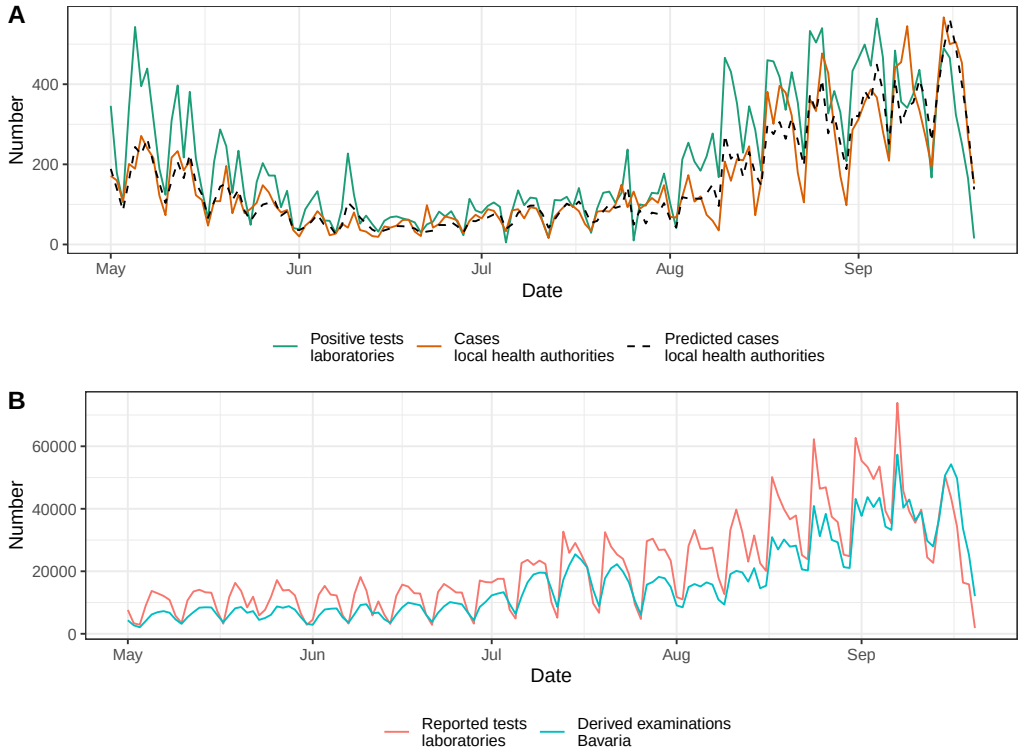

Figure 2: Results of time-varying regression model for number of reported COVID-19 cases in Bavaria based on the reported number of positive PCR-tests by the laboratories (Panel A). Panel B shows the daily numbers of reported PCR-tests by the laboratories, as well as the derived daily numbers of relevant COVID-19 examinations in Bavaria per day.

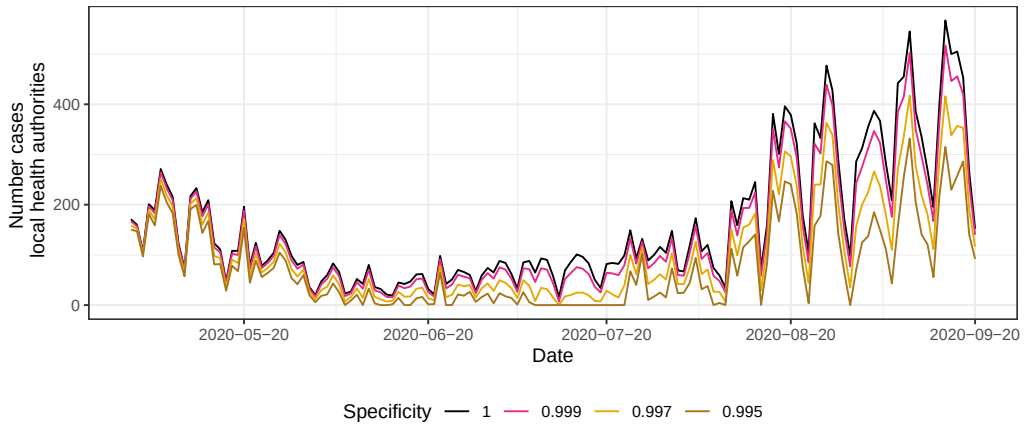

Figure 3: Daily number of new cases reported to local health authorities as given in official COVID-19 case data by LGL and adjusted using an assumption of a specificity of 0.999, 0.997, 0.995 (corresponding to a probability of 0.1%, 0.3%, and 0.5% false positive COVID-19 diagnostics for persons that get tested but are not infected).

of disease onsets and the unadjusted results. This ratio can be interpreted in the sense of a positive predictive value of a reported case. Independent of the concrete assumptions regarding the specificity of the person-specific COVID-19 diagnostic it can be seen, that the effect of false positive classifications was biggest around beginning of July, when the performed number of tests started to increase and case counts were lowest. With rising case counts, the relative share of false positives is decreasing again (due to the additive effect of the false positives). In all plausible scenarios ($0.995 \leq \text{spec} \leq 1$), the increase in testing and the resulting numbers of potential false positives are not big enough to fully explain the recent rise of the epidemic curve.

Figure 5 shows the estimated time-varying effective reproduction number $\hat{R}_e(t)$ utilizing the nowcasting results under different assumptions for the specificity smaller one. The effective reproduction number characterizes the short-term dynamics of the epidemic by estimating the average number of individuals that are infected by an individual with disease onset on a given day. If $R_e(t)$ is smaller than one, case numbers are decreasing, if it is bigger than one, case numbers are increasing within the following days. Depending on the assumption with respect to the specificity, the estimated $\hat{R}_e(t)$ differs biggest in the beginning and mid of July. As already discussed above, the effect of the false positive adjustment is biggest in this period. Note that at beginning of July, the adjusted $\hat{R}_e(t)$ was smaller one when assuming a specificity of 0.995 or 0.997 and around one when not adjusting for false positives. During this period, actual case numbers may have decreased, but this might have been covered up in the reported case numbers due to increased testing and false positives. However, shortly thereafter, the estimated $\hat{R}_e(t)$ is clearly bigger than one for all considered assumptions for the specificity, and the estimated disease dynamics with increasing case numbers appear very similar in August and September.

Figure 6 shows the effect of a sensitivity smaller 1 (i.e., false negative test results) on the estimated epidemic curve under the assumption of a diagnostic specificity of 0.997. We show results for a sensitivity of 0.9 and 0.8 (i.e., we assume problems in the collection and/or handling

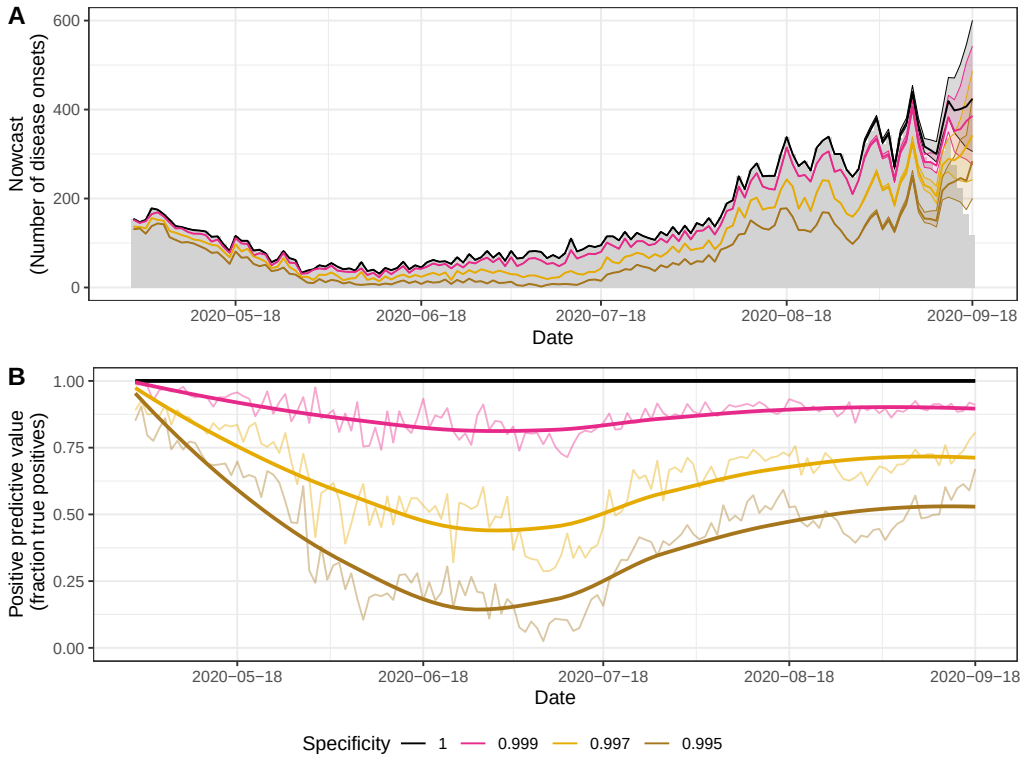

Figure 4: Daily number of new disease onsets in Bavaria based on nowcasting utilizing the official COVID-19 case data as well as different assumptions regarding the specificity of person-specific COVID-19 diagnostics and a sensitivity of one. Shown are daily point estimates of the Bayesian nowcasting (median posterior) as well as 95%-prediction intervals (Panel A). Panel B shows the positive predictive value of the COVID-19 diagnostic (i.e., the probability of a true infection given a registration as case), which corresponds to the fraction of the adjusted number of disease onsets and the unadjusted number of disease onsets when assuming a sensitivity of one.

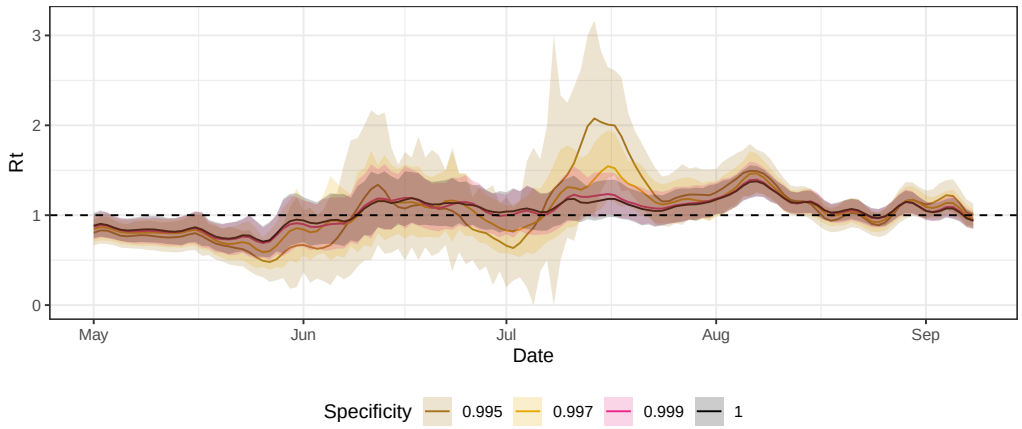

Figure 5: Estimated time-varying effective reproduction number $R_e(t)$ and associated 95%− confidence intervals based on the nowcasting results adjusted for for different assumptions for the specificity smaller one.

of 10% or 20% of the collected specimens). One can see the (point-wise) multiplicative increase in the estimated numbers of disease onsets when assuming a sensitivity < 1. This effect does not depend on the number of tested individuals. Assuming a specificity of 0.997 and a sensitivity of 0.8 (i.e., adjusting for false positive and false negative tests), we see that the absolute number of disease onsets is quite similar to the unadjusted analysis in August and September.

6 Discussion

Interpreting the epidemic curve estimated from daily reported COVID-19 case numbers is common, but has several drawbacks. Problems can occur if the number of performed tests changes over time and it is especially problematic, if the case definition or the diagnostic procedures change over time. In Bavaria, the number of COVID-19 tests was substantially increased during August 2020.

In this work, we focused on the effect of misclassification in COVID-19 diagnostics on the estimation of the epidemic curve and have shown that such misclassification can yield a bias in the estimated curve. When the person-specific sensitivity of the diagnostic procedure is smaller one, this leads to an under-estimation of the number of cases by a constant factor. When the specificity is smaller one, this leads to an over-estimation of the number of cases by an additive term that depends on the number of tested individuals and corresponds to the number of *false positive* cases. This effect of misclassification can become especially problematic when the number of tested individuals changes over time. In this case, the number of *false positive* cases changes over time as well, which can distort the apparent dynamics of the epidemic. We have quantified the impact of potential false positive case numbers in the Bavarian data and found that, under realistic assumptions, the adjusted epidemic curve is still increasing since Mid-July, but the increase might be smaller than the unadjusted curve suggests.

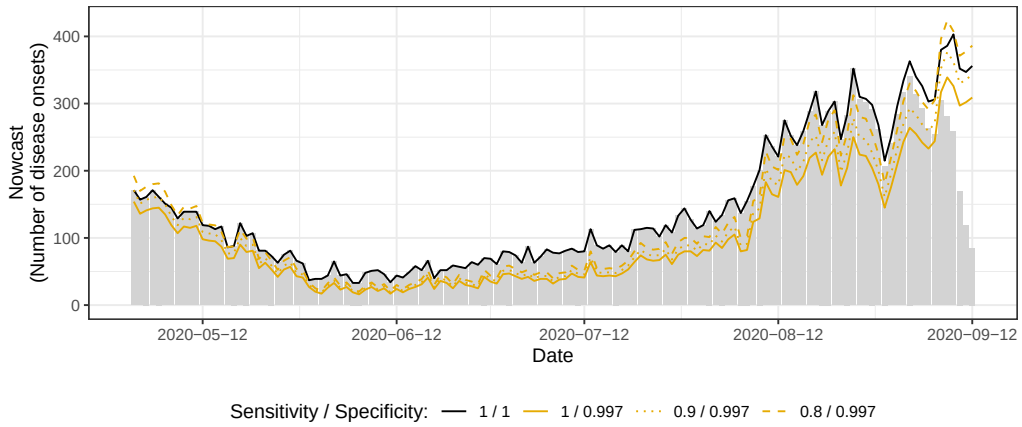

Figure 6: Daily number of new disease onsets in Bavaria based on nowcasting utilizing the official COVID-19 case data assuming a specificity of 0.997 and a sensitivity of 1.0, 0.9, and 0.8.

When addressing the effect of misclassification, the effect of false positives has to be considered jointly with the occurrence of false negatives (i.e., when the sensitivity of the person-specific COVID-19 diagnostic is smaller than 1). In this situation the curve of reported onsets on a given day estimated from the observed data has to be adjusted by a multiplicative factor bigger than one. The size of the adjustment depends on the true number of positives among the tested on a given day. Our analyses show that when considering both false positive and false negatives, the adjusted curves can be almost equivalent to the unadjusted curve during August and September. However, this statement is only true in the low incidence rate setting of that period together with a very high specificity and a sensitivity of around 80%.

Our analysis has some assumptions and limitations. The adjustment for misclassification in COVID-19 diagnostics depends on accurate information with respect to the number of examined individuals. Such information is not directly available for the Bavarian data and we rely on a model based approach for relating the reported case numbers to the reported number of positive tests from Bavarian laboratories. The results appear plausible but can not be directly validated. Another assumption are constant misclassification probabilities with respect to the person-specific COVID-19 diagnostics over time. This assumption might be violated in case of changes in the diagnostic procedures, changes in workload for the laboratories, additional laboratories that perform parts of the testing, and changes or improvements in operating processes. Indeed there is evidence that at least one of the laboratories newly entrusted with no-fee testing at the Bavarian southern borders was reporting single target positivity as positive test results. There is, however, still very few direct information on the quality of the PCR-testing available under field conditions and especially no comprehensive information on temporal changes. In consequence, misclassification bias due to imperfect specificity may vary regionally with respect to the area served by different laboratories and may well have been at the upper range of the assumptions for false positive results locally. Since nowcasting is performed on the level of aggregated case numbers in whole Bavaria, local (short-term) changes could, however, even out and we think that our calculations over ranges of plausibel assumption for sensitivity and speci-

ficity can give a realistic overview about potential biases and effects of misclassification on the estimated epidemic curve.

As stated in the introduction, potential misclassification in COVID-19 diagnostics is not the only problem related to the interpretation of reported case numbers. It is well known that a diagnostic is not always performed for COVID-19 infected individuals. This leads to a relevant difference between the number of infected individuals and the number of registered infections. This problem is, however, not COVID-19-specific, but also occurs with other diseases when they only cause mild symptoms in some cases - as typically described by the so-called surveillance pyramid (see e.g., Gibbons et al. (2014)). The extent of under-reporting can be characterized by the case detection ratio (number of registered cases divided by number of infected persons). There a first results of serological studies (e.g., Streeck et al. (2020), Santos-Hövener et al. (2020)) that aim to quantify the case detection ratio by comparing the numbers of reported cases, i.e., detected via positive PCR-tests, and all seropositive persons,i.e., persons with COVID-19 antibodies based on serological analyses. Those studies report case detection ratios in three different regions between approximately 0.2 and 0.4 for the first phase of the epidemic. This corresponds to a considerable fraction of unreported cases in the official surveillance data. If the case detection ratio is constant over time, the epidemic curve estimated from reported case data is reduced by a constant factor, but its structure remains the same. However, the case detection ratio might be influenced by the testing strategy and the expansion of testing could increase the detection ratio. Therefore, one has to be careful when comparing absolute number of cases over longer periods of time. More specifically, the absolute numbers of reported cases during March-April is most probably not directly comparable to the numbers during August and September. In our analysis and figures, we focused on the epidemic curve starting in May 2020, i.e., after the first phase of the epidemic. Nevertheless, the (remaining) increase of the (misclassification-adjusted) epidemic curve during August-September might still partly be driven by an increasing case detection ratio. This question can not be answered based on the real-time analysis of daily case numbers and additional information has to be considered, e.g., numbers of hospital admissions, deaths and longitudinal data on antibody prevalence.

Still, we believe that our real-time estimation and analysis of the epidemic curve based on nowcasting is a valuable tool to monitor short-term changes and the evolution of the epidemic situation. Based on the analysis presented in this manuscript, we rebut the statement that the recent increase in cases during August and September in Bavaria is only driven by false positive cases. These data-based results fit into considerations based on further knowledge about the epidemic situation: Bavarian school vacation ended on September 9, 2020, with a peak of test numbers in the first week of September and a decline thereafter, while case numbers remain elevated beyond this date. Hence, the increase in case numbers observed in September is unlikely a result of increased testing and false positive cases only. Moreover, local outbreaks following the return of citizens especially from high risk regions with subsequent positively tested contact persons validate an increase in truly infected reported cases. At the same time, an uncoupling between reported cases and hospitalizations and deaths is likely to be due to effective protective measures in high risk groups, such as the institutionalized elderly with relevant comorbidities, and could also be partly explained by an increased case detection ratio due to increased testing.

The aspect of false testing results is a relevant issue for both epidemiologic surveillance and decision making in public health and has been quantified for different scenarios for the first time in this article.

References

Echtermann, A. (2020). PCR-Test auf SARS-CoV-2: Warum in der Praxis falsch-positive Ergebnisse selten sind. `https://correctiv.org/faktencheck/hintergrund/2020/09/09/pcr-test-auf-sars-cov-2-warum-in-der-praxis-falsch-positive-ergebnisse-selten-sind`.

European Commission (2020). Current performance of COVID-19 test methods and devices and proposed performance criteria, 16 April 2020. Technical report, European Commission.

Gibbons, C. L., Mangen, M.-J. J., Plass, D., Havelaar, A. H., Brooke, R. J., Kramarz, P., Peterson, K. L., Stuurman, A. L., Cassini, A., Fèvre, E. M., et al. (2014). Measuring underreporting and under-ascertainment in infectious disease datasets: a comparison of methods. *BMC public health*, 14(1):147.

Günther, F., Bender, A., Katz, K., Küchenhoff, H., and Höhle, M. (2020). Nowcasting the COVID-19 pandemic in Bavaria. *medRxiv*.

Hastie, T. and Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 55(4):757–779.

Kucirka, L. M., Lauer, S. A., Laeyendecker, O., Boon, D., and Lessler, J. (2020). Variation in false-negative rate of reverse transcriptase polymerase chain reaction–based sars-cov-2 tests by time since exposure. *Annals of Internal Medicine*.

Obadia, T., Haneef, R., and Boëlle, P.-Y. (2012). The R0 package: a toolbox to estimate reproduction numbers for epidemic outbreaks. *BMC Medical Informatics and Decision Making*, 12(1):147.

Padhye, N. S. (2020). Reconstructed diagnostic sensitivity and specificity of the rt-pcr test for covid-19. *medRxiv*.

Pepe, M. (2004). *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Robert Koch Institut (2020). Hinweise zur Testung von Patienten auf Infektion mit dem neuartigen Coronavirus SARS-CoV-2: Direkter Erregernachweis durch RT-PCR (Stand: 18.09.2020). `https://www.rki.de/DE/Content/InfAZ/N/Neuartiges_Coronavirus/Vorl_Testung_nCoV.html`.

Rogan, W. and Gladen, B. (1978). Estimating prevalence from results of a screening-test. *American Journal of Epidemiology*, 107(1):71–76.

Santos-Hövener, C., Busch, M. A., Koschollek, C., Schlaud, M., Hoebel, J., Hoffmann, R., Wilking, H., Haller, S., Allen, J., Wernitz, J., et al. (2020). Seroepidemiologische Studie zur Verbreitung von SARS-CoV-2 in der Bevölkerung an besonders betroffenen Orten in Deutschland–Studienprotokoll von CORONA-MONITORING lokal. *Journal of Health Monitoring*, 5.

Seifried, J., Böttcher, S., Albrecht, S., Stern, D., Willrich, N., Zacher, B., Mielke, M., Rexroth, U., and Hamouda, O. (2020). Erfassung der SARS-CoV-2-Testzahlen in Deutschland (Stand 9.9.2020). *Epid Bull*, 37:12 – 15.

Stan Development Team (2020). RStan: the R interface to Stan. R package version 2.21.2.

Streeck, H., Schulte, B., Kuemmerer, B., Richter, E., Hoeller, T., Fuhrmann, C., Bartok, E., Dolscheid, R., Berger, M., Wessendorf, L., Eschbach-Bludau, M., Kellings, A., Schwaiger, A., Coenen, M., Hoffmann, P., Noethen, M., Eis-Huebinger, A.-M., Exner, M., Schmithausen, R., Schmid, M., and Hartmann, G. (2020). Infection fatality rate of SARS-CoV-2 infection in a German community with a super-spreading event. *medRxiv*.

Watson, J., Whiting, P. F., and Brush, J. E. (2020). Interpreting a covid-19 test result. *Bmj*, 369.

Woloshin, S., Patel, N., and Kesselheim, A. S. (2020). False Negative Tests for SARS-CoV-2 Infection — Challenges and Implications. *New England Journal of Medicine*, 383(6):e38.

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B)*, 73(1):3–36.

World Health Organization (2020). Laboratory testing for coronavirus disease (COVID-19) in suspected human cases: interim guidance, 19 March 2020. Technical report, World Health Organization.