



SOFTWARE TOOL ARTICLE

REVISED Bayes Lines Tool (BLT): a SQL-script for analyzing diagnostic test results with an application to SARS-CoV-2-testing [version 3; peer review: 2 approved]

Wouter Aukema ¹, Bobby Rajesh Malhotra ², Simon Goddek³,
Ulrike Kämmerer ⁴, Peter Borger ⁵, Kevin McKernan ⁶,
Rainer Johannes Klement ⁷

¹Independent Data and Pattern Scientist, Hoenderloo, 7351BD, The Netherlands

²Department for Digital Arts, University for Applied Arts Vienna, Vienna, 1030, Austria

³Independent Scientist, Ede, 6711 VS, The Netherlands

⁴Department of Obstetrics and Gynaecology, University Hospital of Würzburg, Würzburg, 97080, Germany

⁵The Independent Research Initiative on Information & Origins, Loerrach, 79540, Germany

⁶Medical Genomics, Beverly, MA, 01915, USA

⁷Department of Radiation Oncology, Leopoldina Hospital Schweinfurt, Schweinfurt, 97422, Germany

V3 First published: 10 May 2021, 10:369
<https://doi.org/10.12688/f1000research.51061.1>

Second version: 18 Jun 2021, 10:369
<https://doi.org/10.12688/f1000research.51061.2>











Latest published: 16 Feb 2022, 10:369
<https://doi.org/10.12688/f1000research.51061.3>


Abstract

The performance of diagnostic tests crucially depends on the disease prevalence, test sensitivity, and test specificity. However, these quantities are often not well known when tests are performed outside defined routine lab procedures which make the rating of the test results somewhat problematic. A current example is the mass testing taking place within the context of the world-wide SARS-CoV-2 crisis. Here, for the first time in history, laboratory test results have a dramatic impact on political decisions. Therefore, transparent, comprehensible, and reliable data is mandatory. It is in the nature of wet lab tests that their quality and outcome are influenced by multiple factors reducing their performance by handling procedures, underlying test protocols, and analytical reagents. These limitations in sensitivity and specificity have to be taken into account when calculating the real test results. As a resolution method, we have developed a Bayesian calculator, the Bayes Lines Tool (BLT), for analyzing disease prevalence, test sensitivity, test specificity, and, therefore, true positive, false positive, true negative, and false negative numbers from official test outcome reports. The calculator performs a simple SQL (Structured Query Language) query and can easily be implemented on any system supporting SQL. We provide an example of influenza test results from California, USA, as well

Open Peer Review

Approval Status  

	1	2
version 3		
(revision)		
16 Feb 2022		
version 2		
(revision)		
18 Jun 2021		
version 1		
10 May 2021		

- Phillip M. Bentley**, European Spallation Source ESS, Lund, Sweden
- Mariska M G Leeflang** , Amsterdam UMC, University of Amsterdam, Amsterdam, The Netherlands

Any reports and responses or comments on the article can be found at the end of the article.

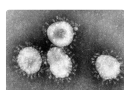
as two examples of SARS-CoV-2 test results from official government reports from The Netherlands and Germany-Bavaria, to illustrate the possible parameter space of prevalence, sensitivity, and specificity consistent with the observed data. Finally, we discuss this tool's multiple applications, including its putative importance for informing policy decisions.

Keywords

Bayes, COVID19, PCR Test, SARS-CoV-2; SQL



This article is included in the **Emerging Diseases and Outbreaks** gateway.



This article is included in the **Coronavirus** collection.

Corresponding authors: Wouter Aukema (wouter@aukema.org), Rainer Johannes Klement (rainer_klement@gmx.de)

Author roles: **Aukema W:** Conceptualization, Formal Analysis, Investigation, Methodology, Resources, Software, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Malhotra BR:** Conceptualization, Formal Analysis, Methodology, Project Administration, Resources, Validation, Visualization, Writing – Review & Editing; **Goddek S:** Conceptualization, Project Administration, Resources, Supervision, Validation, Writing – Review & Editing; **Kämmerer U:** Conceptualization, Resources, Supervision, Validation, Writing – Review & Editing; **Borger P:** Conceptualization, Resources, Supervision, Validation, Writing – Review & Editing; **McKernan K:** Conceptualization, Project Administration, Resources, Supervision, Validation, Writing – Review & Editing; **Klement RJ:** Conceptualization, Formal Analysis, Investigation, Methodology, Project Administration, Resources, Software, Supervision, Validation, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: The author(s) declared that no grants were involved in supporting this work.

Copyright: © 2022 Aukema W *et al.* This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Aukema W, Malhotra BR, Goddek S *et al.* **Bayes Lines Tool (BLT): a SQL-script for analyzing diagnostic test results with an application to SARS-CoV-2-testing [version 3; peer review: 2 approved]** F1000Research 2022, **10**:369 <https://doi.org/10.12688/f1000research.51061.3>

First published: 10 May 2021, **10**:369 <https://doi.org/10.12688/f1000research.51061.1>

REVISED Amendments from Version 2

The new version includes some further explanations about the usage of the BLT calculator and how results should be interpreted.

Any further responses from the reviewers can be found at the end of the article

1. Introduction

In December 2019, a cluster of patients with pneumonia of unknown origin was associated with the emergence of a novel beta-coronavirus,¹ first named 2019-nCoV² and later specified as severe acute respiratory syndrome-coronavirus-2 (SARS-CoV-2).³ This outbreak led to the rapid development of reverse transcriptase - quantitative polymerase chain reaction (RT-qPCR) tests to identify SARS-CoV-2 RNA in specimens obtained from patients.^{2,4}

After sporadic SARS-CoV-2 positive cases in January^{5,6} to the end of February 2020 worldwide cases of the SARS-CoV-2-associated disease ‘COVID-19’ began to accumulate, causing policymakers in many countries to introduce countermeasures. These non-pharmaceutical interventions predominantly started worldwide around March 2020 while the virus was characterized as a pandemic on 11 March, 2020.^{6,7} As a result, for almost two years now, large parts of the world are in a COVID-19 crisis-mode with daily reporting of SARS-CoV-2 cases in dashboards worldwide.⁸ The definition of ‘cases’ and ‘prevalence estimates’ was based on RT-qPCR testing, independent of the clinical diagnosis. Thereby, a person is considered a case (i.e., infected), once a test turns out positive.⁹

Like all laboratory tests, however, the SARS-CoV-2 RT-qPCR tests are not flawless. This is because sensitivity and specificity depend on a multiplicity of confounding factors. These factors cover the test design, the lab application, and possible contaminations with substances/nucleic acids interfering with the reaction.^{10,11} Consequently, both false-negative and false-positive results have been reported.^{12,13} Nevertheless, the test system’s limitations are rarely discussed in scientific publications and public health systems despite their crucial role for making inferences about the possible infection status of a tested person.¹⁴ Many more or less defined commercial and laboratory ‘in house’ tests are now routinely being used,¹⁵ often without standardised guidelines, which leads to entirely unknown test performance specifications.¹⁶ The few studies aiming to estimate sensitivity and specificity of SARS-CoV-2 RT-qPCR tests have reported sensitivities and specificities in the ranges $\geq 30\%$ and $\geq 80\%$, respectively - therefore, the communicated data seldom can offer precise distinctions.¹⁴

Given the critical role that dashboards and graphs based on SARS-CoV-2 test results play for policymakers, health professionals, and the general public,⁸ our objective was to develop a Bayesian calculator that could calculate test quantities and prevalence solely based on officially reported numbers of total and positive tests, i.e., without making any *a priori* assumptions. In this way, time trend estimates and country-to-country comparisons of these test performance measures as well as disease prevalence estimates become possible, producing in-depth insights, making projections/simulations possible, and providing a more holistic understanding of the daily incoming data in general.

2. Methods**2.1 General description of the calculator**

The Bayes Lines Tool (BLT) calculator is based on Bayes’ theorem and estimates the true and false positive, and true and false negative numbers at a given time point for which the total number of tests performed and the number of positive test results is known. These data are usually reported and published by official government bodies daily and/or weekly. Thus, the model uses the following information:

- Publishing date or report identifier of the test data
- Number of performed tests (#tests)
- Number of reported positive results (#positives)

The model takes this information as a given fact and uses it to make inferences about the test performance parameters (sensitivity and specificity) as well as the prevalence (also known as the base rate) - these inferences are essential for estimating the number of true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN). It is assumed that there is no knowledge of either the prevalence or the sensitivity and specificity of the tests used. Instead, the model explores all possible combinations of two of these three parameters within reasonable ranges specified by the user; for each of these combinations, the third parameter can then be calculated using the dependencies through Bayes’ theorem. Finally, all parameter combinations that result in TP+FP estimates consistent with the known number of positive tests are selected and stored as confusion matrices.

Table 1. A confusion matrix for a SARS-CoV-2 test containing absolute numbers of true (TP) and false (FP) positives and true (TN) and false (FN) negatives derived from equations (2)-(5).

Actual infection status	Test result positive	Test result negative
INFECTED	TP	FN
NOT INFECTED	FP	TN

A single confusion matrix contains TP, FP, TN, and FN in absolute numbers (Table 1). For a given prevalence, sensitivity, and specificity these are derived from Bayes' theorem:

$$P(I|T) = \frac{P(T|I) \times P(I)}{P(T)} \quad (1)$$

Here, T denotes the hypothesis that a test comes out positive ($\neg T$ its denial) and I the hypothesis that an individual is infected, so that $P(I)$ is the prevalence and $P(T|I)$ is the test sensitivity. $P(T)$ is the marginal probability of a positive test, which we estimate as the frequency of positive test results, whereas $P(I|T)$ is the probability of being infected given that the test came out positive. With the normalizing constant $P(T)$ estimated as $P(T) = \frac{\#positives}{\#tests}$ and $P(I|T)$ estimated as the proportion of infected individuals among those in which the test came out positive, equation (1) becomes:

$$TP = P(I|T) \times \#positives = \text{sensitivity} \times \text{prevalence} \times \#tests \quad (2)$$

Equation (2) thus shows that the number of TPs depends on the prevalence, test sensitivity and total number of tests performed. Using $P(\neg T|\neg I) = \text{specificity}$ and $\#negatives = \#tests - \#positives$, an analogous derivation leads to

$$TN = P(\neg I|\neg T) \times \#negatives = \text{specificity} \times (1 - \text{prevalence}) \times \#tests \quad (3)$$

From Equations (2) and (3), FP and FN follow as

$$FP = \#positives - TP \quad (4)$$

$$FN = \#tests - \#positives - TN \quad (5)$$

2.2 Implementation

For the implementation presented here, the two parameters which varied are as follows:

- Sensitivity from 0.005 to 1 with 0.005 increments.
- Specificity from 0.005 to 1 with 0.005 increments.

For a given sensitivity and specificity as well as number of tests and positives, the prevalence can then be computed as

$$\text{prevalence} = \frac{\left(\frac{\#positives}{\#tests} + \text{specificity} - 1\right)}{\text{sensitivity} + \text{specificity} - 1} \quad (6)$$

Hereby, calculations for combinations of sensitivity and specificity that add to ≤ 1 are omitted, and cases in which prevalence turns out negative or larger than 1 are discounted as unphysical.

We developed an SQL query that generates all possible Bayesian confusion matrices for a series of diagnostic test results, without making assumptions about prevalence, sensitivity, or specificity.

The code in PostgreSQL is given as follows (Code 1):

```
with tests as (
  select
    :reg :: text as region_name,
    :rid :: text as report_id,
```

```

        :tst :: float as tests,
        :pos :: float as positives
    ),
permutations as (
    select
        sens :: float as sensitivity,
        spec :: float as specificity
    from
        generate_series(0.005, 1.000, 0.005) as sens,
        generate_series(0.005, 1.000, 0.005) as spec
    ),
prevalences as (
    select
        (positives/tests + specificity - 1) :: float /
        (sensitivity + specificity - 1) :: float as prevalence,
    *
    from
        permutations,
        tests
    where
        sensitivity + specificity > 1
    ),
matrices as (
    select
        (tests * prevalence * sensitivity) :: float as true_positives,
        (tests * (1 - prevalence) * specificity) :: float as true_negatives,
    *
    from
        prevalences
    where
        prevalence between 0 and 1
    ),
results as (
    select
        positives - true_positives as false_positives,
        (tests - positives) - true_negatives as false_negatives,
    *
    from
        matrices
    )
select
    region_name,
    report_id,
    (tests) :: int as tests_performed,
    (positives) :: int as positives_reported,
    (tests * prevalence) :: int as has_disease,
    (tests * (1 - prevalence)) :: int as hasnot_disease,
    (true_positives) :: int as true_positives,
    (false_positives) :: int as false_positives,
    (true_negatives) :: int as true_negatives,
    (false_negatives) :: int as false_negatives,
    sensitivity :: numeric(4,3),

```

```

specificity :: numeric(4,3),
prevalence :: numeric(4,3)
from
  results
where
  (false_positives + true_positives) :: int = positives :: int

```

Given the test results published in the databases and given all generated permutations and consequently all possible confusion matrices, only those are returned that match the positive test results. With only the resulting confusion matrices for which TP+FP match the positives reported in the input data, we are able to identify patterns that provide additional insights for further investigation.

In order to produce confusion matrices for a series of reports, such as daily test result numbers, several approaches are possible. In this manuscript we describe a practical application for using a Batch/Script approach. The Script is used on Apple OSX, the example below using COVID-19 data from the Netherlands (Code 2):

```

psql -h localhost -d postgres -U postgres -A --set=rid='\20200601\' --set=reg=
\'Netherlands_GGD\' --set=tst=1552--set=pos=73 -f BLTV3.sql >> Netherlands_GGD.
txt
psql -t -h localhost -d postgres -U postgres -A --set=rid='\20200602\' --set=reg=
\'Netherlands_GGD\' --set=tst=6819 --set=pos=203 -f BLTV3.sql >> Netherlands_GGD.
txt
psql -t -h localhost -d postgres -U postgres -A --set=rid='\20200603\' --set=reg=
\'Netherlands_GGD\' --set=tst=8867 --set=pos=165 -f BLTV3.sql &gt;&gt; Nether-
lands_GGD.txt
psql -t -h localhost -d postgres -U postgres -A --set=rid='\20200604\' --set=reg=
\'Netherlands_GGD\' --set=tst=9339 --set=pos=171 -f BLTV3.sql &gt;&gt; Nether-
lands_GGD.txt
psql -t -h localhost -d postgres -U postgres -A --set=rid='\20200605\' --set=reg=
\'Netherlands_GGD\' --set=tst=9464 --set=pos=135 -f BLTV3.sql &gt;&gt; Nether-
lands_GGD.txt
psql -t -h localhost -d postgres -U postgres -A --set=rid='\20200606\' --set=reg=
\'Netherlands_GGD\' --set=tst=7843 --set=pos=125 -f BLTV3.sql >> Netherlands_GGD.
txt
psql -t -h localhost -d postgres -U postgres -A --set=rid='\20210224\' --set=reg=
\'Netherlands_GGD\' --set=tst=52551 --set=pos=4374 -f BLTV3.sql >> Nether-
lands_GGD.txt

```

2.3 Data

For the examples demonstrated in the Results section below, we extracted test data from:

- A hypothetical scenario used for assessing the performance of BLT and demonstrating the so-called spectrum effect^{17,18}
- Influenza data for the Californian Bay Area obtained from the California Open Data Portal at: https://data.ca.gov/dataset/influenza-surveillance/resource/d2207905-14eb-4264-9a02-8b6ac15ddc39?inner_span=True
- The Netherlands/Dutch Corona Dashboard database, used as examples for a daily report and a time trend analysis: <https://coronadashboard.rijksoverheid.nl/landelijk/positief-geteste-mensen>
- The German LGL Bayern database, derived from RKI (Robert Koch Institute) data: https://www.lgl.bayern.de/gesundheit/infektionsschutz/infektionskrankheiten_a_z/coronavirus/karte_coronavirus/

3. Results

In the following section examples are provided that demonstrate the application of our calculator for the data referenced in Section 2.3.

3.1 A hypothetical scenario

Consider the following hypothetical scenarios displayed in [Table 2](#) that we used for a general check of BLT’s performance. In scenarios 1 and 2, we consider a disease which has a prevalence of 20% in two different subpopulations (e.g. young and old people, respectively). The prevalence was chosen for illustrative purposes only; in most real-world situations, much lower disease prevalence values would be encountered. Each subpopulation has its own test characteristics: In subpopulation 1, test sensitivity is 95% and specificity 75%, while in subpopulation 2, sensitivity is 75% and specificity 95%. Consider that 10,000 tests have been performed in the total population. In scenario 1, the total population consists of an equal mix of both subpopulations, while in scenario 2 the total population consists of 75% subpopulation 1. The different mixture of subpopulations leads to a different number of positive test results, and hence a different input for BLT. The overall test performance measures (sensitivity and specificity) are a weighted average between the subpopulation test performance measures. This is called the spectrum effect.¹⁷

Now consider a different scenario, in which the total population is a mix between two subpopulations with different susceptibility towards the disease, and hence different prevalence, but the test performs equally well in both subpopulations. In scenario 3, each subpopulation contributes 50% to the overall population, while in scenario 4, the less susceptible population contributes 80% (8,000 tests). Now the overall prevalence is the weighted average of the subpopulation prevalence values, and overall test sensitivity and specificity are equal to those of the subpopulations.

[Figure 1](#) displays all solutions that BLT delivers for scenarios 1-4, with the known solutions of the overall and subpopulations highlighted. It is visible that the spectrum effect observed in [Table 2](#) is also visible in [Figure 1](#), as it translates into the percentages of TPs, TNs, FPs and FNs. What is critical is the fact that BLT, which only works with the total number of tests and positives obtained, would not be able to distinguish between scenarios 1, 3 and 4. All three are compatible with the output set of confusion matrices. One should thus keep in mind for the interpretation of BLT’s output that the solution corresponding to reality is determined by the mix of subpopulations being tested, which in turn might have their own specific subpopulation prevalence, sensitivity and specificity values. In other words, one should be aware of the spectrum effect.^{17,18} If possible, one should thus use knowledge about prevalence and test performance measures to filter out the confusion matrices consistent with what is known about “the reality”.

3.2 California/USA (diagnostic Influenza-testing)

[Figure 2](#) shows the results of applying BLT to weekly influenza test data from the Californian Bay Area, USA. The upper panel displays the number of positive tests reported over time, where the estimated number of TPs is overlaid in small dots (confusion matrices) whose color represents the estimated prevalence (see legend on the right of [Figure 2](#)). Filters

Table 2. A hypothetical testing scenario.

Estimation	TP	TN	FP	FN	Prevalence [%]	Sensitivity [%]	Specificity [%]
Scenario 1: Balanced distribution of population 1 and 2							
Overall	1700	6800	1200	300	20	85	85
Population 1	950	3000	1000	50	20	95	75
Population 2	750	3800	200	250	20	75	95
Scenario 2: Unbalanced distribution of populations: 75% population 1							
Overall	1800	6400	1600	200	20	90	80
Population 1	1425	4500	1500	75	20	95	75
Population 2	375	1900	100	125	20	75	95
Scenario 3: Balanced distribution of populations with different prevalence							
Overall	2410	7075	490	25	24.35	99.0	93.5
Susceptible	1900	2880	200	20	38.4	99.0	93.5
Less susceptible	510	4195	290	5	10.3	99.0	93.5
Scenario 4: Unbalanced distribution of populations with different prevalence							
Overall	2117	7046	783	54	21.7	97.5	90.0
Susceptible	780	1080	120	20	40.0	97.5	90.0
Less susceptible	1337	5966	663	34	17.1	97.5	90.0

Total number of tests is 10,000 in all scenarios. Note that in scenarios 1, 3 and 4, the total number of positives is 2900. TP: True positive number; TN: True negative number; FP: False positive number; FN: False negative number.

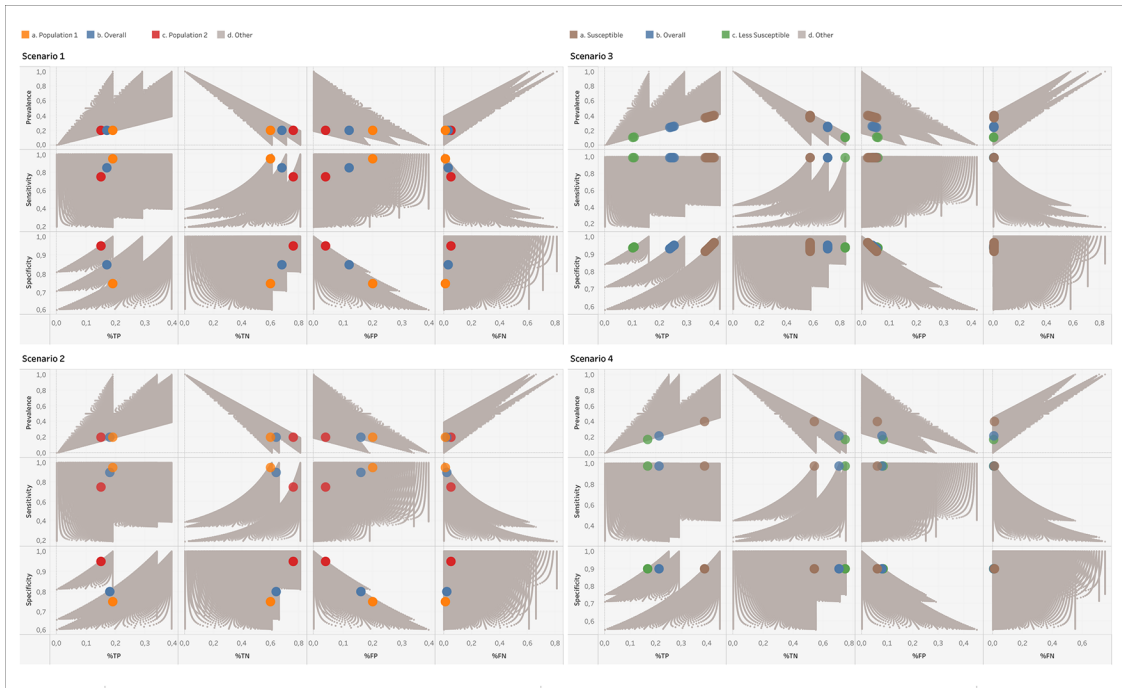


Figure 1. Results of running BLT on the four scenarios given in Table 2. The correct solutions corresponding to the overall and subpopulations of these scenarios are highlighted as large colored points, while all other solutions compatible with the number of tests and positives are shown in grey. For scenario 3, no exact match of prevalence, sensitivity and specificity to the TP and FP numbers could be obtained with the step sizes used in Code 1, so that we display the closest matches. %TP, %TN, %FP, %FN: Percentages of TP, TN, FP and FN numbers relative to the total number of tests performed.

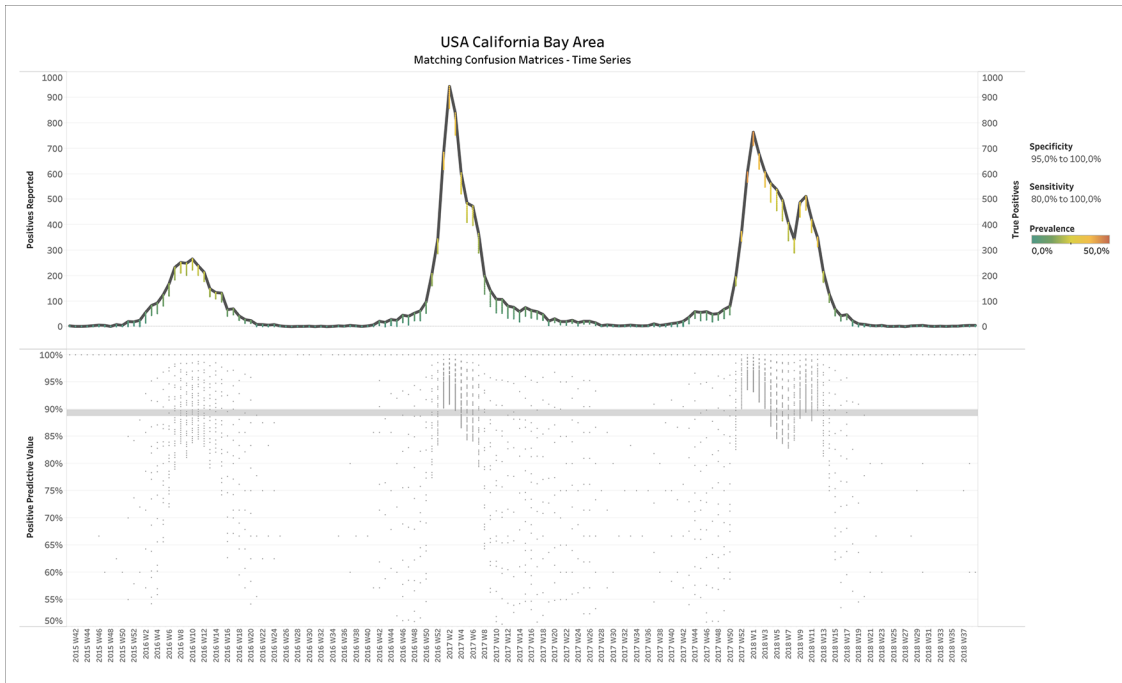


Figure 2. Report ID day vs. positives reported and true positives (upper panel) or positive predictive value (lower panel) for USA-CA-BayArea influenza test data. Upper panel: Color shows details about Prevalence. The view is filtered on specificity and sensitivity. The specificity filter ranges from 95.0% to 100.0%. The sensitivity filter ranges from 80.0% to 100.0%.

have been applied on specificity (95.0% - 100.0%) and sensitivity (80.0% - 100.0%). One could see that the number of TP tests is close to number of positives reported, except for some deviations during the spring and summer months when prevalence was estimated correctly as low.

The lower panel shows the positive predictive value (PPV), for each confusion matrix, defined as $PPV = \frac{TP}{TP+FP}$, which confirms a high accuracy of the tests: The median PPV of all confusion matrices over time was almost 90%.

3.3 The Netherlands (diagnostic COVID-19 testing)

269 daily reports were downloaded from the Dutch government Corona dashboard and processed with the SQL-query. This resulted in 809,830 confusion matrices matching the daily reports from June 1st, 2020 until Feb 24th, 2021. The upper panel of Figure 3 plots the median PPV, with the corresponding number of performed and positive tests plotted in the lower panel. Note that the left and right y-axes in the lower panel are on different scales.

It can be observed that in contrast to the influenza example (Figure 2), the PPVs are now much lower, with a median average around 50%. For this estimation, no filters were applied on sensitivity, specificity or prevalence. When *a posteriori* knowledge is available about the diagnostic tests and/or the circumstances in which they were performed, different scenarios can be applied to the output. This is exemplarily visualized in Figure 4, in which some reasonable filters for a SARS-CoV-2 testing environment have been applied. Notice how the PPV started to increase sharply from a median around 50% before mid-September 2020 to 80-90% during the fall and winter.

Finally, Figure 5 shows the negative predictive value (NPV) for the Netherlands data with similar filters as in Figure 4, except for choosing a less optimistic sensitivity range of 60-80%, which is consistent with some clinical data. It can be noticed that NPV remains relatively high throughout the entire time range. Median NPV over time does not drop below 90%, even after reducing the range for sensitivity to as low as 60-80%. We also tested the impact of this lower sensitivity range on the PPV, but could not detect any visible impact, consistent with the finding that low-specificity tests cannot distinguish between the hypotheses that a positively tested individual is infected with SARS-CoV-2 or not regardless of sensitivity.¹⁴

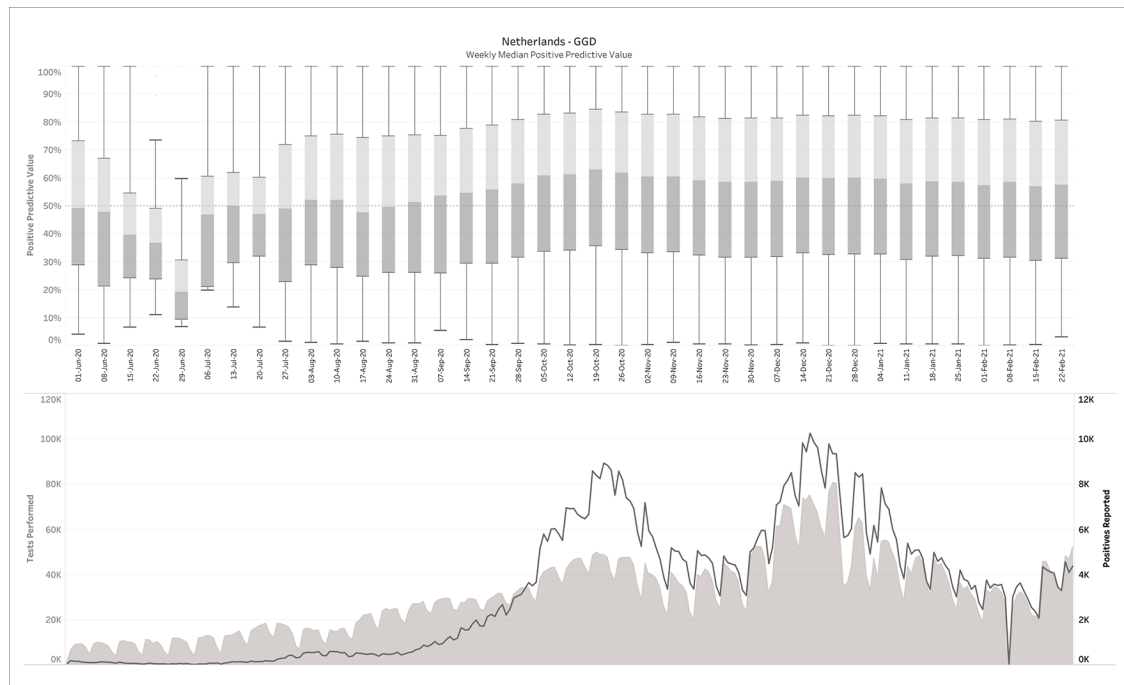


Figure 3. The Netherlands - June 1 2020-Feb 24 2021, weekly median positive predictive value (upper panel), in comparison with tests performed and positive tests (lower panel). No filters on prevalence, specificity or sensitivity were applied here.

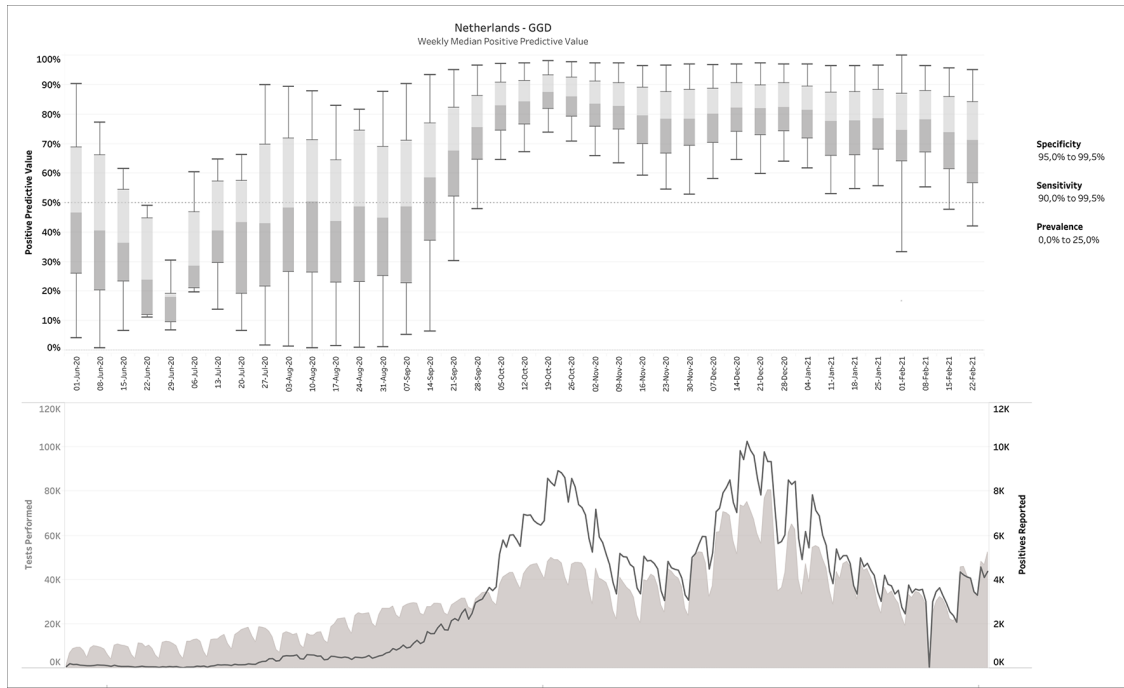


Figure 4. Same as Figure 3, but now with filters applied. The example above shows the 40,200 possible confusion matrices that fit the given report, for $90.0\% \leq \text{sensitivity} \leq 99.9\%$ and $95.0\% \leq \text{specificity} \leq 99.5\%$ and $0 \leq \text{prevalence} \leq 20\%$.

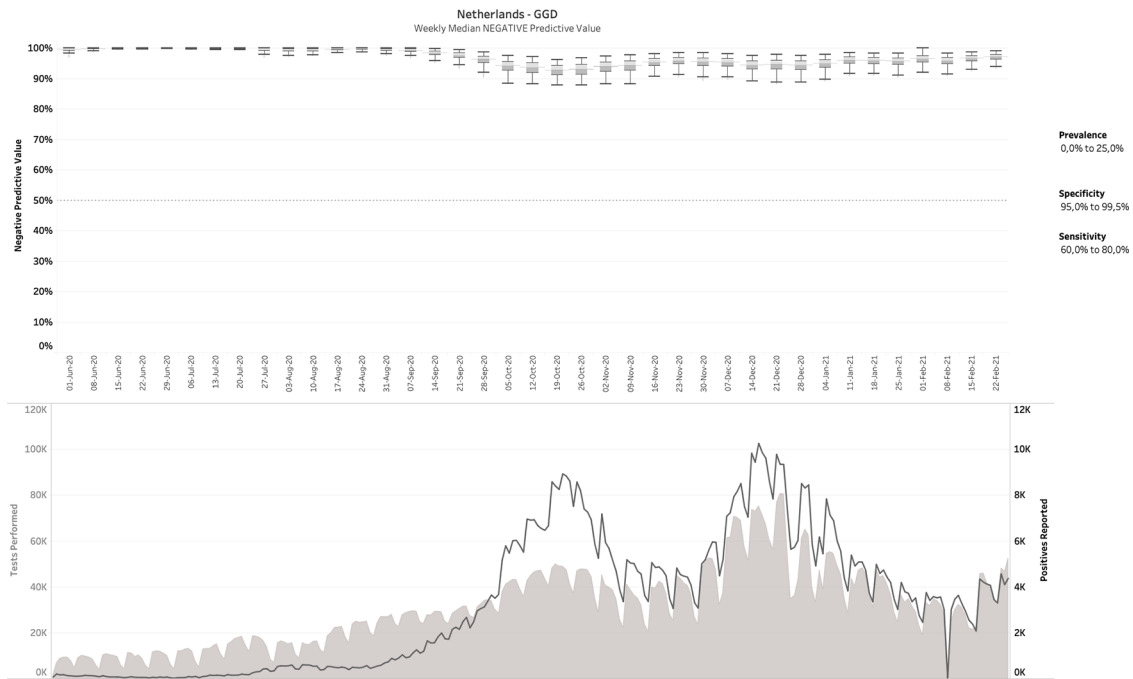


Figure 5. Same as Figure 4, but now plotting NPV and changing the range for sensitivity to 60-80%.

3.4 Germany - Bavaria (diagnostic COVID-19 testing)

Figure 6 shows the output of BLT applied to weekly SARS-CoV-2 testing data from Bavaria in Germany. The thick grey line displays the number of positive tests reported over time, while the colored batches show the solutions of BLT for the TP numbers according to prevalence. Note that in low prevalence scenarios, the TPs do usually not come close to the

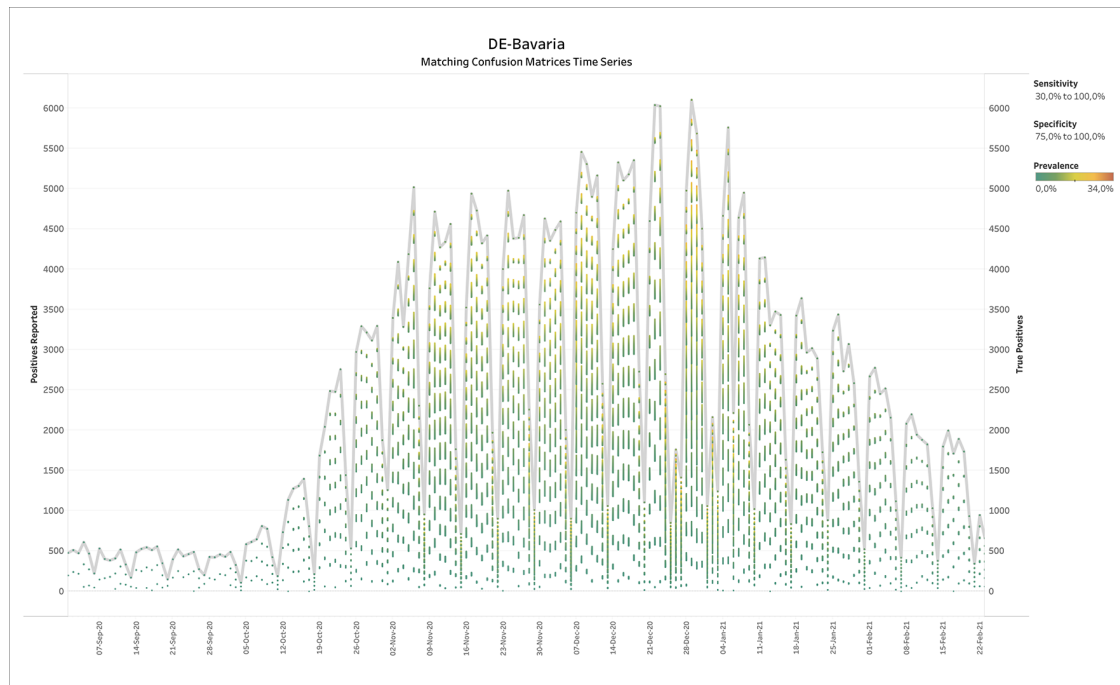


Figure 6. Bavaria, Germany - Weekly reports with positives reported and true positives calculated by BLT. For the true positives, the color shows details about prevalence. The Specificity filter ranges from 75.0% to 100.0%. The Sensitivity filter ranges from 30.0% to 100.0%. Reports range from week ending 26th February 2020 until week ending 17th February 2021.

reported number of positives. At the end of the summer, the prevalence values compatible with the official test reports suggested low prevalence, but also a discrepancy between the number of positive tests and TPs, suggesting a large number of FPs.

4. Discussion

The developed Bayesian calculator tool allows the estimation of possible values for the essential variables' prevalence, sensitivity, and specificity for a specific period of time (e.g., daily or weekly, depending on the input data the user supplies). The solutions provided by BLT are derived from Bayes's theorem (Equation 1) under the assumption that $P(T) = \frac{\#positives}{\#tests}$ and $P(I|T) = \frac{TP}{\#positives}$. In cases of low total and positive test numbers, these assumptions might not hold exactly, but BLT should nevertheless find close solutions to the actual test performance measures. As our applied examples show, the strength of BLT lies in its application to mass testing scenarios such as those conducted during the SARS-CoV-2 crisis.

The BLT calculations are unbiased in the sense that they use all possible and sensible combinations of prevalence, sensitivity, and specificity, and let Bayes' theorem decide which combinations match the actually observed data. The result for a given matching combination of these three particular parameters is provided in the form of a confusion matrix which contains the TP, TN, FP, and FN numbers. In the case where more than one combination is compatible with the given input data, the user may start simulating different scenarios, e.g., by applying prior knowledge regarding the expected prevalence range on a given date and test sensitivity and specificity estimates. This enables the user to further constrain the combinatorial possibilities of the output variables. For example, if disease prevalence in our hypothetical examples given in Figure 1 would have been known to range around 20%, lower and upper bounds for the TP, TN, FP, and FN percentages could be readily obtained from this graph. Thus, one would learn that a positive test result should not be trusted with high probability, but a negative test result would be very reliable. It is important to emphasize that there is no "wrong" output of the BLT calculator, since the output logically follows from the laws of probability; it is the responsibility of the user to decide which output possibilities best apply to the real situation under which the test had been performed.

Prevalence is a crucial factor for any inferences based on diagnostic tests, even though it is often not taken into account in practice. This results in the so-called base-rate fallacy.¹⁹ Our calculator may result in several possible prevalence values that are compatible with the observed data. In this case, knowledge about the population that has been tested should be

used to constrain the possibilities. In 2020, for instance, prevalence-values in the range 12-15% were estimated for German hotspot regions,^{20,21} while prevalence was zero in an asymptomatic German mother-and-child population tested in April 2020.²² In an early COVID-19 related publication which compared RT-qPCR to chest computer tomography in 1014 COVID-19 patients from the Tongji hospital in Wuhan, China, prevalence appeared to be very high: in total 830 patients were described to be confirmed or highly likely to have COVID-19, and of those 580 were diagnosed by chest CT and RT-qPCR and another 250 by CT and clinical decision. These results suggest a prevalence of 81.9% in these patients. A preprint publication²³ aimed at estimating the sensitivity and specificity of the Chinese RT-qPCR tests by a Bayesian model incorporating information from both chest CT and clinical decision classification. The author obtained sensitivity of 0.707 (95% CI range: 0.668-0.749) and specificity of 0.851 (95% CI range: 0.774-0.941). Applying BLT to these data and assuming that only the cases in which both chest CT and RT-qPCR came out positive (i.e., filtering on 580 TPs), our model reveals a sensitivity of 65.3% and specificity ranging from 83.1%-83.6%, not too different from the estimates of the more complex analysis.²³

During the SARS-CoV-2 crisis an unprecedented mass testing not only of symptomatic, but also asymptomatic cases emerged as a strategy. One would expect the prevalence to be substantially higher in the former than in the latter population. As our scenarios 3 and 4 from section 3.1 shows, if there is a mixture of two populations with very different prevalence values, the resulting overall prevalence is a weighted average, provided that the sensitivity and specificity of the tests is similar in both populations.

Our results display the known dependence of a test predictive value from the disease prevalence. For example, the World Health Organization (WHO) stated “that disease prevalence alters the predictive value of test results; as disease prevalence decreases, the risk of false positive increases”.²⁴ This means that the probability that a person who has a positive result (SARS-CoV-2 detected) is truly infected with SARS-CoV-2 decreases as prevalence decreases, irrespective of the claimed specificity of the test system.²⁴ This statement may be more accurately described as the number of TPs decreasing relative to a constant FP rate so the ‘risk of false positives’ only increases relative to the TP numbers, but the FP frequency is assumed to remain constant across a given number of tests. However, multiple modes of error may be in play. We should not assume FPs are independent of contamination from TP samples. There are higher risks of contamination in rapidly growing laboratories. Contamination of samples in the low disease prevalence seasons (summer) will go unnoticed as they do not produce a qPCR signal. Contamination prone methods may only become evident in the form of elevated and perhaps falsely assumed TPs once the disease prevalence increases in the winter.

In light of the above WHO statement, the rationale for mass testing strategies implemented during periods of low prevalence (e.g., summer) appears questionable. Furthermore, mass testing increases the risk of poor sample handling and laboratory contamination which might partly explain the high FP numbers our calculator predicts. For example, Patrick *et al.* argued that besides intrinsic test performance, amplicon contamination due to high throughput processing of samples within a laboratory would be the best explanation for an increased rate of FP detections made during an outbreak of the human coronavirus HCoV-OC43 in a Canadian facility.²⁵

While much attention has been placed on population frequency of disease and its impact on false positives, it is critical to understand the role of false negatives and the impact these can have on track and trace systems. The nasal swabs are known to vary tremendously in RNaseP Ct values suggesting highly variable sampling or limited RNA stability in the testing reagent chain.²⁶ Woloshin *et al.* demonstrate 27-40% FNs with nasopharyngeal and throat swabs respectively and underscore the importance of understanding pre-test probabilities when interpreting qPCR results.²⁷ These FN numbers are probably not due to the PCR itself, but are related to handling issues and the above discussed problems, as well as the time point within the course of infection that the sample is taken. In a meta-analysis of clinical data, Kucirka *et al.* found that the probability of a FN test was 100% at day 1 of an infection with SARS-CoV-2 (prior to symptom onset), and then decreased to 38% (95% credible interval 18-65%) at the day of symptom onset down to its minimum of 20% (12-30%) three days after symptom onset, after which it rose again to 66% (54-77%) three weeks after the infection.²⁸ Hence, according to these numbers, even in infected individuals sensitivities below 30% are possible, a range that we excluded in our analysis consistent with Klement and Bandyopadhyay.¹⁴ This points to additional problems when testing asymptomatic individuals, because in case that they are truly infected, a high number of FNs is going to result.

With the script presented here, we can think of many variations when it comes to the range of sensitivity and specificity, their step-sizes (granularity) and the ‘where’ clause as well as the strictness of matching TP+FP against the reported positives. For example, one could also increment prevalence on a log-scale to account for the fact that prevalence in many settings of diseases is very low.¹⁴

We are aware that choices made in these areas have a significant impact on the number of matching confusion matrices. An impact/sensitivity analysis was not performed, although we suspect that such analysis might reveal additional insights. However, we think that the amount of matching confusion matrices per result that the above query produces delivers sufficient material to make useful observations.

Future research with different data-repositories, for instance ECDC/TESSy-data would be very beneficial to identify a solid balance between precision (step-size in the permutations), number of matching confusion matrices, and overall query performance.

5. Conclusions

We have developed an easy-to-use Bayesian calculator (Bayes Lines Tool, BLT) to estimate prevalence, sensitivity, and specificity, and therefore TP, TN, FP, and FN numbers, from official test outcome numbers. With typical reports - especially as produced for SARS-CoV-2 tests - revealing just the number of positives and number of tests performed, the BLT SQL implementation generates confusion matrices that fit within the boundaries of a typical simplified report, based on permutations of sensitivity and specificity. Its implementation is thereby not limited to SQL but can be applied on any platform of choice.

The ability to assess posterior probability independent of the circumstances in which diagnostic tests are performed, reveals a wide spectrum of opportunities for new applications both for the scientific community as well as for health professionals and policy makers around the globe. This is especially relevant for the mass testing taking place within the containment strategies of worldwide governments against the SARS-CoV-2. The BLT SQL query for the first time allows one to display a real estimation of the SARS-CoV-2 situation against the background of testing volume and quality and thus will provide a valuable tool for decision makers to monitor the test strategy and the effect of interventional procedures.

This tool will not only allow official institutions to survey the test situation and obtain a better basis for planning their interventions, but also allows for individuals who got tested to use the confusion matrices as an aid for interpreting their test results in view of the population they were tested in.

Data availability

Underlying data

All data underlying the results is linked in section 2.3 of the article. The hypothetical example is given in [Table 2](#). No additional source data is required.

Software availability

Zenodo:

Bayes Lines Tool (BLT) - A SQL-script for analyzing diagnostic test results with an application to SARS-CoV-2-testing, <http://doi.org/10.5281/zenodo.4594210>.²⁹

Code is available under the terms of the [Creative Commons Attribution 4.0 International license](#) (CC-BY 4.0).

The SQL-code and an example implementation in Excel and a Tableau work-book file can be downloaded at <https://bayeslines.org/>.

Acknowledgements

We thank Michiel Maandag for bringing down-to-earth counterweight and alignment to the team. We wish to thank Dimitri Georganas for his support during the initial development of the model. Finally, we thank Andreas Macher for sharing his expertise in the optimisation of the SQL query.

References

- Ren LL, Wang YM, Wu ZQ, *et al.*: **Identification of a novel coronavirus causing severe pneumonia in human: a descriptive study.** *Chin Med J (Engl)*. 2020; **133**(9): 1015–24.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Zhu N, Zhang D, Wang W, *et al.*: **A novel coronavirus from patients with pneumonia in China, 2019.** *N Engl J Med*. 2020; **382**(8): 727–33.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Gorbalenya AE, Baker SC, Baric RS, *et al.*: **The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2.** *Nat Microbiol*. 2020; **5**(4): 536–44.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Corman VM, Landt O, Kaiser M, *et al.*: **Detection of 2019 novel coronavirus (2019-nCoV) by real-time RT-PCR.** *Euro Surveill*. 2020; **25**(3): 1–8.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Wölfel R, Corman VM, Guggemos W, *et al.*: **Virological assessment of hospitalized patients with COVID-2019.** *Nature*. 2020; **581**(7809): 465–9.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Hua J, Shaw R: **Corona Virus (COVID-19) “Infodemic” and Emerging Issues through a Data Lens: The Case of China.** *Int J Environ Res Public Health*. 2020; **17**(7): 2309.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- World Health Organization: **WHO Director-General’s opening remarks at the media briefing on COVID-19-11 March 2020.** 2020 [cited 2021 Feb 6].
[Reference Source](#)
- Everts J: **The dashboard pandemic.** *Dialogues Hum Geogr*. 2020; **10**(2): 260–4.
[Publisher Full Text](#)
- European Centre for Disease Prevention and Control: **Case definition for coronavirus disease 2019 (COVID-19), as of 3 December 2020.** 2020 [cited 2021 Jan 22].
[Reference Source](#)
- van Zyl G, Maritz J, Newman H, *et al.*: **Lessons in diagnostic virology: expected and unexpected sources of error.** *Rev Med Virol*. 2019; **29**(4): 1–7.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Younes N, Al-SAdeq DW, Al-Jighefee H, *et al.*: **Challenges in Laboratory Diagnosis of the Novel.** *Viruses*. 2020; **12**(6): 582.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Wernike K, Keller M, Conraths FJ, *et al.*: **Pitfalls in SARS-CoV-2 PCR diagnostics.** *Transbound Emerg Dis*. 2020.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Arevalo-Rodriguez I, Buitrago-Garcia D, Simancas-Racines D, *et al.*: **False-negative results of initial RT-PCR assays for COVID-19: A systematic review.** *PLoS One*. 2020; **15**(12): e0242958.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Klement RJ, Bandyopadhyay PS: **The Epistemology of a Positive SARS-CoV-2 Test.** *Acta Biotheor*. 2020.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Mascuch SJ, Fakhretaha-Aval S, Bowman JC, *et al.*: **A blueprint for academic laboratories to produce SARS-cov-2 quantitative RT-PCR test kits.** *J Biol Chem*. 2020; **295**(46): 15438–53.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Zhou H, Liu D, Ma L, *et al.*: **A SARS-CoV-2 Reference Standard Quantified by Multiple Digital PCR Platforms for Quality Assessment of Molecular Tests.** *Anal Chem*. 2020; **93**(2): 715–21.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Mulherin SA, Miller WC: **Spectrum Bias or Spectrum Effect? Subgroup Variation in Diagnostic.** *Ann Intern Med*. 2002; **137**(7): 598–602.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Goehring C, Perrier A, Morabia A: **Spectrum bias: A quantitative and graphical analysis of the variability of medical diagnostic test performance.** *Stat Med*. 2004; **23**(1): 125–35.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Bar-Hillel M: **The base-rate fallacy in probability judgments.** *Acta Psychol (Amst)*. 1980; **44**(3): 211–33.
[Publisher Full Text](#)
- Streeck H, Schulte B, Kümmerer BM, *et al.*: **Infection fatality rate of SARS-CoV2 in a super-spreading event in Germany.** *Nat Commun*. 2020; **11**(1): 1–12.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Santos-Hövenner C, Neuhauser HK, Rosario AS, *et al.*: **Serology- And PCR-based cumulative incidence of SARS-cov-2 infection in adults in a successfully contained early hotspot (CoMoLo study), Germany, May to June 2020.** *Euro Surveill*. 2020; **25**(47): 1–8.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Reisinger EC, Von Possel R, Warnke P, *et al.*: **Screening of Mothers in a COVID-19 Low-Prevalence Region: Determination of SARS-CoV-2 Antibodies in 401 Mothers from Rostock by ELISA and Confirmation by Immunofluorescence.** *Dtsch Medizinische Wochenschrift*. 2020; **145**(17): E96–100.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Padhye NS: **Reconstructed diagnostic sensitivity and specificity of the RT-PCR test for COVID-19.** *medRxiv*. 2020.
[Publisher Full Text](#)
- World Health Organization: **WHO Information Notice for IVD Users 2020/05.** 2021 [cited 2021 Jan 22].
[Reference Source](#)
- Patrick DM, Petric M, Skowronski DM, *et al.*: **An outbreak of human coronavirus OC43 infection and serological cross-reactivity with SARS coronavirus.** *Can J Infect Dis Med Microbiol*. 2006; **17**(6): 330–6.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Dahdouh E, Lázaro-Perona F, Romero-Gómez MP, *et al.*: **Ct values from SARS-CoV-2 diagnostic PCR assays should not be used as direct estimates of viral load.** *J Infect*. 2020.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Woloshin S, Patel N, Kesselheim AS: **False Negative Tests for SARS-CoV-2 Infection — Challenges and Implications.** *N Engl J Med*. 2020; **383**(6): e38.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Kucirka L, Lauer S, Laeyendecker O, *et al.*: **Variation in False-Negative Rate of Reverse Transcriptase Polymerase Chain Reaction—Based SARS-CoV-2 Tests by Time Since Exposure.** *Ann Intern Med*. 2020; **173**(4): 262–7.
- Aukema W, Kämmerer U, Borger P, *et al.*: **Bayes Lines Tool (BLT) - A SQL-script for analysing diagnostic test results with an application to SARS-CoV-2-testing (Version 4.2).** *Zenodo*. 2021, March 10.
[Publisher Full Text](#)

Open Peer Review

Current Peer Review Status:  

Version 3

Reviewer Report 02 March 2022

<https://doi.org/10.5256/f1000research.121230.r123945>

© 2022 Leeflang M. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

 **Mariska M G Leeflang** 

Department of Epidemiology and Data Science, Amsterdam UMC, University of Amsterdam, Amsterdam, The Netherlands

I have nothing to add - my concerns were already addressed in a previous version.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Clinical epidemiology and evaluation of medical tests

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 2

Reviewer Report 04 October 2021

<https://doi.org/10.5256/f1000research.57358.r91450>

© 2021 Leeflang M. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

 **Mariska M G Leeflang** 

Department of Epidemiology and Data Science, Amsterdam UMC, University of Amsterdam, Amsterdam, The Netherlands

The authors developed a what they call 'Bayesian calculator' to estimate false and true positives and negatives from reported test results. They state that this will be a useful tool for health

professionals and policy makers, and even for individuals to interpret their own test results.

As I am not familiar with SQL-code and would have no idea how to implement this calculator on my own computer, I thought that the authors partly provided "sufficient details of the code, methods (...) to allow replication of the software development". So I have not checked for mistakes in the code, or whether the calculator actually works. It would be really helpful if someone could do that.

Technically speaking, the formulas and the explanations all seem to be in order. I have little to comment on that. However, there are a few semantics-issues that may be resolved, and the rationale of a separate calculator is not entirely clear to me.

Some more specific comments:

1. Using the phrase 'Bayesian calculator', implies to me that Bayesian statistics have been used, and that the factors the authors mention in their abstract and introduction (sample handling, underlying test protocols etc.) have been taken into account to go from a prior belief about sensitivity/specificity/prevalence to a posterior (after accounting for other factors) belief. However, when reading the manuscript, it turns out that the BLT is nothing more than a huge number of permutations given a starting value of positive and negative results. It provides a range of possible true values, without providing an indication of how realistic all these possibilities are. Therefore, I think the authors are overselling a relatively simple calculator and that the new thing of this BLT is actually only the way the data are presented. I can do these calculations in Excel as well, but that would give me a headache to provide the right figures and graphs. So maybe the whole article should tone down the novelty a bit.
2. In my previous comment I mentioned the lack of information about how realistic some predictions may be. Would it be possible to add this information?
3. If I were a policy maker and I would get Figure 1 out of this program, how should I interpret the results and how should I implement the information in my policy making? I am missing the link with practice. Could the authors maybe provide some instructions about what the results mean and what their implications for practice maybe?
4. I find abbreviations and acronyms confusing, although that may be a personal thing. For example, CM for confusion matrix is only one word less in the word count every time CM is being used. But using the full term is much more informative and easier to read.
5. I am not sure whether the examples given (20% prevalence in the general population) are realistic. I think the true prevalence of SARS-CoV-2 infections has been much lower at any given point in time.
6. The authors stated that: "Our results confirm the recent World Health Organization (WHO) statement "that disease prevalence alters the predictive value of test results; as disease prevalence decreases, the risk of false positive increases"." However, this is an inherent given for predictive values. It is in their calculation. So this statement follows from logic, while the way it was written here, it implies that the authors have proven this WHO statement to be correct. And it implies that this is a recent finding. Both are not true. So

please rephrase it. Maybe using language explaining that your findings follow the premise that prevalence alters the predictive value.

7. I think it is a missed opportunity that the authors have not performed a sensitivity or impact analysis. Now it is just showing us how the SQL code works, but it does not provide us further insights into where results may go wrong, or when they become more or less reliable.

Is the rationale for developing the new software tool clearly explained?

Partly

Is the description of the software tool technically sound?

Yes

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?

Partly

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?

Partly

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Clinical epidemiology and evaluation of medical tests

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Reviewer Report 27 August 2021

<https://doi.org/10.5256/f1000research.57358.r87900>

© 2021 Bentley P. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



Phillip M. Bentley

European Spallation Source ESS, Lund, Sweden

The article has been amended according to review 1 and should be indexed in its current form.

Is the rationale for developing the new software tool clearly explained?

Yes

Is the description of the software tool technically sound?

Yes

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?

Yes

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?

Yes

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Physics, data analysis, simulations, computer modelling, game theory, strategy.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 1

Reviewer Report 18 May 2021

<https://doi.org/10.5256/f1000research.54169.r84995>

© 2021 Bentley P. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



Phillip M. Bentley

European Spallation Source ESS, Lund, Sweden

The authors have done a great job and this article should be indexed after addressing the issues below:

- Equation 6 has a typing mistake.
- The authors state on page that "These FN numbers are probably not due to the PCR itself,

for which sensitivity is almost 100% (<https://www.finddx.org/covid-19-old/sarscov2-eval-molecular/>), but a matter of handling issues and the above-discussed problems." This is incorrect. The false negative rate of the PCR test is documented as a function of time (see ref 1),¹ and is mostly related to low early virus shedding initially. The middle phase, when the test has the lowest false negative rate, is as the authors describe, but then the lack of virus particles as the patient recovers becomes the major factor. Laboratory-based validation of PCR testing, as cited by the authors, differs from the clinical FN as seen in the present study due to potential flaws in the entire sample collection, handling, and processing chain. It is important to consider the whole chain in evaluating error rates. This is perhaps the most important aspect of test errors, since FP results in some inconvenience and/or worry whilst FN provides a dangerous false sense of security.

- The analysis of Bavaria (figure 5) is roughly consistent with the numbers reported by Kucirka, but those of the Netherlands (figure 4) suggest a sensitivity range that is highly optimistic. The specificity on the other hand could be very good depending on the regional expertise.
- The authors need to address the difference between their data and what would be expected from Kucirka *et al.*
- The negative predictive value is very important for governments/regions to release healthy people with confidence. The authors should present this number as they have done for the positive predictive value.

References

1. Kucirka L, Lauer S, Laeyendecker O, Boon D, et al.: Variation in False-Negative Rate of Reverse Transcriptase Polymerase Chain Reaction–Based SARS-CoV-2 Tests by Time Since Exposure. *Annals of Internal Medicine*. 2020; **173** (4): 262-267 [Publisher Full Text](#)

Is the rationale for developing the new software tool clearly explained?

Yes

Is the description of the software tool technically sound?

Yes

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?

Yes

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?

Yes

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Physics, data analysis, simulations, computer modelling, game theory, strategy.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 15 Jun 2021

Rainer Klement, Leopoldina Hospital Schweinfurt, Schweinfurt, Germany

We thank Dr. Bentley for his time to evaluate our article and constructive comments that we think have led to further improvements. We hope that the revised version can now be approved.

Below, we reply to each of Dr. Bentley's points. The revisions to the text have been marked with track changes in MS Word.

- Equation 6 has a typing mistake.

Answer: Thank you for pointing it out, we advised the F1000 team to correct it.

- The authors state on page that "These FN numbers are probably not due to the PCR itself, for which sensitivity is almost 100% (<https://www.finddx.org/covid-19-old/sarscov2-eval-molecular/>), but a matter of handling issues and the above-discussed problems." This is incorrect. The false negative rate of the PCR test is documented as a function of time (see ref 1),¹ and is mostly related to low early virus shedding initially. The middle phase, when the test has the lowest false negative rate, is as the authors describe, but then the lack of virus particles as the patient recovers becomes the major factor. Laboratory-based validation of PCR testing, as cited by the authors, differs from the clinical FN as seen in the present study due to potential flaws in the entire sample collection, handling, and processing chain. It is important to consider the whole chain in evaluating error rates. This is perhaps the most important aspect of test errors, since FP results in some inconvenience and/or worry whilst FN provides a dangerous false sense of security.

Answer: Thank you for pointing out the time factor relative to the time of infection and the study by Kucirka et al. We have changed the text into the following: "These FN numbers are probably not due to the PCR itself, but are related to handling issues and the above discussed problems, as well as the time point within the course of infection that the sample is taken. In a meta-analysis of clinical data, Kucirka et al. found that the probability of a FN test was 100% at day 1 of an infection with SARS-CoV-2 (prior to symptom onset), and then decreased to 38% (95% credible interval 18-65%) at the day of symptom onset down to its minimum of 20% (12-30%) three days after symptom onset, after which it rose again to 66% (54-77%) three weeks after the infection. Hence, according to these numbers, even in infected individuals sensitivities below

30% are possible, a range that we excluded in our analysis consistent with Klement and Bandyopadhyay. This points to additional problems when testing asymptomatic individuals, because in case that they are truly infected, a high number of FNs is going to result.”

We try to avoid judgements such as yours (“since FP results in some inconvenience and/or worry whilst FN provides a dangerous false sense of security”), which may invoke subjective arguments regarding “inconveniences”. We acknowledge that this is the general perception about FN versus FP and will elaborate further on this, with your last comment / point. We appreciate that you point this out and will add an additional figure to prevent any suggestion of bias in the article.

- The analysis of Bavaria (figure 5) is roughly consistent with the numbers reported by Kucirka, but those of the Netherlands (figure 4) suggest a sensitivity range that is highly optimistic. The specificity on the other hand could be very good depending on the regional expertise.

Answer: When reducing sensitivity towards a less optimistic range of 60-80%, we see no visible impact on PPV in figure 4, probably because sensitivity has no significant influence on PPV. This is now described in words in Section 3.3.

- The authors need to address the difference between their data and what would be expected from Kucirka *et al.*

Answer: A brief discussion has been added, in line with the answer to your point above: “Hence, according to these numbers, even in infected individuals sensitivities below 30% are possible, a range that we excluded in our analysis consistent with Klement and Bandyopadhyay. This points to additional problems when testing asymptomatic individuals, because even if they are infected, the resulting FN numbers may provide a false sense of security.”

- The negative predictive value is very important for governments/regions to release healthy people with confidence. The authors should present this number as they have done for the positive predictive value.

Answer: Thank you for pointing this out. We understand and agree that NPV should be presented as prominently as PPV and will do so by adding a new figure for the Netherlands, taking in consideration your remarks about the sensitivity range being set too optimistically. Notice how NPV remains relatively high throughout the entire time range. Median NPV over time does not drop below 90%, even after we reduce the range for sensitivity to as low as 60-80%.

Competing Interests: No competing interests exist.

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com

F1000Research