Correcting COVID-19 PCR Prevalence for False Positives in the Presence of Vaccination Immunity

Michael Halem[1]
BeCare Link LLC
7 April 2021

Abstract

Many public health authority reports on COVID-19 cases confound positive test results with population prevalence. As the population prevalence approaches the PCR test false positive rate (FPR), for example during a vaccination campaign, it is necessary to adjust the raw test results for the false positive rate. This paper provides a technique for estimating the test false positive rate and making the correction to test population prevalence in the absence of accurate and definitive specificity.

Using current data providing by the Public Health England (PHE) as of the most recent complete data, a false positive rate of 1.16% (95% CI 1.09 - 1.23%) was found for the PHE PCR test for the period 1 January through 29 March 2021. During this period, the test population prevalence is decreasing, starting at a decay rate estimated as 3.0% per day (CI 2.79 - 3.14%). This rate of decay increased to an estimated 14.7% by the end of the period (CI 13.30 - 16.16%) Finally, mean test population prevalence was estimated at 14.3% (CI 13.75 - 14.87%) on 1 January and is estimated to have declined significantly to 0.06% (CI 0.00 - 0.13%). If PCR test positivity are used without the application of the false positive rate, the percent positive PCR tests will eventually "flatline" at the false positive rate, and produce a false positive bias even if test population prevalence should fall to zero.

Keywords: Epidemiology, PCR, False Positive Rate, FPR

Introduction

Many public health authority reports on COVID-19 cases confound positive test results with population prevalence. As the population prevalence approaches the PCR (polymerase chain reaction) test false positive rate (FPR), for example during a vaccination campaign, it is necessary to adjust the raw test results for the false positive rate. This paper provides a technique for estimating the test false positive rate and making the correction to test population prevalence in the absence of accurate and definitive specificity.

Methods: Data

For the analysis, it is necessary to have a relatively clean data stream with biases removed. For most public reporting, the PCR and LFD (lateral flow device) test results are either combined to give case counts; or if kept separate, the PCR test results may be biased using reported LFD positives without adjustment for the total LFD tests conducted in the denominator. Fortunately, Public Health England (PHE) has recently provided separate daily data fields for PCR tests and LFD tests, including for LFD tests that have been confirmed by PCR. Therefore, it is relatively simple to construct an unbiased PCR test positivity from the ratio of from their data field "newCasesPCROnlyBySpecimenDateRollingSum" which specifically excludes PCR tests that have been previously confirmed by a prior positive LFD test. Further, the denominator of total PCR tests can be adjusted by subtracting out the total positive LFD tests

[1] Correspondence: michael.halem@becare net

that have been submitted to PCR for testing, which while a relatively small number (about 800 compared to 240,000 tests per day) adds a small amount of accuracy.

Theoretically the technique could be used for a combined LFD and PCR testing stream. However, this is only true if the ratio of LFD to PCR tests remained constant such that an overall FPR could be computed. A more sophisticated method (not presented) allows these to be computed together. However, it is the author's belief that the PHE LFD test data is biased due to non-reporting of negative tests from home testing of students and staff in the England education system. Analysis of such potentially biased data is beyond the scope of this paper.

Seven day rolling total PCR, LFD, and vaccination data are downloaded from PHE [1, 2] as csv files for the following fields:

Item	PHE API Field	Description
1	newCasesPCROnlyBySpecimenDateRollingSum	7 Day Rolling Sum of Positive PCR tests. Excludes retests of positive LFD which would bias upward the positivity.
2	newPCRTestsByPublishDateRollingSum	7 Day Rolling Sum of all PCR tests (positive, negative, or rejected), a denominator for the PCR percentage positive.
3	newCasesLFDConfirmedPCRBySpecimenDateRollingSum	7 Day Rolling Sum of LFD Positive Tests Confirmed by PCR by Specimen Date
4	newCasesLFDOnlyBySpecimenDateRollingSum	7 Day Rolling Sum of LFD Positive Tests Not Confirmed by PCR by Specimen Date
5	cumPeopleVaccinatedFirstDoseByPublishDate	Total People Vaccinated with First Dose
6	cumPeopleVaccinatedSecondDoseByPublishDate	Total People Vaccinated with Second Dose

Table 1 - Downloaded Data Fields

It should be noted that PHE provides the LFD and PCR test numerators (the new positive "cases") only by specimen date. The PCR denominator is generated from the "newPCRTestsByPublishDateRollingSum" field on the specimen date. For this field, it is estimated that publish date data represents specimens that were taken on average 2 days previously. Informal spot checks of other PHE data using linear regression of total cases by specimen date against publish date, shifted between 0 and 7 days confirms this assumption is reasonable. Further, the total PCR test rolling sum is relatively stable with less than a 4 percent day to day change. Rolling sum data is divided by 7 to give a (7 day) daily moving average. The percentage positivity is found by dividing the PCR positive tests unrelated to a prior LFD test (field 2) by the total PCR tests less the LFD positive LFD tests which are assumed to have been submitted to PCR for confirmation. Without excluding confirmed LFD tests in the PCR numerator, the positive PCR testing would have a large bias towards a greater positivity because positive LFD tests have pre-screened the PCR tests. Without the removing LFD tests that have been submitted to PCR for confirmation, the denominator would have a slightly larger bias.

Vaccination population immunity is estimated from item's 5 and 6 using the population of England as the denominator, assuming an empirically determined 7 day delay between vaccination and immunity, and assuming a reasonable 80% immunity from dose one [3], and an additional 20% immunity from dose two.

Methods: Modelling

An inspection of the current test data will show that there is a decline similar to exponential decay as a large portion of the English population became vaccinated. The precise decay dynamics is not needed as the exponential model is sufficiently parsimonious and consistent with solutions to SIR family of models when the susceptible population has been depleted sufficiently for the reproduction number R to be significantly below 1 over short time periods. Generally, a population's prevalence exponential decay (or growth) can written as:

$$C(t) = I_0 \, e^{rt} \qquad \textbf{Eq. 1}$$

where $C(t)$ is the percentage prevalence (i.e. infectious, or previously infected depending on the authorities definition and testing strategy), I_0 is the initial prevalence at the start of the measurement period, t is the days from the start, and r is the daily decay rate if negative. The exponential solution follows directly from the differential equation for the change of infection with respect to time in the classic SIR model and similar models [4]

$$dI/dt = (\beta \, S/N - \gamma) \, I \qquad \textbf{Eq. 2}$$

where I is the time varying infection, t is time (by convention in days), β is the transmissibility (assumed constant), S/N is the percentage of population that is susceptible, and γ is the recovery rate from the infected group to the recovered group. Using the exponential distribution, $\gamma = 1/\tau$, where τ is the mean time of infection: for COVID-19 somewhere between 5 and 10 days. It can be seen by substitution that Eq. 1, the exponential increase or decay, is the solution to Eq. 2 when all parameters in the parenthesis are constant. Further it can be seen that the rate of exponential decay $r = \beta \, S/N - \gamma$. From inspection it is obvious that the decay rate r is a linear function of the susceptible percentage of the total population S/N such that over shorter time periods and with (assuming transmissibility β is constant) and where new natural infections are small relative to vaccinations, the decay rate is a linear function of the vaccinations that have removed susceptible people from the population.

This allows us to write a general revised version of r for a regression model:

$$r(t) = -[r_0 + k_{vax} \, (v_{mean} - v_0) \, t] \qquad \textbf{Eq. 3}$$

where r_0 is the initial decay rate at time t_0 (i.e. $t = 0$), $v_{mean}$is the mean percentage with vaccination immunity between for the time interval t_0 and t, v_0 is the percentage with vaccination immunity at t_0, and k_{vax} is a regression solved constant showing the change in decay rate per excess vaccinations days over the initial vaccination rate.

The algebra for computing test positivity from population prevalence is well known. For a complete derivation please see [5] mathematics appendix, where Equation 30A is presented here as Eq. 4 under the simplification that $N = 1$ so that all units are in percentage of the population. Using the definition of specificity as $1 - f$, where f is the false positive rate (FPR) then

$$T_+ = I \, (S - f) + f \qquad \textbf{Eq. 4}$$

where I is the population prevalence, T_+ is the test positivity, and S the test sensitivity. With no reference standard, the sensitivity is not estimated and is set arbitrarily to 1 (i.e. perfect):

$$T_+ = I \, (1 - f) + f \qquad \textbf{Eq. 5}$$

An overall model, suitable to solve with empirical test data using R's non-linear least squares error minimization technique [6, 7] is created combining equations Eq. 1, 3 and 4:

$$T_+ \;=\; I_0\, e^{-[r_0 + k_{vax}\,(v_{mean} - v_0)\,t\,]t}\,(1 - f) + f + \epsilon \qquad \textbf{Eq. 6}$$

where ϵ is the residual error that the non-linear least squares model is minimizing.

Results

The result is summarized below:

```
Formula: log(pctPCRpos) ~ log((1 - FPR) * I0 * exp(-(r0 + kvax * (vax - vax0) * t) * t) + FPR)

Parameters:
     Estimate Std. Error t value Pr(>|t|)
I0   0.143074   0.002797    51.2   <2e-16 ***
r0   0.029668   0.000888    33.4   <2e-16 ***
FPR  0.011608   0.000326    35.6   <2e-16 ***
kvax 0.002780   0.000234    11.9   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.0541 on 84 degrees of freedom

Algorithm "port", convergence message: relative convergence (4)

Period 2021-01-01 to 2021-03-29: 88 days
---------------------------------------------------------
Initial Prevalence            I0 =  14.307% (CI  13.748 -  14.867%)
Initial Daily Decay Rate      r0 =   2.967% (CI   2.789 -   3.144%)
Geom Avg End Decay Rate     rGeo =   6.401% (CI   5.797 -   7.006%)
End Period Decay Rate       rEnd =  14.729% (CI  13.299 -  16.158%)
False Positive Rate          FPR =   1.161% (CI   1.096 -   1.226%)
Final Prevalence            Iend =   0.055% (CI   0.000 -   0.126%)
PCR Fit Statistics var =  0.679481 SD resid =  0.054067 rSquared =  0.995698 rSquared Adj =
0.995544
```

Table 2 - Model Fit Result

The estimates and their confidence intervals (+/- 2 standard errors) have been extracted to 6 decimal places of precision, with an ad hoc R^2 (percent of the variance of the independent variable explained) so as to be roughly comparable to an ordinary least squares linear regression.

The regression estimates $log(T_+)$ via the R predict.nls() function (part of the base R stats package). T_+ is estimated as $e^{log(T_+)}$ produce the test population prevalence by I rearranging Eq. 5:

$$I = (T_+ - f)/(1 - f) \qquad \textbf{Eq. 7}$$

Confidence intervals were obtained when directly estimated by the R nls() function, by doubling the nls() summary standard errors. The the geometric average decay rate $(rGeo)$ and the ending decay rate (r_{end}) were estimated assuming that variables were random, normally distributed, and uncorrelated using the simplifying formula for the variance of the addition of two such random variables, i.e. ${\sigma_{X+Y}}^2 = {\sigma_X}^2 + {\sigma_Y}^2$.

In the case of the end of period population prevalence estimate I_{end}, the regressions estimate of $log(T_+)$ has a residual standard error, in this case 0.054. The CI for T_+ is thus $e^{log(T_+)\pm 2SE_T}$. The confidence interval was calculated by substituting into Eq. 7 the respective worst case standard errors for T_+ and f (i.e. flip the sign for the plus or minus):

$$I_{CI} = [e^{log(T_+)\pm 2SE_T} - (f \mp SE_f)]/[1 - (f \mp SE_f)] \qquad \textbf{Eq. 8}$$

Of note, because the PCR sensitivity is set to 1, it may well find positive "cases" that are not infectious, i.e. viral fragments from prior infections.

The results are summarized in the below graph which show the data, the fit and a short projection for both test positivity and test population prevalence. The exponential like decay to the false positive rate can be clearly seen, as can the large difference between test population prevalence (the red line) and test positivity (the black X's and the black line. (Of note is that a similar graph can be obtained using Israel Ministry of Health data [8], but that the Israel data lacks documentation to separate PCR and LFD.)

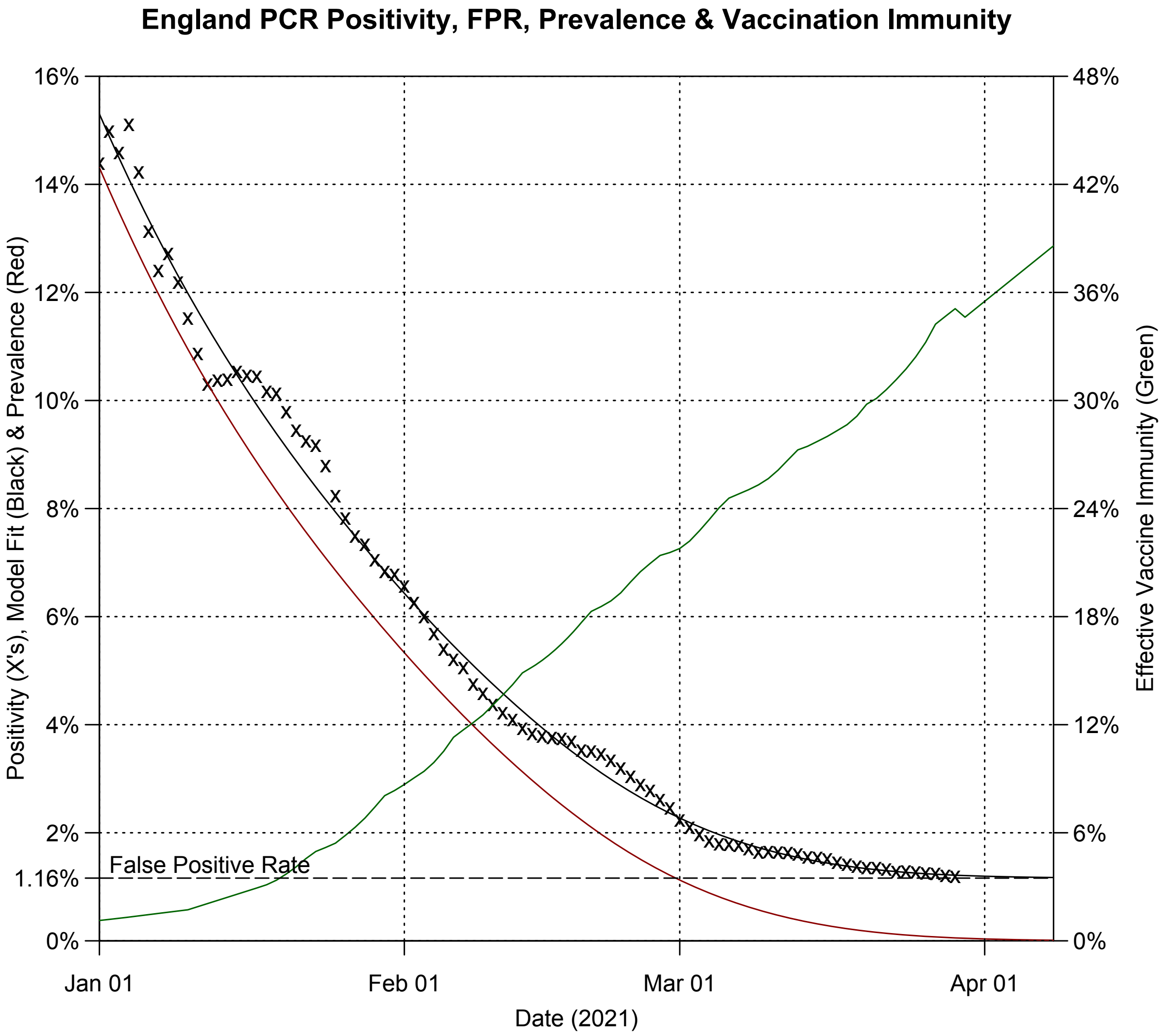


Chart 1 - Data and Model Fit

Discussion and Conclusion

A technique is presented to extract PCR false positive rates (i.e. specificity). The exponential like decay to a false positive rate floor visually speaks for itself. Using this data, test population prevalence may be extracted. For this PHE data (2021 to date), the population prevalence is well below the estimated false positive rate and is rapidly decaying. Without the application of the false positive rate to the current PCR test data, at this stage of the epidemic, the PCR test positivity will "flatline" with time, and continue to produce false positives even if the population prevalence has fallen to zero.

Limitations

This paper is not peer reviewed. While the author exercised reasonable care in presenting the results, hidden mistakes may be contained therein. The statistical techniques used within this paper may not be statistically robust.

While the Public Health England data is relatively clean, it is an amalgamation of multiple PCR testing sites, each which may change its test parameters at any time, resulting in a different false positive rate. The technique performs best when the testing parameters and the tests used are consistent and homogeneous. For example, some public health authorities may mix lateral flow test results with PCR test results; or may bias test results by screening first, or dropping negative results. Such mixed, changed, or biased data makes the ability of the technique to discriminate trends less reliable.

A constant transmissibility β (i.e. social distancing) is assumed. The change in naturally acquired immunity (but not the absolute level) over the period is assumed to be relatively insignificant.

The technique works during an epidemic curve period when the total infections are falling rapidly in a consistent manner, so that the false positive rate floor can be detected. During other periods, the false positive floor may not be discernible.

The reported validation of many COVID-19 PCR tests indicate that the specificity rate is 100% (i.e. false positive rate is zero). [9, 10]. The results presented here are in contradiction to those reports. The author suggests that real-world testing of large populations has a different specificity than the laboratory and small scale validations. Of note, is that the UK Government Office of Statistics stated in April 2020 that the operational false positive rate was unknown [11] and that an April 2020 review [12] of COVID-19 PCR false positives published test results found a wide range (0.3% to 6.3%) of false positive rates for both COVID-19 PCR and other comparable non-COVID-19 PCR tests.

Code and Data Availability

All code is available on line via MedRxiv or by request to the author. All data is downloaded from Public Health England using the function fetchEngland() contained within the code, or alternatively can be downloaded manually using the url's contained within the code's comments. The most recent data download files are included for reference.

[1] Public Health England, Coronavirus (COVID-19) in the UK, Data Download, https://coronavirus.data.gov.uk/details/download, as accessed March and April 2021.

[2] Public Health England, Coronavirus (COVID-19) in the UK, About the Data, https://coronavirus.data.gov.uk/details/about-data, as accessed 31 March 2021.

[3] Thompson MG, Burgess JL, Naleway AL, et al. Interim Estimates of Vaccine Effectiveness of BNT162b2 and mRNA-1273 COVID-19 Vaccines in Preventing SARS-CoV-2 Infection Among Health Care Personnel, First Responders, and Other Essential and Frontline Workers — Eight U.S. Locations, December 2020–March 2021. MMWR Morb Mortal Wkly Rep 2021;70:495–500. DOI: http://dx.doi.org/10.15585/mmwr.mm7013e3external icon.

[4] Wikipedia contributors. Compartmental models in epidemiology. Wikipedia, The Free Encyclopedia. April 3, 2021, 21:46 UTC. https://en.wikipedia.org/w/index.php?title=Compartmental_models_in_epidemiology&oldid=1015841752. Accessed April 6, 2021.

[5] M. Halem. Calibrating an epidemic compartment model to seroprevalence survey data. medRxiv, page 2020.05.27.20110478, 01 2020, https://doi.org/10.1101/2020.05.27.20110478 (p. 17)

[6] R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/. Note: the nls (Nonlinear Least Squares) function is part of the R stats package, itself part of the R core.

[7] The PORT Mathematical Subroutine Library, Phyllis A. Fox, Editor, 1984, AT&T Bell Telephone Laboratories, Inc. http://www.netlib.org/portt/ (Used as an option within the R nls function.)

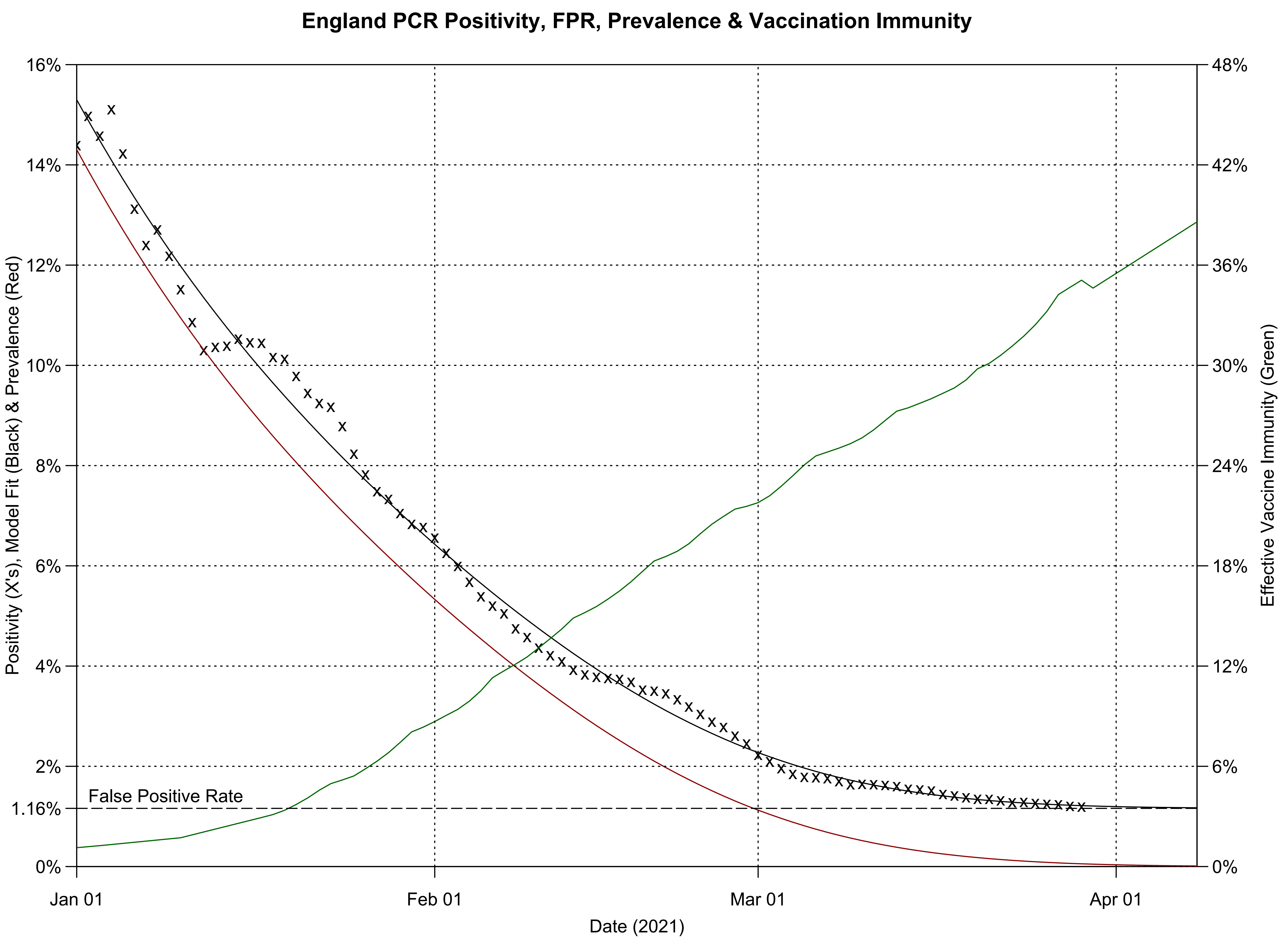
[8] Israel Ministry of Health, COVID-19 Dashboard, Downloaded 2 April 2021, https://datadashboard.health.gov.il/COVID-19/general

[9] United States Food and Drug Administration, Emergency Use Authorization (EUA) Summary, COVID-19 RT-PCR Test (Laboratory Corporation of America), 9 December 2020, https://www.fda.gov/media/136151/download

[10] Comparison of 12 molecular detection assays for SARS-CoV-2, Yasufumi Matsumura, Tsunehiro Shimizu, Taro Noguchi, Satoshi Nakano, Masaki Yamamoto, Miki Nagao, bioRxiv 2020.06.24.170332; doi: https://doi.org/10.1101/2020.06.24.170332

[11] Carl Mayers and Kate Baker, Government Office of Statistics (GOS) for the Scientific Advisory Group for Emergencies (SAGE), Impact of false-positives and false-negatives in the UK's COVID-19 RT-PCR testing programme, 3 June 2020, https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/895843/S0519_Impact_of_false_positives_and_negatives.pdf

[12] Andrew N. Cohen, Bruce Kessel, Michael G. Milgroom, Diagnosing SARS-CoV-2 infection: the danger of over-reliance on positive test results, medRxiv 2020.04.26.20080911; doi: https://doi.org/10.1101/2020.04.26.20080911 , version 4, 28 September 2020.


England PCR Positivity, FPR, Prevalence & Vaccination Immunity
Positivity (X's), Model Fit (Black) & Prevalence (Red)
Effective Vaccine Immunity (Green)
16%
14%
12%
10%
8%
6%
4%
2%
1.16%
0%
48%
42%
36%
30%
24%
18%
12%
6%
0%
False Positive Rate
Jan 01
Feb 01
Mar 01
Apr 01
Date (2021)