Eur. Phys. J. Plus (2021) 136:481
https://doi.org/10.1140/epjp/s13360-021-01448-2

THE EUROPEAN PHYSICAL JOURNAL PLUS

Regular Article

A study on the possible merits of using symptomatic cases to trace the development of the COVID-19 pandemic

Gianluca Bonifazi[1,2], **Luca Lista**[3,4], **Dario Menasce**[5], **Mauro Mezzetto**[6,a], **Daniele Pedrini**[5], **Roberto Spighi**[2], **Antonio Zoccoli**[2,7]

[1] Università Politecnica delle Marche, Ancona, Italy
[2] INFN Sezione di Bologna, Bologna, Italy
[3] Università degli Studi di Napoli Federico II , Naples, Italy
[4] INFN Sezione di Napoli, Naples, Italy
[5] INFN Sezione di Milano Bicocca, Milan, Italy
[6] INFN Sezione di Padova , Padua, Italy
[7] Alma Mater Studiorum Università di Bologna, Bologna, Italy

Received: 5 January 2021 / Accepted: 16 April 2021

Abstract In a recent work, we introduced a novel method to compute the effective reproduction number R_t and we applied it to describe the development of the COVID-19 outbreak in Italy. The study is based on the number of daily positive swabs as reported by the Italian Dipartimento di Protezione Civile. Recently, the Italian Istituto Superiore di Sanità made available the data relative of the symptomatic cases, where the reporting date is the date of beginning of symptoms instead of the date of the reporting of the positive swab. In this paper, we will discuss merits and drawbacks of this data, quantitatively comparing the quality of the pandemic indicators computed with the two samples.

1 Introduction

The worldwide data about the development of the COVID-19 outbreak is always reported as daily number of positive swabs, see for instance [1]. In a recent paper, we introduced a novel method to compute the effective reproduction number R_t based upon the counting of daily positive swabs [2]. Positive swabs data suffers from several problems, since they can be biased by different strategies and response time for swab data taken in different regions and different periods of time. Data collection is affected by strong weekend effects in recording the values, due to reduced capacity of processing swabs on Saturdays and Sundays. Furthermore, the reporting of a positive swab introduces a variable delay between the date of contagion and the date of reporting.

Potentially, the reporting of symptomatic cases, together with the date of symptom onset, could attenuate most of these issues. In principle, symptomatic cases should suffer less from different strategies of swab data taking, being the most urgent cases to be treated, and the date

[a] e-mail: mauro.mezzetto@pd.infn.it (corresponding author)

of symptom onset should be less influenced by weekend effects and should not be affected by additional delays introduced by the processing and reporting of a molecular swab.

On the other hand, the sample of symptomatic cases is a subset of the total number of cases, consequently the size of the sample is an issue for relatively small populations, like Italian regions or provinces. Furthermore, a bias could be introduced if the true fraction of symptomatic cases changes during the pandemic because of a modification of the age distribution of infected people.

From December 6th 2020, the numbers of symptomatic cases, associated to the date of symptom onset, are made available in Italy by the daily bulletin of the Istituto Superiore di Sanità (ISS) [3].[1] The published data contains the history of all the symptomatic cases on a national basis, while for regions and provinces are reported only the cumulative data of positive swabs.

Data about positive swabs are published, since the beginning of the outbreak, by the Italian Dipartimento di Protezione Civile (DPC) [4]. It becomes then possible to compare the information that can be extracted from the full sample of positive swabs with the one from the sub-sample of symptomatic cases.

In the following, we will work out several indicators to compare the merits and the differences of the two samples.

2 The data

We show in Fig. 1 the daily data of the symptomatic cases[2] and positive swab samples. We perform a fit to the data with the sum of four derivatives of Gompertz functions, $g(t; a, b, c)$:

$$g(t; a, b, c) = a\, e^{-b\,t} e^{-c\, e^{-b\,t}} \tag{1}$$

in order to take care, respectively, of the first phase of the outbreak in the period March–April, the increase of August, the third phase in October–December with the main peak at November and another local maxima at the end of December, as reported in the same Fig. 1.

The main parameters of the fits are reported in Table 1.

By comparing the dates of the second peak we conclude that the positive swab sample is delayed by 7.9 ± 0.2 days with respect to the sympotmatics one, which can be considered the average delay between the appearance of the symptoms and the reporting of a positive swab. This number takes into account that asymptomatic cases are mostly detected by a tracing of the symptomatic cases (which is the cause of delay) and not by a generic screening of the population. The fact that the delay at the first peak was larger by 3.9 days with respect to the delay at the second peak could be understood as more efficient procedures for swab processing developed along the outbreak.

[1] It should be noted that the collection of molecular swabs began on February 24, 2020 and the reported symptomatic cases reported before this date refer to positive swabs only. For this reason, the symptomatic cases reported from January 28 to February 24 represent an incomplete sample and we do not consider them in the following for the whole pandemic period.

[2] ISS reports two set of data for the symptomatic cases, called "casi_inizio_sintomi" and "casi_inizio_sintomi_sint". This latter set of data is described as "number of cases of confirmed SARS-CoV-2 virus infection for which a symptom onset date is indicated except for cases declared as asymptomatic". In absence of any better explanation we use this sample in the following.

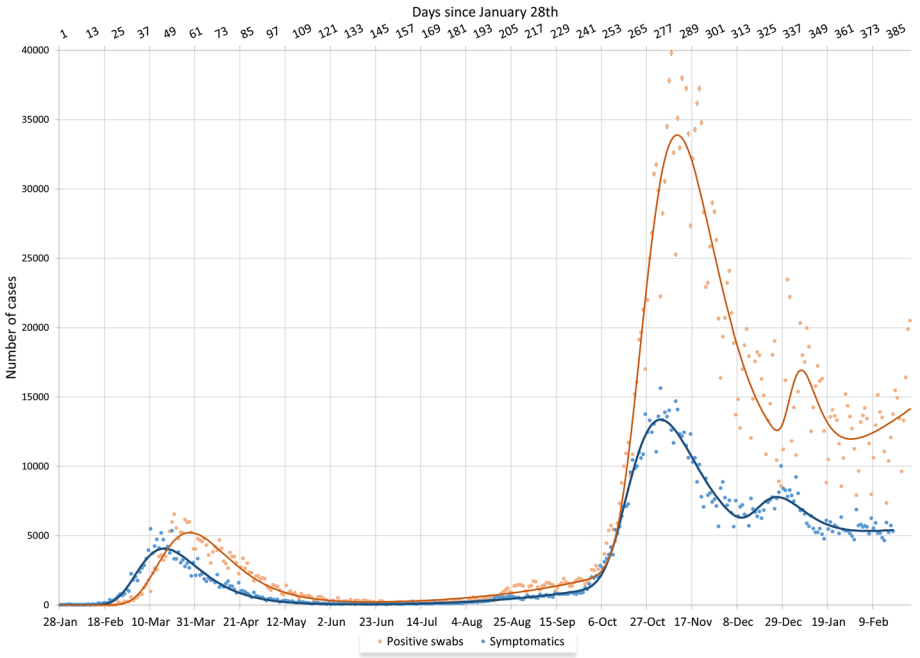

Fig. 1 Distribution of the symptomatic cases (light blue) and of the positive swabs (orange) as reported by ISS and DPC, respectively. Poisson errors are drawn but are of the same size of the bullets. The continuous lines are the fits to these data as described in the text. The upper horizontal axis displays days since January 28th, 2020

Table 1 For the symptomatics and positive swabs samples we display the fit values of the peak positions expressed in days since January 28th, the values of the Full Width Half Maximum (FWHM) and the standard deviations of the pull distributions for both the first and second peak. The errors of the peak position and of the FWHM are computed according to the covariance matrix of the global fit. Pulls and FWHM are discussed in Sects. 2.1 and 2.2, respectively

		Position (days)	FWHM (days)	Pulls (st. dev.)
Symptomatics	First peak	49.4 ± 0.1	35.0 ± 0.1	7.2
	Second peak	279.2 ± 0.1	51.0 ± 0.2	7.1
Positive swabs	First peak	61.2 ± 0.3	39.4 ± 0.5	8.4
	Second peak	287.1 ± 0.2	49.9 ± 0.3	16.2

2.1 Pulls

We can compare the amount of dispersion present in the two samples by computing the pulls of the curves. Pulls are defined as the difference of the fitted point fit function with the data point, divided by the Poisson error of the data point. Would the Poisson error be the only source of errors, we would have to find a distribution of pulls with a standard deviation of 1, if the adopted model was the correct one.

Pulls distributions are displayed in Fig. 2. Limiting the analysis to the first peak, where the size of the two samples is similar, we obtain a standard deviation of the residual distributions equal to 7.2 for the symptomatic cases and of 8.4 for the positive swabs.

A contribution to these high values could be due to a non-perfect parameterization of the data or to an underestimation of the quoted errors that does not take into account additional systematic contributions.

The pulls distribution of the positive swabs considerably worsen at the second peak. We note that in this period antigenic tests initiated to be used by several Italian regions to complement molecular swabs, and from January 15th, 2021 antigenic tests are accounted together with the molecular ones. This change of strategy, with the introduction of tests with different efficiency and processing time, could have contributed to the randomization of the data.

We can conclude that the sample of symptomatic cases does not significantly reduce the dispersion of the data around the central values with respect to the positive swab sample. Changes on the strategy of swab collection can anyway introduce important additional fluctuations in the positive swab sample.

2.2 Full Width Half Maximum

Another quantity that could be influenced by additional fluctuations present in the positive swabs sample is the width of the peaks. If, for instance, the delay between the date of appearance of symptoms and the date of reporting of a positive swab would follow a broad distribution, this could affect the width of the fitted peaks.

We compute the Full Width Half Maximum (FWHM) of the peaks as the distance between half peak position in the rising and descending parts and of the Gompertz curves. According to the values reported in Table 1, we found a FWHM of 35.0 and 39.4 days at the first peak

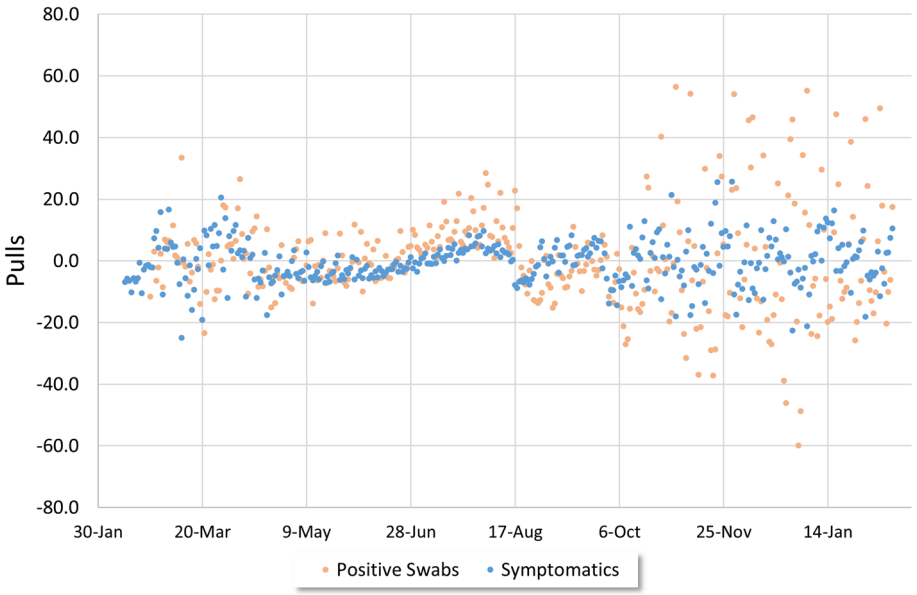

Fig. 2 Distribution of the pulls for the four Gompertz fits to the positive swab and symptomatics samples

and 51.0 and 49.9 days at the second peak for the symptomatic and positive swab samples, respectively.

We do not take into consideration the second peak for this computation because it's described by the sum of three Gomperz curves and in the case of the symptomatics the distribution restarts for a third local maximum before the second peak reaches the half height, biasing the FWHM.

We consider the differences at the first maximum as an indication of a significant contribution of the dispersion of swab reporting times to the distribution of positive cases. We can quantify this contribution, by taking into account that in the gaussian approximation the FWHM is equal to 2.35 standard deviations. If we attribute the increase of FWHM, from the value of the symptomatics first peak ($\mathrm{FWHM_{sym}} = 35.0$ days, see Table 1) to the value of the positive swabs first peak ($\mathrm{FWHM_{ps}} = 39.4$ days), to a gaussian dispersion of the positive swabs reporting times σ_{swabs}, this dispersion results to be:

$$\sigma_{\mathrm{swabs}} = \sqrt{\left(\frac{\mathrm{FWHM_{sym}}}{2.35}\right)^2 - \left(\frac{\mathrm{FWHM_{ps}}}{2.35}\right)^2} = 7.7 \text{ days}. \tag{2}$$

2.3 Fraction of asymptomatic cases

A side result of these comparisons is the distribution of the fraction of asymptomatic cases in the positive swabs sample along the outbreak. If we anticipate by 8 days the positive swabs distribution, according to the above discussion, we can compute day by day the difference of

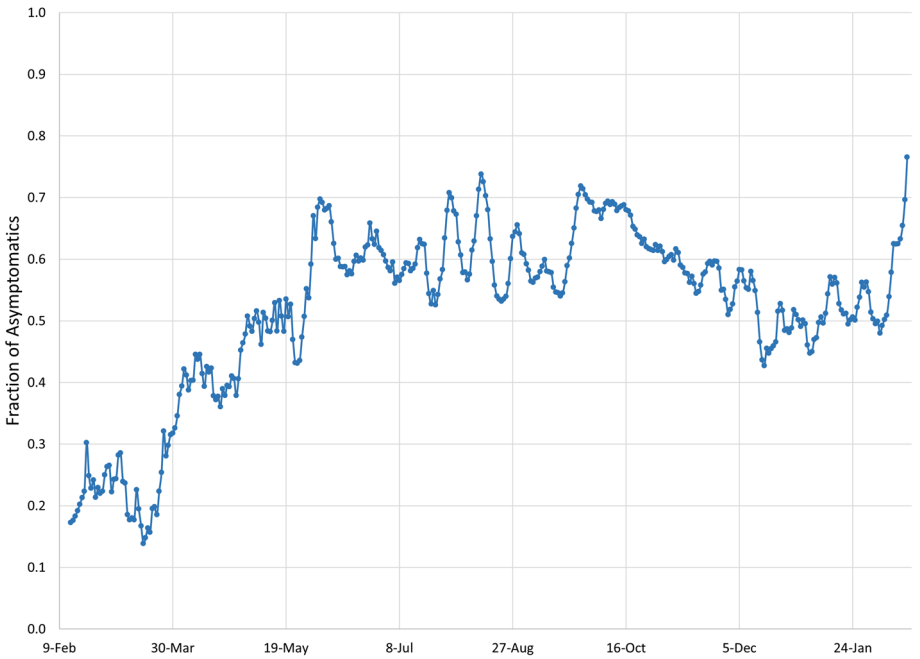

Fig. 3 Fraction of asymptomatic cases in the positive swab sample. Dates are adjusted following those of the symptomatic cases, as discussed in the text

the two data and from this the fraction of asymptomatic cases in the positive swabs sample. The result is displayed in Fig. 3.

We observe that at the beginning of the pandemic the fraction of asymptomatic cases became as small as 0.15 during the first peak of the pandemics. It then grew to about 0.6 at the end of first peak, and remained stable until the second peak was reached, when the total number of swabs was probably insufficient to guarantee a proper tracing and the fraction of asymptomatics decreased to 0.5.

2.4 Stability of the symptomatic sample

The number of symptomatic cases needs time to stabilize, since they are collected after the reporting of a positive swab and it takes time to execute, process and report a molecular swab after the onset of symptoms. To check the stability of the sample we considered the data as they were collected on December 16th, 2020 and as they were later updated on March 1st, 2021; Fig. 4 shows the variation, day by day, between the original data with respect to its updated values, up to December 16th, 2020.

ISS does not use the last 14 days of data for the computation of R_t, [3], and indeed from Fig. 4 it's evident that the last 14 days undergo to huge variations, some additional days are probably needed for a better stabilization of the data. We note however that also all previous days are affected by the recounting of data, with variations as big as 10%. This means that the symptomatic sample never fully stabilizes, and the derived values of R_t are in this way subject to continuous revisions.

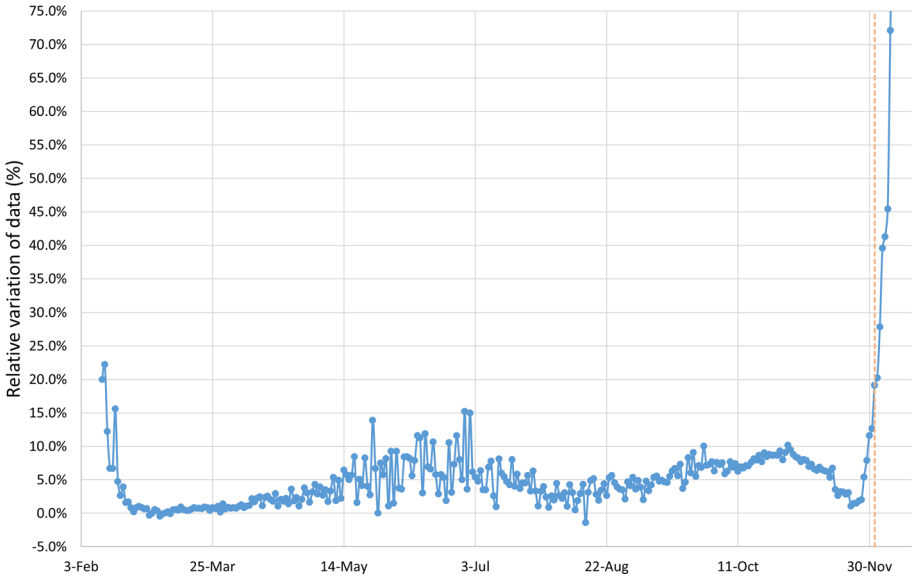

Fig. 4 Relative difference, in percentage, of the counts of systematic cases as reported on December 16th, 2020 (n_i) and on March 1st, 2021 (n_i') expressed as $(n_i' - n_i)/n_i$ (%). The vertical dashed line indicates the 14th day to the end of the period. The vertical scale is truncated, so not all the values of the last 14 days are visible

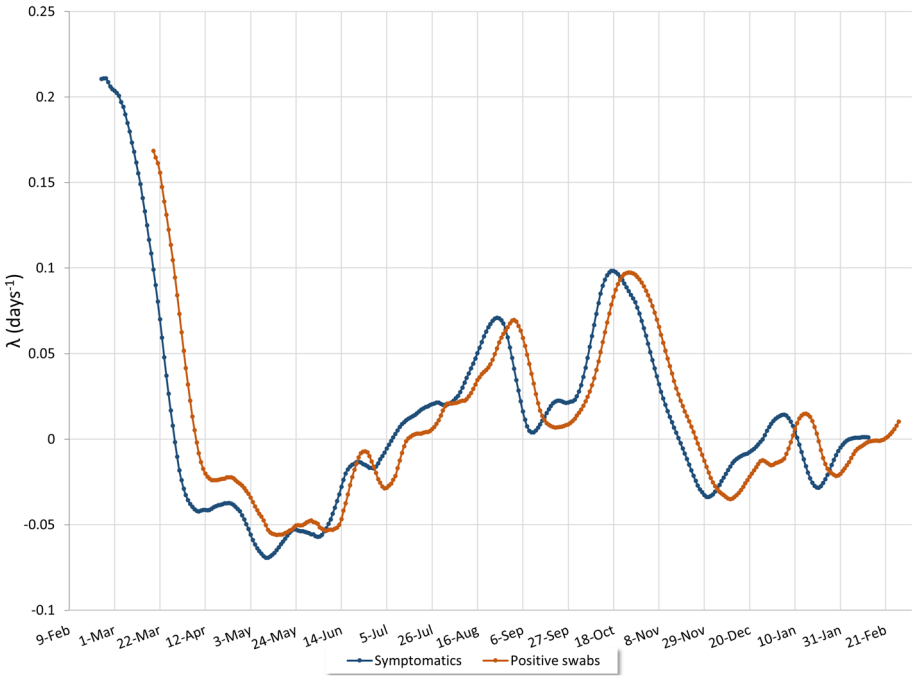

Fig. 5 Growth rate λ of the exponential fit to reported cases in the last 14 days for the symptomatic sample (blue) and the positive swabs sample (red)

3 The indicators of the development of the outbreak

We discussed in [2] that the growth rate $\lambda = \log 2/t_2$, where t_2 is the doubling time of an exponential fit to the data in the last "n" days, is as good an indicator as R_t for the description of the development of the outbreak. λ is computed via an exponential fit to the last 14 days and we display its moving average along 14 days. Results are displayed in Fig. 5.

The values of λ computed with the symptomatics and positive swabs samples have almost the identical behaviour with the characteristic delay of the positive swabs curve.

We display the same result in terms of the more familiar R_t in Fig. 6, computed using the algorithm published in [2]. In this case, for a better comparison, we anticipate the positive swabs R_t by 8 days, according to the conclusions of Sect. 2.

The two R_t estimations are in good agreement within errors, both in shape and absolute values. The agreement demonstrates that the computation of R_t from the positive swab sample is robust against the strong variations of the fraction of asymptomatics happened during the first peak as well as the significant increase of the dispersion of collected data happened during the second peak.

For the sake of completeness we repeat the same comparison with four of the most common algorithms used in literature to evaluate R_t: Wallinga and Teunis [5], Cori et al. [6], both computed thanks to the public package EpiEstim [7], Bettencourt-Ribeiro [8], computed following the indications of [9], and Robert Koch Institute (RKI) [10]. The plots are reported in Fig. 7 and show the identical behaviour of the plot of Fig. 6. According to ISS [3], their official value of R_t is computed with the Cori et al. algorithm.

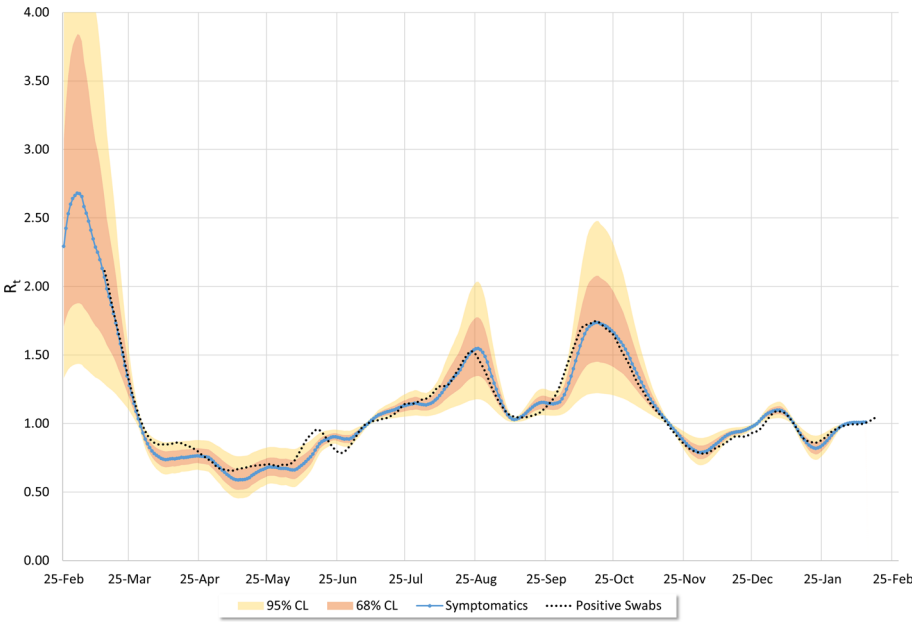

Fig. 6 R_t computed for the symptomatic sample (blue) with the uncertainty band up to 68% confidence level (orange) and up to 95% confidence level (yellow). Superimposed is R_t computed with the positive swabs sample (black dotted), moved to the left by 8 days according with the conclusions of the above discussions. R_t is computed with the algorithm published in [2]

4 Conclusions

We have compared the information that can be extracted about the development of the COVID-19 outbreak in Italy by using the daily new cases reported for the infected with symptoms along with the total sample of positive swabs. The symptomatics is a particularly valuable control sample because it suffers of less systematic effects than the positive swabs sample.

We observe a modest reduction of the dispersion of the data with the sample of symptomatic cases and a better definition of the peaks in the distribution of the daily positive cases. The differences between the two curves lie mostly in a delay between the appearance of the symptoms and the date of the reported positive swab, amounting to about 8 days. By applying this correction, the two samples are comparable and the corresponding two extracted distributions of R_t turn out to be in good agreement within errors.

One could believe that the symptomatic sample, being preempted by about 8 days, could identify in advance the trends of the outbreak, however, as discussed in Sect. 2.4, it needs 14 days to be correctly determined. Considering this effect, the symptomatic cases sample, as a real-time estimator, is retarded with respect to the positive swabs sample. The sample of positive swabs provides real-time evaluations of R_t that are faster and more stable.

We conclude that the sample of the positive swabs can be safely used to monitor the development of the COVID-19 outbreak.

We publish daily estimates of R_t in real time, together with more information about the development of the Italian outbreak in [11]. Daily values for the major world countries are also reported.

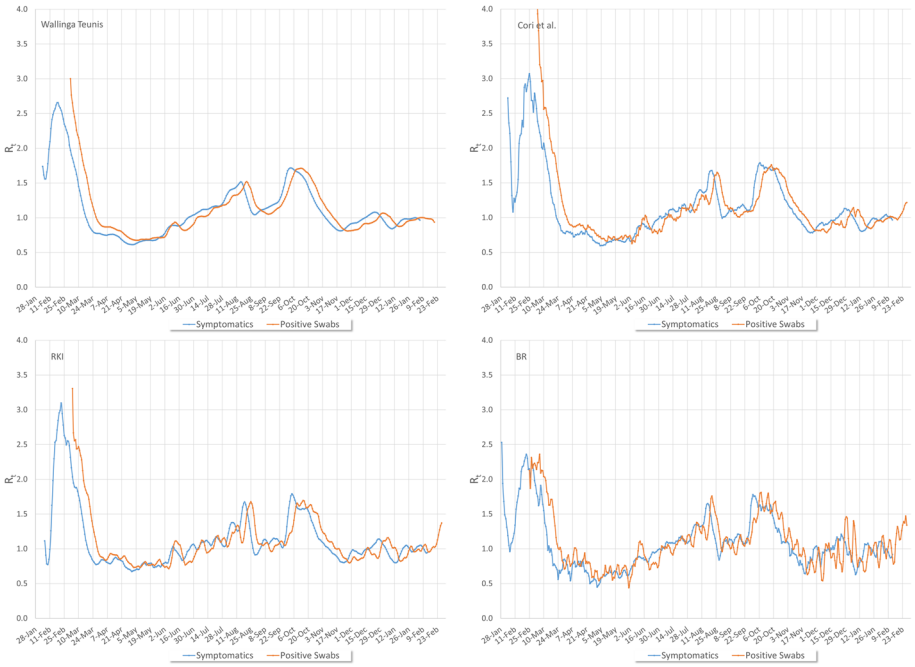

Fig. 7 R_t computed for the symptomatic sample (blue) and the positive swabs sample (orange) with four different algorithms (see the text). From upper left, clockwise: Wallinga-Teunis, Cori et al., Bettencourt-Ribeiro and RKI

Acknowledgements We acknowledge the effort of ISS of making public the data of the symptomatic cases of COVID-19. The present work has been done in the context of the INFN CovidStat project that produces an analysis of the public Italian COVID-19 data. The results of the analysis are published and updated daily on the website `covid19.infn.it/`. The project has been supported in various ways by a number of people from different INFN Units. In particular, we wish to thank, in alphabetic order: Stefano Antonelli (CNAF), Fabio Bredo (Padova Unit), Luca Carbone (Milano-Bicocca Unit), Francesca Cuicchio (Communication Office), Mauro Dinardo (Milano-Bicocca Unit), Paolo Dini (Milano-Bicocca Unit), Rosario Esposito (Naples Unit), Stefano Longo (CNAF), and Stefano Zani (CNAF). We also wish to thank Prof. Domenico Ursino (Università Politecnica delle Marche) for his supportive contribution.

Data Availability Statement This manuscript has associated data in a data repository. [Authors' comment: All data included in this manuscript are available upon request by contacting with the corresponding author.]

References

1. E. Dong, H. Du, L. Gardner, An interactive web-based dashboard to track COVID-19 in real time. Lancet Infect. Dis. **20**, 533–534 (2020). https://doi.org/10.1016/S1473-3099(20)30120-1
2. G. Bonifazi et al., A simplified estimate of the Effective Reproduction Number Rt using its relation with the doubling time and application to Italian COVID-19 data. Eur. Phys. J. Plus, **136**(4), 1–14 (2021). https://doi.org/10.1140/epjp/s13360-021-01339-6
3. https://www.epicentro.iss.it/coronavirus/sars-cov-2-dashboard
4. Dipartimento della Protezione Civile, *Dati COVID-19 Italia*, https://github.com/pcm-dpc/COVID-19
5. J. Wallinga, P. Teunis, Different Epidemic Curves for Severe Acute Respiratory Syndrome Reveal Similar Impacts of Control Measures. Am. J. Epidemiol. **160**(6), 509–516 (2004). https://doi.org/10.1093/aje/kwh255

6. A. Cori, N.M. Ferguson, C. Fraser, S. Cauchemez, A New Framework and Software to Estimate Time-Varying Reproduction Numbers During Epidemics. Am. J. Epidemiol. **178**(9), 1505–1512 (2013). https://doi.org/10.1093/aje/kwt133
7. EpiEstim: *Estimate Time Varying Reproduction Numbers from Epidemic Curves*, https://cran.r-project.org/web/packages/EpiEstim/index.html
8. L.M.A. Bettencourt, R.M. Ribeiro, Real Time Bayesian Estimation of the Epidemic Potential of Emerging Infectious Diseases. PLoS ONE **3**(5), e2185 (2008). https://doi.org/10.1371/journal.pone.0002185
9. K. Systrom, *The Metric We Need to Manage COVID-19*. R_t: *the effective reproduction number*, 2020, http://systrom.com/blog/the-metric-we-need-to-manage-covid-19/
10. Robert Koch Institut, *Erläuterung der Schätzung der zeitlich variierenden Reproduktionszahl* (2020), https://www.rki.de/DE/Content/InfAZ/N/Neuartiges_Coronavirus/Projekte_RKI/R-Wert-Erlaeuterung.pdf
11. CovidStat INFN, https://covid19.infn.it/