Supplementary Information

Contact tracing procedure

The Local Health Authority provided data on confirmed SARS-CoV-2 cases in Chieti province. In particular, data quality was improved by means of cleaning, validation, and updating of the original information. Coherence between the possible dates of infection in infector-infected pairs was evaluated. Then, information on address of cases was used to find spatial coordinates through OpenStreetMap and geocoding procedures developed using the R package tidygeocoder (Cambon & Hernangómez, 2021). Network analysis techniques (by using visNetwork, dplyr and igraph libraries in R environment, version 1.4.1106) were used to generate a relational spatial-temporal network data set, considering each positive case as “node” and the epidemiological or spatial-proximity contact between them as “link”. For each node, a reference data was given, calculated as the earliest date between that of clinical onset and the date in which the positive sample was taken. A trace forward analysis was performed, using each positive case as seed and a time window of 66 days starting from 01/12/2020, to identify all possible temporally valid infectious contact chains generated by each positive case. Cases were grouped into clusters based on reported related cases, additional information, or spatial proximity.

Estimating the key epidemiological parameters

Methods for the estimation of Re(t) are based on the calculation of the incidence of symptomatic cases in conjunction with the estimation, using the clinical onset dates, of the serial interval distribution S( ) as a proxy of the generation interval G( ) (Cori et al., 2013; Gostic et al., 2020; Hens et al., 2012; Thompson et al., 2019). Knight & Mishra (Knight & Mishra, 2020) derived the G( ) distribution from S( ) to evaluate the differences in the corresponding Re(t), and showed how the use of S( ) may result in a biased estimation of Re(t), assuming both non negative or negative-permitting serial interval distribution (Knight & Mishra, 2020). This approach may result in biased estimations of Re(t), allowing the possibility of negative serial interval distributions. A more recent study by Gostic et al. (Gostic et al., 2020) demonstrated the importance of generation interval misspecification, the effects of delays around incidence and the shift in the estimated Re(t) caused
by smoothing windows. The authors concluded that the model in Cori et al. (Cori et al., 2013) ensures good performances, but adjustments are needed. Because of the lack of detailed data in real situations, estimating Re(t) and correctly inferring its timing of changes may be challenging.

Sampling procedure of exposure times based on the transmission chain

We based our approach on the principle that a complete (or almost complete) transmission chain can correctly drive the exposure time sampling within a plausible time window. We assumed that each date of exposure could be bounded. When the date of symptoms onset was available, the exposure was assumed to have occurred within 15 days before, in line with the incubation time (Ht) reported by Lauer et al. (Lauer et al., 2020), and considering that less than 1‰ of exposures fall outside this time frame. When the date of symptoms onset was not available, the exposure was assumed to fall within a period between the date of the first positive diagnostic test and the lower bound of the infector (the corresponding primary case). SARS-CoV-2 transmission may occur before the onset of clinical signs (Nishiura et al., 2020; Tindale et al., 2020), and its incubation period varies greatly, and, therefore, it is theoretically possible for an infected person (secondary case) to show symptoms even before his/her infector. Therefore, we applied an iterative procedure using the following rules to further narrow the possible period of exposure:

I. The upper bound of the exposure period of the infector must be earlier than the minimum of all the upper bounds of linked infected people;

II. The lower bound of the exposure of an infected person must be later than the lower bound of his/her infector.

Starting from the index case (seed) of each epi-cluster, we sampled the exposure time within its interval and updated the lower bound of the linked infected persons accordingly. We used a uniform sampling for all cases except the seeds, for whom we used the H(t) as in Lauer et al. (Lauer et al., 2020). We repeated the procedure to the whole transmission chain for the 14 epi-clusters and generated 10,000 infection trees. A schematic representation of the procedure above described is reported in Figure S1.
Figure S1. How the infector-infected pairs and dates of clinical onset help to identify the most plausible time interval during which the exposure took place. Left panel: an example of infection chain is presented. A first transmission chain is composed by person s, who infected i1, i2 and i3. Then a second chain involved person i3, who infected j1 and j2. Panel A. For each case, the date of clinical onset is reported as the arrowheads and the earliest date of exposure (based on the incubation period) is reported as a circle. Panel B. Considering the infector-infected pair s – i2, it must be assumed that the exposure time for s (infector) was before the appearance of clinical symptoms in i2 (infected) and therefore we can reduce the exposure window of s taking into account the date of clinical onset of i2 (blue line). Similarly, it must be assumed that the exposure date for i2 was after that of s, and therefore we can reduce the exposure window of i2 taking into account the earliest date of infection of s (green line). The same considerations can be made for each infector-infected pair, allowing to restrict the potential time windows of exposure (diamonds represent the updated bounds of the exposure time window). Panel C. The sampled exposure date (marked with a star) of an infector affects the sampling of the same date of the relative infected in order to assure temporal coherence between the two dates.

Evaluation of data analysis procedure

We generated a synthetic dataset to evaluate the correctness of the procedure used for the estimation of the generation interval\(^1\), considering that the temporal trend of Re(t) is known (true Re(t)), as well as the gamma discretized distributions of incubation times (s() ~ Gamma(5.81, 0.95) (Lauer et al., 2020)) and generation intervals (g() ~ Gamma(2, 1.3), arbitrary chosen).

The dataset is generated under the following assumptions:

- Infinite population composed by susceptible people only;
- All those exposed become symptomatic;
- All symptoms are detected (this implies that the actual infection contact chain coincides with the observed one).

It is considered that starting from a case (v) at time t, this can infect a number of people i ~ Poisson(Re(t)). Newly infected (i) are characterized by an exposure time Ti ~ t + infection date g() \(^1\) Generation interval defined as the time between exposures in each pair of infector and infected person.
and a symptom onset $S_i \sim t + s()$. The generated tree consists of 968 infectors-infected person pairs, and 969 overall cases.

We used the procedure described in this study to generate 5,000 infection trees sampled from the infector-infected person pairs and symptom onset dates of the synthetic dataset to:

1. Evaluate the goodness of the simulated infection incidence compared with the real one;
2. Evaluate the trend of the $R_e(t)$ case-reproduction number compared with the real trend of $R_e(t)$;
3. Compare $R_e(t)$ calculated following the method described by Cori et al. (2013) on the synthetic tree with the real trend of $R_e(t)$;
4. Evaluate the generation interval distribution estimated from the simulated trees with respect to the real distribution $g$;
5. Evaluate the $R_e(t)$, calculated following the method described by Cori et al. (2013), adjusted according to the incidence and transmission values from the simulated trees and applying a shift of the reference time.

The workflow that was followed in the study is illustrated in Figure S2.

**Figure S2. The workflow followed in the study.** Boxes: known value/distribution (black), methods (blue), estimate/result by method (blue, red contoured) and observed data, real or synthetic (orange). Arrows: flow (blue, single) and comparison (yellow, double). Sets: method evaluation on synthetic data (dark blue), Guardiagrele study (green) and proposed method (red).
1. Infection incidence

The infection incidence was estimated by the rounded average of the incidences of infection of each simulated tree, and compared with the correction/adjusted methods described in Gostic et al. (Gostic et al., 2020). The comparison of the deconvolve (calculated considering three iterations), shift, and the infection incidence simulated methods in relation with the real infection incidence is reported in Figure S3, and described in terms of root mean squared error (RMSE).

As reported in the work of Gostic et al. (Gostic et al., 2020), the deconvolve method identifies epidemic peaks better than the other methods and presents a lower RMSE value than the shift method applied on the symptoms onset date. The infection incidence calculated from the simulated trees has a smoother trend; it captures the epidemic peaks worse than the deconvolve method, but presents a minor error than the other methods.

Figure S3. Infection incidence comparison. The dashed, green line represents the real infection incidence; the orange, blue marine and violet lines represent the deconvolve, shift, and infection incidence simulated (SimTrees) methods respectively. Upper each panel is reported the related RMSE value.
2. Case-reproduction number Re(t)

The average case-reproduction number was calculated as the average number of secondary cases per each infected case on the estimated day of exposure t for each generated tree, similarly to Hens et al. (Hens et al., 2012). The average case-reproduction number of each tree has then been furtherly averaged over a 7-day floating window (similarly to what will be done for Re(t) with Epiestim), and centered on the first day, considering a perspective time orientation. Its distribution (from the 10,000 generated trees) has been used estimation of the Re(t) distribution. Figure S4 shows the estimated case Re(t) distribution (CI 95%) compared to the true Re(t) distribution (CI 95%).

Figure S4. Estimated case Re(t) distribution compared to true Re(t) distribution. The green line represents the true Re(t) distribution; the dashed blue line represents the median of the estimated case Re(t) distribution; the dark band the 95% CI and the lighter band the maximum and minimum values respectively; the dashed, red line is Re(t)=1.

3. Re(t) model according to Cori et al. (2013)

Re(t) was calculated using symptom onset dates from infector - infected person pairs with positive intervals (~ 80% of links in the synthetic dataset). The posterior distribution of the serial interval (SI) was calculated by Markov chain Monte Carlo (MCMC) for coarse data using the interval of symptoms onset date + 1. Figure S5 shows the sampled distributions (SI), compared to the generation interval (g) and the incubation period (s).
Figure S5. Sampled distributions and real distribution of the generation interval (g) and the incubation period (s). The bundle of grey lines represents the SI sampled distributions (the red line is its median distribution); the green line is the generation interval (g); the dashed, blue line is the incubation period (s).

In Figure S6 the trend of Re(t) calculated following Cori et al. (Cori et al., 2013) differs substantially from the true Re(t) for several reasons:

- The incidence of symptoms is lagged compared to the incidence of infection;
- The misspecification of the generation interval;
- The floating window of 7 days used (necessary when the data are few) is centered on the last day of the interval.

Figure S6. Re(t) calculated trend. The green line represents the true Re(t) distribution, the dashed blue line represents the median of the estimated Re(t) Cori (Cori model using the serial interval distribution), the dark band represents the 95% CI.
4. Generation interval

The generation interval was determined as the median value of the set of the generation intervals estimated for each infector-infected pair of each generated transmission tree (corresponding to 5,000 intervals for each link), obtaining 705 values.

Generation interval distributions were derived sampling from the posterior of a discrete gamma distribution obtaining 500 distributions sampled. Figure S7 shows the posterior distributions sampled (GI) based on the intervals between the presumed exposure dates, compared to the real generation interval distribution (g) and the incubation period (s).

It is noteworthy that the generation interval distributions calculated by the simulated trees reflect the real generation interval distribution. Therefore, this approach could be followed for the estimation of the infection incidence for the $\text{Re}(t)$ calculation according to Cori et al. (Cori et al., 2013).

Figure S7. Posterior distributions sampled compared to the real generation interval distribution and the incubation period. The grey lines represent the sampled generation interval distributions (GI), the red line is the median distribution, the dashed, green line is the real generation interval (g) and the dashed blue line is the incubation period (s).
5. Re(t) adjusted model according to Cori et al. (2013)

The Re(t) was calculated using the Cori et al. model and the incidence and generation interval derived from the simulated trees, with a shift back of the floating window. As expected, using the adjustments suggested by Gostic et al., the Re(t) adjusted estimate has improved significantly, both in terms of timing and uncertainty with respect to the use of the standard Re(t) model by Cori et al. (Figure S8).

Figure S9 shows the comparison of the case Re(t) and Re(t) Cori adjusted distributions.

![Figure S8. True Re(t) and estimated Re(t) adjusted distribution. The green line represents the true Re(t) distribution, the dashed blue line represents the median of the estimated Re(t) adjusted distribution, and the dark band the 95% CI.](image)
In conclusion, the use of the estimate of Re(t) by EpiEstim cannot disregard the adjustments highlighted in Gostic et al.. The use of the implemented procedure and the calculation of the Re(t) deriving from the case-reproduction number leads to estimates that well reflect the real trend both in terms of the extent of that timing. The illustrated procedure, therefore, allows for an estimation of the generation interval distribution, the infection incidence, and the case-reproduction number on the incubation time distribution (a maximum window is established within which the sample is then uniformly sampled). The expected variability in the estimates, due to a completely non-informative sampling, however, does not affect the uncertainty of the estimates, perhaps due to the information present in the links that bind the sampling. Moreover, the procedure has the advantage of being able to include in the study non-symptomatic cases (although belonging to the epidemic cluster), so the positive test date (upper limit) is known by setting the lower limit according to its infector.
Phylogenetic analysis

Figure S10. Maximum likelihood tree of B.1.1.7 sequences. Maximum likelihood tree of 1972 full-genome SARS-Cov-2 sequences belonging to the B.1.1.7 lineage. Global sequences are shown in grey, Italian sequences in blue. The node containing Italian sequences is shown and this node is used in Fig. 4 (966 tips). Branch length shows nucleotide substitutions per site.

Acknowledgments tables for the Italian sequences shared via GISAID are attached after the References.
References – Supplementary information

Cambon, J., & Hernangómez, D. (2021). Tidygeocoder: Geocoding Made Easy (R package version 1.0.2). https://doi.org/10.5281/zenodo.3981510

Cori, A., Ferguson, N. M., Fraser, C., & Cauchemez, S. (2013). A new framework and software to estimate time-varying reproduction numbers during epidemics. American Journal of Epidemiology, 178(9), 1505–1512. https://doi.org/10.1093/aje/kwt133

Gostic, K. M., McGough, L., Baskerville, E. B., Abbott, S., Joshi, K., Tedijianto, C., Kahn, R., Niehus, R., Hay, J. A., De Salazar, P. M., Hellewell, J., Meakin, S., Munday, J. D., Bosse, N. I., Sherratt, K., Thompson, R. N., White, L. F., Huisman, J. S., Scire, J., … Cobey, S. (2020). Practical considerations for measuring the effective reproductive number, Rt. PLoS Computational Biology, 16(12), 1–21. https://doi.org/10.1371/journal.pcbi.1008409

Hens, N., Calatayud, L., Kurkela, S., Tamme, T., & Wallinga, J. (2012). Robust reconstruction and analysis of outbreak data: Influenza A(H1N1)v transmission in a school-based population. American Journal of Epidemiology, 176(3), 196–203. https://doi.org/10.1093/aje/kws006

Knight, J., & Mishra, S. (2020). Estimating effective reproduction number using generation time versus serial interval, with application to COVID-19 in the Greater Toronto Area, Canada. Infectious Disease Modelling, 5, 889–896. https://doi.org/10.1016/j.idm.2020.10.009

Lauer, S. A., Grantz, K. H., Bi, Q., Jones, F. K., Zheng, Q., Meredith, H. R., Azman, A. S., Reich, N. G., & Lessler, J. (2020). The incubation period of coronavirus disease 2019 (CoVID-19) from publicly reported confirmed cases: Estimation and application. Annals of Internal Medicine, 172(9), 577–582. https://doi.org/10.7326/M20-0504

Nishiura, H., Linton, N. M., & Akhmetzhanov, A. R. (2020). Serial interval of novel coronavirus (COVID-19) infections. International Journal of Infectious Diseases, 93, 284–286. https://doi.org/10.1016/j.ijid.2020.02.060

Thompson, R. N., Stockwin, J. E., van Gaalen, R. D., Polonsky, J. A., Kamvar, Z. N., Demarsh, P. A., Dahlqwist, E., Li, S., Miguel, E., Jombart, T., Lessler, J., Cauchemez, S., & Cori, A. (2019). Improved inference of time-varying reproduction numbers during infectious disease outbreaks. Epidemics, 29(August). https://doi.org/10.1016/j.epidem.2019.100356

Tindale, L. C., Stockdale, J. E., Coombe, M., Garlock, E. S., Lau, W. Y. V., Saraswat, M., Zhang, L., Chen, D., Wallinga, J., & Colijn, C. (2020). Evidence for transmission of covid-19 prior to symptom onset. ELife, 9, 1–34. https://doi.org/10.7554/eLife.57149
Submitting laboratories where the genome data were generated and shared via GISAID, on which this research is based. We gratefully acknowledge the following Authors from the Originating laboratories responsible for obtaining the specimens, as well as the following National and Regional Authorities responsible for obtaining the specimens, as well as the following National and Regional Authorities responsible for obtaining the specimens.
