Supplementary Appendix

This appendix has been provided by the authors to give readers additional information about their work.

Supplement to:

Mohamed Abbas, Tomás Robalo Nunes, Anne Cori, Samuel Cordey, Florian Laubscher, Stephanie Baggio, Thibaut Jombart, Anne Iten, Laure Vieux, Daniel Teixeira, Monica Perez, Didier Pittet, Emilia Frangos, Christophe E. Graf, Walter Zingg, Stephan Harbarth. **Explosive nosocomial outbreak of SARS-CoV-2 in a rehabilitation clinic: the limits of genomics for outbreak reconstruction.**

Table of contents

Supplementary Material

Microbiological Methods

Unbiased high-throughput sequencing (HTS) analysis

All nasopharyngeal swabs (NPS) with cycle threshold (Ct) values ≤ 30 or < 35 obtained for the E gene (Cobas 6800 SARS CoV2 RT-PCR and the Charite assays) and the S gene (BD SARS-CoV2 reagent kit for BD Max system), respectively, and for which sufficient volume remained, were selected for unbiased HTS analysis (n=56).

HTS analysis was performed using the RNA protocol previously published [1]. Briefly, for each NPS, 220 µl were centrifuged at 10,000 × g for 10 min to remove cells. Then, two-hundred µl of cell-free supernatant were treated with 40U of Turbo DNAse (Ambion, Rotkreuz, Switzerland), according to the manufacturer's instructions. Nucleic acids were extracted with TRIzol (Invitrogen, USA). Ribosomal RNA depletion (Ribo-Zero Gold depletion kit (Illumina, USA) was done before libraries preparation (TruSeq total RNA preparation protocol (Illumina)). Libraries concentrations and sizes were analysed using the Qubit (Life Technologies, USA) and the 2200 TapeStation instruments (Agilent, USA), respectively. Libraries were multiplexed (1:4, 1:5 or 1:6) on the HiSeq 4000 platform (Illumina) using the 2x100-bp protocol with dual-indexing.

Duplicate reads were removed using cd-hit (v4.6.8). Low-quality and adapter sequences were trimmed out using Trimmomatic (v0.33). Reads were then mapped against the reference sequence MN908947 using snap-aligner (v1.0beta.18). Consensus for sequences with at least 10-fold coverage were then generated using custom script. In parallel, all raw data were analysed using a bioinformatic pipeline [2] that used virusscan 1.0 (https://github.com/sib-swiss/virusscan) to map reads against a respiratory viruses restricted database from Virosaurus [3] for the detection of co-infections.

NPS with either Ct values > 30 or ≥ 35 obtained for the E gene (Cobas 6800 SARS CoV2 RT-PCR and the Charite assays) and the S gene (BD SARS-CoV2 reagent kit for BD Max system), respectively, or for which low SARS-CoV-2 genome coverage were obtained (i.e. < 80%) by the unbiased HTS method, were sequenced with an amplicon-based sequencing method if a sufficient volume remained. Thus, nucleic acids were

extracted using the NucliSENS easyMAG (bioMérieux, Geneva, Switzerland) and them sequenced using an updated version of the nCoV-2019 sequencing protocol (https://www.protocols.io/view/ncov-2019-sequencing-protocol-bbmuik6w) (Microsynth, Balgach, Switzerland) on a MiSeq instrument (Illumina) with a 2x250-bp protocol.

Phylogenetic analysis

Sequence alignment was performed with MUSCLE (v3.8.31). The Evolutionary analyses were conducted in MEGA X [4] using the Maximum Likelihood method and Tamura 3-parameter model [5]. The tree includes also all SARS-CoV-2 complete genomes sequenced by our laboratory and submitted to GISAID from respiratory samples from COVID-19 positive patients presenting to our institute or others medical centre in Geneva, Switzerland, during the same period.

Implementation of the `outbreaker2` models

We combined epidemiologic and genetic data using the `outbreaker2` package in the R software, which has been used successfully in the reconstruction of the 2003 SARS-CoV-1 outbreak in Singapore [6, 7]. The model uses a Bayesian framework, which combines information on the generation time (time between infections in an infector/infectee pair), with a model of sequence evolution to probabilistically reconstruct the transmission tree.

As dates of onset of symptoms are known, and because dates of infection (i.e. acquisition) are not known with certainty, we imputed serial intervals from estimates from the work by Ali et al. [8], who showed that the serial interval decreased from the early stages of the pandemic due to improved control using non-pharmaceutical measures. For the primary analysis we used a short serial interval (mean 3.0, standard deviation [SD] 4.1), under the assumption of swift isolation of patients following symptoms. We used the incubation period as estimated by Bi et al., which follows a lognormal distribution with parameters mu of 1.57 and sigma 0.65 (corresponding to a mean of 5.95 days and SD 4.31) [9]. Where dates of onset were unavailable, we imputed them by using the median of the difference between symptom onset and date of swab.

The outbreaker package is designed to use contact tracing data to inform who infected whom. These data were unavailable for our outbreak, but we made a series of assumptions to generate a matrix of possible contacts between cases. Initially we aimed to construct this matrix using dates of presence in the hospital/ward for patients based on administrative data, and based on human resources shift rota for HCWs. However we identified potential inconsistencies in the latter, in particular stemming from multiple (including last-minute) changes as a result of many HCWs self-isolating due to possible or confirmed COVID-19. We therefore used these data in a sensitivity analysis, but for our main analysis, we reverted to a simpler set of assumptions to build our contact matrix. In our main analysis we assumed that HCWs were present in the hospital every day until the date of their first positive swab, included. We further assumed that patients only interacted with patients in their own ward, except for frail patients who were assumed to not have contact with any other patients. We assumed that all HCWs and administrative staff were able to infect each other,

but only HCWs could have contact with patients, and only in the wards they were attributed to; HCWs such as physical therapists or doctors who worked across multiple wards could have significant contact with all HCWs and patients. Under these assumptions, we are likely to capture many contacts which did not happen, and it is possible that we miss a few contacts which in fact did happen. However outbreaker does account for imperfect sensitivity and specificity of contact data, the levels of which are estimated as part of the model.

We conducted a number of different analyses, including a base scenario and several (n = 4) one-way sensitivity analyses (Supplementary Table 1):

- Sensitivity analysis 1:
  - We did not make any assumptions about contact patterns, and therefore all cases in the outbreak could potentially infect all other cases.
- Sensitivity analysis 2:
  - We used a longer serial interval (mean 5.2 days, SD 4.7) to allow for potentially slower isolation of symptomatic patients.
- Sensitivity analysis 3:
  - We used the HCW shift data from the human resources department, with minor corrections (removing HCWs that were mislabelled as "present" after date of positive RT-PCR). For both patients and HCWs we categorised days of "susceptibility" (5th percentile of the cumulative incubation period from Bi et al. [9]) and days of "infectiousness" (2 days before symptom onset based on the study by He et al. [10]). The last day of "infectiousness" was the date of swab for HCWs.
- Sensitivity analysis 4:
  - We assumed that isolation precautions prescribed for patients on date of positive RT-PCR were effective, and that from that date patients were no longer infectious.

For each model, we used a uniform prior between 0.75 and 1 for the reporting probability ("pi"). Indeed, we had a comprehensive screening and testing strategy, including of asymptomatic cases, and are therefore confident that we captured a near-total proportion of cases. We obtained sequences for 82% of all identified cases, and these are the cases used in the model. The lower bound of the prior for "pi" thus allows us to have missed 6 cases in addition to the 14 that were not sequenced. Posterior estimates for "pi" were compared to our prior choice to assess the validity of this assumption.

Given our high sampling rate, we allowed a single unobserved case on a transmission chain between any two observed cases (maximum "kappa" of 2 in outbreaker). This allows for identification of missed cases.

We used the default priors for the mutation rate ("mu") for all models (uninformative expoential prior with mean 1), and, where relevant, those for non-infectious contact rate ("lambda") and contact reporting coverage ("eps"), which uniform on [0, 1] [7, 11]. We used the default likelihoods for all models, except for the model without contact data where this was disabled [7, 11].

We ran each outbreaker model over 500,000 iterations of the MCMC, with a thinning of 1 in 500, in order to obtain 1000 posterior likelihood estimates, after a burn-in of 500 iterations. Convergence was assessed visually and through the Gelman-Rubin convergence diagnostic (using the `gelman.diag` function in the R package `coda v0.19-4`) [12], concluding that the chains converged appropriately if the upper limit of the confidence interval was < 1.1.

We reconstructed who infected whom within a Bayesian framework, while simultaneously estimating the dates of infection and mutation rate, as well as the sensitivity and specificity of reported contacts. We were also able to determine whether there were multiple importation events and missed cases. We examined the transmission tree with the highest posterior probability (which was also the iteration with the highest posterior likelihood).

Supplementary Tables

Supplementary Table 1. Composition of baseline `outbreaker2` model and different sensitivity analyses.

Scenario type	**Onset of symptoms**	**Genetic data**	**Contact data**	**Short serial interval**	**Longer serial interval**
Baseline scenario	X	X	X	X	
Sensitivity analyses					
1.	X	X		X	
2.	X	X	X		X
3.	X	X	X[a]	X	
4.	X	X	X[b]	X	

[a] For this model, we used the HR data for healthcare workers (with some corrections)

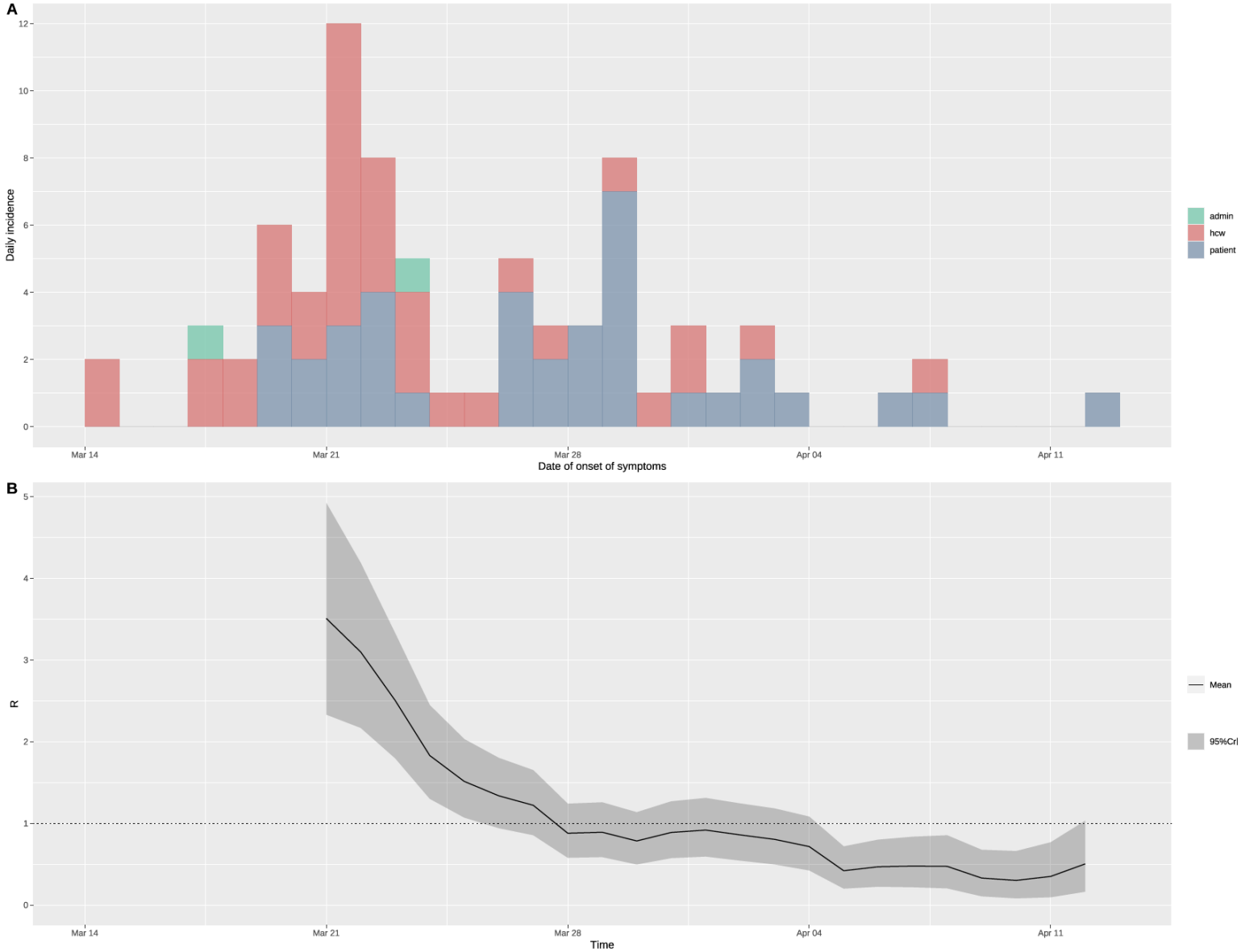
[b] For this model, we assumed that patients were no longer infectious after the date of positive RT-PCR

Supplementary Table 2. Proportions of infections attributed to HCWs (f_{HCW}) derived from a sample of 999 reconstructed transmission trees, for main analysis and all sensitivity analyses.
The Wilcoxon-Mann-Witney test compares the reported mean f_{HCW} to random expectations given the proportion of HCWs cases (46·8%).

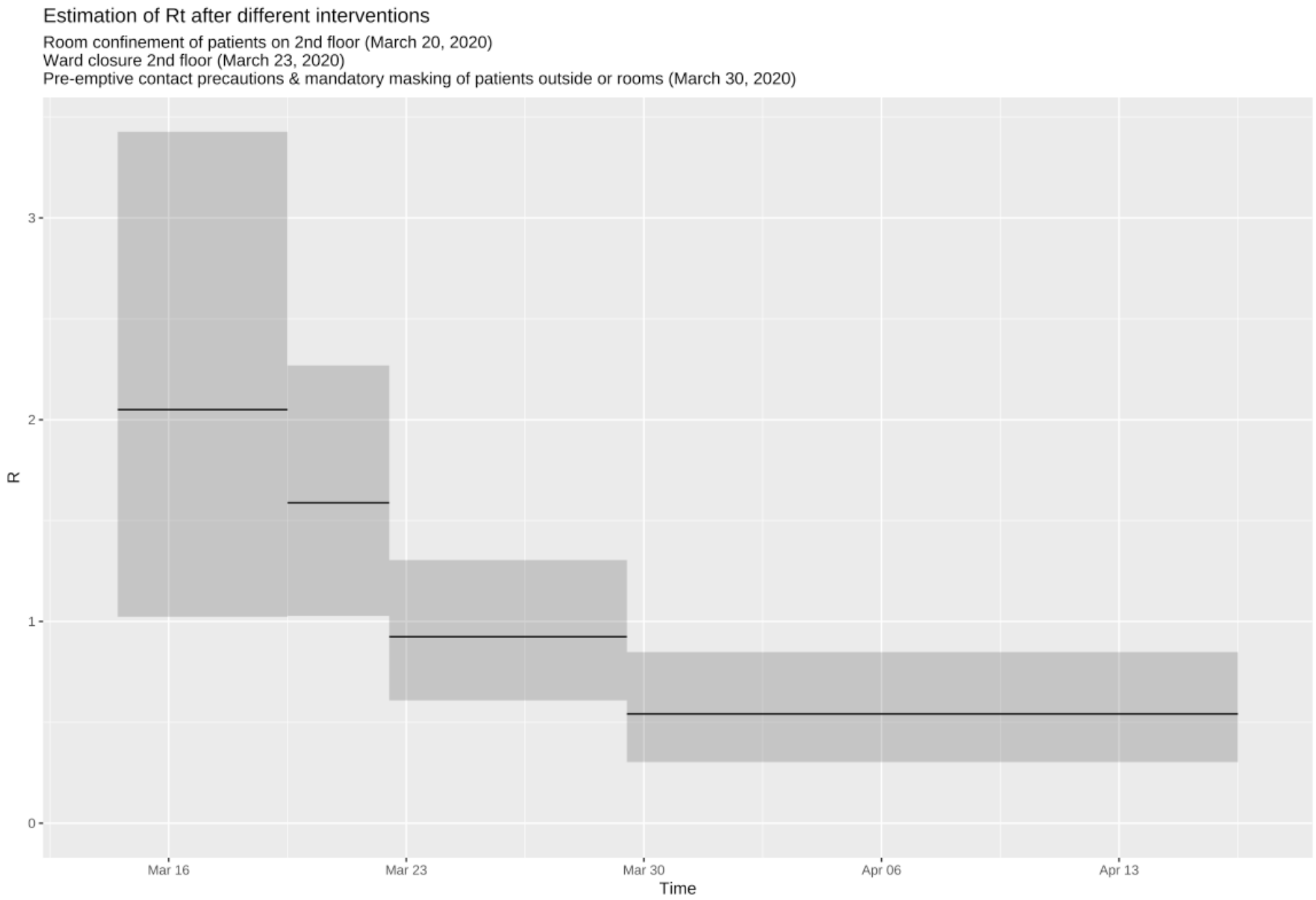
Type of infected case	**Mean f_{HCW}**	**95% CI**	**Relative excess of HCW infections**	**Wilcoxon-Mann-Whitney test**	**p-value**
Main analysis					
All cases	70·7%	70·4% - 71·2%	+ 51·3%	V = 499500	p < 2·2e-16
Only HCWs	61·1%	60·5% - 61·7%	+ 30·6%	V = 492607	p < 2·2e-16
Only patients	79·0%	78·5% - 79·5%	+ 68·5%	V = 499500	p < 2·2e-16
Only frail patients	82·3%	81·8% - 82·9%	+ 76·1%	V = 499495	p < 2·2e-16
Sensitivity analysis 1					
All cases	54·9%	54·4% - 55·3%	+ 17·3%	V = 480614	p < 2·2e-16
Only HCWs	55·0%	54·4% - 55·6%	+ 17·6%	V = 444150	p < 2·2e-16
Only patients	54·6%	54·0% - 55·1%	+ 16·6%	V = 458552	p < 2·2e-16
Only frail patients	54·7%	54·1% - 55·4%	+ 17·0%	V = 434052	p < 2·2e-16
Sensitivity analysis 2					
All cases	72·6%	72·2% - 72·9%	+ 55·2%	V = 499478	p < 2·2e-16
Only HCWs	63·0%	62·4% - 63·6%	+ 34·6%	V = 494900	p < 2·2e-16
Only patients	80·8%	80·3% - 81·3%	+ 72·7%	V = 499486	p < 2·2e-16
Only frail patients	84·2%	83·6% - 84·7%	+ 79·9%	V = 499491	p < 2·2e-16
Sensitivity analysis 3					
All cases	71·8%	71·4% - 72·1%	+ 53·4%	V = 499500	p < 2·2e-16
Only HCWs	63·4%	62·9% - 63·9%	+ 35·6%	V = 498370	p < 2·2e-16
Only patients	78·8%	78·4% - 79·2%	+ 68·5%	V = 499500	p < 2·2e-16
Only frail patients	82·9%	82·4% - 83·3%	+ 77·1%	V = 499500	p < 2·2e-16
Sensitivity analysis 4					
All cases	68·6%	68·2% - 69·1%	+ 46·7%	V = 499495	p < 2·2e-16
Only HCWs	61·5%	60·9% - 62·1%	+ 31·5%	V = 492946	p < 2·2e-16
Only patients	74·6%	74·0% - 75·2%	+ 59·5%	V = 499484	p < 2·2e-16
Only frail patients	77·4%	76·7% - 78·1%	+ 65.6%	V = 499365	p < 2·2e-16

Supplementary Figures

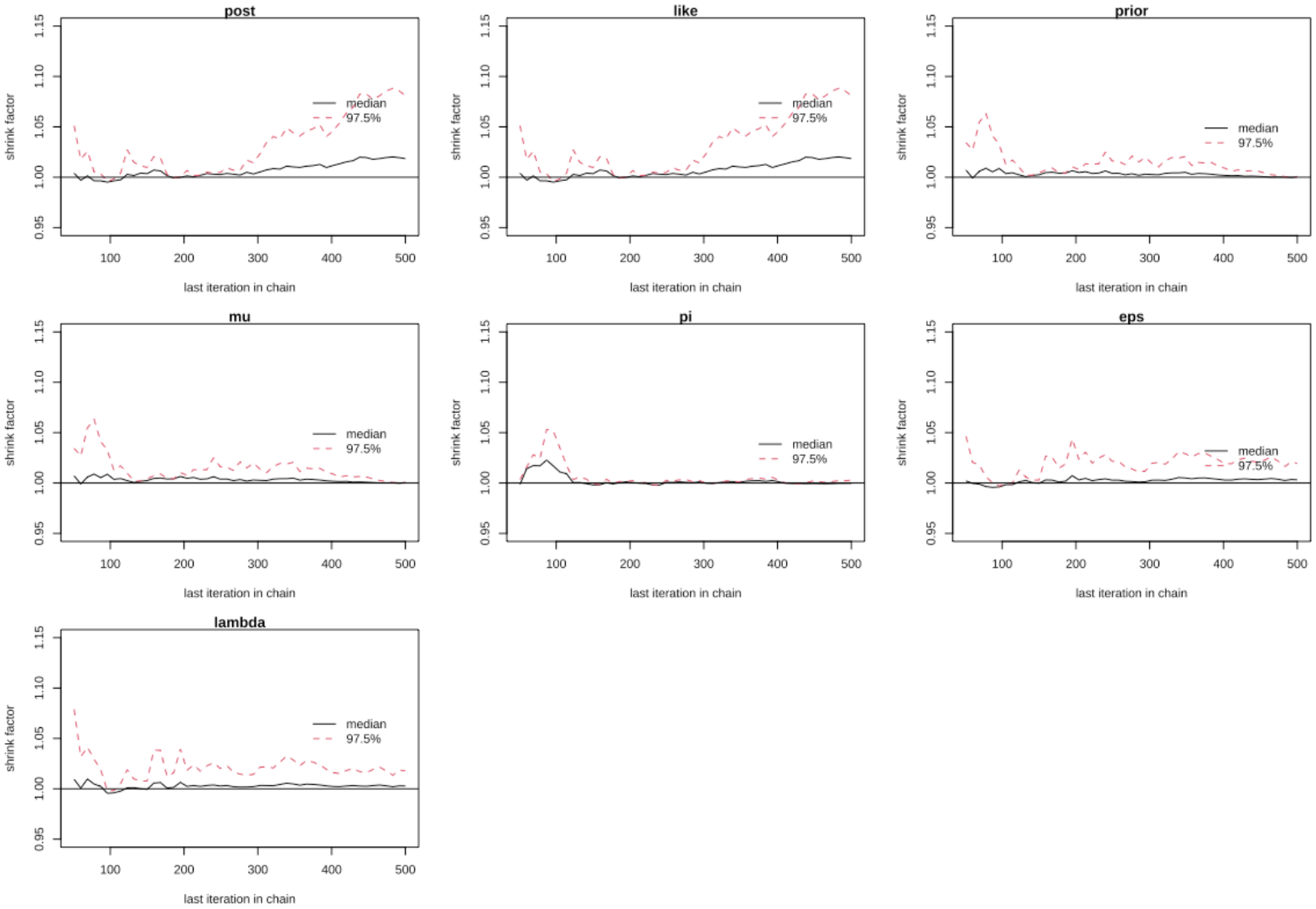
Supplementary Figure 1. Sensitivity analysis of time-varying reproduction number R_t.
Sensitivity analysis of estimated time-varying reproduction number Rt (panel B) using a longer serial interval (mean 5.2 days, SD 4.7) [9]. The epidemic curve (panel A) is shown above for ease of interpretation.



Supplementary Figure 2. Change in time-varying reproduction number after interventions.



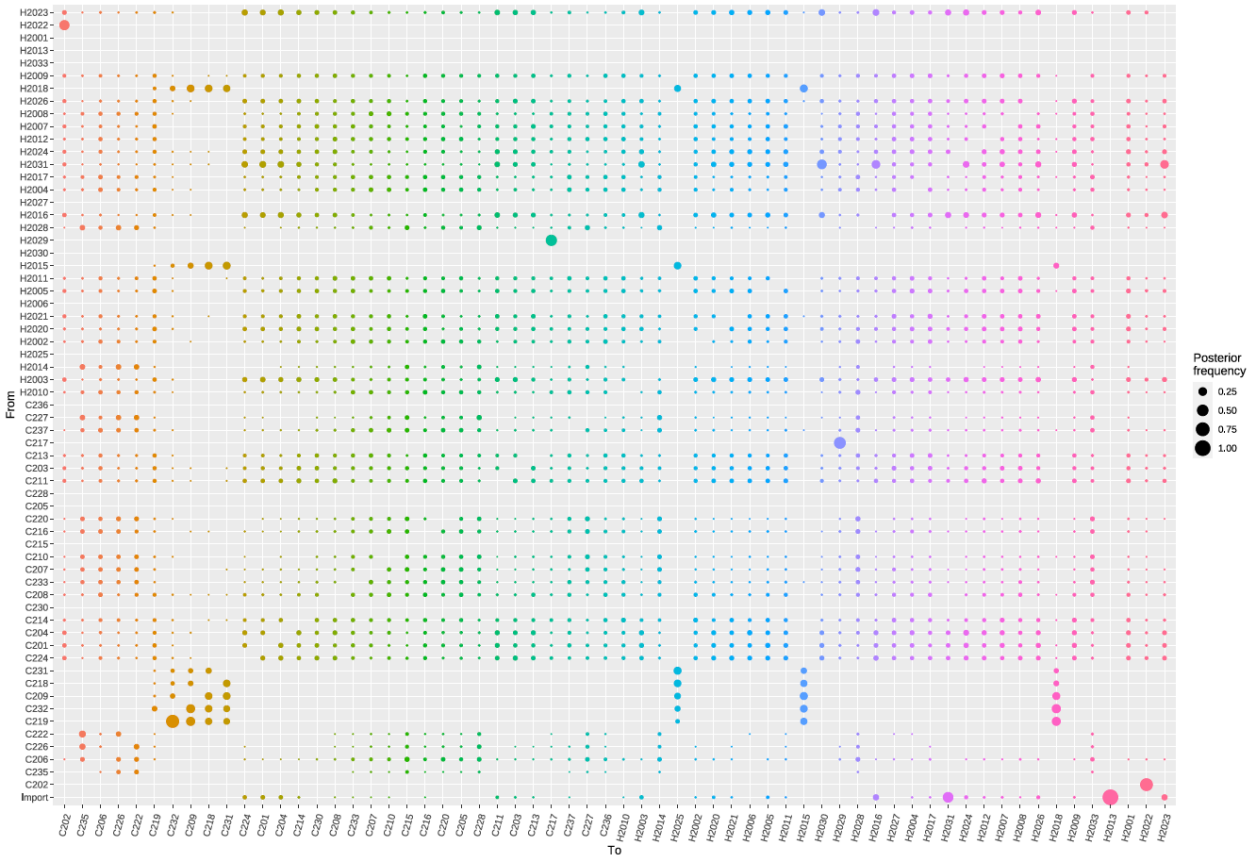
Supplementary Figure 3. Gelman-Rubin diagnostic test of convergence for the Markov-Chain Monte-Carlo iterations.
The values of the upper limit of the 95% confidence interval of the potential scale reduction factors were all <1.1, indicating good convergence.



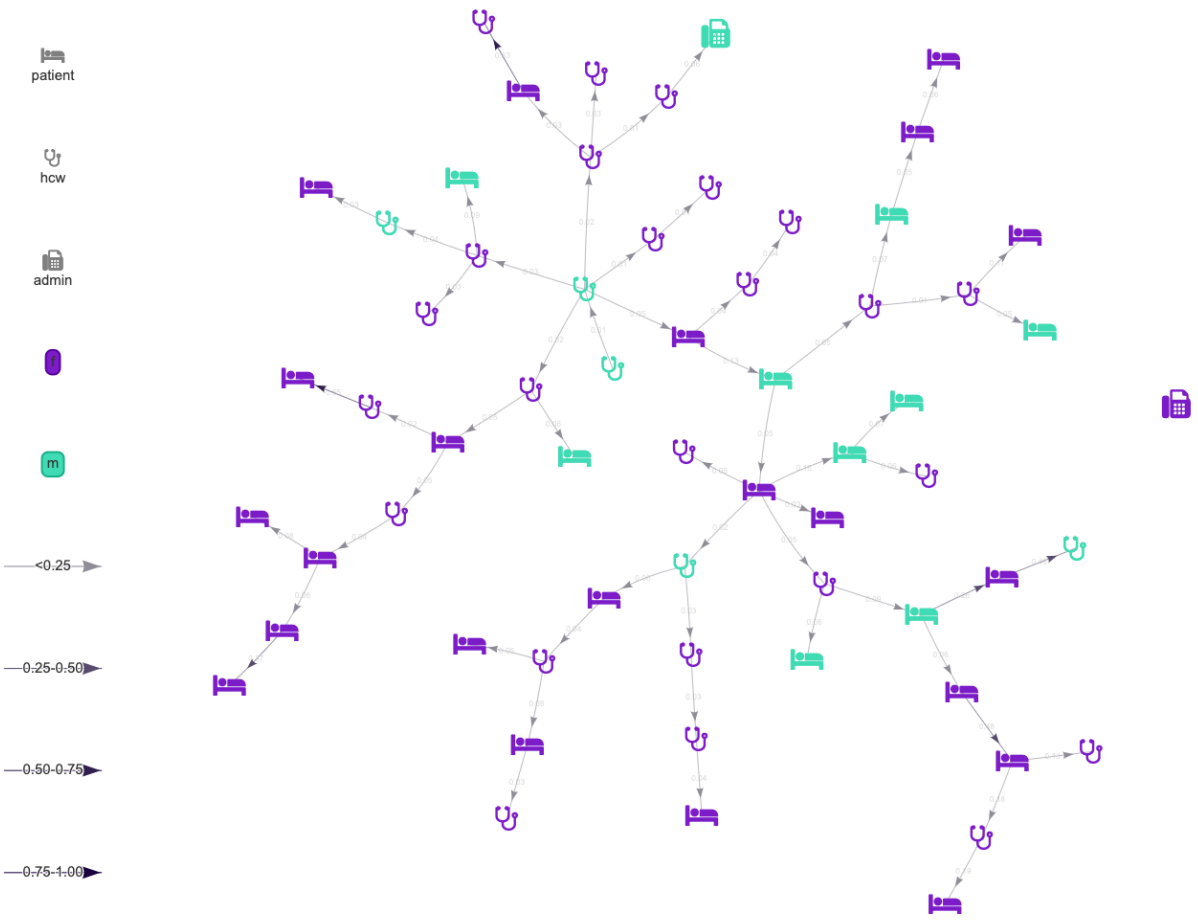
post, log-posterior values; like, log-likelihood values; prior, log-value of the prior; pi, reporting probability; mu, mutation rate; pi, reporting probability, eps, contact reporting coverage; lambda, non-infectious contact rate

Supplementary Figure 4. Sensitivity analysis of outbreaker model with absence of contact data.
A. ancestry reconstruction, B. transmission tree from Markov-Chain Monte-Carlo iteration with highest likelihood.

A.

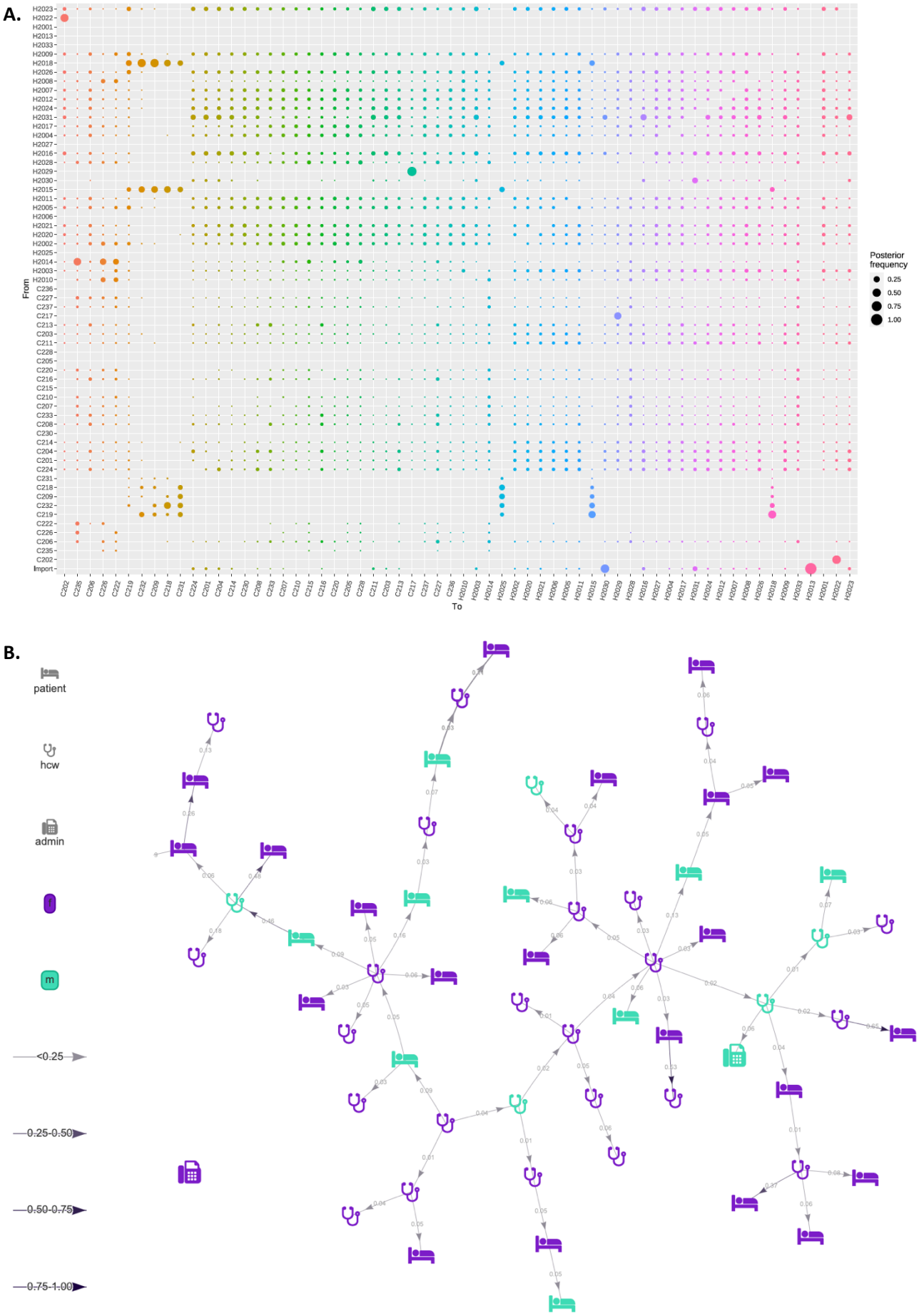


B.

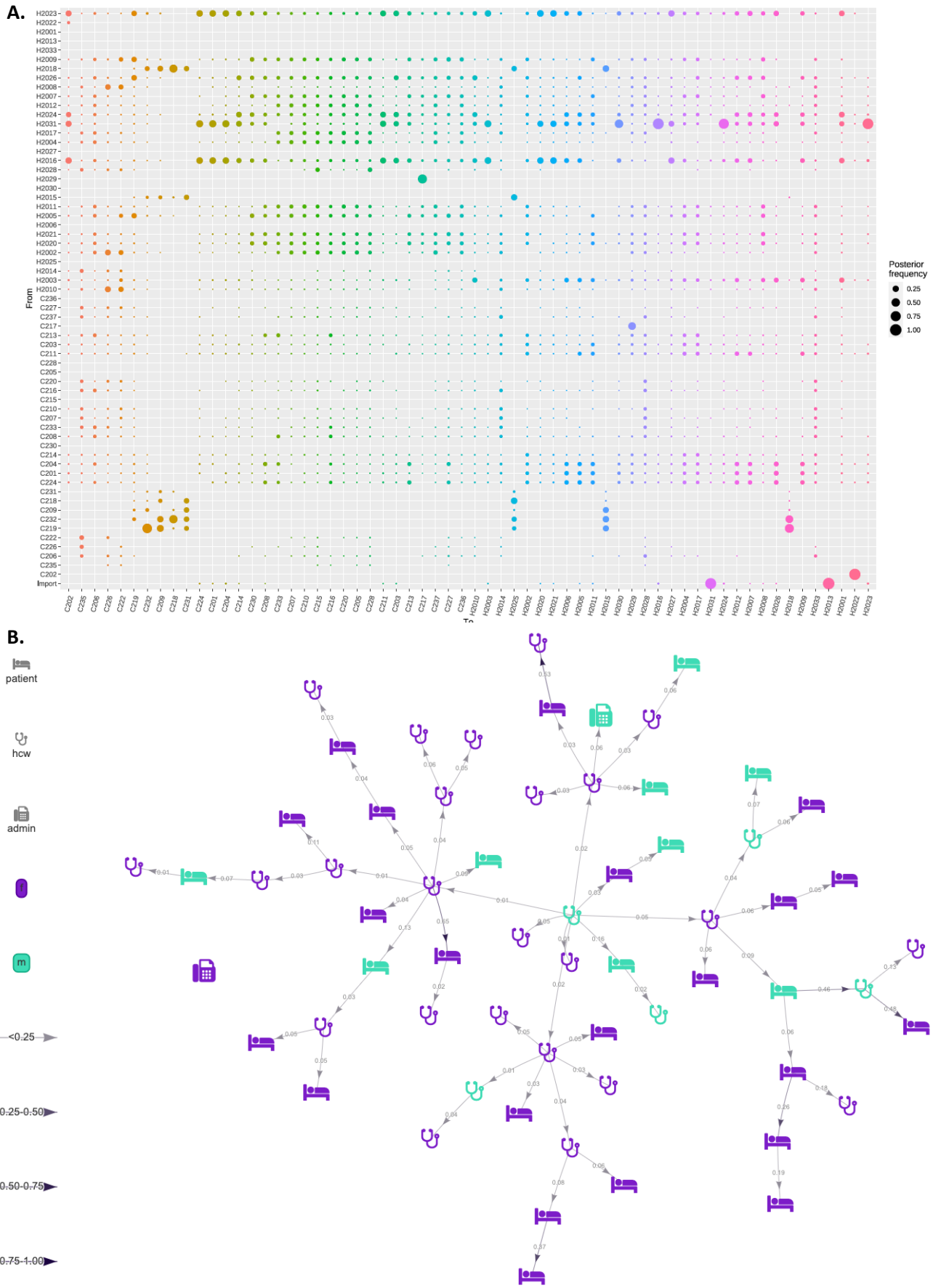


Supplementary Figure 5. Sensitivity analysis of outbreaker model with longer serial interval (mean 5.2 days, SD 4.7) [9].

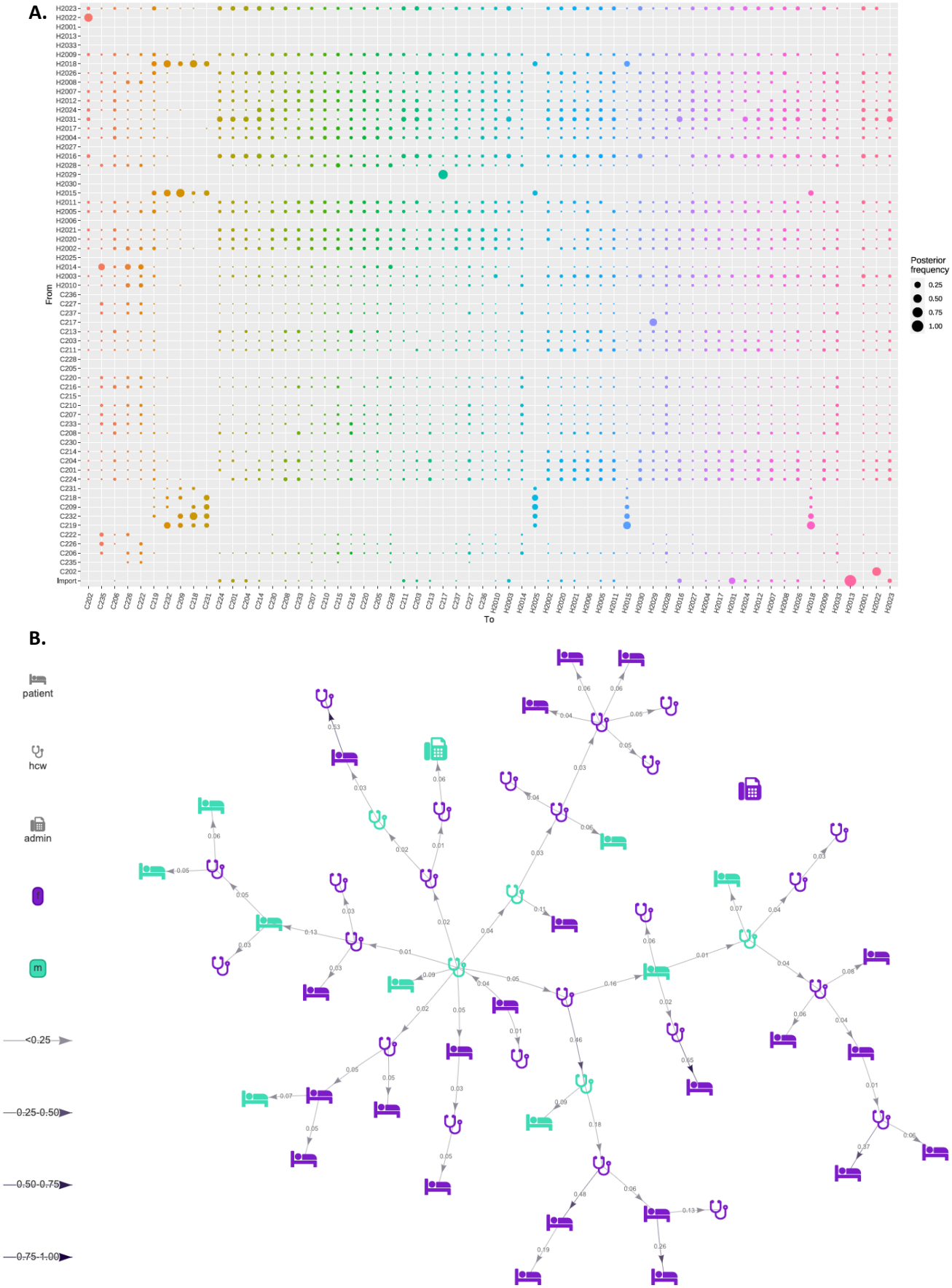
A. ancestry reconstruction, B. transmission tree from Markov-Chain Monte-Carlo iteration with highest likelihood.



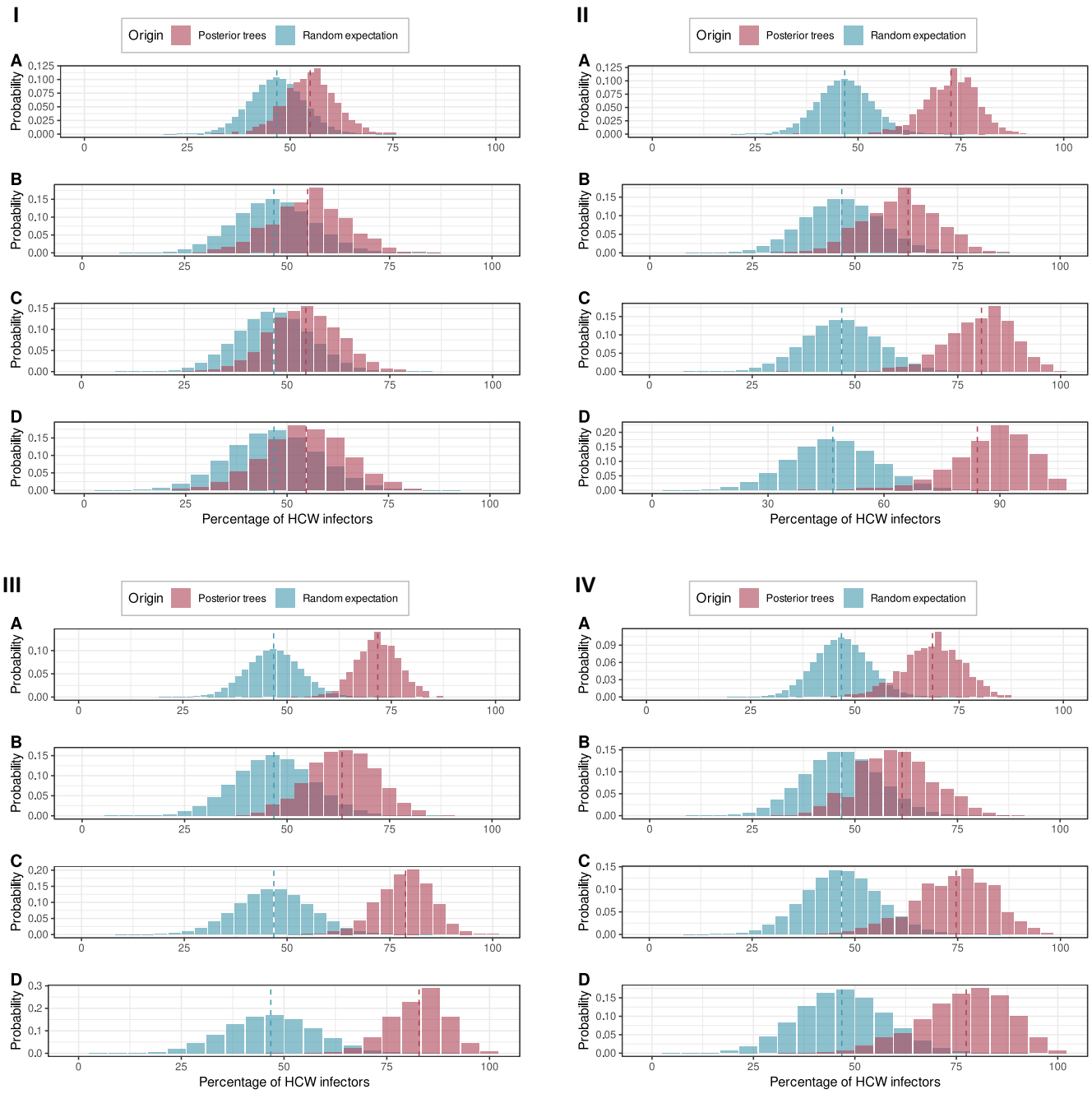
Supplementary Figure 6. Sensitivity analysis of outbreaker model where contacts were based on human resources data for HCWs and on infectious and susceptible periods.
A. ancestry reconstruction, B. transmission tree from Markov-Chain Monte-Carlo iteration with highest likelihood.



Supplementary Figure 7. Sensitivity analysis of outbreaker model where contacts were based on human resources data for HCWs and on infectious and susceptible periods.
A. ancestry reconstruction, B. transmission tree from Markov-Chain Monte-Carlo iteration with highest likelihood.



Supplementary Figure 8. Proportions of transmissions attributed to HCWs (f_{HCW}) for each sensitivity analysis (I-IV).
The blue histograms indicate the expected Binomial distributions of $f_{HCW,}$ given the proportion of HCWs amongst cases. The red histograms show the distribution of $f_{HCW,}$ across 999 transmission trees reconstructed by *outbreaker2*. Dotted lines indicate the mean estimate of the proportion. **A**. All cases. **B.** Transmission to HCWs only. **C.** Transmission to patients only. **D.** Transmission to frail patients only.



References

1. Petty TJ, Cordey S, Padioleau I, Docquier M, Turin L, Preynat-Seauve O, et al. Comprehensive human virus screening using high-throughput sequencing with a user-friendly representation of bioinformatics analysis: a pilot study. Journal of clinical microbiology. 2014;52(9):3351-61.
2. Fernandes JF, Laubscher F, Held J, Eckerle I, Docquier M, Grobusch MP, et al. Unbiased metagenomic next-generation sequencing of blood from hospitalized febrile children in Gabon. Taylor and Francis Ltd.; 2020. p. 1242-4.
3. Gleizes A, Laubscher F, Guex N, Iseli C, Junier T, Cordey S, et al. Virosaurus A Reference to Explore and Capture Virus Genetic Diversity. Viruses. 2020;12(11).
4. Kumar S, Stecher G, Li M, Knyaz C, Tamura K. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. Mol Biol Evol. 2018;35(6):1547-9.
5. Tamura K. Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G+C-content biases. Mol Biol Evol. 1992;9(4):678-87.
6. Jombart T, Cori A, Didelot X, Cauchemez S, Fraser C, Ferguson N. Bayesian reconstruction of disease outbreaks by combining epidemiologic and genomic data. PLoS Comput Biol. 2014;10(1):e1003457.
7. Campbell F, Didelot X, Fitzjohn R, Ferguson N, Cori A, Jombart T. outbreaker2: a modular platform for outbreak reconstruction. BMC Bioinformatics. 2018;19(Suppl 11):363.
8. Ali ST, Wang L, Lau EHY, Xu XK, Du Z, Wu Y, et al. Serial interval of SARS-CoV-2 was shortened over time by nonpharmaceutical interventions. Science. 2020;369(6507):1106-9.
9. Bi Q, Wu Y, Mei S, Ye C, Zou X, Zhang Z, et al. Epidemiology and transmission of COVID-19 in 391 cases and 1286 of their close contacts in Shenzhen, China: a retrospective cohort study. Lancet Infect Dis. 2020;20(8):911-9.
10. He X, Lau EHY, Wu P, Deng X, Wang J, Hao X, et al. Temporal dynamics in viral shedding and transmissibility of COVID-19. Nat Med. 2020;26(5):672-5.
11. Campbell F, Cori A, Ferguson N, Jombart T. Bayesian inference of transmission chains using timing of symptoms, pathogen genomes and contact data. PLoS Comput Biol. 2019;15(3):e1006930.

12. Gelman A, Rubin DB. Inference from Iterative Simulation Using Multiple Sequences. Statistical Science. 1992;7(4):457-72.