PLOS ONE

RESEARCH ARTICLE

Sorting out assortativity: When can we assess the contributions of different population groups to epidemic transmission?

Cyril Geismar[1,2]*, **Peter J. White**[1,2,3], **Anne Cori**[1,4☯], **Thibaut Jombart**[1,2☯]

1 MRC Centre for Global Infectious Disease Analysis, Imperial College School of Public Health, London, United Kingdom, **2** NIHR Health Protection Research Unit in Modelling and Health Economics, Imperial College School of Public Health, London, United Kingdom, **3** Modelling & Economics Unit, UK Health Security Agency, London, United Kingdom, **4** Abdul Latif Jameel Institute for Disease and Emergency Analytics, Imperial College School of Public Health, London, United Kingdom

☯ These authors contributed equally to this work.
* c.geismar21@imperial.ac.uk

Citation: Geismar C, White PJ, Cori A, Jombart T (2024) Sorting out assortativity: When can we assess the contributions of different population groups to epidemic transmission? PLoS ONE 19(12): e0313037. https://doi.org/10.1371/journal.pone.0313037

Editor: Pablo Martin Rodriguez, Federal University of Pernambuco: Universidade Federal de Pernambuco, BRAZIL

Received: August 16, 2024

Accepted: October 14, 2024

Published: December 2, 2024

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pone.0313037

Data Availability Statement: The R package developed to estimate transmission assortativity from transmission chain data can be accessed at:

Abstract

Characterising the transmission dynamics between various population groups is critical for implementing effective outbreak control measures whilst minimising financial costs and societal disruption. While recent technological and methodological advances have made individual-level transmission chain data increasingly available, it remains unclear how effectively this data can inform group-level transmission patterns, particularly in small, rapidly saturating outbreak settings. We introduce a novel framework that leverages transmission chain data to estimate group transmission assortativity; this quantifies the extent to which individuals transmit within their own group compared to others. Through extensive simulations mimicking nosocomial outbreaks, we assessed the conditions under which our estimator performs effectively and established guidelines for minimal data requirements in small outbreak settings where saturation may occur rapidly. Notably, we demonstrate that detecting and quantifying transmission assortativity is most reliable when at least 30 cases have been observed in each group, before reaching their respective epidemic peaks.

Introduction

Understanding the heterogeneous contributions of population groups to disease transmission is crucial for developing effective targeted interventions whilst minimising financial costs and societal disruption. Individuals can be categorised by age [1–3], occupation [4], vaccination status [5], sexual preferences [6–8] and other characteristics relevant to the disease context [9, 10]. These group dynamics are not only determined by distinct contact patterns, as revealed by large-scale contact surveys studies, such as POLYMOD [11] and CoMix [12], but also by varying infectiousness and susceptibility levels [13, 14]. Heterogeneous transmission is particularly salient in healthcare settings, where the confined hospital environment and frequent interactions between healthcare workers (HCWs) and vulnerable patients in various wards [15] create a complex transmission landscape, potentially increasing the risk of infection compared to the general population [16, 17]. Nosocomial outbreaks not only pose significant risks to global

https://github.com/CyGei/linktree. Additionally, an R package for simulating outbreaks involving multiple groups with various transmission assortativity coefficients is available at: https://github.com/CyGei/o2groups. The code used for the analysis presented in this manuscript can be found at: https://github.com/CyGei/o2groups-analysis. Package and analysis code have been archived on Zenodo (analysis: https://zenodo.org/doi/10.5281/zenodo.10616176, package: https://zenodo.org/doi/10.5281/zenodo.10616155).

Funding: CG is supported by a PhD studentship at Imperial College London funded by the National Institute for Health Research (NIHR) Health Protection Research Unit (HPRU) in Modelling and Health Economics, which is a partnership between the UK Health Security Agency (UKHSA), Imperial College London, and the London School of Hygiene & Tropical Medicine (grant code NIHR200908). AC, PJW, TJ are supported by the HPRU in Modelling and Health Economics. This work was supported by the UK Medical Research Council (MRC) Centre for Global Infectious Disease Analysis (grant number MR/X020258/1); this award comes under the Global Health EDCTP3 Joint Undertaking. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

healthcare systems but also worsen patient outcomes and mortality while straining hospital resources and operational capacities [18]. Between March and July 2020, Evans *et al*. estimated that nosocomial transmission in the UK accounted for 20% and 73% of SARS-CoV-2 infections amongst inpatients and HCWs, respectively [19]. Cooper *et al*. further estimated from June 2020 to March 2021 that 1–2% of hospital admissions in England likely acquired SARS--CoV-2 while hospitalised, primarily driven by patient-to-patient transmission [18].

To characterise these group-specific transmission dynamics, modellers have traditionally relied on contact survey data [11, 12], combined with information about the relative infectiousness and/or susceptibility of each group (*e.g*. obtained from epidemiological or serological investigations) [13, 14]. However, this survey data can be biased, have limited sample size or representativeness, and may not be generalisable across different epidemic contexts [20].

Worby *et al*. introduced the Relative Risk (RR) statistic as an alternative approach to evaluate group contributions to epidemic transmission [1, 2]. This metric compares the proportion of cases attributed to a particular group before and after the epidemic peak relative to the total number of cases, offering insights into the group's relative depletion of susceptibles during the epidemic's ascent. While the RR statistic may be useful to determine which group to prioritise for vaccination in large outbreaks with synchronous peaks, it has significant limitations in smaller settings such as nosocomial outbreaks involving rapidly spreading respiratory pathogens such as coronaviruses, influenza, respiratory syncytial virus, or rhinovirus. These outbreaks often feature a small number of cases, variations in group sizes, numbers of imports, contacts and transmission patterns, resulting in asynchronous epidemic peaks. In these settings, Worby's premise that the depletion of susceptible individuals within a group reflects its role in driving the epidemic may not hold.

Given the constraints of conventional methods, innovative approaches are essential to address the challenges posed by nosocomial outbreaks. Healthcare settings are particularly suited for employing advanced outbreak reconstruction tools due to the relatively small size of nosocomial outbreaks, regular surveillance and data collection, and access to whole genome sequencing [21–26]. Such tools typically leverage pathogen genetic sequence data, symptom onset or test collection dates, and contact data within a Bayesian framework to probabilistically reconstruct transmission events, generating posterior sets of transmission chains [27–29]. Research has shown that a significant proportion of HCW SARS-CoV-2 infections are often attributable to their colleagues [21–25], whereas Cook *et al*. found that patient-to-patient and patient-to-HCW transmissions were comparatively more common [26]. However, these studies typically report the proportion of specific transmission types relative to the total number of transmissions (*e.g*., 70% of transmissions were patient-to-patient). Without comparing these observations to expected frequencies, these approaches do not elucidate the underlying transmission dynamics that drive the outbreak.

To address this limitation, Abbas *et al*. developed a statistical test to detect non-random group transmission patterns using the outputs of Bayesian outbreak reconstruction tools [21, 23]. From the reconstructed chains of transmission, the authors estimated the proportion of infections caused by each case type and compared it to an expected proportion based on that type's prevalence amongst all cases [21, 23]. In a nosocomial SARS-COV-2 outbreak at a rehabilitation clinic, they identified that HCWs transmitted more frequently than expected [21]. However, the method's reliance on case prevalence alone neglects the process from exposure to infection and overlooks factors that shape the observed epidemic. Varying contact rates, susceptibility, and distinct mixing patterns can lead to a disproportionate number of a given case type from the outset, making prevalence-based expectations potentially misleading.

Furthermore, existing methods often fail to account for the rapid depletion of susceptibles typical in small outbreaks, such as those in healthcare settings. This saturation effect can

significantly alter transmission patterns, as a saturated group cannot sustain further transmission within the group. Thus analysing proportions of transmission types across an entire outbreak, beyond the point of saturation, may not accurately reflect the underlying baseline transmission patterns.

This paper introduces a novel framework for evaluating transmission patterns amongst distinct groups during an outbreak, addressing the limitations of previous methods. Our approach quantifies group-specific transmission assortativity, from known or probabilistically reconstructed transmission chains, while accounting for group sizes. We evaluate the performance of our estimator through diverse simulations, mimicking nosocomial outbreaks where the populations are fully susceptible at the outbreak's onset. Our aim is to provide guidelines on the minimum data collection requirements and the optimal estimation timeframe, thereby informing IPC strategies in small outbreaks where rapid saturation occurs.

Methods

A new estimator of transmission assortativity

Assortativity has been amply described for social mixing patterns, with homogeneous mixing referring to random contacts between individuals, and heterogeneous mixing denoting interactions characterised by distinct (non-random) patterns depending on group memberships [13, 30]. Heterogeneous mixing can be either *assortative*, where individuals tend to interact more within their own group (*e.g.* social contacts by age [11, 31, 32]), or *disassortative*, where individuals interact preferentially with members of other groups (*e.g.* sexual contacts [33]). Here we use these definitions to characterise the patterns of transmission rather than contact. The resulting transmission patterns thus reflect not only mixing patterns but also differences in infectiousness and susceptibility amongst groups.

To quantify transmission assortativity, we examine the person-to-person transmission patterns. We consider G groups of relative sizes $f_1,\ldots,f_G$ defined as:

$$f_a = \frac{N_a}{\sum_{g=1}^{G} N_g} \quad \forall a = 1,\ldots,G \tag{1}$$

where N_a is the number of individuals in group a. We denote $\beta_{b\leftarrow a}$ the person-to-person transmission rate from an individual in group a to an individual in group b, that is the force of infection that any one infected individual in group a exerted on any one susceptible individual in group b. It follows that, in a fully susceptible population, the expected number of secondary cases in group b generated by one infectious individual in group a is proportional to $\beta_{b\leftarrow a} N_b \propto \beta_{b\leftarrow a} f_b$.

We make the following assumptions:

1. $\beta_{b\leftarrow a}$ is the same for all $b \neq a$, *i.e.*, $\beta_{b\leftarrow a} = \psi$ if $a \neq b$ (S3 in S1 File).
2. $\beta_{a\leftarrow a} = \gamma_a \psi$, where γ_a is the assortativity coefficient for group a.

The assortativity coefficient, γ_a, is defined as the excess probability of a secondary infection taking place within group a compared to random expectation. γ values range from 0 (fully disassortative, *i.e.* no within-group transmissions) to ∞ (fully assortative, *i.e.* transmissions occur exclusively within the group), with 1 indicating homogeneous patterns. For instance, $\gamma_a = 2$ indicates that an infected individual from group a is twice as likely to infect an individual from the same group compared to infecting an individual from another group. Conversely, a γ_a of 1/2 means that an infected individual from group a is twice as likely to infect an individual from another group compared to infecting an individual from the same group.

We derive $\pi_{b\leftarrow a}$, the proportion of secondary cases in group b amongst those generated by an infectious individual in group a, where $a \neq b$, as:

$$\pi_{b\leftarrow a} = \frac{\beta_{b\leftarrow a} f_b}{\sum_{g=1}^{G} \beta_{g\leftarrow a} f_g} = \frac{\beta_{b\leftarrow a} f_b}{\sum_{\substack{g=1 \\ g\neq a}}^{G} \beta_{g\leftarrow a} f_g + \beta_{a\leftarrow a} f_a} = \frac{\beta_{b\leftarrow a} f_b}{\psi(1-f_a) + \gamma_a \psi f_a} = \frac{\psi f_b}{\psi(1-f_a) + \gamma_a \psi f_a}$$

$$= \frac{f_b}{(1-f_a) + \gamma_a f_a} \tag{2}$$

We derive $\pi_{a\leftarrow a}$, the proportion of secondary cases in group a amongst those generated by an infectious individual from the same group, as:

$$\pi_{a\leftarrow a} = \frac{\beta_{a\leftarrow a} f_a}{\psi(1-f_a) + \gamma_a \psi f_a} = \frac{\gamma_a \psi f_a}{\psi(1-f_a) + \gamma_a \psi f_a} = \frac{\gamma_a f_a}{(1-f_a) + \gamma_a f_a} \tag{3}$$

We can obtain γ_a by rewriting Eq 3 as:

$$\gamma_a = \frac{\pi_{a\leftarrow a} \cdot (1-f_a)}{f_a \cdot (1-\pi_{a\leftarrow a})} \tag{4}$$

The proportion of within-group transmission, $\pi_{a\leftarrow a}$ (Eq 3), can be directly calculated from known transmission chain data. It is calculated by dividing the number of observed within-group transmission pairs, $\tau_{a\leftarrow a}$, by the total number of transmissions originating from group a, $\tau_{\cdot\leftarrow a}$. Thus the proportion, $\pi_{a\leftarrow a} = \tau_{a\leftarrow a}/\tau_{\cdot\leftarrow a}$, ranges between 0 and 1 (included).

To simplify interpretation, we introduce a rescaled parameter δ, ranging between -1 (fully disassortative) and 1 (fully assortative), with 0 corresponding to a homogeneous transmission pattern (Fig 1.1 in S1 File) such that:

$$\delta = \begin{cases} 1 & \text{if} \gamma = \infty \\ \dfrac{\gamma - 1}{\gamma + 1} & \text{if} \gamma \neq \infty \end{cases} \tag{5}$$

The formula for δ_a can thus be written as:

$$\delta_a = \frac{\frac{\tau_{a\leftarrow a}}{\tau_{\cdot\leftarrow a}} - f_a}{\frac{\tau_{a\leftarrow a}}{\tau_{\cdot\leftarrow a}} + f_a\left(1 - 2\frac{\tau_{a\leftarrow a}}{\tau_{\cdot\leftarrow a}}\right)} \tag{6}$$

where:

- $\tau_{a\leftarrow a}$: represents the number of transmissions from group a towards group a.
- $\tau_{\cdot\leftarrow a}$: refers to the total number of transmissions emitted by group a.
- $\tau_{a\leftarrow a}/\tau_{\cdot\leftarrow a}$: Denotes the proportion of within-group transmissions for group a denoted as $\pi_{a\leftarrow a}$ in Eq 3.
- f_a: Represents the proportion of the total population that belongs to group a. It is a value between 0 and 1, exclusive of the endpoints as there must be more than 1 group in the population.

The relationship between δ_a, $\pi_{a\leftarrow a}$ (*i.e.* $\tau_{a\leftarrow a}/\tau_{\cdot\leftarrow a}$), and f_a can is visually represented in supplementary Fig 1.2 in S1 File. We can obtain a confidence interval (CI) on $\pi_{a\leftarrow a}$ for various significance (α) levels using the Clopper-Pearson binomial interval method [34] (S1.1 in S1 File). Feeding estimates of $\pi_{a\leftarrow a}$ into Eq 6 provides estimates of δ_a with confidence intervals (S1.1 in S1 File).

All our results are presented using δ rather than γ.

Simulation study

We simulated small outbreaks under various contexts to assess the estimator's performance in scenarios relevant to person-to-person transmission of healthcare-acquired pathogens in a fully susceptible population. We constructed 10,000 sets of input parameters, referred to as 'scenarios', by randomly sampling parameters from pre-defined distributions (Section 1.2 and Fig 2 in S1 File). To account for stochasticity, we conducted 100 simulations for each unique scenario resulting in a total of 1,000,000 simulated outbreaks.

The simulation employed a discrete time branching process modelling individual infections spreading in successive generations. Simulations were specified with: i) group-level parameters including the size of each group, their assortativity coefficients (δ), initial introductions, basic reproduction numbers (R_0) and ii) epidemic level parameters such as the number of groups, the pathogen generation time (w) and incubation period distributions (both assumed the same across groups). The simulation outputs a transmission tree that includes, for each infected individual, their symptom onset date, their group affiliation, and the id of their infector. Using this data for all infected individuals with symptoms up to time t, we calculated $\tau_{a\leftarrow a}$ (the number of within-group transmission pairs) and $\tau_{\cdot\leftarrow a}$ (the total number of transmissions originating from each group). These values, along with the relative sizes of the groups (f_a), were input into Eq 6 to estimate the assortativity coefficients for each group.

In our branching process model, the force of infection (FOI) generated by individual j from group a at time t, towards the whole of group b is defined as:

$$\lambda^j_{b\leftarrow a}(t) = w(t - s^j_a)R_{0a}\pi_{b\leftarrow a} \quad \begin{matrix} \forall a, b = 1, \ldots, G \\ \forall j = 1, \ldots, N_a \end{matrix} \tag{7}$$

where:

- s^j_a is the time of infection of individual j in group a
- R_{0a} is the basic reproduction number of individuals in group a
- w is the probability mass function of the generation time distribution

The total FOI that all individuals in group b collectively receive from all individuals across all groups at time t is obtained as:

$$\lambda_b(t) = \sum_{a=1}^{G}\sum_{j=1}^{N_a}\lambda^j_{b\leftarrow a}(t) \quad \forall b = 1, \ldots, G \tag{8}$$

Hence, the FOI that one individual from group b is exposed to is $\frac{\lambda_b(t)}{N_b}$.

The probability of infection for each individual in group b at time t is then calculated as:

$$p_b(t) = 1 - e^{-\frac{\lambda_b(t)}{N_b}} \tag{9}$$

At time $t + 1$, the number of new cases in group b, $X_b(t + 1)$, is drawn from a binomial distribution:

$$X_b(t+1) \sim \text{Binom}(S_b(t), p_b(t)) \tag{10}$$

where $S_b(t)$ is the number of susceptible individuals in group b at time t.

New cases are allocated at random amongst susceptible individuals. The simulation progresses in discrete daily time steps for 365 days. Nearly all simulations (99.99%) finished with the last infection occurring before day 300. Note that we assume that individuals who have been infected become fully immune.

Assuming that b^i (i^{th} individual in group b) was infected at time t+1, their infector α_{b^i} is drawn across all infected individuals in all groups from a multinomial distribution with probabilities:

$$p(\alpha_{b^i} = a^j)(t+1) = \frac{\lambda^j_{b \leftarrow a}(t)}{\lambda_b(t)} \tag{11}$$

Where a^j is the j^{th} individual in group a.

To assess the performance of our estimator, we computed 4 different performance metrics for each scenario:

- *Bias*: defined as the average difference between the true δ value and its estimate ($\hat{\delta}$) across 100 simulations. It is a measure of the estimator's systematic error and inaccuracy and should be close to 0. Bias is positive when δ is underestimated, indicating underestimation of assortativity or overestimation of disassortativity. Conversely, negative bias occurs when δ is overestimated, indicating overestimation of assortativity or underestimation of disassortativity.
- *Coverage (at significance level α)*: defined as the proportion of simulations (out of 100) where the true δ value is within the estimated CI corresponding to α. We evaluate 4 significance levels 0.05, 0.1, 0.25 and 0.5. Assessing coverage helps determine the reliability of the confidence intervals generated by the estimator. Coverage should approximate 1-α, and the coverage error, which measures the deviation from this target, should be close to 0. A positive coverage error suggests underestimation of uncertainty, while a negative coverage error indicates overestimation.
- *Sensitivity* (true positive rate): defined as the proportion of simulations (out of 100) where the estimator correctly identifies a significant assortative or disassortative effect (*i.e.* the $\hat{\delta}$ CI doesn't contain 0). Sensitivity should be close to 1 (100%).
- *Specificity* (true negative rate): defined as the proportion of simulations (out of 100) where the estimator correctly identifies no significant assortative or disassortative effect (*i.e.* the $\hat{\delta}$ CI contains 0). Specificity should be close to 1 (100%).

We evaluated the estimator's performance at various stages of the outbreak, defined in relation to the group's epidemic peak, i.e. the day with the highest symptom onset incidence following the first case. We hypothesise that in the early stages of an outbreak, up to the group's epidemic peak, the depletion of susceptibles is not substantial enough to significantly alter transmission dynamics. Denoting T the date of the group's peak incidence, we define the *analysis time window* as the time period from the first case of the group to day T x ε, where ε represents any non-negative real number and is referred to as the "peak coefficient". A peak coefficient value of ε = 1 implies analysis until the group's peak, while values above or below 1 imply analysis using data up to before or after the peak respectively (Section 1.3 and Fig 3 in S1 File). Additionally, we introduce the term 'peak asynchronicity', calculated as the standard deviation of peak dates T across groups, to measure heterogeneity in the groups' peak dates.

To assess the impact of the scenario parameters on the performance metrics, separate regressions were conducted with each performance metric as a dependent variable and

scenario parameters as independent variables (S1.4 in S1 File). These regressions provide coefficients that quantify the impact of key parameters, while the (adjusted) R-squared statistic informs on the proportion of variance explained by the model.

Results

Fig 1 presents the estimator's performance across all epidemic scenarios considered.

Bias decreased as the analysis time window expanded, achieving near-zero levels once the group had reached its epidemic peak ($\varepsilon = 1$), with no substantial further improvements at later epidemic stages ($\varepsilon > 1$, Fig 1A).

Coverage performance was contingent upon the significance (α) level and the stage of the group's epidemic (ε) (Fig 1B). Halfway before the epidemic peak (peak coefficient $\varepsilon = 0.5$), coverage at α levels up to 25% was too low, with average errors of 0.22, 0.18 and 0.07 for α levels of 5, 10, and 25%, respectively. In contrast, the 50% coverage was too high with an average error of -0.10. Around the epidemic peak (ε 0.7–1.3), coverage for $\alpha = 5$–10% was good, whilst coverage for $\alpha = 25$–50% was too high (average error -0.14). At later epidemic stages (ε 1.5–5),

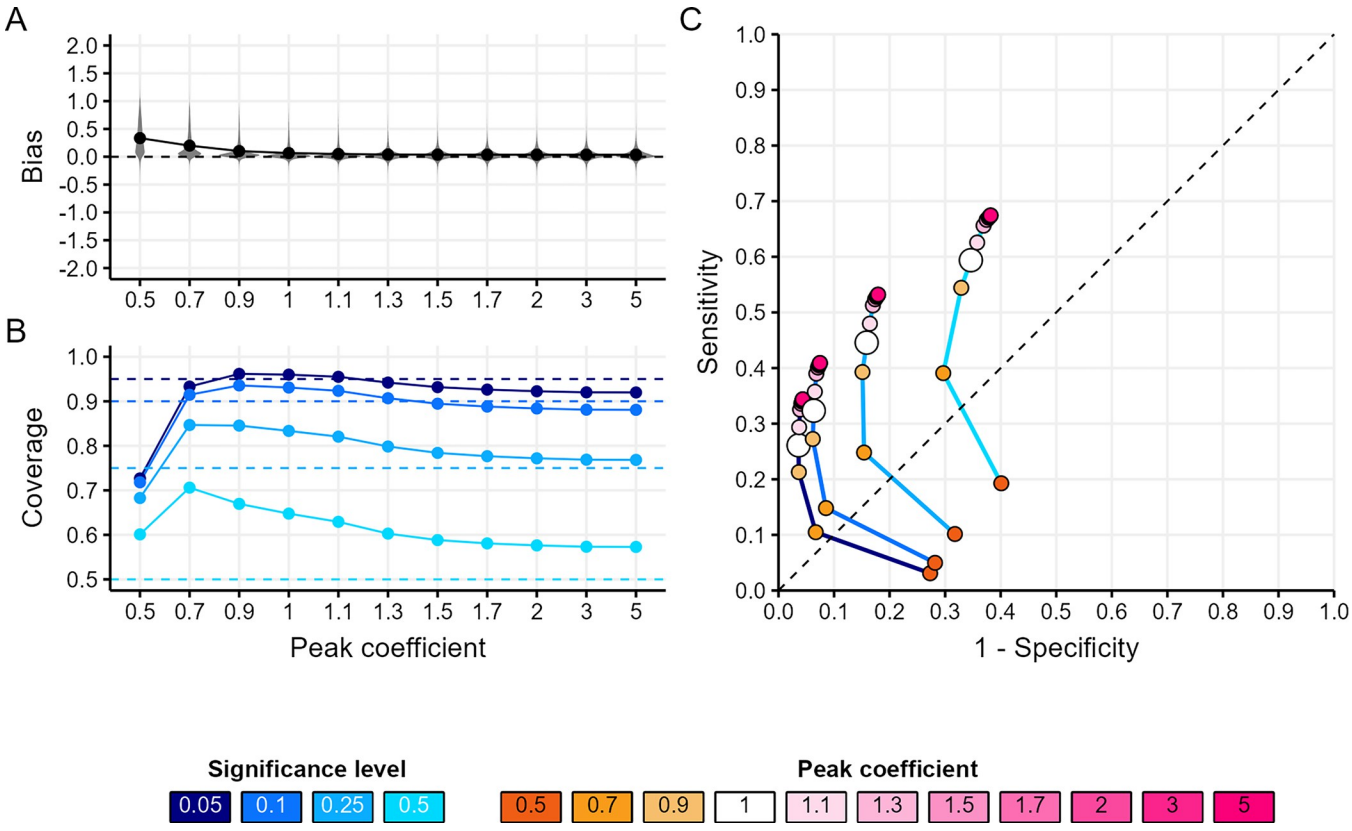

Fig 1. Estimator's performance across all epidemic scenarios. (A) Distribution of bias (the mean difference between the true assortativity δ value and its estimate) by peak coefficient. The peak coefficient (ε) is a non-negative real number used to define the *analysis time window* in relation to the group's epidemic peak. It determines the analysis period from the first case to the day $T\varepsilon$, where T is the date of peak incidence for the group. A value of $\varepsilon = 1$ indicates analysis up to the group's peak date, while values above or below 1 extend the analysis to data after or before the group's peak date, respectively. The peak coefficient serves as a proxy to inform on group-level saturation, past the peak the significant depletion of susceptibles is likely to influence the underlying baseline transmission patterns. (B) Mean coverage (proportion of simulations where the true δ value is within the estimated CI) by peak coefficient for each significance level (blue shades). (C) The Receiver Operating Characteristic (ROC) (the trade-off between sensitivity and specificity) curves by peak coefficient (orange-pink points) for each significance level (blue shaded lines). In panel (A), each point shows the mean metric value across all scenarios for a given peak coefficient. In panels (B) and (C), each point shows the mean metric value across all scenarios for a given peak coefficient and significance level. Dashed lines refer to the metric's target value for (A) and (B) and represent a random classifier's ROC performance for (C).

https://doi.org/10.1371/journal.pone.0313037.g001

coverage was good across most significance levels, although the 50% coverage remained high across all epidemic stages.

Sensitivity and specificity were contingent upon the CI significance level α and the stage of the group's epidemic (ε) (Fig 1C). Larger α values enhanced sensitivity at the expense of specificity, irrespective of the epidemic stage. And, regardless of α, analysing transmission chains later in the epidemic (*i.e.* increasing ε) also enhanced sensitivity, although this improvement was marginal past a peak coefficient of 1.5. However, the gain in sensitivity relative to the loss in specificity induced by delaying the analysis varied with α, with more pronounced tradeoffs for larger α values.

Fig 2 presents the relationship between various epidemic characteristics (columns) and the estimator's performance metrics (rows), for a peak coefficient of 1 and a significance level of 0.05. Additional configurations are shown in supplementary materials (Fig 6 in S1 File).

Our estimator maintained consistent unbiased performance across the entire assortativity range (δ from -1 to 1) (Fig 2 column A row 1). Coverage consistently met the 95% target for $\delta < 0.5$, with a slight decrease in coverage performance for $\delta > 0.5$, although coverage remained close to the target, averaging at 0.91 (sd = 0.10) (Fig 2A2). This decrease in coverage in highly assortative scenarios could be due to a saturation effect: high assortativity will accelerate the

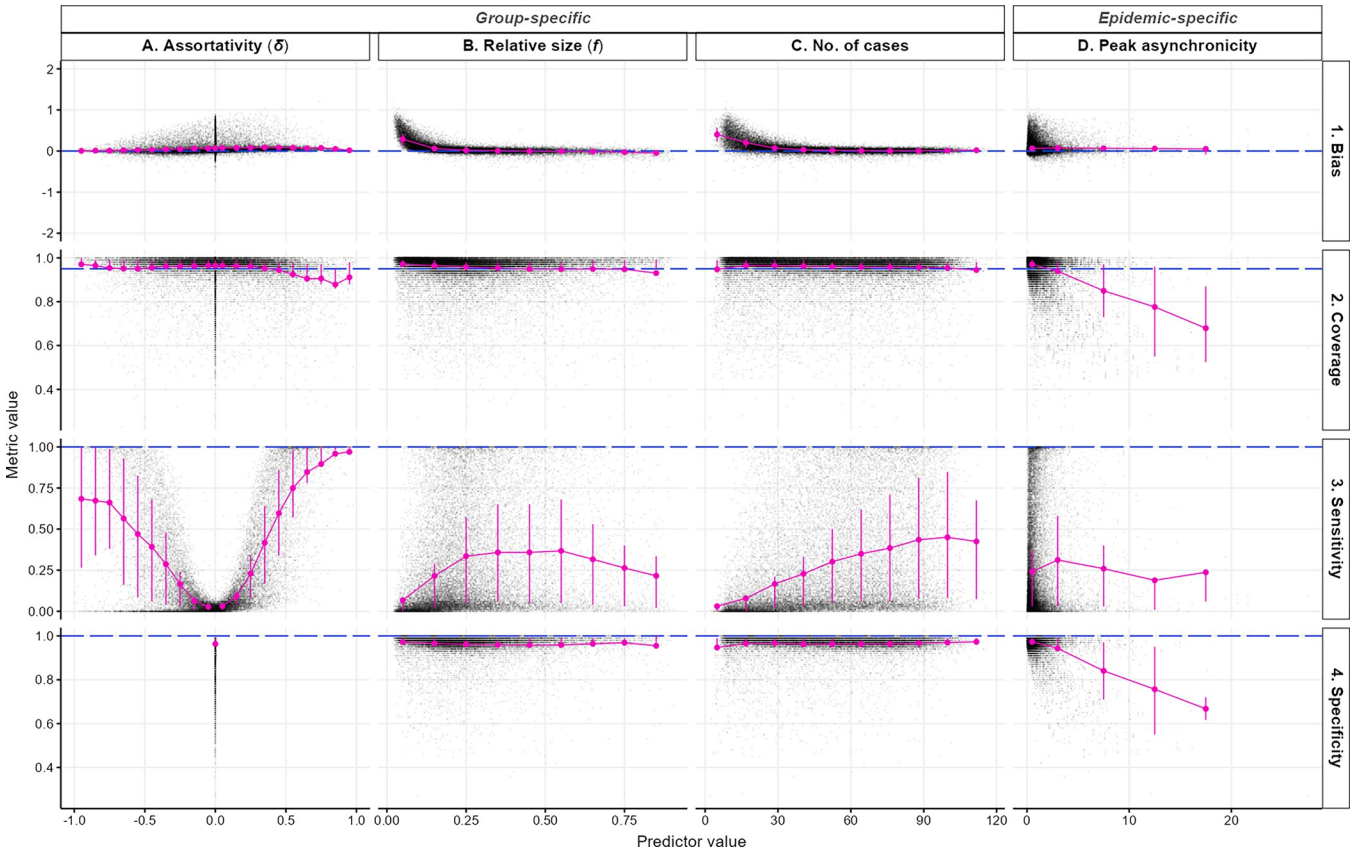

Fig 2. Estimator's performance across scenario parameters and epidemic characteristics. Each row corresponds to one performance indicator and each column corresponds to one simulation parameter or epidemic characteristic. In each panel, the scatter plot depicts the univariate relationship between simulation parameter or epidemic characteristic (x-axis) and the performance metric (y-axis), where each black dot represents the average observation from 100 simulations for each group in every scenario. The pink points and error bars indicate the mean and interquartile range, calculated across different bin widths: 0.1 for δ (A.) and relative group size (B.), 12.5 for the number of cases in the group (C.) and 5 days for the standard deviation of peak date (D.). Dashed blue lines indicate target metric values. Transmission chains were analysed up to the group's epidemic peak ($\varepsilon = 1$), with a significance level of 0.05.

https://doi.org/10.1371/journal.pone.0313037.g002

depletion of susceptibles in the group, eventually resulting in lower observed assortativity compared to the true value (Fig 4 in S1 File). Although the assortativity coefficient δ only had a small effect on bias or coverage, it had a substantial impact on sensitivity, which was higher for larger absolute values of δ. However, sensitivity rose more gradually as $|\delta|$ increased on the disassortative scale compared to the assortative scale (Fig 2A3, Table 1.1 in S1 File), reaching an average of 82% for $\delta \geq 0.5$ compared to 55% for $\delta \leq -0.5$, suggesting a better ability to detect assortative than disassortative transmission. Indeed, assortative transmission implies that transmissions propagate within the same group across multiple generations, consequently increasing the sample size ($\tau_{\cdot \leftarrow a}$ in Eq 6) compared to disassortative transmission, and thus narrowing the CI, thereby enhancing sensitivity. Our linear regression suggested that the assortativity coefficient explained nearly 60% of the variance observed in sensitivity (Table 1.1 in S1 File).

Increasing the number of cases substantially reduced bias (Fig 2C1, Table 2 in S1 File), and increased sensitivity (Fig 2C3, Table 1.2 in S1 File) but had little effect on specificity or coverage (Fig 2C4 and 2C2). Bias was negligible (mean: 0.04, sd: 0.07) once the group reached 30 to 40 cases. Sensitivity was positively correlated with the number of cases: controlling for δ, the odds of detecting an assortative or disassortative pattern increased by 4% with each additional case (Table 1.2 in S1 File).

The relative size of the group had a substantial effect on bias (Fig 2B1, Table 2 in S1 File) and sensitivity (Fig 2B3, Table 1.2 in S1 File) but no effect on specificity (Fig 2B4) nor coverage (Fig 2B2). When groups comprised 10% or more of the total population size, bias was close to 0 (Fig 2B1), and the odds of detecting an assortative pattern increased fourfold, compared to smaller groups (odds ratios (OR) = 4.15, 95% CI = 4.07–4.24) (Fig 2B3, Table 1.2 in S1 File). Relative size and the number of cases jointly accounted for 72% of the variation in bias (Table 2 in S1 File), and contributed to a 42% increase in the pseudo R-squared for the linear regression on sensitivity (from 0.566 in Table 1.1 to 0.805 in Table 1.2 in S1 File).

Diverse transmission dynamics emerge from numerous groups, varying group sizes, reproduction numbers, and/or assortativity coefficients (Fig 5 in S1 File). This diversity results in varying saturation levels between groups over time, affecting transmission patterns within and between groups. Peak asynchronicity, a measure of heterogeneity in epidemic peak timing across groups was negatively associated with coverage (OR = 0.78, 95% CI = 0.78–0.78) and specificity (OR = 0.76, 95% CI = 0.76–0.76), explaining 18% and 24% of the variance, respectively (Tables 3 and 4 in S1 File, Fig 2D2 and 2D4). These results suggest a decrease in our estimator's performance with increasing heterogeneity between groups. However, our estimates remained unbiased (Fig 2D1) and with consistent sensitivity (Fig 2D3) irrespective of that heterogeneity.

In summary, analysing transmission chains at least up to the group's epidemic peak generally improved all performance metrics. Near the group's epidemic peak, coverage with significance levels of 5 or 10% yielded good performance, while levels of 25 and 50% were a bit too high, improving after the peak. Specificity was higher at lower significance levels, while sensitivity was higher at larger significance levels. Increased cases and relative group size contributed to improved estimator accuracy, reduced bias, and heightened sensitivity, with no significant impact on coverage nor specificity. Complex epidemic settings, measured through peak asynchronicity, did not significantly affect sensitivity or bias but were associated with a reduction in coverage and specificity.

Discussion

We developed a method to detect and quantify the transmission assortativity of different groups based on transmission chains. We performed an extensive simulation study covering a

range of epidemic scenarios compatible with viral respiratory nosocomial outbreaks to assess the performance of our approach.

Our results indicate that the estimator's performance is influenced by assortativity patterns, relative group sizes, number of cases, and peak dates asynchronicity.

Generally, under the various settings considered—characterised by small group sizes and rapid saturation -, analysing transmission chains too early in the outbreak, before the group's epidemic peak, results in poor performance across all metrics considered. On the other hand, delaying assortativity coefficient estimation poses challenges for timely policy implementation. Choosing when exactly in the epidemic to analyse transmission chains, and what significance level to use for estimating the assortativity coefficients, will also depend on the objective. For instance, minimising bias and maximising sensitivity is best achieved later in the epidemic, past the group's peak, and using larger significance levels. Conversely, improving coverage and maximising specificity is easiest before the group's epidemic peak and using lower significance levels. Nevertheless, estimating assortativity at a target time before or at the peak requires accurate prediction of the group's peak date which can be very challenging.

As a rule of thumb, we suggest analysing all available transmission chain data up to the group's epidemic peak with a significance level of 0.05. Under this setting, our estimator provides a generally accurate measure of assortativity with reliable coverage and specificity albeit lower sensitivity.

Detecting non-homogeneous transmission patterns (sensitivity) in the presence of relatively small groups (*i.e.* a group constituting less than 10% of the total population), with groups having fewer than 30 cases is challenging, particularly when assortative or disassortative patterns are mild ($-0.5 \leq \delta \leq 0.5$). Importantly, it is considerably easier to detect assortativity than disassortativity, given that assortativity yields more transmission events within the group considered (where most new infections appear) compared to disassortativity (where new infections tend to appear in other groups, by definition). Hence, all other things being equal, larger sample sizes are more easily achieved in assortative groups.

Our approach complements traditional survey-based methods when transmission chains are available. Worby *et al.*'s relative risk estimation [2], measuring each group's proportional change in infection incidence before and after the peak, and Abbas *et al.*'s assessment method [21], comparing actual and expected proportions of infections across groups, do not consider the influence of group size. By integrating group size into our approach, we account for variations in the pool of susceptible individuals within each group, offering a more comprehensive understanding of transmission dynamics. Consequently, our approach should provide novel insights into the impact of group dynamics when estimating transmission patterns.

The main limitation of our approach pertains to the assumption that transmission chains are perfectly known. Although transmission trees can be reconstructed from data, such reconstruction effort comes with inherent uncertainty, which we have not considered here. Conventional epidemiological investigations may provide reliable transmission chains but require intensive labour for contact tracing, data collection and analysis, and may be prone to error [35]. Statistical approaches have been developed to reconstruct who infected whom using data on contacts, symptoms onset dates, and pathogen genome sequences [29], but in some contexts even these prove insufficient to precisely reconstruct transmission trees [21, 23, 36]. Our study underscores the challenges of inferring group contributions in some scenarios, even in the hypothetical instance where transmission trees are perfectly known. Nevertheless, our approach is adaptable and can be extended to reconstructed transmission chains, for example, by estimating the assortativity coefficient over all posterior transmission trees. Future research should delve into understanding how uncertainty surrounding these transmission trees further impacts our ability to infer transmission patterns.

Another limitation of our approach includes that our estimator requires, and is quite sensitive to (Fig 1.2 in S1 File), information on group sizes which may be difficult to obtain in real-life settings, however various methods exist for population size estimation [37]. Our simulations also assumed that individuals who have been infected become permanently immune, an assumption which is typically valid over short time frames but may be unrealistic over longer time horizons. Finally, characterising transmission through assortativity implies that a group's transmission patterns are identical towards all other groups, with only the within-group pattern being distinct. While this approach is fully representative in a two-group scenario, it is limiting when additional groups are involved. Nevertheless, this simplification aligns with established research on social networks and disease transmission dynamics suggesting that assortativity coefficients alone can effectively capture the essence of contact and transmission patterns across various contexts [30, 32, 38].

Despite these limitations, this study provides valuable insights into when the role of different groups in infectious disease transmission can be reliably identified in small outbreak settings, such as nosocomial outbreaks. We provide a framework for estimating group-specific transmission patterns that can be adapted to reconstructed transmission chains for real-world applications. By establishing the conditions under which these patterns become discernible, our findings can guide the timing and applicability of targeted control policies in these critical early-stage scenarios.

Supporting information

S1 File. Supplementary material. Supplementary materials for the manuscript.
(DOCX)

Acknowledgments

Simulations, analyses and visualisations were performed using the R software version 4.4.0 (https://www.R-project.org/) and the ggplot2 package (https://ggplot2.tidyverse.org/).

Author Contributions

Conceptualization: Cyril Geismar, Peter J. White, Anne Cori, Thibaut Jombart.

Data curation: Cyril Geismar.

Formal analysis: Cyril Geismar.

Funding acquisition: Peter J. White, Anne Cori, Thibaut Jombart.

Investigation: Cyril Geismar, Peter J. White, Anne Cori, Thibaut Jombart.

Methodology: Cyril Geismar, Peter J. White, Anne Cori, Thibaut Jombart.

Project administration: Cyril Geismar.

Resources: Cyril Geismar, Peter J. White, Anne Cori, Thibaut Jombart.

Software: Cyril Geismar.

Supervision: Peter J. White, Anne Cori, Thibaut Jombart.

Validation: Cyril Geismar, Anne Cori, Thibaut Jombart.

Visualization: Cyril Geismar.

Writing – original draft: Cyril Geismar.

Writing – review & editing: Cyril Geismar, Anne Cori, Thibaut Jombart.

References

1. Worby CJ, Kenyon C, Lynfield R, Lipsitch M, Goldstein E. Examining the role of different age groups and of vaccination during the 2012 Minnesota pertussis outbreak. Scientific Reports 2015; 5:13182. https://doi.org/10.1038/srep13182 PMID: 26278132
2. Worby CJ, Chaves SS, Wallinga J, Lipsitch M, Finelli L, Goldstein E. On the relative role of different age groups in influenza epidemics. Epidemics 2015; 13:10–6. https://doi.org/10.1016/j.epidem.2015.04.003 PMID: 26097505
3. Griggs EP, Flannery B, Foppa IM, Gaglani M, Murthy K, Jackson ML, et al. Role of Age in the Spread of Influenza, 2011–2019: Data From the US Influenza Vaccine Effectiveness Network. American Journal of Epidemiology 2022; 191:465–71. https://doi.org/10.1093/aje/kwab205 PMID: 34274963
4. Beale S, Hoskins S, Byrne T, Fong WLE, Fragaszy E, Geismar C, et al. Differential Risk of SARS-CoV-2 Infection by Occupation: Evidence from the Virus Watch prospective cohort study in England and Wales. Journal of Occupational Medicine and Toxicology 2023; 18:5. https://doi.org/10.1186/s12995-023-00371-9 PMID: 37013634
5. Metcalf CJE, Hampson K, Tatem AJ, Grenfell BT, Bjørnstad ON. Persistence in Epidemic Metapopulations: Quantifying the Rescue Effects for Measles, Mumps, Rubella and Whooping Cough. PLOS ONE 2013; 8:e74696. https://doi.org/10.1371/journal.pone.0074696 PMID: 24040325
6. Staras SAS, Cook RL, Clark DB. Sexual Partner Characteristics and Sexually Transmitted Diseases Among Adolescents and Young Adults. Sex Transm Dis 2009;36: https://doi.org/10.1097/OLQ.0b013e3181901e32 https://doi.org/10.1097/OLQ.0b013e3181901e32. PMID: 19265739
7. Endo A, Murayama H, Abbott S, Ratnayake R, Pearson CAB, Edmunds WJ, et al. Heavy-tailed sexual contact networks and monkeypox epidemiology in the global outbreak, 2022. Science 2022; 378:90–4. https://doi.org/10.1126/science.add4507 PMID: 36137054
8. Beyrer C, Baral SD, van Griensven F, Goodreau SM, Chariyalertsak S, Wirtz AL, et al. Global epidemiology of HIV infection in men who have sex with men. The Lancet 2012; 380:367–77. https://doi.org/10.1016/S0140-6736(12)60821-6.
9. Shaweno D, Karmakar M, Alene KA, Ragonnet R, Clements AC, Trauer JM, et al. Methods used in the spatial analysis of tuberculosis epidemiology: a systematic review. BMC Medicine 2018; 16:193. https://doi.org/10.1186/s12916-018-1178-4 PMID: 30333043
10. Ypma RJF, Jonges M, Bataille A, Stegeman A, Koch G, van Boven M, et al. Genetic Data Provide Evidence for Wind-Mediated Transmission of Highly Pathogenic Avian Influenza. The Journal of Infectious Diseases 2013; 207:730–5. https://doi.org/10.1093/infdis/jis757 PMID: 23230058
11. Mossong J, Hens N, Jit M, Beutels P, Auranen K, Mikolajczyk R, et al. Social contacts and mixing patterns relevant to the spread of infectious diseases. PLoS Medicine 2008; 5:e74. https://doi.org/10.1371/journal.pmed.0050074 PMID: 18366252
12. Jarvis CI, Van Zandvoort K, Gimma A, Prem K, Auzenbergs M, O'Reilly K, et al. Quantifying the impact of physical distance measures on the transmission of COVID-19 in the UK. BMC Medicine 2020; 18:124. https://doi.org/10.1186/s12916-020-01597-8 PMID: 32375776
13. Anderson RM, May RM. Infectious diseases of humans: dynamics and control. Oxford university press; 1991.
14. Wallinga J, Teunis P, Kretzschmar M. Using Data on Social Contacts to Estimate Age-specific Transmission Parameters for Respiratory-spread Infectious Agents. American Journal of Epidemiology 2006; 164:936–44. https://doi.org/10.1093/aje/kwj317 PMID: 16968863
15. Shirreff G, Huynh B-T, Duval A, Pereira LC, Annane D, Dinh A, et al. Assessing respiratory epidemic potential in French hospitals through collection of close contact data (April–June 2020). Sci Rep 2024; 14:3702. https://doi.org/10.1038/s41598-023-50228-8 PMID: 38355640
16. Nguyen LH, Drew DA, Graham MS, Joshi AD, Guo C-G, Ma W, et al. Risk of COVID-19 among frontline health-care workers and the general community: a prospective cohort study. The Lancet Public Health 2020; 5:e475–83. https://doi.org/10.1016/S2468-2667(20)30164-X PMID: 32745512
17. Temime L, Gustin M-P, Duval A, Buetti N, Crépey P, Guillemot D, et al. A Conceptual Discussion About the Basic Reproduction Number of Severe Acute Respiratory Syndrome Coronavirus 2 in Healthcare Settings. Clinical Infectious Diseases 2021; 72:141–3. https://doi.org/10.1093/cid/ciaa682 PMID: 32473007
18. Cooper BS, Evans S, Jafari Y, Pham TM, Mo Y, Lim C, et al. The burden and dynamics of hospital-acquired SARS-CoV-2 in England. Nature 2023; 623:132–8. https://doi.org/10.1038/s41586-023-06634-z.
19. Evans S, Agnew E, Vynnycky E, Stimson J, Bhattacharya A, Rooney C, et al. The impact of testing and infection prevention and control strategies on within-hospital transmission dynamics of COVID-19 in English hospitals. Philosophical Transactions of the Royal Society B: Biological Sciences 2021; 376:20200268. https://doi.org/10.1098/rstb.2020.0268 PMID: 34053255

20. Hoang T, Coletti P, Melegaro A, Wallinga J, Grijalva CG, Edmunds JW, et al. A systematic review of social contact surveys to inform transmission models of close-contact infections. Epidemiology (Cambridge, Mass) 2019; 30:723. https://doi.org/10.1097/EDE.0000000000001047 PMID: 31274572

21. Abbas M, Nunes TR, Cori A, Cordey S, Laubscher F, Baggio S, et al. Explosive nosocomial outbreak of SARS-CoV-2 in a rehabilitation clinic: the limits of genomics for outbreak reconstruction. Journal of Hospital Infection 2021; 117:124–34. https://doi.org/10.1016/j.jhin.2021.07.013 PMID: 34461177

22. Gordon CL, Trubiano JA, Holmes NE, Chua KYL, Feldman J, Young G, et al. Staff to staff transmission as a driver of healthcare worker infections with COVID-19. Infect Dis Health 2021; 26:276–83. https://doi.org/10.1016/j.idh.2021.06.003 PMID: 34344634

23. Abbas M, Cori A, Cordey S, Laubscher F, Robalo Nunes T, Myall A, et al. Reconstruction of transmission chains of SARS-CoV-2 amidst multiple outbreaks in a geriatric acute-care hospital: a combined retrospective epidemiological and genomic study. eLife 2022; 11:e76854. https://doi.org/10.7554/eLife.76854 PMID: 35850933

24. Lindsey BB, Villabona-Arenas CJ, Campbell F, Keeley AJ, Parker MD, Shah DR, et al. Characterising within-hospital SARS-CoV-2 transmission events using epidemiological and viral genomic data across two pandemic waves. Nat Commun 2022; 13:671. https://doi.org/10.1038/s41467-022-28291-y PMID: 35115517

25. Haanappel CP, Oude Munnink BB, Sikkema RS, Voor in 't holt AF, de Jager H, de Boever R, et al. Combining epidemiological data and whole genome sequencing to understand SARS-CoV-2 transmission dynamics in a large tertiary care hospital during the first COVID-19 wave in The Netherlands focusing on healthcare workers. Antimicrobial Resistance & Infection Control 2023; 12:46. https://doi.org/10.1186/s13756-023-01247-7.

26. Cook KF, Beckett AH, Glaysher S, Goudarzi S, Fearn C, Loveson KF, et al. Multiple pathways of SARS-CoV-2 nosocomial transmission uncovered by integrated genomic and epidemiological analyses during the second wave of the COVID-19 pandemic in the UK. Front Cell Infect Microbiol 2023;12. https://doi.org/10.3389/fcimb.2022.1066390 PMID: 36741977

27. Jombart T, Cori A, Didelot X, Cauchemez S, Fraser C, Ferguson N. Bayesian Reconstruction of Disease Outbreaks by Combining Epidemiologic and Genomic Data. PLOS Computational Biology 2014; 10:e1003457. https://doi.org/10.1371/journal.pcbi.1003457 PMID: 24465202

28. Didelot X, Fraser C, Gardy J, Colijn C. Genomic Infectious Disease Epidemiology in Partially Sampled and Ongoing Outbreaks. Molecular Biology and Evolution 2017; 34:997–1007. https://doi.org/10.1093/molbev/msw275 PMID: 28100788

29. Duault H, Durand B, Canini L. Methods Combining Genomic and Epidemiological Data in the Reconstruction of Transmission Trees: A Systematic Review. Pathogens 2022; 11:252. https://doi.org/10.3390/pathogens11020252 PMID: 35215195

30. Newman MEJ. Assortative Mixing in Networks. Phys Rev Lett 2002; 89:208701. https://doi.org/10.1103/PhysRevLett.89.208701 PMID: 12443515

31. Kiss IZ, Green DM, Kao RR. The effect of network mixing patterns on epidemic dynamics and the efficacy of disease contact tracing. J R Soc Interface 2008; 5:791–9. https://doi.org/10.1098/rsif.2007.1272 PMID: 18055417

32. Nishiura H, Cook AR, Cowling BJ. Assortativity and the probability of epidemic extinction: A case study of pandemic influenza A (H1N1-2009). Interdisciplinary Perspectives on Infectious Diseases 2011;2011. https://doi.org/10.1155/2011/194507 PMID: 21234337

33. Li J, Luo J, Liu H. Disassortative mixing patterns of drug-using and sex networks on HIV risk behaviour among young drug users in Yunnan, China. Public Health 2015; 129:1237–43. https://doi.org/10.1016/j.puhe.2015.07.020 PMID: 26298584

34. Clopper CJ, Pearson ES. The use of confidence or fiducial limits illustrated in the case of the binomial. Biometrika 1934; 26:404–13.

35. Faye O, Boëlle P-Y, Heleze E, Faye O, Loucoubar C, Magassouba N, et al. Chains of transmission and control of Ebola virus disease in Conakry, Guinea, in 2014: an observational study. The Lancet Infectious Diseases 2015; 15:320–6. https://doi.org/10.1016/S1473-3099(14)71075-8 PMID: 25619149

36. Campbell F, Strang C, Ferguson N, Cori A, Jombart T. When are pathogen genome sequences informative of transmission events? PLOS Pathogens 2018; 14:e1006885. https://doi.org/10.1371/journal.ppat.1006885 PMID: 29420641

37. Gutreuter S. Comparative performance of multiple-list estimators of key population size. PLOS Global Public Health 2022; 2:e0000155. https://doi.org/10.1371/journal.pgph.0000155 PMID: 35928219

38. Badham J, Stocker R. The impact of network clustering and assortativity on epidemic behaviour. Theoretical Population Biology 2010; 77:71–5. https://doi.org/10.1016/j.tpb.2009.11.003 PMID: 19948179