

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

Cyril Geismar<sup>1\*</sup>, Peter J White<sup>1,2</sup>, Anne Cori<sup>1†</sup>, Thibaut Jombart<sup>1†</sup>

<sup>1</sup> MRC Centre for Global Infectious Disease Analysis and NIHR Health Protection Research Unit in Modelling and Health Economics, Department of Infectious Disease Epidemiology, School of Public Health, Imperial College London, London, UK.

<sup>2</sup> Modelling & Economics Unit, UK Health Security Agency, London, UK.

\*Corresponding Author: [c.geismar21@imperial.ac.uk](mailto:c.geismar21@imperial.ac.uk)

† These authors contributed equally

## 43 Abstract

44 Characterising the transmission dynamics between various population groups is critical  
45 for implementing effective outbreak control measures whilst minimising financial costs and  
46 societal disruption.

47 Traditionally, mathematical models have primarily relied on assumptions of contact  
48 patterns to characterise transmission between groups. Thanks to technological and  
49 methodological advances, transmission chain data is increasingly available, providing  
50 information about individual-level transmission. However, it remains unclear how  
51 effectively and under what conditions such data can inform on transmission patterns  
52 between groups.

53 In this paper, we introduce a novel metric that leverages transmission chain data to  
54 estimate group transmission assortativity; this quantifies the extent to which individuals  
55 transmit within their own group compared to others. Through extensive simulations, we  
56 assessed the conditions under which our estimator performs effectively and established  
57 guidelines for minimal data requirements. Notably, we demonstrate that detecting and  
58 quantifying transmission assortativity is most reliable when groups have reached their  
59 epidemic peaks, consist of at least 30 cases each, and represent at least 10% of the total  
60 population each.

61

62

63

64

65

66

## 67 **Author Summary**

68 Efficient outbreak control relies on understanding how infection spreads between affected groups,  
69 such as healthcare workers and patients or specific age groups. Policies and interventions may differ  
70 substantially depending on how much transmission is within groups or between them. However,  
71 assessing transmission patterns between groups is challenging as these patterns are not only  
72 influenced by social contacts but also by variations in individual susceptibility and infectiousness,  
73 which changes over time. To address this challenge, we developed an estimator that utilises  
74 information on transmission chains (who infected whom), enabling the identification and  
75 quantification of transmission patterns between groups. Through extensive simulations, we  
76 assessed the conditions under which our estimator performs effectively and established guidelines  
77 for minimal data requirements. Our results suggest that inferring transmission patterns is most  
78 reliable when groups have reached their respective epidemic peaks, contain at least 30 cases each  
79 and constitute at least 10% or more of the total population, each.

80

81

82

83

84

85

86

87

88

89

90

91

92

## 93 Introduction

94 In response to the COVID-19 pandemic, governments across the world implemented nationwide  
95 lockdowns, later transitioning to targeted pharmaceutical and non-pharmaceutical interventions  
96 based on factors such as location, age, and vaccination status. However, these measures could not  
97 always be optimised in real-time due to a lack of quantitative understanding regarding the varying  
98 roles different groups played in disease transmission. For example, the decision to close schools  
99 was initially based on the assumption that children were significant drivers of transmission [1,2]. Yet,  
100 subsequent research suggested that children may be less susceptible to infection and that schools  
101 may not have played a major role in the transmission of SARS-CoV-2 [3–6].

102 The ability to detect and quantify the contributions of different groups to transmission during an  
103 outbreak is essential for implementing effective control measures. Not only does it enhance our  
104 comprehension of transmission dynamics within a population, but it may also lead to better  
105 predictions of the epidemic's future trajectory and enables the development of evidence-based public  
106 health strategies tailored to the outbreak's characteristics.

107 Broadly, two approaches have been employed to assess the contributions of different groups to  
108 epidemic transmission. First, dedicated surveys have been conducted to measure the frequency of  
109 contact between different groups; combined with information about the relative infectiousness and  
110 susceptibility of each group (e.g. obtained from epidemiological or serological investigations), these  
111 data can be used by transmission models to estimate transmission assortativity [7,8]. Unfortunately,  
112 the underlying contact data can be biased, have limited sample size or representativeness, and may  
113 not be generalisable across different epidemic contexts [9–11].

114 Alternatively, transmission assortativity can be directly assessed from observed transmission  
115 patterns e.g. by measuring the proportion of cases in different groups [12,13] or by reconstructing  
116 the transmission chains [14,15]. These approaches have their own limitations. For instance,  
117 accurately reconstructing transmission chains is challenging [16] and even with perfectly known  
118 transmission chains, transmission assortativity estimation may be impeded by differences in group  
119 sizes and group-level saturation (*i.e.* the depletion of susceptibles).

120 This paper introduces a novel framework for evaluating transmission patterns among distinct groups

121 during an outbreak, utilising known transmission chains to quantify group-specific assortativity. We  
122 evaluate the performance of our estimator through simulations across diverse outbreak scenarios  
123 and offer guidance on the minimum data collection requirements and the optimal estimation  
124 timeframe to inform policy.

125

## 126 **Methods**

### 127 **A new estimator of transmission assortativity**

128 Assortativity has been amply described for social mixing patterns, with homogeneous mixing  
129 referring to random contacts between individuals, and heterogeneous mixing denoting interactions  
130 characterised by distinct (non-random) patterns depending on group memberships [7].  
131 Heterogeneous mixing can be either *assortative*, where individuals tend to interact more within their  
132 own group (e.g. social contacts by age [9,17,18]), or *disassortative*, where individuals interact  
133 preferentially with members of other groups (e.g. sexual contacts [19]). Here we use these definitions  
134 to characterise the patterns of transmission rather than contact.

135 To quantify transmission assortativity, we examine the person-to-person transmission patterns. We  
136 denote  $\beta_{b \leftarrow a}$  the person-to-person transmission rate from an individual in group  $a$  to an individual in  
137 group  $b$ . We assume that  $\beta_{a \leftarrow a}$  can be expressed as  $\beta_{a \leftarrow a} = \gamma_a \beta_{b \leftarrow a}$  (with  $a \neq b$ ), where  $\gamma_a$  is the  
138 assortativity coefficient for group  $a$ .  $\gamma_a$  is defined as the excess probability of a secondary infection  
139 taking place within group  $a$  compared to random expectation.  $\gamma$  values range from 0 (fully  
140 disassortative) to  $\infty$  (fully assortative), with 1 indicating homogeneous patterns. For instance,  $\gamma_a = 2$   
141 indicates that an infected individual from group  $a$  is twice as likely to infect an individual from the  
142 same group compared to infecting an individual from another group. Conversely, a  $\gamma_a$  of 1/2 means  
143 that an infected individual from group  $a$  is twice as likely to infect an individual from another group  
144 compared to infecting an individual from the same group.

145 We consider  $G$  groups of relative sizes  $f_1, \dots, f_G$  defined as:

$$f_a = \frac{N_a}{\sum_{g=1}^G N_g} \quad \forall a = 1, \dots, G \quad (1)$$

146 where  $N_a$  is the number of individuals in group  $a$ .

147 For an infectious individual in group  $a$ , the proportion  $\pi_{b \leftarrow a}$  of secondary cases who are expected in  
148 group  $b$  is (details in supplementary materials S1.1):

$$\pi_{b \leftarrow a} = \frac{f_b}{\gamma_a f_a + (1 - f_a)} \quad \forall a, b = 1, \dots, G; \quad a \neq b \quad (2)$$

149 and

$$\pi_{a \leftarrow a} = \frac{\gamma_a f_a}{\gamma_a f_a + (1 - f_a)} \quad \forall a = 1, \dots, G \quad (3)$$

150

151 We can obtain  $\gamma_a$  by rewriting equation 3 as:

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

152

153 Here we assume that transmission chains are known. Among infections generated by infected  
154 individuals in group  $a$ , the proportion of secondary cases in group  $a$ ,  $\pi_{a \leftarrow a}$ , can therefore be  
155 calculated as:

$$\pi_{a \leftarrow a} = \frac{\tau_{a \leftarrow a}}{\tau_{\cdot \leftarrow a}} \quad (5)$$

156 where  $\tau_{a \leftarrow a}$  is the number of observed within-group transmission pairs and  $\tau_{\cdot \leftarrow a}$  is the total number of  
157 infections coming from group  $a$ . We can obtain a confidence interval (CI) on  $\pi_{a \leftarrow a}$  for various  
158 significance ( $\alpha$ ) levels using the Clopper-Pearson binomial interval method [20] (S1.2). Feeding  
159 estimates of  $\pi_{a \leftarrow a}$  from equation 5 into equation 4, provides estimates of  $\gamma_a$  with confidence intervals.

160 To simplify interpretation, we introduce a rescaled parameter  $\delta$ , ranging between -1 (fully  
161 disassortative) and 1 (fully assortative), with 0 corresponding to a homogeneous transmission  
162 pattern (Figure S1), defined as:

$$\delta = \begin{cases} 1 & \text{if } \gamma = \infty \\ \frac{\gamma-1}{\gamma+1} & \text{if } \gamma \neq \infty \end{cases} \quad (6)$$

163 All our results are presented using  $\delta$  rather than  $\gamma$ .

## 164 **Simulation study**

165 We simulated numerous outbreaks under various contexts to assess the estimator's performance.

166 The simulation employed a discrete time branching process modelling individual infections spreading

167 in successive generations. Simulations were specified with: i) group-level parameters including the

168 size of each group, their assortativity coefficients ( $\delta$ ), initial introductions, basic reproduction

169 numbers ( $R_0$ ) and ii) epidemic level parameters such as the number of groups, the pathogen

170 generation time ( $w$ ) and incubation period ( $v$ ) distributions (both assumed the same across groups).

171 The simulation outputs the transmission tree of the infected individuals including their group and that

172 of their infector, their date of infection and date of symptom onset. We constructed 10,000 sets of

173 input parameters, referred to as 'scenarios', by randomly sampling parameters from pre-defined

174 distributions (S1.3, Figure S2). To account for stochasticity, we conducted 100 simulations for each

175 unique scenario resulting in a total of 1,000,000 simulated outbreaks.

176 In our branching process model, the force of infection (FOI) generated by individual  $j$  from group  $a$

177 at time  $t$ , towards each individual in group  $b$  is defined as :

178

$$\lambda_{b \leftarrow a}^j(t) = w(t - s_a^j) R_{0a} \pi_{b \leftarrow a} \quad \begin{array}{l} \forall a, b = 1, \dots, G \\ \forall j = 1, \dots, N_a \end{array} \quad (7)$$

179 where:

180 •  $s_a^j$  is the time of infection of individual  $j$  in group  $a$

181 •  $R_{0a}$  is the basic reproduction number of individuals in group  $a$

182 The total FOI that group  $b$  receives from all groups at time  $t$  is obtained as:

$$\lambda_b(t) = \sum_{a=1}^G \sum_{j=1}^{N_a} \lambda_{b \leftarrow a}^j(t) \quad \forall b = 1, \dots, G \quad (8)$$

183

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

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

185

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

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

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

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

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

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

194 Where  $a^j$  is the  $j^{\text{th}}$  individual in group  $a$ .

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

- 197
- *Bias*: defined as the average difference between the true  $\delta$  value and its estimate ( $\hat{\delta}$ ) across  
198 100 simulations. It is a measure of the estimator's systematic error and inaccuracy and should  
199 be close to 0. Bias is positive when  $\delta$  is underestimated, indicating underestimation of  
200 assortativity or overestimation of disassortativity. Conversely, negative bias occurs when  $\delta$  is  
201 overestimated, indicating overestimation of assortativity or underestimation of



202 disassortativity.

- 203 • *Coverage (at significance level  $\alpha$ )*: defined as the proportion of simulations (out of 100) where  
204 the true  $\delta$  value is within the estimated CI corresponding to  $\alpha$ . We evaluate 4 significance  
205 levels 0.05, 0.1, 0.25 and 0.5. Assessing coverage helps determine the reliability of the  
206 confidence intervals generated by the estimator. Coverage should approximate  $1-\alpha$ , and the  
207 coverage error, which measures the deviation from this target, should be close to 0. A positive  
208 coverage error suggests underestimation of uncertainty, while a negative coverage error  
209 indicates overestimation.
- 210 • *Sensitivity (true positive rate)*: defined as the proportion of simulations (out of 100) where the  
211 estimator correctly identifies a significant assortative or disassortative effect (i.e. the  $\hat{\delta}$  CI  
212 doesn't contain 0). Sensitivity should be close to 1 (100%).
- 213 • *Specificity (true negative rate)*: defined as the proportion of simulations (out of 100) where  
214 the estimator correctly identifies no significant assortative or disassortative effect (i.e. the  $\hat{\delta}$   
215 CI contains 0). Specificity should be close to 1 (100%).

216

217 We evaluated the estimator's performance at various stages of the outbreak, defined in relation to  
218 the group's epidemic peak, i.e. the day with the highest symptom onset incidence following the first  
219 case. Denoting  $T$  the date of the group's peak incidence, we define the *analysis time window* as the  
220 time period from the first case of the group to day  $T \times \epsilon$ , where  $\epsilon$  represents any non-negative real  
221 number and is referred to as the "peak coefficient". A peak coefficient value of  $\epsilon=1$  implies analysis  
222 until the group's peak, while values above or below 1 imply analysis using data up to before or after  
223 the peak respectively (S1.4, Figure S3). Additionally, we introduce the term 'peak asynchronicity',  
224 calculated as the standard deviation of peak dates  $T$  across groups, to measure heterogeneity in the  
225 groups' peak dates.

226

227 To assess the impact of the scenario parameters on the performance metrics, separate regressions  
228 were conducted with each performance metric as a dependent variable and scenario parameters as  
229 independent variables (S1.5).

## 230 Results

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

232

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

236

237 Coverage performance was contingent upon the significance ( $\alpha$ ) level and the stage of the group's  
238 epidemic ( $\epsilon$ ) (Figure 1B). Halfway before the epidemic peak (peak coefficient  $\epsilon=0.5$ ), coverage at  $\alpha$   
239 levels up to 25% was too low, with average errors of 0.22, 0.18 and 0.07 for  $\alpha$  levels of 5, 10, and  
240 25%, respectively. In contrast, the 50% coverage was too high with an average error of -0.10. Around  
241 the epidemic peak ( $\epsilon$  0.7-1.3), coverage for  $\alpha = 5$ -10% was good, whilst coverage for  $\alpha = 25$ -50%  
242 was too high (average error -0.14). At later epidemic stages ( $\epsilon$  1.5-5), coverage was good across  
243 most significance levels, although the 50% coverage remained high across all epidemic stages.

244

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

251

252 Figure 2 presents the relationship between various epidemic characteristics (columns) and the  
253 estimator's performance metrics (rows), for a peak coefficient of 1 and a significance level of 0.05.  
254 Additional configurations are shown in supplementary materials (Figure S6).

255

256 Our estimator maintained consistent unbiased performance across the entire assortativity range ( $\delta$   
257 from -1 to 1) (Figure 2 column A row 1). Coverage consistently met the 95% target for  $\delta < 0.5$ , with  
258 a slight decrease in coverage performance for  $\delta > 0.5$ , although coverage remained close to the

259 target, averaging at 0.91 (sd = 0.10) (Figure 2A2). This decrease in coverage in highly assortative  
260 scenarios could be due to a saturation effect: high assortativity will accelerate the depletion of  
261 susceptibles in the group, eventually resulting in lower observed assortativity compared to the true  
262 value (Figure S4). Although the assortativity coefficient  $\delta$  only had a small effect on bias or coverage,  
263 it had a substantial impact on sensitivity, which was higher for larger absolute values of  $\delta$ . However,  
264 sensitivity rose more gradually as  $|\delta|$  increased on the disassortative scale compared to the  
265 assortative scale (Figure 2A3, Table S1.1), reaching an average of 82% for  $\delta \geq 0.5$  compared to  
266 55% for  $\delta \leq -0.5$ , suggesting a better ability to detect assortative than disassortative transmission.  
267 Indeed, assortative transmission implies that transmissions propagate within the same group across  
268 multiple generations, consequently increasing the sample size ( $\tau \leftarrow a$  in equation 5) compared to  
269 disassortative transmission, and thus narrowing the CI, thereby enhancing sensitivity. Our linear  
270 regression suggested that the assortativity coefficient explained nearly 60% of the variance observed  
271 in sensitivity (Table S1.1).

272

273 Increasing the number of cases substantially reduced bias (Figure 2C1, Table S2), and increased  
274 sensitivity (Figure 2C3, Table S1.2) but had little effect on specificity or coverage (Figure 2C4 and  
275 2C2). Bias was negligible (mean: 0.04, sd: 0.07) once the group reached 30 to 40 cases. Sensitivity  
276 was positively correlated with the number of cases: controlling for  $\delta$ , the odds of detecting an  
277 assortative or disassortative pattern increased by 4% with each additional case (Table S1.2).

278

279 The relative size of the group had a substantial effect on bias (Figure 2B1, Table S2) and sensitivity  
280 (Figure 2B3, Table S1.2) but no effect on specificity (Figure 2B4) nor coverage (Figure 2B2). When  
281 groups comprised 10% or more of the total population size, bias was close to 0 (Figure 2B1), and  
282 the odds of detecting an assortative pattern increased fourfold, compared to smaller groups (odds  
283 ratios (OR) = 4.15, 95% CI = 4.07 – 4.24) (Figure 2B3, Table S1.2). Relative size and the number  
284 of cases jointly accounted for 72% of the variation in bias (Table S2), and contributed to a 42%  
285 increase in the pseudo R-squared for the linear regression on sensitivity (from 0.566 in Table S1.1  
286 to 0.805 in Table S1.2).

287

288 Diverse transmission dynamics emerge from numerous groups, varying group sizes, reproduction  
289 numbers, and/or assortativity coefficients (Figure S5). This diversity results in varying saturation  
290 levels between groups over time, affecting transmission patterns within and between groups. Peak  
291 asynchronicity, a measure of heterogeneity in epidemic peak timing across groups was negatively  
292 associated with coverage (OR = 0.78, 95% CI = 0.78-0.78) and specificity (OR = 0.76, 95% CI =  
293 0.76-0.76), explaining 18% and 24% of the variance, respectively (Table S3 and S4, Figure 2D2 and  
294 2D4). These results suggest a decrease in our estimator's performance with increasing  
295 heterogeneity between groups. However, our estimates remained unbiased (Figure 2D1) and with  
296 consistent sensitivity (Figure 2D3) irrespective of that heterogeneity.

297

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

## 306 Discussion

307 We developed a method to detect and quantify the transmission assortativity of different groups  
308 based on transmission chains. We performed an extensive simulation study covering a range of  
309 epidemic scenarios to assess the performance of our approach.

310

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

313 Generally, analysing transmission chains too early in the outbreak, before the group's epidemic  
314 peak, results in poor performance across all metrics considered. On the other hand, delaying  
315 assortativity coefficient estimation poses challenges for timely policy implementation. Choosing

316 when exactly in the epidemic to analyse transmission chains, and what significance level to use for  
317 estimating the assortativity coefficients, will also depend on the objective. For instance, minimising  
318 bias and maximising sensitivity is best achieved later in the epidemic, past the group's peak, and  
319 using larger significance levels. Conversely, improving coverage and maximising specificity is  
320 easiest before the group's epidemic peak and using lower significance levels. Nevertheless,  
321 estimating assortativity at a target time before or at the peak requires accurate prediction of the  
322 group's peak date which can be very challenging.

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

326  
327 Detecting non-homogeneous transmission patterns (sensitivity) in the presence of relatively small  
328 groups (*i.e.* a group constituting less than 10% of the total population), with groups having fewer  
329 than 30 cases is challenging, particularly when assortative or disassortative patterns are mild ( $-0.5$   
330  $\leq \delta \leq 0.5$ ). Importantly, it is considerably easier to detect assortativity than disassortativity, given that  
331 assortativity yields more transmission events within the group considered (where most new  
332 infections appear) compared to disassortativity (where new infections tend to appear in other groups,  
333 by definition). Hence, all other things being equal, larger sample sizes are more easily achieved in  
334 assortative groups.

335  
336 Our approach complements traditional survey-based methods when transmission chains are  
337 available. Worby *et al.*'s relative risk estimation [12], measuring each group's proportional change in  
338 infection incidence before and after the peak, and Abbas *et al.*'s assessment method [15], comparing  
339 actual and expected proportions of infections across groups, do not consider the influence of group  
340 size. By integrating group size into our approach, we account for variations in the pool of susceptible  
341 individuals within each group, offering a more comprehensive understanding of transmission  
342 dynamics. Consequently, our approach should provide novel insights into the impact of group  
343 dynamics when estimating transmission patterns.

344

345 The main limitation of our approach pertains to the assumption that transmission chains are perfectly  
346 known. Although transmission trees can be reconstructed from data, such reconstruction effort  
347 comes with inherent uncertainty, which we have not considered here. Conventional epidemiological  
348 investigations may provide reliable transmission chains but require intensive labour for contact  
349 tracing, data collection and analysis, and may be prone to error [21]. Statistical approaches have  
350 been developed to reconstruct who infected whom using data on contacts, symptoms onset dates,  
351 and pathogen genome sequences [22], but in some contexts even these prove insufficient to  
352 precisely reconstruct transmission trees [14,23]. Our study underscores the challenges of inferring  
353 group contributions in some scenarios, even in the hypothetical instance where transmission trees  
354 are perfectly known. Nevertheless, our approach is adaptable and can be extended to reconstructed  
355 transmission chains, for example, by estimating the assortativity coefficient across all posterior  
356 transmission trees in the setting of Abbas *et al.* [15]. Future research should delve into understanding  
357 how uncertainty surrounding these transmission trees further impacts our ability to infer transmission  
358 patterns.

359

360 Another limitation of our approach includes that our estimator requires information on group sizes  
361 which may be difficult to obtain in real-life settings, however various methods exist for population  
362 size estimation [24]. Our simulations also assumed that individuals who have been infected become  
363 permanently immune, an assumption which is typically valid over short time frames but may be  
364 unrealistic over longer time horizons.

365

366 Despite these limitations, this study provides a valuable first step towards evaluating the  
367 contributions of different groups to the transmission of infectious diseases and informing targeted  
368 control policy.

369

370

## 371 **Data Availability**

372 The analysis code is freely available on a GitHub repository: [https://github.com/CyGei/o2groups-](https://github.com/CyGei/o2groups-analysis)  
373 [analysis](https://github.com/CyGei/o2groups-analysis). An R package has been developed for simulating outbreak scenarios and is also available  
374 on GitHub at: <https://github.com/CyGei/o2groups>.

375 Package and analysis code have been archived on Zenodo ( analysis:  
376 <https://zenodo.org/doi/10.5281/zenodo.10616176>, package:  
377 <https://zenodo.org/doi/10.5281/zenodo.10616155>)

## 378 **Acknowledgements**

379 CG is supported by a PhD studentship at Imperial College London funded by the National Institute  
380 for Health Research (NIHR) Health Protection Research Unit (HPRU) in Modelling and Health  
381 Economics, which is a partnership between the UK Health Security Agency (UKHSA), Imperial  
382 College London, and the London School of Hygiene & Tropical Medicine (grant code NIHR200908).  
383 AC, PJW are supported by the HPRU in Modelling and Health Economics. This work was supported  
384 by the UK Medical Research Council (MRC) Centre for Global Infectious Disease Analysis (grant  
385 number MR/X020258/1); this award comes under the Global Health EDCTP3 Joint Undertaking.

## 386 **Author Contributions**

387 Conceptualization: CG, AC, TJ, PJW.

388 Methodology: CG, AC, TJ, PJW.

389 Software: CG.

390 Validation: AC, TJ, PJW, CG.

391 Formal analysis: CG, AC, TJ, PJW.

392 Data Curation: CG.

393 Original Draft: CG.

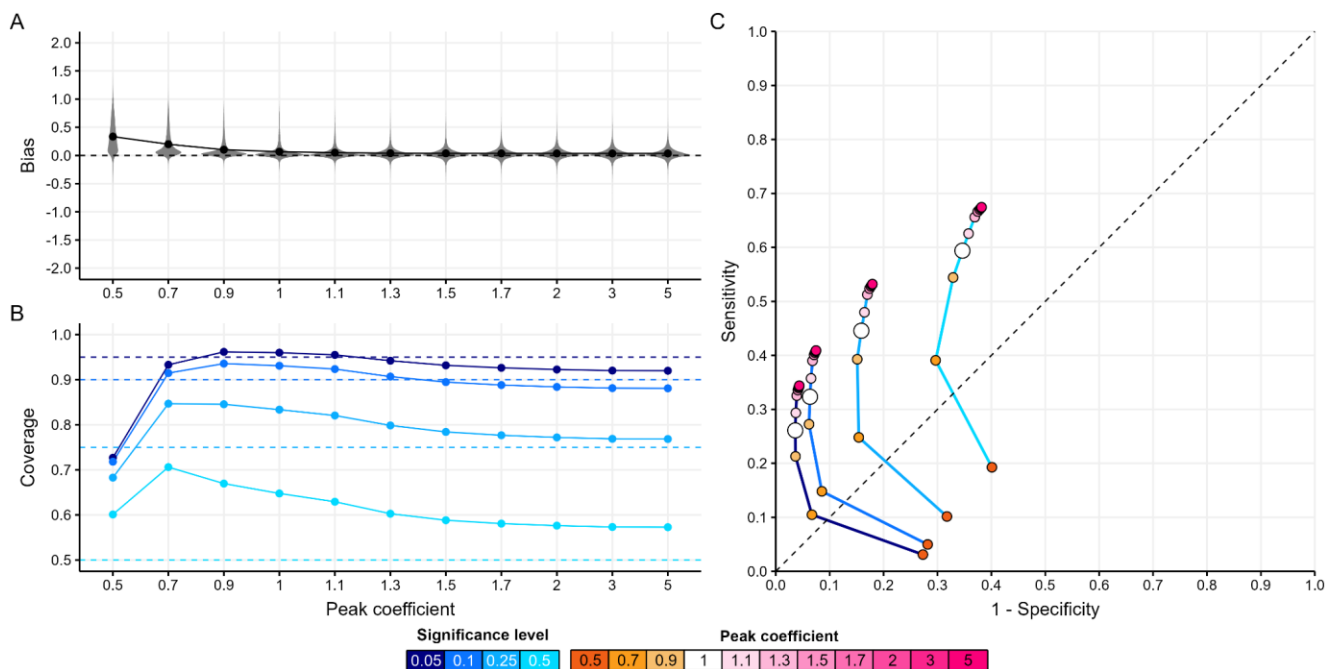
394 Writing: CG, TJ, PJW, AC.

395 Visualisation: CG.

396 Supervision: AC, TJ, PJW.

397 **Figures & Tables**

398



399

400 **Figure 1: Estimator's performance across all epidemic scenarios.**

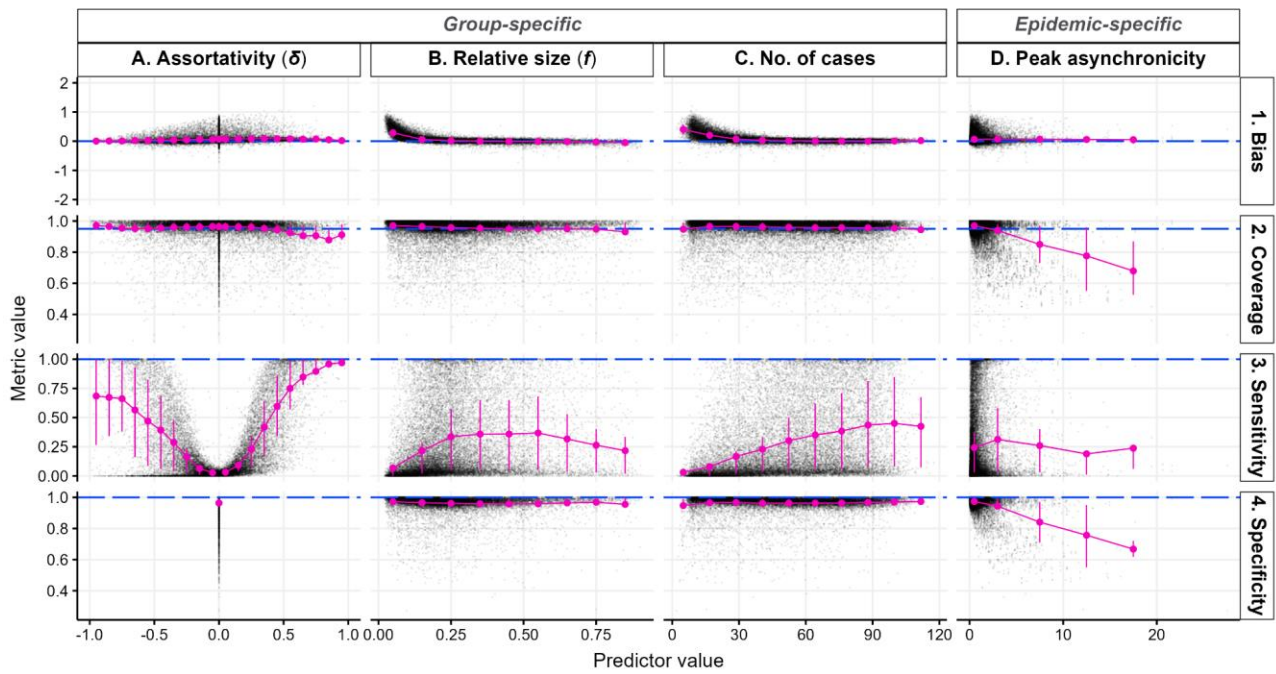
401 A. Distribution of bias (the mean difference between the true assortativity  $\delta$  value and its estimate)  
 402 by peak coefficient. The peak coefficient ( $\epsilon$ ) is a non-negative real number used to define the *analysis*  
 403 *time window* in relation to the group's epidemic peak. It determines the analysis period from the first  
 404 case to the day  $T\epsilon$ , where  $T$  is the date of peak incidence for the group. A value of  $\epsilon=1$  indicates  
 405 analysis up to the group's peak date, while values above or below 1 extend the analysis to data after  
 406 or before the group's peak date, respectively.

407 B. Mean coverage (proportion of simulations where the true  $\delta$  value is within the estimated CI) by  
 408 peak coefficient for each significance level (blue shades).

409 C. The Receiver Operating Characteristic (ROC) (the trade-off between sensitivity and specificity)  
 410 curves by peak coefficient (orange-pink points) for each significance level (blue shaded lines).

411 In panel A, each point shows the mean metric value across all scenarios for a given peak coefficient.  
 412 In panels B and C, each point shows the mean metric value across all scenarios for a given peak  
 413 coefficient and significance level. Dashed lines refer to the metric's target value for A and B and  
 414 represent a random classifier's ROC performance for C.





415

416 **Figure 2: Estimator's performance across scenario parameters and epidemic characteristics.**

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

426

427

428

429

430

431

## References

- 433 [1] Tseng Y-J, Olson KL, Bloch D, Mandl KD. Smart Thermometer–Based Participatory  
434 Surveillance to Discern the Role of Children in Household Viral Transmission During the  
435 COVID-19 Pandemic. *JAMA Netw Open* 2023;6:e2316190–e2316190.
- 436 [2] Park YJ, Choe YJ, Park O, Park SY, Kim Y-M, Kim J, et al. Contact Tracing during Coronavirus  
437 Disease Outbreak, South Korea, 2020. *Emerg Infect Dis* 2020;26:2465–8.  
438 <https://doi.org/10.3201/eid2610.201315>.
- 439 [3] Wu JT, Mei S, Luo S, Leung K, Liu D, Lv Q, et al. A global assessment of the impact of school  
440 closure in reducing COVID-19 spread. *Philos Trans R Soc Math Phys Eng Sci*  
441 2021;380:20210124. <https://doi.org/10.1098/rsta.2021.0124>.
- 442 [4] Viner RM, Russell SJ, Croker H, Packer J, Ward J, Stansfield C, et al. School closure and  
443 management practices during coronavirus outbreaks including COVID-19: a rapid systematic  
444 review. *Lancet Child Adolesc Health* 2020;4:397–404.
- 445 [5] Heavey L, Casey G, Kelly C, Kelly D, McDarby G. No evidence of secondary transmission of  
446 COVID-19 from children attending school in Ireland, 2020. *Eurosurveillance* 2020;25:2000903.
- 447 [6] Cordery R, Reeves L, Zhou J, Rowan A, Watber P, Rosadas C, et al. Transmission of SARS-  
448 CoV-2 by children to contacts in schools and households: a prospective cohort and  
449 environmental sampling study in London. *Lancet Microbe* 2022;3:e814–23.
- 450 [7] Anderson RM, May RM. *Infectious diseases of humans: dynamics and control*. Oxford  
451 university press; 1991.
- 452 [8] Wallinga J, Teunis P, Kretzschmar M. Using Data on Social Contacts to Estimate Age-specific  
453 Transmission Parameters for Respiratory-spread Infectious Agents. *Am J Epidemiol*  
454 2006;164:936–44. <https://doi.org/10.1093/aje/kwj317>.
- 455 [9] Mossong J, Hens N, Jit M, Beutels P, Auranen K, Mikolajczyk R, et al. Social contacts and  
456 mixing patterns relevant to the spread of infectious diseases. *PLoS Med* 2008;5:e74.
- 457 [10] Hoang T, Coletti P, Melegaro A, Wallinga J, Grijalva CG, Edmunds JW, et al. A systematic  
458 review of social contact surveys to inform transmission models of close-contact infections.  
459 *Epidemiol Camb Mass* 2019;30:723.
- 460 [11] CMMID COVID-19 working group, Jarvis CI, Van Zandvoort K, Gimma A, Prem K, Klepac P, et  
461 al. Quantifying the impact of physical distance measures on the transmission of COVID-19 in  
462 the UK. *BMC Med* 2020;18:124. <https://doi.org/10.1186/s12916-020-01597-8>.
- 463 [12] Worby CJ, Chaves SS, Wallinga J, Lipsitch M, Finelli L, Goldstein E. On the relative role of  
464 different age groups in influenza epidemics. *Epidemics* 2015;13:10–6.
- 465 [13] Worby CJ, Kenyon C, Lynfield R, Lipsitch M, Goldstein E. Examining the role of different age  
466 groups and of vaccination during the 2012 Minnesota pertussis outbreak. *Sci Rep*  
467 2015;5:13182.
- 468 [14] Abbas M, Nunes TR, Cori A, Cordey S, Laubscher F, Baggio S, et al. Explosive nosocomial  
469 outbreak of SARS-CoV-2 in a rehabilitation clinic: the limits of genomics for outbreak  
470 reconstruction. *J Hosp Infect* 2021;117:124–34.
- 471 [15] Abbas M, Cori A, Cordey S, Laubscher F, Robalo Nunes T, Myall A, et al. Reconstruction of  
472 transmission chains of SARS-CoV-2 amidst multiple outbreaks in a geriatric acute-care  
473 hospital: a combined retrospective epidemiological and genomic study. *eLife* 2022;11:e76854.  
474 <https://doi.org/10.7554/eLife.76854>.
- 475 [16] Duault H, Durand B, Canini L. Methods Combining Genomic and Epidemiological Data in the  
476 Reconstruction of Transmission Trees: A Systematic Review. *Pathogens* 2022;11:252.
- 477 [17] Nishiura H, Cook AR, Cowling BJ. Assortativity and the probability of epidemic extinction: A  
478 case study of pandemic influenza A (H1N1-2009). *Interdiscip Perspect Infect Dis* 2011;2011.
- 479 [18] Kiss IZ, Green DM, Kao RR. The effect of network mixing patterns on epidemic dynamics and  
480 the efficacy of disease contact tracing. *J R Soc Interface* 2008;5:791–9.  
481 <https://doi.org/10.1098/rsif.2007.1272>.
- 482 [19] Li J, Luo J, Liu H. Disassortative mixing patterns of drug-using and sex networks on HIV risk  
483 behaviour among young drug users in Yunnan, China. *Public Health* 2015;129:1237–43.
- 484 [20] Clopper CJ, Pearson ES. The use of confidence or fiducial limits illustrated in the case of the  
485 binomial. *Biometrika* 1934;26:404–13.
- 486 [21] Faye O, Boëlle P-Y, Heleze E, Faye O, Loucoubar C, Magassouba N, et al. Chains of  
487 transmission and control of Ebola virus disease in Conakry, Guinea, in 2014: an observational

- 488 study. *Lancet Infect Dis* 2015;15:320–6.
- 489 [22] Jombart T, Cori A, Didelot X, Cauchemez S, Fraser C, Ferguson N. Bayesian Reconstruction  
490 of Disease Outbreaks by Combining Epidemiologic and Genomic Data. *PLOS Comput Biol*  
491 2014;10:e1003457. <https://doi.org/10.1371/journal.pcbi.1003457>.
- 492 [23] Geismar C, Nguyen V, Fragaszy E, Shrotri M, Navaratnam AM, Beale S, et al. Bayesian  
493 reconstruction of SARS-CoV-2 transmissions highlights substantial proportion of negative  
494 serial intervals. *Epidemics* 2023;44:100713.
- 495 [24] Gutreuter S. Comparative performance of multiple-list estimators of key population size. *PLOS*  
496 *Glob Public Health* 2022;2:e0000155. <https://doi.org/10.1371/journal.pgph.0000155>.