

1 **A standardised differential privacy framework for epidemiological modelling with mobile**
2 **phone data**

3 Merveille Koissi Savi¹, Akash Yadav², Wanrong Zhang³, Navin Vembar⁴, Andrew Schroeder²,
4 Satchit Balsari⁵, Caroline O. Buckee⁶, Salil Vadhan³, Nishant Kishore^{6*}

5

6 ¹ Dana Farber Cancer Institute, Harvard School of Medicine, Department of Medical Oncology,
7 MA, USA

8 ²Direct Relief, Santa Barbara, CA, USA

9 ³Department of Computer Sciences, Harvard John A. Paulson School of Engineering & Applied
10 Sciences, MA, USA

11 ⁴Camber Systems, Washington, DC, USA

12 ⁵Department of Emergency Medicine, Harvard Medical School, Boston, USA

13 ⁶Department of Epidemiology, Harvard TH Chan School of Public Health, Boston, USA

14

15

16 *Corresponding author: nish.kishore@gmail.com

17 **Abstract**

18 During the COVID-19 pandemic, the use of mobile phone data for monitoring human mobility
19 patterns has become increasingly common, both to study the impact of travel restrictions on
20 population movement and epidemiological modelling. Despite the importance of these data, the
21 use of location information to guide public policy can raise issues of privacy and ethical use.
22 Studies have shown that simple aggregation does not protect the privacy of an individual, and there
23 are no universal standards for aggregation that guarantee anonymity. Newer methods, such as
24 differential privacy, can provide statistically verifiable protection against identifiability but have
25 been largely untested as inputs for compartment models used in infectious disease epidemiology.
26 Our study examines the application of differential privacy as an anonymisation tool in
27 epidemiological models, studying the impact of adding quantifiable statistical noise to mobile
28 phone-based location data on the bias of ten common epidemiological metrics. We find that many
29 epidemiological metrics are preserved and remain close to their non-private values when the true
30 noise state is less than 20, in a count transition matrix, which corresponds to a privacy-less
31 parameter $\epsilon = 0.05$ per release. We show that differential privacy offers a robust approach to
32 preserving individual privacy in mobility data while providing useful population-level insights for
33 public health. Importantly, we have built a modular software pipeline to facilitate the replication
34 and expansion of our framework.

35 **Author Summary**

36 Human mobility data has been used broadly in epidemiological population models to better
37 understand the transmission dynamics of an epidemic, predict its future trajectory, and evaluate
38 potential interventions. The availability and use of these data inherently raises the question of how
39 we can balance individual privacy and the statistical utility of these data. Unfortunately, there are

40 few existing frameworks that allow us to quantify this trade-off. Here, we have developed a
41 framework to implement a differential privacy layer on top of human mobility data which can
42 guarantee a minimum level of privacy protection and evaluate their effects on the statistical utility
43 of model outputs. We show that this set of models and their outputs are resilient to high levels of
44 privacy-preserving noise and suggest a standard privacy threshold with an epsilon of 0.05. Finally,
45 we provide a reproducible framework for public health researchers and data providers to evaluate
46 varying levels of privacy-preserving noise in human mobility data inputs, models, and
47 epidemiological outputs.

48 **Introduction**

49 The use of private mobile phone data for various applications in public health, urban planning, and
50 response to natural disasters has been steadily growing for more than a decade [1–3] . The COVID-
51 19 pandemic has accelerated this trend, and the use of mobility data has increased, following the
52 need to monitor and make policy decisions related to travel restrictions and lockdowns. These data
53 were incorporated into epidemiological models during the pandemic to monitor or forecast SARS-
54 COV-2 transmission.

55 Mobility data from mobile phones allow us to quantify changes in human movement, identify how
56 social contacts cluster, evaluate where cases come into contact with others, and predict the
57 probability of geographic spread [4]. Data acquired from cell phone metadata recorded for billing
58 purposes or from digital platforms are aggregated and shared with researchers, who can then get
59 significant information from mobility patterns [5–7]. Such studies have been used to explain the
60 seasonal pattern of dengue in Pakistan and rubella in Kenya, for example [5,7]. These models are
61 predominantly metapopulation models in which mobility data are used to determine the impact of

62 human migration on the trajectory of infectious diseases. During the COVID-19 pandemic, the use
63 of mobility data increased around the world, and metapopulation models were used to understand
64 the relationship between human mobility and the spread of the epidemic, predict the dynamics of
65 the epidemic, and estimate the effectiveness of nonpharmaceutical interventions such as
66 lockdowns, reopenings, and social distancing, based on other work modelling the spatial dynamics
67 of pathogens [4–6].

68 Despite the statistical utility of these datasets, important privacy concerns remain about the sharing
69 of personal data, even if they are deidentified and aggregated. Standardised approaches are
70 currently lacking for data-sharing agreements and guidelines on the appropriate ways to protect
71 individual privacy while using mobility data for public health. As big data, the semantic web, the
72 interconnectedness of digital technology, and the "Internet of Things" (IoT) increase the volume
73 and velocity of data, it becomes easier to reanonymise such aggregated data [8].

74 Several privacy frameworks have been developed to address the trade-off between privacy and
75 utility for statistical analyses [9–15]. Amongst these frameworks, *differential privacy* (DP) has
76 become the leading approach to balance this trade-off [16]. DP is a parameterized privacy concept,
77 where the privacy parameter ϵ allows for a smooth trade-off between privacy and utility for
78 statistical analyses [17]. Informally, an algorithm that is ϵ -differentially private ensures that any
79 particular output of the algorithm is at most e^ϵ more likely when we arbitrarily change one data
80 entry. In DP, observations are perturbed by adding noise coming from a carefully chosen
81 distribution [17]. A DP mechanism applied to a mobility matrix of travel between different
82 locations will prevent disclosing the exact number of movements and will also keep the private
83 information of the individual (home and work location, etc.) hidden.

84 DP is considered the gold standard of statistical privacy, as its application can be proven to
85 preserve privacy while quantifying the trade-off between privacy and the utility of the released
86 statistics [16]. The trade-off between privacy and utility is important because the noisier the output,
87 the less useful it may be for inference. Increasingly, DP is used for the public release of data sets
88 by industries and governments such as Google, Apple, Microsoft, Facebook (Meta), Uber, and the
89 US Census Bureau, but it remains unclear how DP should be used in the context of mobility data
90 for epidemiological frameworks [18].

91 In this paper, we examine how differential privacy can be applied to infectious disease modelling
92 and analyse the impact of different levels of privacy on the reconstruction of epidemic features
93 through simulation. Our method is based on a previously validated epidemiological
94 metapopulation model, and we investigate the effect of the addition of privacy-preserving noise
95 on key epidemiological outputs of interest. We used real-world mobility data from New York State
96 during the early stages of the COVID-19 pandemic in the United States and show that the
97 application of differential privacy can bias certain epidemiological metrics. We propose that
98 differential privacy offers a rigorous and quantifiable approach to safely using mobile phone data
99 during epidemics for modelling purposes.

100 **Results**

101 *Mobility data*

102 The mobility matrices included data from August 15 to November 15, 2020, and contained a total
103 of 812,587 transitions made between sixty-two counties of New York State, with a mean of 9,029
104 transitions a day. The observed daily transitions ranged from a minimum of 600, occurring in
105 Hamilton County, to a maximum of 77,131 in Suffolk County. The maximum transition between

106 counties occurred between Queens and Kings counties, with 5,262, whereas we counted 14
107 combinations of zero transitions during the selected windows. After applying DP, the absolute
108 number of transitions was affected, but the relative rank of the intercounty routes with respect to
109 the volume of travel remained the same. We initiated a variety of common scenarios to assess the
110 effect of added noise on bias and variability in our epidemiological parameters of interest.

111 *Scenarios with initial outbreaks in large and small regions*

112 We first address the impact of starting epidemics in large versus small counties to determine
113 whether DP would have systematic impacts on the dynamics overall. Kings and Queens are the
114 largest counties in New York State with an approximate population of 2 million individuals each
115 [19]. Allegany and Essex are the smallest counties in New York state, with populations of
116 approximately 46,000 and 37,000 individuals, respectively. In each of these counties (first the two
117 largest, and then the two smallest), we seeded 20 infectious individuals to spark an epidemic. In
118 the scenario with large counties, we observed epidemics that started around the 50th day and
119 peaked around the 75th day, reaching approximately 1% of the population living in these areas. In
120 the smaller counties, the epidemic began around the 60th day and peaked on the 150th day,
121 reaching approximately 5% of the population (Suppl.).

122 We evaluated the metrics of interest over 1,000 iterations for each combination of scenarios and
123 noise. We observed that when the epidemic is seeded in Queens and Kings, the epidemic size and
124 the proportion of counties with at least one case are higher compared to an outbreak seeded in
125 smaller counties (Fig. 1A). When noise is above 20, the values for the epidemic size for observed,
126 asymptomatic, and symptomatic infected, the size at the peak of the epidemic, and the proportion
127 of counties with one case are lower than those obtained when the mobility matrix is not perturbed.

128 However, the values obtained for the rate of spread, effective reproductive rate, risk of importation,
129 probability of importation, and mean importation rate are higher than those obtained for the non-
130 perturbed dataset (Fig. 1).

131 *Scenarios with Epidemics in Well- and Poorly Connected Regions*

132 To address how the effect of DP on network connectivity would impact predicted disease
133 dynamics, we simulated an outbreak in three pairs of counties with varying levels of connectivity
134 to Kings County. The first simulation in Monroe and Saratoga counties was designed to assess the
135 impact of low connectivity (less than 20% of transitions during the period) on the disease dynamic.
136 The second scenario targeted counties in the median of transitions, such as Putman and
137 Westchester counties, to assess the dose-response effect of the epidemiological model. The third
138 scenario was simulated in Schoharie and Lewis counties (no transition to Kings County during the
139 period) to assess the impact in places that were isolated in the larger mobility network. When the
140 outbreak is simulated in Monroe and Saratoga (Scenario 3), the epidemic begins around the 60th
141 day and the number of infected persons reaches the maximum around the 150th day, with less than
142 1% of the total population living in this area infected. When the outbreak is seeded in medium
143 connectivity areas such as Putnam and Westchester (Scenario 4), less than 0.6% of the population
144 became infected around the 75th day after the epidemic peaks around the 40th day. When the
145 outbreak is seeded in an area with low connectivity to Kings County, i.e., connectivity close to 0
146 such as Lewis and Schoharie (scenario 5), less than 0.07% of the population is infected around the
147 200th day since the epidemic only starts around the 90th day (Suppl.).

148 We found that regardless of network connectivity, epidemiological metrics degraded as noise
149 increased (Fig. 1B). As such in the three scenarios addressing the change in the network of

150 mobility, namely when i) the epidemic is sparked in two random counties having less than 20%
151 transition to Kings County, ii) the epidemic is sparked in two random counties with a median
152 transition to Kings County, and iii) the epidemic is sparked in a county with no transition to Kings
153 County; we observed a similar pattern in the distribution of the metric to what we observed when
154 there was an outbreak in small counties (scenario 2). Specifically, the size of the epidemic, the day
155 that the epidemic peaks, the fraction of counties with at least one case, the size of the epidemic,
156 the average exposure time, the maximum exposure time, and the minimum exposure time are
157 smaller than the baseline. The spread rate, the effective reproductive rate, the importation risk, the
158 mean importation risk rate, and the probability of infection are higher than the baseline, especially
159 when the noise is above 33.33 (Fig. 1B). We observed a significant change in the epidemiological
160 metrics only when the value of noise added to perturb the transition matrix is above those of the
161 scenario targeting the location of the first cases (small versus large county) (Fig. 1B).

162 *Scenarios with varying epidemiological parameters*

163 To address the nature of the epidemic, we simulated three changes in the trajectory of the epidemic
164 in Kings and Queens counties. Specifically, we simulated i) a faster epidemic through the increase
165 of the transmission rate, ii) a heavy load of asymptomatic individuals, and ii) an absence of
166 asymptomatic individuals in the population. When the transmission rate increases (scenario 6) we
167 can observe that the epidemic starts around the 40th day and reaches its peak around the 75th with
168 almost 3% of the population infected. When the fraction of symptomatic individuals increases, the
169 size of the epidemic also increases and reaches 1.5% of the population around the 75th day since
170 the epidemic starts around the 40th day after the first case (scenario 7). When the fraction of
171 documented infection decreases (scenario 8), there is no declared epidemic, as only asymptomatic
172 people are recorded in the population, reaching a fraction of 0.008% after the 100th day.

173 When the transmission rate increases, the epidemic spreads quickly (Fig. 1C). When the
174 asymptomatic rate increases, the probability of infection will subsequently increase. The trajectory
175 of the epidemic is similar to the non-perturbed dataset. However, above the noise of 33.33,
176 epidemiological metrics are either more conservative (lower than those of the baseline) or more
177 volatile (higher than those of the baseline) (Fig. 1C). Furthermore, we found that the fraction of
178 counties with at least one case is not affected by the change in i) the transmission rate and ii) the
179 fraction of symptomatic individuals (Fig. 1C)

180 [insert Fig. 1 here]

181 **Discussion**

182 Several metapopulation models were developed throughout the SARS-CoV2 pandemic to inform
183 decision making, predict the trajectory of the disease and identify weaknesses in the healthcare
184 system [20–23]. The mobility data used to parameterize these models provided information on
185 geographic and behavioral heterogeneity between populations, but these data could theoretically
186 be used to identify individuals or their unique travel behavior, which warrants privacy preservation
187 measures [24]. Our study shows that in metapopulation models that use mobility data, the
188 application of privacy-preserving noise results in unbiased estimates of metrics of interest at a wide
189 range of noise values with an upper limit that allows for a significant privacy-preserving budget.

190 We found that mobility matrices that are infused with noise values below 20, that is, loss of privacy
191 loss $\epsilon = 0.05$ per matrix, can help protect the privacy of individuals who contribute their data,
192 while limiting bias in the estimation of public health measures of interest when used for
193 epidemiological modelling. Intuitively, adding noise to these mobility matrices may result in
194 newly created connections between locations that would not otherwise be connected, strengthening
195 connections that would otherwise be weak, or vice versa. In some cases, we may even see the
196 removal of connections on specific days. Predictions of the spread of the rural area may be more
197 affected than those of the areas connected to urban centers. However, sensitivity analyses could
198 be performed to provide robustness, and the purpose and geographic scope of the model will dictate
199 how important this degradation is.

200 As noise increases above 20, estimates such as the epidemic size, the day that the epidemic peaks,
201 and the average epidemic size are biased downwards as the mobility matrix decreases connectivity

202 to large population centers and distributes the epidemic into many smaller locations with lower
203 contact rates. Similarly, estimates such as the rate of spread, the risk of importation, and the
204 effective reproduction rate are biased upwards as mobility between smaller and poorly connected
205 locations increases, leading to greater importation into areas with smaller population sizes. Our
206 study demonstrates that for epidemiological metapopulation models using mobility data, metrics
207 estimates are fairly unbiased up to a noise threshold of 20, which provides greater privacy
208 protection than previous studies [23,25].

209 Although our pipeline only evaluated a specific combination of mobility data, metapopulation
210 model, and metrics, it provides a " *plug-and-play* " interface for researchers to assess bias using
211 proprietary models and mobility data [26]. As mobility data sets become increasingly available
212 and used in metapopulation models, we provide a flexible framework to identify the evaluation-
213 specific maximum privacy-preserving noise that can be incorporated into these mobility data
214 before they result in biased outputs.

215 **Methods**

216 The pipeline workflow for the next analysis is represented in the following schematic architecture
217 (Fig. 2). This flow diagram shows the preprocessing before and after acquisition of the mobility
218 data, and, most importantly, how synthetic data has been used to parameterize the metapopulation
219 mode.

220 [insert Fig. 2 here]

221 ***Mobility Data***

222 We obtained mobility data from Camber Systems (the provider), a third-party analytics company
223 that purchased advertising technology (ad tech) data from many data brokers. The data covered 90
224 days from August 15 to November 15, 2020, representing between 3-7% of the total American
225 Community Survey (ACS), a county-specific population in New York State. The original data
226 consisted of a log of user global positioning system (GPS) coordinates, sorted and grouped by a
227 unique device identification number. These data have all the identifying information removed,
228 cleaned to remove duplicate entries or unrealistic usage, used to calculate device-specific modal
229 locations, and aggregated at the county level [27]. The key metric of interest used in these analyses
230 was movement between counties in 8 hour increments. Movement was defined as the change in
231 the location of a device from time period $t-1$ to the location of the device at time t . To further
232 guarantee anonymity, the provider used a predefined group of devices per area, removed data that
233 represented small numbers of devices, and applied an initial layer of privacy noise to the data set
234 to ensure that the basic privacy preservation mechanisms were in place before providing access to
235 these data to researchers [28]. We then added an additional layer of postproduction differential
236 privacy (PPDP) (see next section) and aggregated it into 24-hour blocks of time with averaged
237 transitions between counties. The process consists of generating an origin/destination matrix
238 normalised to the ACS population for each county. The matrix was then randomly sampled and
239 replicated 500 times to extend the data set time period.

240 *Application of Differential Privacy*

241 As background, a mechanism M taking a database in a domain H and producing outputs in a
242 domain R $M: H \rightarrow R$ is ϵ -differential private if and only if for every pair of neighboring
243 databases $x, y \in H$, such that they differ in at most one entry, and for any subset of possible outputs
244 $S \subseteq R$, we have

$$Pr[M(x) \in S] \leq e^\epsilon Pr[M(y) \in S], \quad (1)$$

245 where the probability is taken over the randomness of the mechanism M . Equation (1) suggests
246 that if two databases x, y are sufficiently close due to the perturbation, then it becomes difficult
247 for random attackers to uncover the privacy of the observed individuals. This is achieved by
248 perturbing the true observations by adding noise from a carefully chosen distribution. The
249 parameter quantifying the privacy loss ϵ represents the likelihood that an attacker with nearly full
250 information about a database can determine whether their target is in the database. DP offers a
251 quantifiable tradeoff between accuracy and privacy. Mobility data is aggregated data that could
252 display the transmission of small groups of individuals. Our goal is to preserve the privacy of these
253 groups and hide low transitions by applying differential privacy.

254 The Laplace mechanism is a common differential privacy mechanism, which adds Laplace noise
255 to query values in which the noise scales with Δ/ϵ , where Δ is the query sensitivity. DP
256 compositions adaptively allowing us to design a mechanism with several building blocks ensuring
257 efficient protection of privacy achievable using the advanced composition 10.

258 For all $\epsilon, \delta, \delta' > 0$, the class of (ϵ, δ) -differentially private mechanisms satisfies $(\epsilon', k\delta + \delta')$ -
259 differential privacy under k -fold adaptive composition for (Eq. 2):

$$\epsilon' = \sqrt{2k \log\left(\frac{1}{\delta'}\right)\epsilon + k\epsilon(e^\epsilon - 1)} \quad (2)$$

260 To assess the tradeoff between accuracy and utility, we further privatize the synthetic data using
261 the composition theorem with the privacy parameter epsilon ranging from 0.01 to 16 by the means

262 of the Laplace mechanism using the ‘smartnoise sdk’ library [10]. The transition data contains the
263 movements for 8-hour time blocks over 90 days, and using the advanced composition theorem
264 with $k = 270$, the total privacy budget is as follows (Eq. 3):

$$\epsilon' = \sqrt{540 \log\left(\frac{1}{\delta'}\right)} \epsilon + 270 \epsilon(e^\epsilon - 1) = 84.6\epsilon + 270\epsilon(e^\epsilon - 1). \quad (3)$$

265 For $\epsilon = 0.01$, $\delta' = 1.064494$, we have $\epsilon' = 0.8911355$ used to the existing deployment.

266 The rationale for using this range of epsilon lies in the fact that below 0.01 the infused noise is
267 extremely large, compromising the accuracy of the transition matrix, and above 16 the total privacy
268 budget is extremely large, compromising the privacy. More specifically, since the transition matrix
269 used already has privacy noise applied, with a value of $\epsilon = 16$ means, the synthetic transition
270 obtained is similar to the one received from the provider. However, for $\epsilon = 0.01$, the synthetic data
271 is more protective since low transitions are more hidden due to the large amount of noise added
272 through the Laplace mechanism. To simplify interpretation, from here on, we evaluate *noise* which
273 is the inverse of the privacy loss ϵ .

274 ***Metapopulation model***

275 The disease dynamic was modeled with a Susceptible-Exposed-Infected Symptomatic-Infected
276 asymptomatic model as follows (Eq. 4-7).

$$\frac{dS_i}{dt} = -\frac{\beta S_i I_i^r}{N_i} - \frac{\mu \beta S_i I_i^u}{N_i} \quad (4)$$

$$\frac{dE_i}{dt} = \frac{\beta S_i I_i^r}{N_i} + \frac{\mu \beta S_i I_i^u}{N_i} - \frac{E_i}{Z} \quad (5)$$

$$\frac{dI_i^s}{dt} = \alpha \frac{E_i}{Z} - \frac{I_i^s}{D} \quad (6)$$

$$\frac{dI_i^a}{dt} = (1 - \alpha) \frac{E_i}{Z} - \frac{I_i^a}{D} \quad (7)$$

277 where S_i, E_i, I_i^s, I_i^a are the susceptible, exposed, infected symptomatic, infected asymptomatic, and
 278 total population in a county i .

279 The synthetic mobility datasets were integrated into the previous system (Eq.4-7) and documented
 280 [11] by the following equations (Eq. 8-12),

$$\frac{dS_i}{dt} = -\frac{\beta S_i I_i^r}{N_i} - \frac{\mu \beta S_i I_i^u}{N_i} + \theta \sum_j \frac{M_{ij} S_j}{N_j - I_j^r} - \theta \sum_j \frac{M_{ji} S_i}{N_j - I_j^r} \quad (8)$$

$$\frac{dE_i}{dt} = \frac{\beta S_i I_i^r}{N_i} + \frac{\mu \beta S_i I_i^u}{N_i} - \frac{E_i}{Z} + \theta \sum_j \frac{M_{ij} E_j}{N_j - I_j^r} - \theta \sum_j \frac{M_{ji} E_i}{N_j - I_j^r} \quad (9)$$

$$\frac{dI_i^r}{dt} = \alpha \frac{E_i}{Z} - \frac{I_i^r}{D} \quad (10)$$

$$\frac{dI_i^u}{dt} = (1 - \alpha) \frac{E_i}{Z} - \frac{I_i^u}{D} + \theta \sum_j \frac{M_{ij} I_j^u}{N_j - I_j^r} - \theta \sum_j \frac{M_{ji} I_i^u}{N_j - I_j^r} \quad (11)$$

$$N_i = N_i + \theta \sum_j M_{ij} - \theta \sum_j M_{ji} \quad (12)$$

281 where S_i, E_i, I_i^r, I_i^u are the susceptible, exposed, documented infected, undocumented infected, and
 282 total population in a county i .

283 The system of equations (Eq. 8-12) thus took into account both the mobility and the contagion
284 describing the epidemic's evolution on the metapopulations network. We assumed that the
285 randomness in the contagion followed a Poisson distribution and was documented elsewhere ¹².
286 Most specifically, we seeded cases in a specific location, then, for each time t , the disease spread
287 through the metapopulation network according to the transition matrix when people are moving
288 between counties from the first day to the 500th day.

289 [insert Table 1 here]

290 *Epidemiological metrics*

291 In reviewing epidemiological models using mobility data, we identified salient metrics of interest,
292 including probability of infection [12], risk of importation [13], incubation period [15], mean
293 importation rate [14], size at the epidemic peak [29], effective reproduction number [11], epidemic
294 size [22], proportion of counties with at least one case [30], rate of spread [31], timing of the peak
295 [32] and the average size of the peak [33].

296 *Epidemiological scenarios*

297 To assess the effect of noise on these metrics, we evaluated eight scenarios with three salient
298 characteristics and provided a general formula to incorporate more. We evaluated scenarios where
299 the epidemiological metrics of interest were driven by i) the location of the first case, ii) changes
300 in connectivity, and iii) changes in epidemiological parameters (Table 2).

301 [insert Table 2 here]

302 To assess the impact of privacy on the epidemiological metric, we ran each set of parameters
303 through 1000 Monte Carlo iterations and visualised the results.

304 **Data and Code Availability**

305 Data and codes are available at https://github.com/crisisready/DP_Metapopulation

306

307 **References**

- 308 1. Oliver N, Lepri B, Sterly H, Lambiotte R, Deletaille S, De Nadai M, et al. Mobile phone data
309 for informing public health actions across the COVID-19 pandemic life cycle. *Sci Adv.*
310 2020;6: eabc0764. doi:10.1126/sciadv.abc0764
- 311 2. Wu W, Niu X. Influence of Built Environment on Urban Vitality: Case Study of Shanghai
312 Using Mobile Phone Location Data. *J Urban Plan Dev.* 2019;145: 04019007.
313 doi:10.1061/(ASCE)UP.1943-5444.0000513
- 314 3. Yabe T, Jones NKW, Rao PSC, Gonzalez MC, Ukkusuri SV. Mobile phone location data for
315 disasters: A review from natural hazards and epidemics. *Comput Environ Urban Syst.*
316 2022;94: 101777. doi:10.1016/j.compenvurbsys.2022.101777
- 317 4. Grantz KH, Meredith HR, Cummings DAT, Metcalf CJE, Grenfell BT, Giles JR, et al. The
318 use of mobile phone data to inform analysis of COVID-19 pandemic epidemiology. *Nat*
319 *Commun.* 2020;11: 4961. doi:10/ghtqmf
- 320 5. Wesolowski A, Qureshi T, Boni MF, Sundsøy PR, Johansson MA, Rasheed SB, et al. Impact
321 of human mobility on the emergence of dengue epidemics in Pakistan. *Proc Natl Acad Sci.*
322 2015;112: 11887–11892. doi:10/gg2nc8
- 323 6. Fiadino P, Ponce-Lopez V, Antonio J, Torrent-Moreno M, D’Alconzo A. Call Detail Records
324 for Human Mobility Studies: Taking Stock of the Situation in the “Always Connected Era.”
325 Proceedings of the Workshop on Big Data Analytics and Machine Learning for Data
326 Communication Networks. New York, NY, USA: Association for Computing Machinery;
327 2017. pp. 43–48. doi:10.1145/3098593.3098601
- 328 7. Wesolowski A, Metcalf CJE, Eagle N, Kombich J, Grenfell BT, Bjørnstad ON, et al.
329 Quantifying seasonal population fluxes driving rubella transmission dynamics using mobile
330 phone data. *Proc Natl Acad Sci.* 2015;112: 11114–11119. doi:10/f7qntk
- 331 8. Rocher L, Hendrickx JM, de Montjoye Y-A. Estimating the success of re-identifications in
332 incomplete datasets using generative models. *Nat Commun.* 2019;10: 3069.
333 doi:10.1038/s41467-019-10933-3
- 334 9. Dwork C, Roth A. The Algorithmic Foundations of Differential Privacy. *Found Trends@*
335 *Theor Comput Sci.* 2014;9: 211–407. doi:10.1561/04000000042

- 336 10. OpenDP. SmartNoise - OpenDP SmartNoise. Available:
337 <https://docs.smartnoise.org/en/stable/index.html>
- 338 11. Li R, Pei S, Chen B, Song Y, Zhang T, Yang W, et al. Substantial undocumented infection
339 facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2). *Science*. 2020;368:
340 489–493. doi:10/ggn6c2
- 341 12. Schaber KL, Perkins TA, Lloyd AL, Waller LA, Kitron U, Paz-Soldan VA, et al. Disease-
342 driven reduction in human mobility influences human-mosquito contacts and dengue
343 transmission dynamics. *PLOS Comput Biol*. 2021;17: e1008627.
344 doi:10.1371/journal.pcbi.1008627
- 345 13. Chang S, Pierson E, Koh PW, Gerardin J, Redbird B, Grusky D, et al. Mobility network
346 models of COVID-19 explain inequities and inform reopening. *Nature*. 2021;589: 82–87.
347 doi:10/gjhmt2
- 348 14. Tizzoni M, Bajardi P, Decuyper A, King GKK, Schneider CM, Blondel V, et al. On the Use
349 of Human Mobility Proxies for Modeling Epidemics. *PLOS Comput Biol*. 2014;10:
350 e1003716. doi:10.1371/journal.pcbi.1003716
- 351 15. Kraemer MUG, Yang C-H, Gutierrez B, Wu C-H, Klein B, Pigott DM, et al. The effect of
352 human mobility and control measures on the COVID-19 epidemic in China. *Science*.
353 2020;368: 493–497. doi:10.1126/science.abb4218
- 354 16. Yang X, Fienberg SE, Rinaldo A. Differential Privacy for Protecting Multi-dimensional
355 Contingency Table Data: Extensions and Applications. *J Priv Confidentiality*. 2012;4: 101–
356 125.
- 357 17. Dwork C, McSherry F, Nissim K, Smith A. Calibrating Noise to Sensitivity in Private Data
358 Analysis. In: Halevi S, Rabin T, editors. *Theory of Cryptography*. Berlin, Heidelberg:
359 Springer; 2006. pp. 265–284. doi:10.1007/11681878_14
- 360 18. OpenDP. [cited 23 Oct 2021]. Available: <https://opendp.org/home>
- 361 19. U.S. Census Bureau QuickFacts: United States. [cited 11 Mar 2023]. Available:
362 <https://www.census.gov/quickfacts/fact/table/US#>
- 363 20. Calvetti D, Hoover AP, Rose J, Somersalo E. Metapopulation Network Models for
364 Understanding, Predicting, and Managing the Coronavirus Disease COVID-19. *Front Phys*.
365 2020;8. Available: <https://www.frontiersin.org/articles/10.3389/fphy.2020.00261>
- 366 21. Coletti P, Libin P, Petrof O, Willem L, Abrams S, Herzog SA, et al. A data-driven
367 metapopulation model for the Belgian COVID-19 epidemic: assessing the impact of
368 lockdown and exit strategies. *BMC Infect Dis*. 2021;21: 503. doi:10.1186/s12879-021-
369 06092-w

- 370 22. Balcan D, Colizza V, Gonçalves B, Hu H, Ramasco JJ, Vespignani A. Multiscale mobility
371 networks and the spatial spreading of infectious diseases. *Proc Natl Acad Sci.* 2009;106:
372 21484–21489. doi:10.1073/pnas.0906910106
- 373 23. Houssiau F, Rocher L, de Montjoye Y-A. On the difficulty of achieving Differential Privacy
374 in practice: user-level guarantees in aggregate location data. *Nat Commun.* 2022;13: 29.
375 doi:10.1038/s41467-021-27566-0
- 376 24. de Montjoye Y-A, Gambs S, Blondel V, Canright G, de Cordes N, Deletaille S, et al. On the
377 privacy-conscientious use of mobile phone data. *Sci Data.* 2018;5: 180286.
378 doi:10.1038/sdata.2018.286
- 379 25. Bassolas A, Barbosa-Filho H, Dickinson B, Dotiwalla X, Eastham P, Gallotti R, et al.
380 Hierarchical organization of urban mobility and its connection with city livability. *Nat*
381 *Commun.* 2019;10: 4817. doi:10.1038/s41467-019-12809-y
- 382 26. Savi MK, Yavad A, Vembar N, Kishore N. A standardized differential privacy framework
383 for epidemiological modeling with mobile phone data. 2022. Available:
384 https://github.com/crisisready/DP_Metapopulation
- 385 27. Kishore N, Kiang MV, Engø-Monsen K, Vembar N, Schroeder A, Balsari S, et al. Measuring
386 mobility to monitor travel and physical distancing interventions: a common framework for
387 mobile phone data analysis. *Lancet Digit Health.* 2020;2: e622–e628. doi:10/gg96w9
- 388 28. Pereira M, Kim A, Allen J, White K, Ferres JL, Dodhia R. U.S. Broadband Coverage Data
389 Set: A Differentially Private Data Release. *arXiv*; 2021. doi:10.48550/arXiv.2103.14035
- 390 29. Zhou Y, Xu R, Hu D, Yue Y, Li Q, Xia J. Effects of human mobility restrictions on the spread
391 of COVID-19 in Shenzhen, China: a modelling study using mobile phone data. *Lancet Digit*
392 *Health.* 2020;2: e417–e424. doi:10/gg78fk
- 393 30. Souch JM, Cossman JS, Hayward MD. Interstates of Infection: Preliminary Investigations of
394 Human Mobility Patterns in the COVID-19 Pandemic. *J Rural Health.* 2021;37: 266–271.
395 doi:10.1111/jrh.12558
- 396 31. Kishore N, Kahn R, Martinez PP, Salazar PMD, Mahmud AS, Buckee CO. Lockdown related
397 travel behavior undermines the containment of SARS-CoV-2. *medRxiv.* 2020;
398 2020.10.22.20217752. doi:10/ghts2k
- 399 32. Balcan D, Hu H, Goncalves B, Bajardi P, Poletto C, Ramasco JJ, et al. Seasonal transmission
400 potential and activity peaks of the new influenza A(H1N1): a Monte Carlo likelihood analysis
401 based on human mobility. *BMC Med.* 2009;7: 45. doi:10.1186/1741-7015-7-45
- 402 33. Lenormand M, Louail T, Cantú-Ros OG, Picornell M, Herranz R, Arias JM, et al. Influence
403 of sociodemographic characteristics on human mobility. *Sci Rep.* 2015;5: 10075.
404 doi:10.1038/srep10075

405

406 **Author contributions**

407 Conceptualization: C.B., N.K., A.S., S.V., N.V.

408 Data curation & Formal analysis: M.K.S, A.Y, W.Z, N.V, N.K

409 Funding acquisition: S.B, C.B, S.V.

410 Supervision: N.V, A.S, S.B, C.B, S.V, N.K

411 Writing- Original draft: M.K.S, N.K

412 Writing- review & editing: M.K.S, A.Y, W.Z, N.V, A.S, S.B, C.B, S.V, N.K

413 **Conflict of interest**

414 All the authors declare having no competing or conflicting interest.

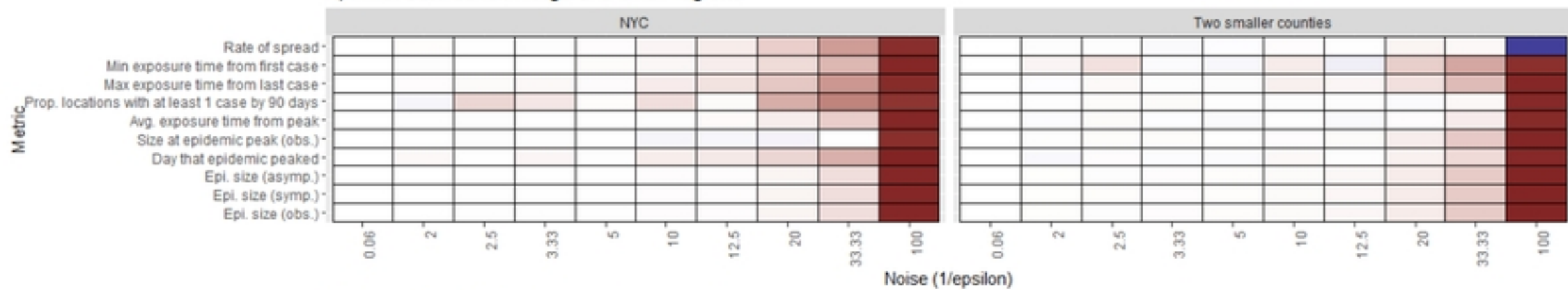
415 **Ethical statements**

416 All methods were performed in accordance with the relevant guidelines and regulations.

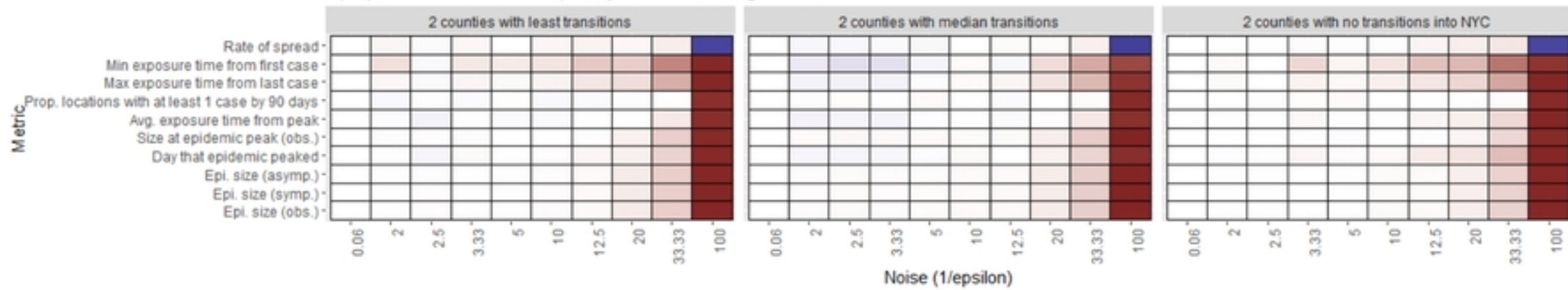
417 **Fundings**

418 This work was supported by a Trust in Science Grant awarded by the Harvard Data Science
419 Initiative.

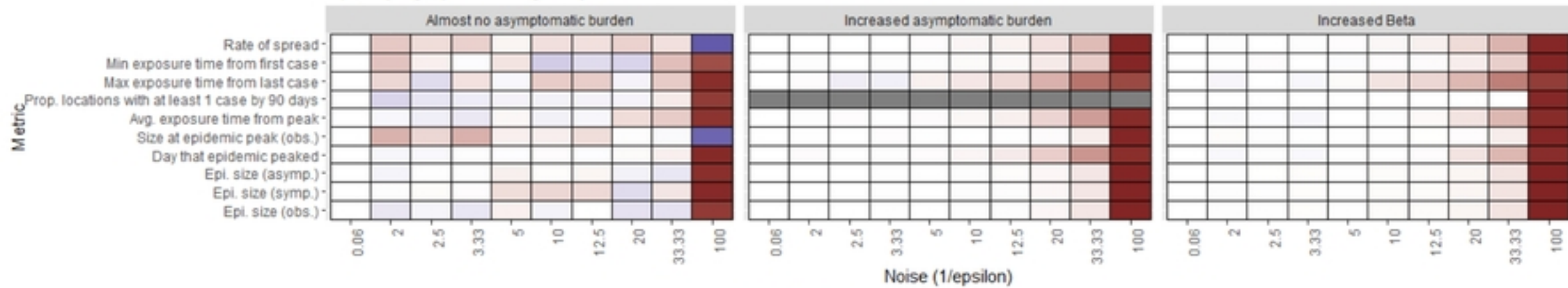
A) Initial outbreaks in large and small regions



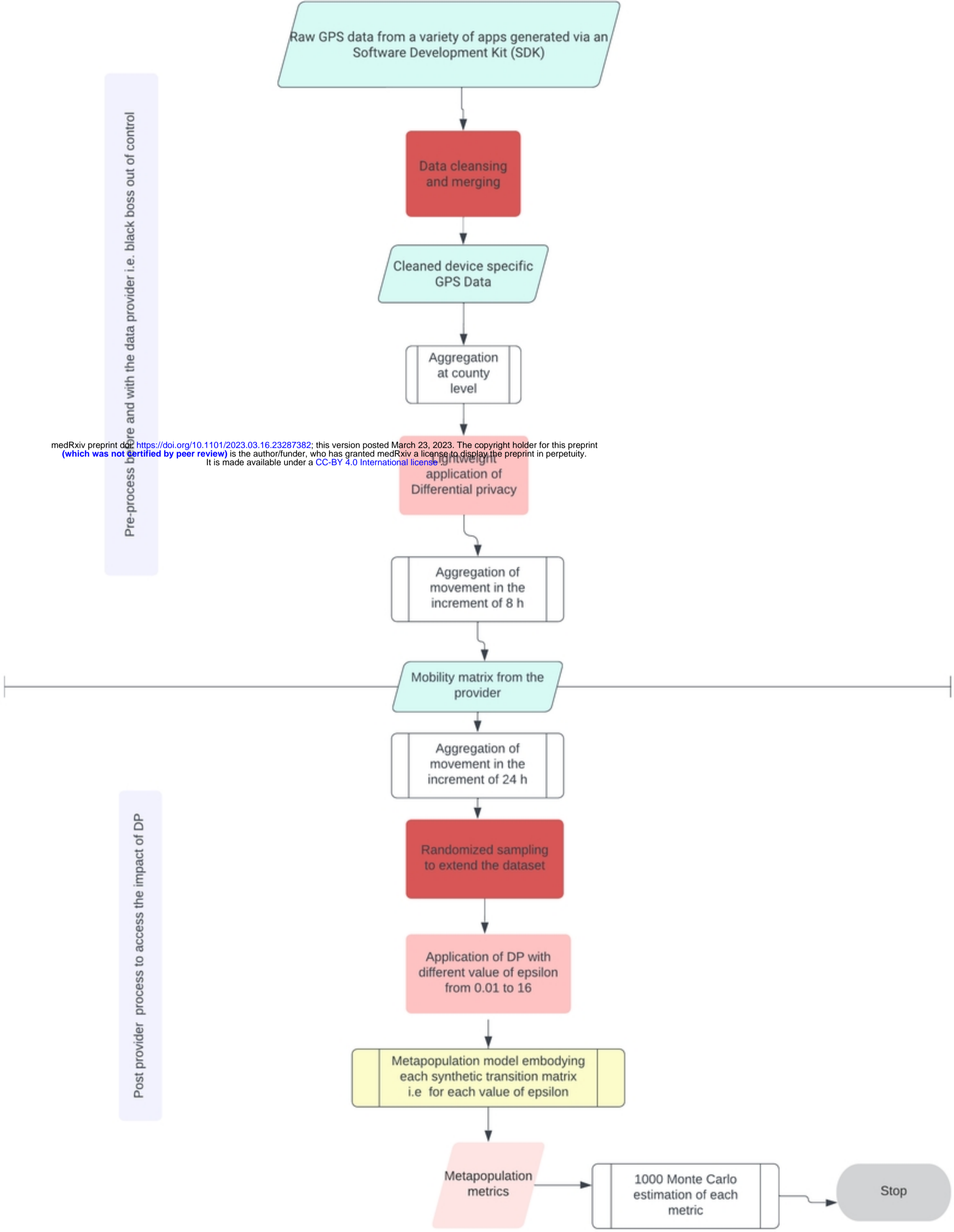
B) Epidemics in well and poorly-connected regions



C) Varying epidemiological parameters



Figure



medRxiv preprint doi: <https://doi.org/10.1101/2023.03.16.23287382>; this version posted March 23, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a [CC-BY 4.0 International license](https://creativecommons.org/licenses/by/4.0/).

Figure