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

Merveille Koissi Savi[1], Akash Yadav[2], Wanrong Zhang[3], Navin Vembar[4], Andrew Schroeder[2], Satchit Balsari[5], Caroline O. Buckee[6], Salil Vadhan[3], Nishant Kishore[6*]

[1] Dana Farber Cancer Institute, Harvard School of Medicine, Department of Medical Oncology, MA, USA

[2]Direct Relief, Santa Barbara, CA, USA

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

[4]Camber Systems, Washington, DC, USA

[5]Department of Emergency Medicine, Harvard Medical School, Boston, USA

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

*Corresponding author: nish.kishore@gmail.com

Abstract

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

Author Summary

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

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

Introduction

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

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

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

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

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

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

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

Results

Mobility data

The mobility matrices included data from August 15 to November 15, 2020, and contained a total of 812,587 transitions made between sixty-two counties of New York State, with a mean of 9,029 transitions a day. The observed daily transitions ranged from a minimum of 600, occurring in 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 75^{th} 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.

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

Scenarios with Epidemics in Well- and Poorly Connected Regions

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

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

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

Scenarios with varying epidemiological parameters

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

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

[insert Fig. 1 here]

Discussion

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

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

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

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

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

Methods

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

[insert Fig. 2 here]

Mobility Data

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

Application of Differential Privacy

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

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

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

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

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

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

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

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

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

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

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

Metapopulation model

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

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

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

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

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

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

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

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

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

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

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

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

where S_i, E_i, I_i^r, I_i^u are the susceptible, exposed, documented infected, undocumented infected, and 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.

Data and Code Availability

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

References

1. Oliver N, Lepri B, Sterly H, Lambiotte R, Deletaille S, De Nadai M, et al. Mobile phone data for informing public health actions across the COVID-19 pandemic life cycle. Sci Adv. 2020;6: eabc0764. doi:10.1126/sciadv.abc0764

2. Wu W, Niu X. Influence of Built Environment on Urban Vitality: Case Study of Shanghai Using Mobile Phone Location Data. J Urban Plan Dev. 2019;145: 04019007. doi:10.1061/(ASCE)UP.1943-5444.0000513

3. Yabe T, Jones NKW, Rao PSC, Gonzalez MC, Ukkusuri SV. Mobile phone location data for disasters: A review from natural hazards and epidemics. Comput Environ Urban Syst. 2022;94: 101777. doi:10.1016/j.compenvurbsys.2022.101777

4. Grantz KH, Meredith HR, Cummings DAT, Metcalf CJE, Grenfell BT, Giles JR, et al. The use of mobile phone data to inform analysis of COVID-19 pandemic epidemiology. Nat Commun. 2020;11: 4961. doi:10/ghtqmf

5. Wesolowski A, Qureshi T, Boni MF, Sundsøy PR, Johansson MA, Rasheed SB, et al. Impact of human mobility on the emergence of dengue epidemics in Pakistan. Proc Natl Acad Sci. 2015;112: 11887–11892. doi:10/gg2nc8

6. Fiadino P, Ponce-Lopez V, Antonio J, Torrent-Moreno M, D'Alconzo A. Call Detail Records for Human Mobility Studies: Taking Stock of the Situation in the "Always Connected Era." Proceedings of the Workshop on Big Data Analytics and Machine Learning for Data Communication Networks. New York, NY, USA: Association for Computing Machinery; 2017. pp. 43–48. doi:10.1145/3098593.3098601

7. Wesolowski A, Metcalf CJE, Eagle N, Kombich J, Grenfell BT, Bjørnstad ON, et al. Quantifying seasonal population fluxes driving rubella transmission dynamics using mobile phone data. Proc Natl Acad Sci. 2015;112: 11114–11119. doi:10/f7qntk

8. Rocher L, Hendrickx JM, de Montjoye Y-A. Estimating the success of re-identifications in incomplete datasets using generative models. Nat Commun. 2019;10: 3069. doi:10.1038/s41467-019-10933-3

9. Dwork C, Roth A. The Algorithmic Foundations of Differential Privacy. Found Trends® Theor Comput Sci. 2014;9: 211–407. doi:10.1561/0400000042

10. OpenDP. SmartNoise - OpenDP SmartNoise. Available: https://docs.smartnoise.org/en/stable/index.html

11. Li R, Pei S, Chen B, Song Y, Zhang T, Yang W, et al. Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2). Science. 2020;368: 489–493. doi:10/ggn6c2

12. Schaber KL, Perkins TA, Lloyd AL, Waller LA, Kitron U, Paz-Soldan VA, et al. Disease-driven reduction in human mobility influences human-mosquito contacts and dengue transmission dynamics. PLOS Comput Biol. 2021;17: e1008627. doi:10.1371/journal.pcbi.1008627

13. Chang S, Pierson E, Koh PW, Gerardin J, Redbird B, Grusky D, et al. Mobility network models of COVID-19 explain inequities and inform reopening. Nature. 2021;589: 82–87. doi:10/ghjmt2

14. Tizzoni M, Bajardi P, Decuyper A, King GKK, Schneider CM, Blondel V, et al. On the Use of Human Mobility Proxies for Modeling Epidemics. PLOS Comput Biol. 2014;10: e1003716. doi:10.1371/journal.pcbi.1003716

15. Kraemer MUG, Yang C-H, Gutierrez B, Wu C-H, Klein B, Pigott DM, et al. The effect of human mobility and control measures on the COVID-19 epidemic in China. Science. 2020;368: 493–497. doi:10.1126/science.abb4218

16. Yang X, Fienberg SE, Rinaldo A. Differential Privacy for Protecting Multi-dimensional Contingency Table Data: Extensions and Applications. J Priv Confidentiality. 2012;4: 101–125.

17. Dwork C, McSherry F, Nissim K, Smith A. Calibrating Noise to Sensitivity in Private Data Analysis. In: Halevi S, Rabin T, editors. Theory of Cryptography. Berlin, Heidelberg: Springer; 2006. pp. 265–284. doi:10.1007/11681878_14

18. OpenDP. [cited 23 Oct 2021]. Available: https://opendp.org/home

19. U.S. Census Bureau QuickFacts: United States. [cited 11 Mar 2023]. Available: https://www.census.gov/quickfacts/fact/table/US#

20. Calvetti D, Hoover AP, Rose J, Somersalo E. Metapopulation Network Models for Understanding, Predicting, and Managing the Coronavirus Disease COVID-19. Front Phys. 2020;8. Available: https://www.frontiersin.org/articles/10.3389/fphy.2020.00261

21. Coletti P, Libin P, Petrof O, Willem L, Abrams S, Herzog SA, et al. A data-driven metapopulation model for the Belgian COVID-19 epidemic: assessing the impact of lockdown and exit strategies. BMC Infect Dis. 2021;21: 503. doi:10.1186/s12879-021-06092-w

22. Balcan D, Colizza V, Gonçalves B, Hu H, Ramasco JJ, Vespignani A. Multiscale mobility networks and the spatial spreading of infectious diseases. Proc Natl Acad Sci. 2009;106: 21484–21489. doi:10.1073/pnas.0906910106

23. Houssiau F, Rocher L, de Montjoye Y-A. On the difficulty of achieving Differential Privacy in practice: user-level guarantees in aggregate location data. Nat Commun. 2022;13: 29. doi:10.1038/s41467-021-27566-0

24. de Montjoye Y-A, Gambs S, Blondel V, Canright G, de Cordes N, Deletaille S, et al. On the privacy-conscientious use of mobile phone data. Sci Data. 2018;5: 180286. doi:10.1038/sdata.2018.286

25. Bassolas A, Barbosa-Filho H, Dickinson B, Dotiwalla X, Eastham P, Gallotti R, et al. Hierarchical organization of urban mobility and its connection with city livability. Nat Commun. 2019;10: 4817. doi:10.1038/s41467-019-12809-y

26. Savi MK, Yavad A, Vembar N, Kishore N. A standardized differential privacy framework for epidemiological modeling with mobile phone data. 2022. Available: https://github.com/crisisready/DP_Metapopulation

27. Kishore N, Kiang MV, Engø-Monsen K, Vembar N, Schroeder A, Balsari S, et al. Measuring mobility to monitor travel and physical distancing interventions: a common framework for mobile phone data analysis. Lancet Digit Health. 2020;2: e622–e628. doi:10/gg96w9

28. Pereira M, Kim A, Allen J, White K, Ferres JL, Dodhia R. U.S. Broadband Coverage Data Set: A Differentially Private Data Release. arXiv; 2021. doi:10.48550/arXiv.2103.14035

29. Zhou Y, Xu R, Hu D, Yue Y, Li Q, Xia J. Effects of human mobility restrictions on the spread of COVID-19 in Shenzhen, China: a modelling study using mobile phone data. Lancet Digit Health. 2020;2: e417–e424. doi:10/gg78fk

30. Souch JM, Cossman JS, Hayward MD. Interstates of Infection: Preliminary Investigations of Human Mobility Patterns in the COVID-19 Pandemic. J Rural Health. 2021;37: 266–271. doi:10.1111/jrh.12558

31. Kishore N, Kahn R, Martinez PP, Salazar PMD, Mahmud AS, Buckee CO. Lockdown related travel behavior undermines the containment of SARS-CoV-2. medRxiv. 2020; 2020.10.22.20217752. doi:10/ghts2k

32. Balcan D, Hu H, Goncalves B, Bajardi P, Poletto C, Ramasco JJ, et al. Seasonal transmission potential and activity peaks of the new influenza A(H1N1): a Monte Carlo likelihood analysis based on human mobility. BMC Med. 2009;7: 45. doi:10.1186/1741-7015-7-45

33. Lenormand M, Louail T, Cantú-Ros OG, Picornell M, Herranz R, Arias JM, et al. Influence of sociodemographic characteristics on human mobility. Sci Rep. 2015;5: 10075. doi:10.1038/srep10075

Author contributions

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

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

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

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

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

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

Conflict of interest

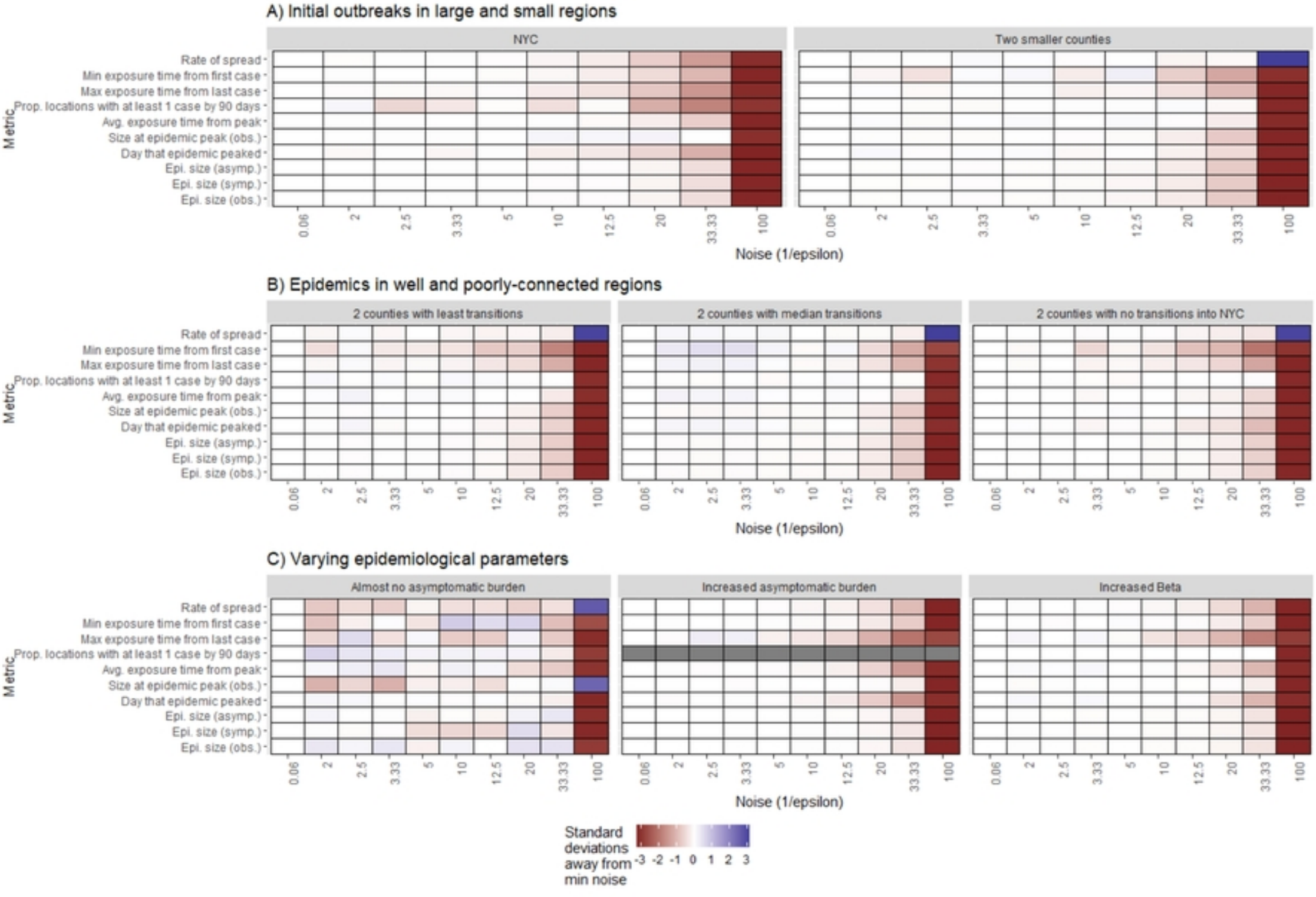
All the authors declare having no competing or conflicting interest.

Ethical statements

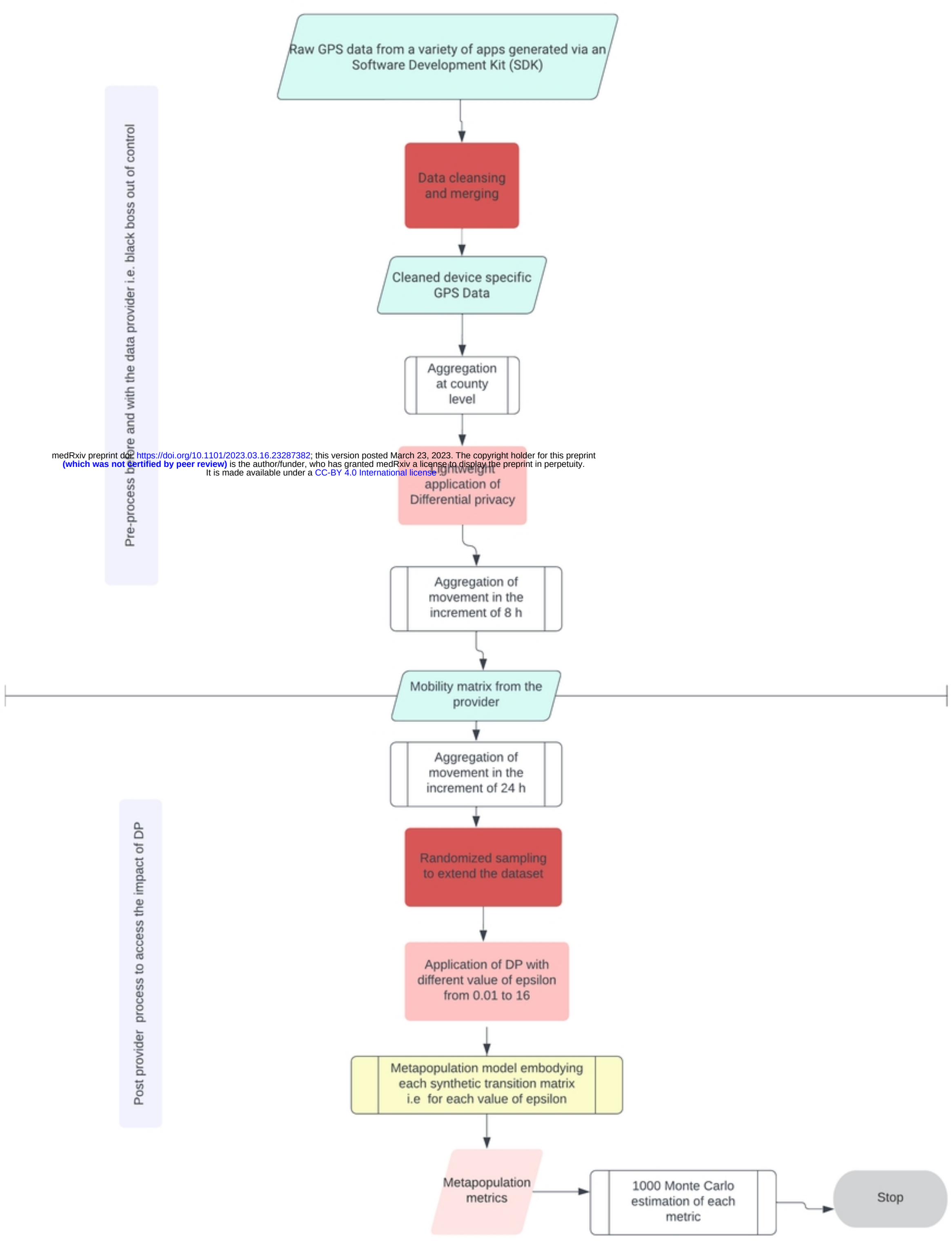
All methods were performed in accordance with the relevant guidelines and regulations.

Fundings

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


A) Initial outbreaks in large and small regions
NYC
Two smaller counties
B) Epidemics in well and poorly-connected regions
2 counties with least transitions
2 counties with median transitions
2 counties with no transitions into NYC
C) Varying epidemiological parameters
Almost no asymptomatic burden
Increased asymptomatic burden
Increased Beta
Metric
Rate of spread
Min exposure time from first case
Max exposure time from last case
Prop. locations with at least 1 case by 90 days
Avg. exposure time from peak
Size at epidemic peak (obs.)
Day that epidemic peaked
Epi. size (asymp.)
Epi. size (symp.)
Epi. size (obs.)
0.06
2
2.5
3.33
5
10
12.5
20
33.33
100
Noise (1/epsilon)
Standard deviations away from min noise
-3 -2 -1 0 1 2 3

Figure



Figure