Data and model needs for generalizable inferences linking human mobility and infectious disease transmission

Natalya Kostandova[+] (1), Ronan Corgel (1), Shweta Bansal (2), Sophie Bérubé (1), Eimear Cleary (3), Chelsea Hansen (4, 5), Matt D.T. Hitchings (6, 7), Bernardo García-Carreras (8), Lauren Gardner (9, 1), Moritz U G Kraemer (10, 11), Shengjie Lai (3), Yao Li (1), Amanda C. Perofsky (4, 5), Giulia Pullano (2), Jonathan M Read (12), Gabriel Ribeiro dos Santos (13), Henrik Salje (13), Saki Takahashi (1), Cécile Viboud (4), Jasmine Wang (1), Derek AT Cummings* (6, 7), Amy Wesolowski* (1)

((1) Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA, (2) Department of Biology, Georgetown University, Washington, District of Columbia, USA, (3) WorldPop, School of Geography and Environmental Science, University of Southampton, Southampton SO17 1BJ, UK, (4) Fogarty International Center, National Institutes of Health, Bethesda, MD, USA, (5) Brotman Baty Institute, University of Washington, Seattle, Washington, USA, (6) Department of Biostatistics, University of Florida, Gainesville, USA, (7) Emerging Pathogens Institute, University of Florida, Gainesville, USA, (8) Department of Biology, University of Florida, Gainesville, FL, USA, (9) Department of Civil and Systems Engineering, Johns Hopkins University, Baltimore, MD, 21218, USA, (10) Department of Biology, University of Oxford, Oxford, United Kingdom, (11) Pandemic Sciences Institute, University of Oxford, Oxford, United Kingdom, (12) Centre for Health Information Computation and Statistics, Lancaster Medical School, Lancaster University, Lancaster, United Kingdom, (13) Department of Genetics, University of Cambridge, Cambridge CB2 3EH, UK

[+]Corresponding author: Natalya Kostandova, nkostan1@jh.edu

Standfirst

Mobility data have been used to reconstruct infectious disease dynamics and tailor control and elimination measures. We describe three challenges in using these data and identify opportunities to leverage existing resources, improving our understanding of human mobility. We propose reporting guidelines to increase the interpretation, standardization, and reuse of existing mobility data sets.

Main Text

Quantifying human mobility is critical to developing a more complete understanding of how travel and contacts between populations facilitate the spread of infectious pathogens. Information about human mobility has been used in reconstructing and predicting transmission dynamics, and determining the effectiveness of control strategies for a wide range of diseases, including influenza, malaria, Ebola, cholera [1], and SARS-CoV-2 [2]. Since the beginning of the COVID-19 pandemic, there has been an increased availability of data describing different aspects of human mobility. However, more work is needed to address the challenges and opportunities of evaluating how, when, where, and for which populations generalizable insights on mobility can be applied to various epidemiological scenarios to prepare for future pandemic and epidemic threats.

Challenge 1: Identifying the most relevant proxies of potentially infectious interactions and risk. Along the spectrum of pathogen transmission states from pre-emergence to elimination, mobility data can provide valuable information about how and where contact potentially occurs between susceptible and infected individuals (Table 1). These data have been applied for multiple purposes, such as using population commuting information to refine estimates of community-level vaccine coverage [3], or using mobile phone data to quantify mobility between administrative units in Brazil in a metapopulation model used to calculate

schistosomiasis prevalence [4]. The specific modeling purpose provides guidance on the type, granularity (e.g., temporal, individual versus population, spatial), and overlap (e.g., from the same locations, time frame, population) of mobility data needed. However, little empirical evidence exists in the literature on mobility datasets that can be considered the gold standard that are validated against observational data on the disease of interest. With no consensus gold standard mobility dataset, understanding when, where, and for whom mobility data are available will be context-specific and difficult to generalize. Evaluating mobility data against measures of disease transmission is further complicated in geographic areas where disease data are sparse, biased, or unavailable. In addition, many of the underlying disease processes, like successful transmission events, are highly stochastic and rarely observed. Proxy measures and ecological analyses are often used to estimate infection status; however, such approaches can mask variability in the association between disease and mobility within the unit of analysis, potentially resulting in ecological bias [5]. While ideal datasets used in analyses depend on the questions being asked and the populations of interest, a better understanding of how to map mobility data to transmission-relevant behaviors is needed.

Challenge 2: Harmonization, integration, and availability of multiple data sources. Often, 'mobility' is considered a catch-all term and is used as a proxy measure for the connectivity between locations, or to encompass behaviors that may expose a susceptible individual to an infectious individual. However, different mobility datasets report a wide range of mobility-related behaviors, including the number of trips between pairs of locations, clustering of individuals in a specific location, percentage of devices staying home, and contact rates, among others. The way data are collected (e.g., self-response on a questionnaire, tracking GPS location on a mobile phone) also varies across data sets. Often, the temporal and spatial aggregation processes used for different datasets vary in their approach, and many datasets are censored based on a minimum population number for a particular spatial resolution to minimize privacy concerns. These differences make it challenging to harmonize and compare across data types. With growing data availability in the public domain, it may be possible to combine different datasets to create single estimates of mobility or for example translate estimates from contact data to flows. However, to date, there has been limited research and evaluation of which statistical and mathematical approaches, including simulation, should be used to integrate and evaluate disparate data sets.

Challenge 3: Accounting for sampling and measurement bias. While data access has greatly expanded, there remain gaps in our understanding of mobility across locations and populations. For example, few data sets include information on the movement of children or the very elderly or describe mobility patterns in many low- and middle-income settings, such as parts of Sub-Saharan Africa [6]. Data collected by mobile phone providers require mobile phone ownership and use, while network penetration rates vary by geographical location. Some mobility data obtained via social media applications rely on a combination of internet use, smartphone access, and opting in to share location information. These may result in sampling bias, when access to and use of technologies used to create the data sets is associated with mobility patterns. When data are available for a population or location, they often lack additional metadata to describe differences in mobility across groups or individuals, such as by age, sex, or socio-economic status [7]. Finally, data may only capture one aspect of travel, such as trip counts, and ignore other aspects, such as trip frequency or duration.

Despite these challenges, researchers have multiple opportunities to leverage existing resources to provide a step change in our understanding of human mobility and ways these insights can be used to understand infectious disease transmission and control measures.

Opportunity 1: Use simulation and modeling to better understand data needs and uses across pathogens. Simulation and modeling of both changes in mobility behavior and transmission dynamics can help us to better understand how various mobility data sets may result in different spatio-temporal disease dynamics. Assessing multiple competing models of mobility in infectious disease models can help evaluate the utility of data over a null model or assumption, i.e., including no connectivity between populations or a basic spatial interaction model [8]. Further evaluation of a hierarchy of model complexity, i.e., from basic assumptions of no mixing to

simple non-parameterized spatial interaction models to parameterized mobility models, can provide insight into how and what aspects of human behavior drive transmission. Finally, this approach can serve as an additional methodology for propagating uncertainty throughout transmission-modeled simulations by allowing for the evaluation of uncertainty due to model misspecification. Using simulation to road-test mobility data and develop better-informed estimates of connectivity driving disease transmission can help build a more generalized understanding of human mobility for future applications. For example, simulations can be used to explore the impact of data censoring and aggregation on modeled transmission dynamics, guiding data needs and requirements for particular use cases. As an illustration, in Figure 1, we display the results from a simple stochastic, discrete-time metapopulation compartmental model, where we explored the impact of demographic bias, censoring, and temporal and spatial aggregation on the dynamics of predicted epidemics. The simulations demonstrate that these biases and processing approaches can have significant impact on estimates of disease arrival times in simulations of disease transmission. Detailed information about the simulation, parameter values, and code used are available on Github. Explorations can be conducted to examine how these factors would impact disease inference using simulation as a guiding principle.

Opportunity 2: Reporting guidelines to increase the interpretation, standardization, and reuse of existing mobility data sets. Interpreting results from models that integrate mobility data, and determining how these data may be relevant for future applications, can benefit from a framework that makes it easy to understand how and what type of data were used. However, there are no established guidelines or systematic structure for reporting how mobility data are pre-processed, incorporated into infectious disease models, and reported in the results of analyses. This makes it difficult not only to understand the way mobility data were used and to compare results across published research, but also to reproduce results. Here, we propose a specific set of reporting guidelines to be included by researchers using mobility data (Supplement 1). These guidelines cover key components of mobility data origin, analysis, and use, including a description of raw mobility data, pre-processing by the data provider, processing of the data carried out by the researcher, how mobility data were integrated into transmission analyses, and the results that should be presented. Use of these guidelines will allow for consistency in presenting mobility data, easier communication of results, and improved understanding of the quality and scope of mobility data used. Guidelines can also increase reproducibility and enable the integration of multiple data sets, which can, in turn, provide opportunities for identifying and accounting for bias due to sampling or measurement between datasets. Building a framework off of the FAIR principles could further help mitigate future problems.

Opportunity 3: Investment in data repositories to ensure continuity of data access. Following a dramatic increase in availability of infection and mobility data from various sources during the SARS-CoV-2 pandemic, data access has since become increasingly limited or costly. For example, updates to Data for Good at Meta: Mobility Data, a public data source providing county-level information on relative travel, were discontinued as of December 31, 2020 [9]. Similarly, Google COVID-19 Mobility Reports are no longer updated as of October 15, 2022, while SafeGraph stopped providing social distancing metrics, including percent time staying at home, in April 2021. There is a need, however, to maintain the availability and ease of access to historical data. The use of repositories and specialized R packages could make facilitating access and availability easier. However, continued access requires advocacy with governments and other decision-makers. This must address the barriers to making the data available, such as funding human resources necessary to collect, maintain, and manage the data, and should use evidence to make the case for why data should continue to be collected and/or collated. To avoid inappropriate allocation of resources, we must identify the most important data to collect, and use a targeted approach in collecting it. Investment in data repositories would allow for better assessment of data gaps to target empirical resources to populations that are under-represented in datasets but key to understanding transmission dynamics. Central repositories could also pave the way to expanding data and resource access beyond a select set of research groups, which has historically been the case due to proprietary access.

Since the beginning of the COVID-19 pandemic, there was an unprecedented increase in access to, and use of, data describing human mobility patterns to evaluate and guide public health interventions. Now, there is a need to leverage what resources were made available during the COVID-19 pandemic to help better plan and respond to future disease threats. To accomplish these goals, many of the challenges of access, transparency, and reproducibility should be addressed, as in genomic data analyses and the push for open and easily accessible data [10]. Beyond access, there are methodological challenges to concretely tie together mobility and transmission that could be strengthened using simulation, reporting guidelines to improve standardization, and sustainable data repositories. These steps are crucial in leveraging the exciting promise of mobility data to better understand and mitigate future disease outbreaks and epidemics in an increasingly connected world.

Table 1. Use cases of mobility data in public health response.

How can information on human mobility be used in public health response?					
Phase	**Main question(s)**	**Mobility data needs**	**How could the mobility data be used**	**Key unknowns**	**Example of use case**
1. Prior to an outbreak or in preparation	Given an outbreak, where will the pathogen spread?	Ideally access to multiple data sets given the high uncertainty in the population at risk or transmission patterns. Priority towards data describing international travel and from locations that are likely sources of initial spread. Data that could describe general movement patterns, since risk factors may not be identified, are preferable to understand how the mobility data could be integrated and used with disease surveillance network information	Mobility could be integrated into modeling frameworks to estimate introduction rates to other locations, given a range of initial starting locations.	What is the risk of importations to other locations?	Assessing risk of dengue by location to guide development of early warning and response systems[1]. Mobility data used: Twitter data on mobility between pairs of neighborhoods Method used: Regression analysis to assess predictors of dengue incidence. Number of cases modeled using Bayesian spatio-temporal modleing framework assuming a Poisson distribution.
2. Emergence	What is the order and timing of locations for the initial spatial	More specific temporal and spatial information for areas already	Using case data where the pathogen has emerged, estimate spatial spread to other	Who is actually infected? Which locations are infected?	Identifying potential secondary hubs for viral transmission of SARS-CoV-2[2].

	spread of a pathogen?	impacted and likely to spread; also requires information on the magnitude and frequency of travel between specific origins and destinations such as location data from mobile phones and social media, or GPS trackers.	locations using the frequency and number of trips to locations at risk.	How will different mobility behaviors drive transmission? How will superspreading or skewed transmission events impact the spatial spread	Mobility data used: Air passenger volumes between international airports / countries Method used: Reconstruction of dispersal patterns of variants of concern using phylogenic and phylogeographic methods. Use of regression to test association between travel volume, COVID-19 case and death counts, and international travel ban on estimated mean monthly exports of the virus.
3. Control	What is the amount of mixing or coupling between different populations to determine how interventions may or may not be impacted? How do different contact rates impact the effectiveness of interventions?	Validated mobility data against historic outbreaks, mapping of behaviors onto transmission inferred from disease data as well as detailed information about interventions that were deployed. Mixing and contact rates between populations, time spent at locations, origin-destination trip counts, duration, impact of	Inference against data to identify which data sources, mobility behaviors are most relevant for different diseases transmission	How do different mobility measures map onto behaviors that continue to drive transmission?	Demonstrating heterogeneity of SARS-CoV-2 transmission in rural and urban prefectures of Japan[3]. Mobility data used: Monthly flow between regions in Japan obtained from location data of mobile phone users Method used: Simulations using a spatial compartmental model to analyze effects of restricting mobility between regions in Japan

		interventions (aimed at mobility) on changes in mobility such as travel surveys, mobile phone and social media data, transportation.			on spatial spread of disease
4. Elimination	Where are locations less likely to be able to eliminate due to importation from endemic areas? Are there priority locations where additional monitoring and surveillance should be conducted?	Multiple types of data sets can be used that ideally would map specific movement patterns to infection status/risk of introduction events (e.g., travel surveys of cases); focus on local and international travel patterns.	Evaluation of various data sets against actual resurgence events and uncertainty in reestablishing transmission	Uncertainty of reestablishing local transmission	Identifying areas for targeted vector control in context of near-elimination of malaria[4]. Mobility data used: Mobile phone data (Haiti); census microdata from El Salvador, Costa Rica, Haiti, and Nicaragua Method used: After validating census data in estimating population movement, used census data to predict exportation and importation of malaria cases.

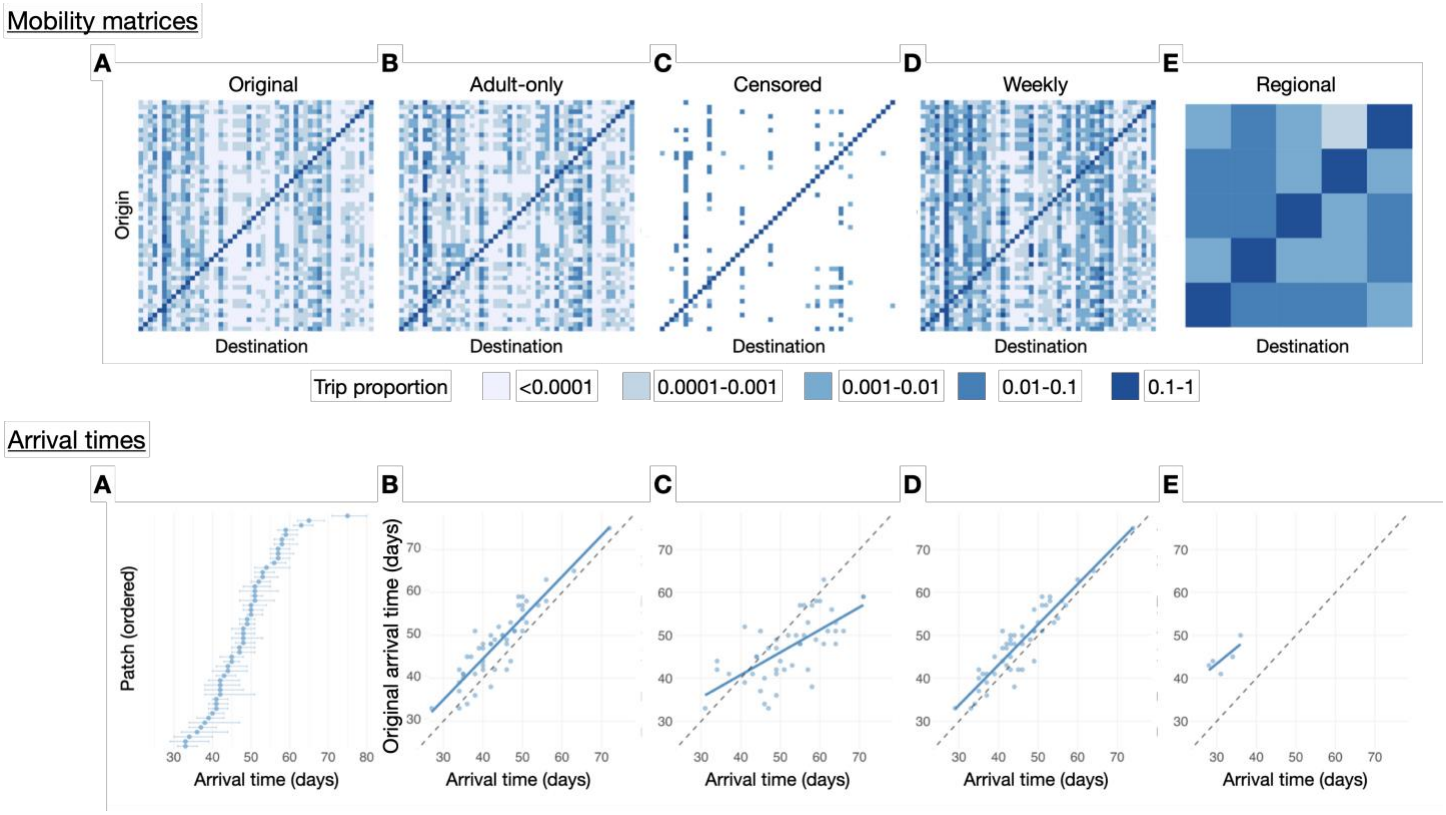
[1] Ramadona, Aditya Lia, Yesim Tozan, Lutfan Lazuardi, and Joacim Rocklöv. "A Combination of Incidence Data and Mobility Proxies from Social Media Predicts the Intra-Urban Spread of Dengue in Yogyakarta, Indonesia." *PLOS Neglected Tropical Diseases* 13, no. 4 (April 15, 2019): e0007298. https://doi.org/10.1371/journal.pntd.0007298.

[2] Tegally, Houriiyah, Eduan Wilkinson, Joseph L.-H. Tsui, Monika Moir, Darren Martin, Anderson Fernandes Brito, Marta Giovanetti, et al. "Dispersal Patterns and Influence of Air Travel during the Global Expansion of SARS-CoV-2 Variants of Concern." *Cell* 186, no. 15 (July 20, 2023): 3277-3290.e16. https://doi.org/10.1016/j.cell.2023.06.001.

[3] Kondo, Keisuke. "Simulating the Impacts of Interregional Mobility Restriction on the Spatial Spread of COVID-19 in Japan." *Scientific Reports* 11, no. 1 (September 23, 2021): 18951. https://doi.org/10.1038/s41598-021-97170-1.

[4] Ruktanonchai, Nick W., Darlene Bhavnani, Alessandro Sorichetta, Linus Bengtsson, Keith H. Carter, Roberto C. Córdoba, Arnaud Le Menach, et al. "Census-Derived Migration Data as a Tool for Informing Malaria Elimination Policy." *Malaria Journal* 15, no. 1 (May 11, 2016): 273. https://doi.org/10.1186/s12936-016-1315-5.

Figure 1. Illustration of the effect of incomplete data availability on dynamics of predicted epidemics. The first row ("Mobility matrices") depicts the origin-destination mobility matrices, or the amount of travel between origin (y-axis) and destination (x-axis) patches or regions. A) Mobility matrix derived from a gravity model for a simulated population, comprised of adults and children with different gravity model parameters ("Original"). B) Adult-only matrix, where we assume only trips by adults were captured. C) Censored mobility matrix, where all travel between origin and destination patches with <900 trips were censored / unreported. D) Weekly mobility matrix. Weekly probability of departure was calculated as 1 – exp(-probability of (original) daily departure * 7), and off-diagonal probabilities were rescaled to maintain the same diffusion patterns. E) Regional mobility matrix, where patches were classified into five regions using k-means clustering. The second row ("Arrival times") presents the results of simulations of disease transmission using different mobility matrices to quantify connectivity between patches. Initial infections were seeded into the most populous patch, and arrival times for each patch were calculated as the time from the beginning of the simulation to the time of arrival in the patch or region. Arrival was defined as at least 5 average cumulative cases in a patch across 100 simulations, or 50 average cumulative cases for the region-level simulation. Detailed information on the model and parameters used are available on Github. Panel A) shows arrival time in days for each of the patches, using the original mobility matrix. Panels B) through E) show arrival time in days using each of the four mobility matrices presented in the first row on the x-axis, compared with the arrival times from the original mobility matrix on the y-axis.



Supplement 1: Reporting guidelines for research utilizing human mobility data

Domain: Methods

Raw data

1. Explicitly describe origin / source of input data, specifying whether the data are open (e.g., Displacement Tracking Matrix) or proprietary (e.g., Call Detail Records).
2. What exactly is the nature of location in the data (e.g., Individuals' home location? Locations where the phone is registered? Location where people spend the most amount of time during a day? Home and workplace location, for commuting data? The furthest location from the origin? The accuracy of location data collected via different approaches such as signaling towers, GPS, internet check-in, wifi, post code, etc.).
3. What is the unit the data measures? I.e., what is in each matrix cell (temporal, spatial scale, duration thresholds).
4. What is the definition of a "connection" and a "stay"?
5. What is the sampling frame of the population? What are known or potential sources of bias in the data (including direction of bias)? Note this could include sampling bias, recall bias, how the data define mobility between locations, definition of home location of a device, etc. Where applicable and available, what is the market share / penetration rates for the network or data provider? How does this vary over time / geographic area?
6. Date/time frequency and spatial scale of collection.
7. Whether data are absolute or relative values; if relative, describe baseline reference.

Data Provider Pre-processing: What (if we know) has been done to the raw data

8. How exactly data were aggregated (spatially, temporally, by group, if applicable).
9. Definition, algorithm, and upper bound used for any censoring. If data were censored, how does this appear in the data set? Is it clear where data were censored, vs. where there was no travel observed?
10. Privacy protection and spatial anonymization methods applied for georeferenced locations.
11. Noise or smoothing.
12. For mobile device GPS location data, which mobile phone applications are used for tracking individual devices?

Researcher Processing data

13. Describe the meaning of the elements of the resulted aggregated matrix i.e. the diagonal vs. off-diagonal elements (e.g., probability of moving from A to B, time spent in A by people living in B, number of people living in A moving in B).
14. Aggregation – describe the spatio-temporal aggregation procedure i.e., were the data aggregated to a particular temporal (weekly, monthly), and spatial (regions, departments, municipalities) scale for modeling purposes? How was this achieved? Were mobility data provided in non-matrix form and aggregated (e.g., aggregate between-location movement at a population level)?
15. Disaggregation – describe if the data were converted to a higher spatio-temporal resolution prior to use (e.g. conversion of weekly data to daily; converting state-level mobility information into county-level information for country-level infection modeling).
16. Imputation – were data 'gaps' or sparsity corrected for by imputation? How was this achieved? (possibly include modeling of mobility data here to fill gaps).
17. Exclusion – were any elements of provided data removed or excluded? (e.g., filtering locations or time periods, removing of outliers. Did the researcher impose any censoring themselves?)
18. Boundaries – how were boundaries treated (e.g. where spatio-temporal unit of data sources don't match the flow in and out of study area)?

19. Combination and comparison – were multiple datasets combined? If yes, how was this achieved? How were incongruous spatial scales treated? Temporal?
20. Adjustments for sampling variation – if the sample size or users contributing to data changed over time. Controlling for sample (panel) size over time and/or space (e.g. the number of devices in the data provider's dataset in a given location and time point).
21. Adjustments to reduce other potential biases in data.
22. Is any estimate of uncertainty included in the data, and if so, how is it propagated through any data manipulation? This includes (especially) any imputation conducted.

Integration of mobility data into transmission modeling analysis

23. Contribution of mobility to the force of infection (the per capita rate at which susceptibles are infected), or describe the mobility process if not embedded into the force of infection. That is, how do individuals in location *i* adjust the force of infection in location *j*? How was this modeled? What is the mechanistic interpretation? Is mobility used to adjust the age contact matrix instead of explicitly modeling fluxes? (note, there are lots of different modeling options, from 'at distance' coupling through to full individual-based models and travel/stays).
24. How is mobility location linked to infection location information in the model? How are connections between home location, infection location, and mobility data modeled?
25. Any use by the researcher of modeling on the mobility data itself (for example, exponential gravity model), including selection of models (which models were considered, how they were evaluated, outliers, fit).
26. How critical are the mobility data to the infection process? Can alternative (simpler) models better explain observed infection patterns? Consider the choice of null models and perform model comparisons.
27. Risk of transmission when people travel together using the same means of transportation (e.g. plane, bus, train) during their trips.

Domain: Results

25. Provide summary statistics of mobility information used in modelling (for example, parameters and their uncertainty from fitted gravity model, or summary of distribution for probability of departure from origin patches).
26. Validation – researchers should consider validating the mobility data used in analysis against alternative (independently collected) information and report the results.
27. Sensitivity analysis – how robust are the primary results to the assumptions and choices made in the aggregation/imputation and other processing of the mobility data? Include assessment of propagated uncertainty.
28. Sharing – can the mobility data, an alternative form of these data, or information to reproduce the data be shared to aid reproducibility? This could include the aggregated mobility matrix used in modeling, global mobility indicators, or coefficients and standard errors from a fitted mobility model. This is particularly important when proprietary mobility data is used.
29. Code – processing and modeling code should be published.

References

1. Barbosa H, Barthelemy M, Ghoshal G, James CR, Lenormand M, Louail T, et al. Human mobility: Models and applications. Phys Rep. 2018;734: 1–74. doi:https://doi.org/10.1016/j.physrep.2018.01.001

2. Zhang M, Wang S, Hu T, Fu X, Wang X, Hu Y, et al. Human mobility and COVID-19 transmission: a systematic review and future directions. Ann GIS. 2022;28: 501–514. doi:10.1080/19475683.2022.2041725

3. Delamater PL, Leslie TF, Yang YT, Jacobsen KH. An approach for estimating vaccination coverage for communities using school-level data and population mobility information. Appl Geogr Sevenoaks Engl. 2016;71: 123–132. doi:10.1016/j.apgeog.2016.04.008

4. Mari L, Gatto M, Ciddio M, Dia ED, Sokolow SH, De Leo GA, et al. Big-data-driven modeling unveils country-wide drivers of endemic schistosomiasis. Sci Rep. 2017;7: 489. doi:10.1038/s41598-017-00493-1

5. Wakefield J. Ecologic studies revisited. Annu Rev Public Health. 2008;29: 75–90. doi:10.1146/annurev.publhealth.29.020907.090821

6. Wardle J, Bhatia S, Kraemer MUG, Nouvellet P, Cori A. Gaps in mobility data and implications for modelling epidemic spread: A scoping review and simulation study. Epidemics. 2023;42: 100666. doi:10.1016/j.epidem.2023.100666

7. 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

8. Pullano G, Alvarez-Zuzek LG, Colizza V, Bansal S. Characterizing US spatial connectivity: implications for geographical disease dynamics and metapopulation modeling. medRxiv; 2023. p. 2023.11.22.23298916. doi:10.1101/2023.11.22.23298916

9. napsgCS. Facebook for Good: Mobility Data (Latest Available) - Overview. [cited 15 Feb 2024]. Available: https://www.arcgis.com/home/item.html?id=48cd4d3e22df4ec2b45c8d9b882aa685

10. International Nucleotide Sequence Database Collaboration. [cited 13 May 2024]. Available: https://www.insdc.org/