Under-Represented in the Population Flow

Chuchu Liu, Petter Holme, Sune Lehmann, Wenchuan Yang, Xin Lu

Submitted to: JMIR Formative Research
on: December 05, 2023

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on it's website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressively prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Under-Represented in the Population Flow

Chuchu Liu[1] PhD; Petter Holme[2] Prof Dr; Sune Lehmann[3] Prof Dr; Wenchuan Yang[1] PhD; Xin Lu[1] Prof Dr

[1]National University of Defense Technology Changsha CN
[2]Aalto University Espoo FI
[3]Technical University of Denmark 2800 Kgs. Lyngby DK

Corresponding Author:
Xin Lu Prof Dr
National University of Defense Technology
College of Systems Engineering, National University of Defense Technology, Changsha, 410073, China
Changsha
CN

Abstract

Background: In recent years, a range of novel smart-phone derived data streams about human mobility have become available on a near real-time basis. These data have been used, for example, to perform traffic forecasting and epidemic modeling. During the COVID-19 pandemic in particular, human travel behavior has been used as a key component of epidemiological modeling to provide more reliable estimates about the volumes of the pandemic's importation and transmission routes, or to identify hotspots.

Objective: However, nearly universally in the literature, the representativeness of these data –how they relate to the underlying real-world human mobility – has been overlooked. This disconnect between data and reality is especially relevant in the case of socially disadvantaged minorities.

Methods: By analyzing travel trajectories extracted from an exceptionally comprehensive sample of 318 million mobile phone users, representing an entire nation, we found a significant difference in the demographic composition of those who travel and the overall population.

Results: We show that this difference strongly impacts outcomes of epidemiological forecasts, which typically assume that flows represent underlying demographics.

Conclusions: Our findings imply that it is necessary to measure and quantify the inherent biases related to non-representativeness for accurate epidemiological surveillance and forecasting.

(JMIR Preprints 05/12/2023:55013)
DOI: https://doi.org/10.2196/preprints.55013

Original Manuscript

Under-Represented in the Population Flow

Chuchu Liu[1,2], Petter Holme[3], Sune Lehmann[4], Wenchuan Yang[2], Xin Lu[2,5]
[1]School of Economics and Management, Changsha University of Science and Technology, Changsha, China.
[2]College of Systems Engineering, National University of Defense Technology, Changsha, China.
[3]Department of Computer Science, Aalto University, Espoo, Finland.
[4]Department of Applied Mathematics and Computer Science, Technical University of Denmark, Denmark.
[5]Department of Global Public Health, Karolinska Institute, Stockholm, Sweden.

Corresponding Author:
Xin Lu, PhD
College of Systems Engineering
National University of Defense Technology
Changsha, 410073
China
Phone: (86) 18627561577
Email: xin.lu.lab@outlook.com

Abstract

Background: In recent years, a range of novel smart-phone derived data streams about human mobility have become available on a near real-time basis. These data have been used, for example, to perform traffic forecasting and epidemic modeling. During the COVID-19 pandemic in particular, human travel behavior has been used as a key component of epidemiological modeling to provide more reliable estimates about the volumes of the pandemic's importation and transmission routes, or to identify hotspots. However, nearly universally in the literature, the representativeness of these data –how they relate to the underlying real-world human mobility – has been overlooked. This disconnect between data and reality is especially relevant in the case of socially disadvantaged minorities.

Objectives: The objective of this study is to illustrate the data non-representativeness in human mobility and the impact of this non-representativeness on modeling dynamics of the epidemic. This paper systematically evaluates how real-world travel flows differ from census-based estimations, especially in the case of socially disadvantaged minorities, such as the elderly and females, and further measures biases introduced by this difference in epidemiological studies.

Methods: To understand the demographical composition of population movements, a nationwide mobility dataset from 318 million mobile phone users in China from Jan. 1, 2020, to Feb. 29, 2020, was collected. Specifically, this paper quantifies the disparity in the population composition between actual migrations and resident-composition according to census data, then shows how this non-representativeness impacts epidemiological modeling by constructing an age-structured SEIR model of COVID-19 transmission.

Results: We found a significant difference in the demographic composition of those who travel and the overall population. In the population flows, 59% of travelers are young and 36% are middle-aged ($P<0.0001$), which is completely different from the overall adult population composition of China (where 36% are youth and 40% are middle-aged). This difference would introduce a striking bias in epidemiological studies: the estimation of maximum daily infections differs nearly three times, and the peak time has a large gap of 46 days.

Conclusions: The difference between actual migrations and resident-composition strongly impacts outcomes of epidemiological forecasts, which typically assume that flows represent underlying demographics. Our findings imply that it is necessary to measure and quantify the inherent biases related to non-representativeness for accurate epidemiological surveillance and forecasting.

Keywords
Human mobility; data representativeness; population composition; COVID-19; epidemiological modeling

Introduction

With large-scale empirical data (e.g., mobile phone records, GPS data, and location-based social network data) becoming available with increasingly fine spatial and temporal resolution [1], quantitative studies on individual and collective mobility patterns have flourished in the past few years [2-6]. These developments have offered advances with respect to understanding migratory flows, traffic forecasting, urban planning, and epidemic modeling [7-10]. The ongoing COVID-19 pandemic has further intensified discussions on how to optimally use human mobility research to support outbreak response as well as non-pharmaceutical interventions (e.g., contact tracing) [11-14].

The representativeness of datasets used to infer real-world human mobility, however, has typically not been explicitly incorporated in such analyses. This is potentially troubling as representativeness is known to be especially poor for socially disadvantaged minorities, such as low-income groups, women, children, and elderly people. For example, it has been confirmed that individuals' probability of travel is not randomly or equally distributed, and there is significant heterogeneity when comparing the travel patterns of different demographic groups [15-17]. For example, women are more localized than men in their movements and visit fewer locations in regions such as Latin America, Bangladesh, and sub-Saharan Africa [18,19]. In the specific case of epidemic outbreaks, low-income individuals are not necessarily able to limit their exposure to a circulating virus by reducing mobility and must continue e.g. commuting behavior to remain employed. Thus, this group is subject to a substantially higher probability of becoming infected in an epidemic than higher-income groups [20]. Further, different contact rates across age groups have been observed in COVID-19 incidence cases [21], and higher COVID-19 infection rates among disadvantaged racial and socioeconomic groups have been observed in multiple [22,23]. It has also been argued that including information about demographic heterogeneity in human mobility patterns, e.g., by combining demographically stratified travel data with disease research would make epidemiological models more robust [24].

While it is widely recognized that rich new data sources can near real-time information about human mobility [25], and provide powerful input into models that estimate imported cases using regional mobility information when modeling pathogen transmission, the state-of-the-art models do not consider data representativeness. This is typically because that, for privacy concerns, most datasets are not disaggregated demographically. Instead, relevant information on demographic features and social relationships is traditionally collected by censuses and other surveys [26-28].

As we argue below, however, simply considering the population demographics at the origin of a trip does not represent the traveling population.

To systematically evaluate how real-world travel flows differ from census-based estimations, we use an aggregated and anonymized dataset collected from 318 million mobile phones. Specifically, we quantify the disparity in the population composition between actual migrations and resident-composition according to census data and find significant differences. We then show how this non-representativeness impacts epidemiological modeling. The aim of this study is to illustrate the data non-representativeness in human mobility and the impact of this non-representativeness on modeling dynamics of the epidemic.

Methods

Data Description

In China, a total of 847 million Chinese people used mobile phones to surf the internet, accounting for 99.1% of the total netizens. The penetration rate of mobile phone usage among the population aged 15–65 years is almost 100%, providing extensive coverage and high representativeness for the national population. To understand the demographical composition of population movements, we collected nationwide mobility data from 318 million mobile phone users in China from Jan 1, 2020 to Feb 29, 2020. All population flow data has been aggregated based on users' geographic locations and demographic characteristics (e.g., gender and age). To enhance the extrapolation and representation of the population, a machine learning method was used to extrapolate the data to all users of the entire network, which also agrees well with the official population statistics *($R^2=0.98$)* (see *Multimedia Appendix 1*).

Epidemical Modeling

To illustrate the impact of data non-representativeness on modeling dynamics of the epidemic, we constructed an age-structured SEIR model of COVID-19 transmission by Prem et al. [29]. In fitting this age-mixing transmission model with heterogeneous contact rates between age groups [30], the differential age composition of traveling people and the overall national population were input as alternative parameters respectively. By comparing model outputs, we measured the bias caused by the data non-representativeness of demographic composition in forecasting epidemic dynamics.

Results

Demographic Heterogeneity in Traveling Individuals

For our analysis, we draw on a unique dataset from China. China is an ideal location to study representativeness in mobile phone data because of its very high smartphone penetration. In China, the penetration rate of mobile phone usage among the population aged 15–65 years is almost 100%, providing extensive coverage and high representativeness for the national population [31]. We estimate the full national mobility at the city level by extrapolating from 318 million mobile phone users (see *Methods* and *Multimedia Appendix 1* for details).

Our comparison reveals a marked difference between the overall population composition and those who travel. Below, we define “young” as individuals in their 20s-30s, “middle-aged” as people in their 40s-50s, and “elderly” as those over 60. Specifically, we find that the majority of population flows within China are generated by men and by young people. Although mobility behavior fluctuates strongly across our observation period (Fig. 1A, 1B) and is affected by temporal factors such as weekdays and holidays, the composition of travelers does not change significantly over different periods (Fig. 1C, 1D). In the population flows, 59% of travelers are young and 36% are middle-aged (as children generally do not have mobile phones, all ratios are calculated with minors younger than 18 years of age excluded). This ratio is completely different from what we find in the overall adult population composition of China (about 941 million in total) [32], where 36% are youth and 40% are middle-aged. Furthermore, daily male travelers constitute approximately 59% of the total number of traveling individuals, which is greater than the overall proportion of men (51.2%). Compared to men, women travel less, but we find that when they travel, women tend to move slightly further than males, with 175 km traveled per person in an average between-city trip, compared to 170 km for men (175 km vs.170 km, $P<0.0001$).

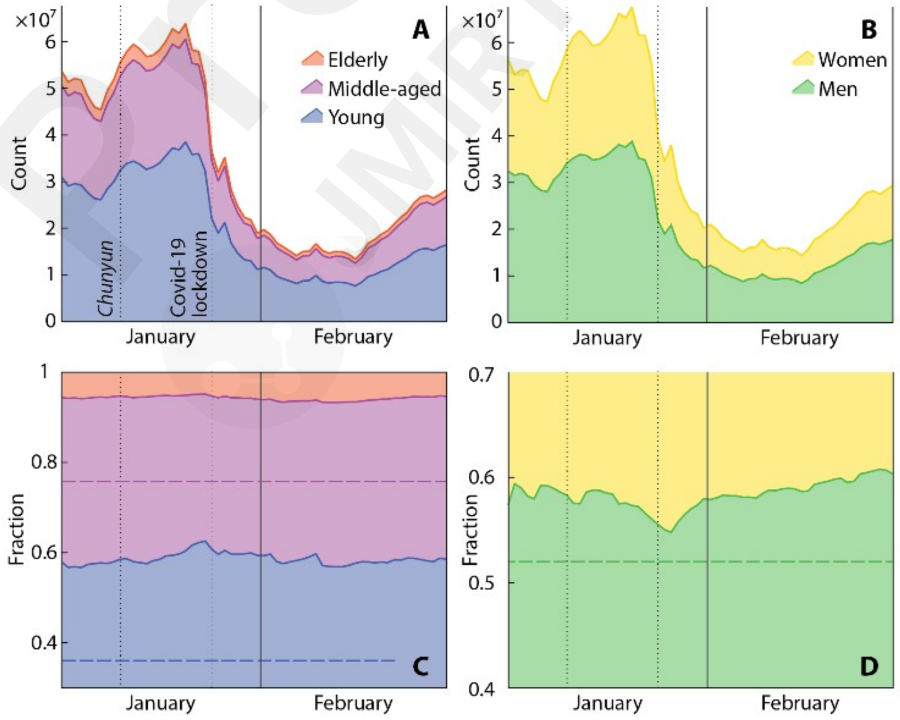


Fig. 1. Profiles of the cross-city movements extracted from mobile phone data during Jan 1 to Feb 29, 2020 in China. (A) and (B) shows the daily travel number for different age and gender groups; (C) and (D) shows the respective ratio. Dashed horizontal lines denote the composition of respective groups in the latest 7th census. As

children generally do not have mobile phones, the proportions of young, middle-aged, and old people add up to 100%. Young: 20s-30s; middle-aged: 40s-50s; elderly: ≥60s.

Bias from Data Non-representativeness

Since individual mobility is the primary reason for the spatial diffusion of an epidemic, it is important to directly explore how the demographic heterogeneity of human migration behaviors impacts our ability to forecast the spatial behavior of epidemics.

By fitting an age-structured transmission model [29,30], and including differential age composition as an input parameter, we measured the possible bias caused by the data non-representativeness of the traveling population in modeling epidemic dynamics. Feeding the composition of travelers and composition from the census data separately into the model, we found the predicted number of infected cases has a striking bias: the maximum daily infections of these two results differ by nearly three times (521 infected cases in a total of 1 million people for the composition of travelers and 1521 cases for composition from the census), and their peak time has a large gap of 46 days. Although the elderly are the most susceptible population, the two infection rates of elder people respectively collected from mobility data and census data deviate strongly (Fig. 2). Further, while the predicted cumulative confirmed cases do gradually stabilize late in the epidemic, the gap between the two model results is non-negligible (with a deviation of around 79.5%) with respect to infection volumes. Failing to include information about age and gender structure in real-world human mobility is thus likely to introduce considerable biases in epidemiological studies, especially in the early phase of an epidemic outbreak caused by imported cases [24].

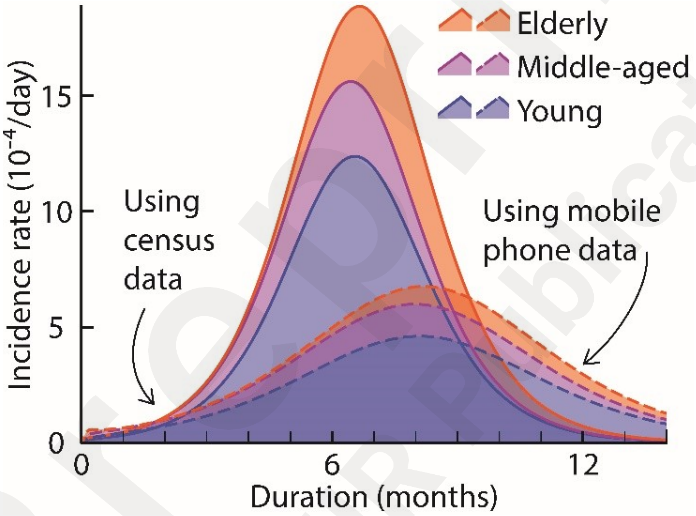


Fig. 2. Dynamics of the incidence rates of different groups predicted by the age-structured model. Solid lines indicate the incidence rates of different age groups from census data, and dashed lines indicate the incidence rates by using traveling data from mobile phones.

Discussion

By comparing mobility traces from mobile phone users to census data, our study has highlighted a number of striking differences in the demographic composition of those who travel with respect to the overall population. For example, we found that 59% of travelers are young and 36% are middle-aged, which is completely different from the adult population composition of China, where 36% are young and 40% are middle-aged. The travel probability and travel distance between males and females also have a significant difference. This realization is especially important in the case of epidemic forecasting and increased awareness of this issue in the scientific community has the potential to improve, not just epidemiological models, but our overall understanding of possible biases when inferring human mobility from cell-phone data, and representational issues of mobility data. It is important to emphasize that while China is an ideal place to study representativeness, our findings about which fraction of individuals compose the population of travelers are specific to China. The fraction of young, middle-aged, old, men, and women who travel is likely to depend on a range of factors and can be expected to be different in different countries.

What generalizes is the realization that understanding the representativeness of mobility data is crucial for epidemic monitoring and forecasting. Thus, our results imply that when generalizing results from population mobility analysis, these differences should be included in the analysis to avoid potential biases caused by data non-representativeness. For example, in the case of COVID-19, as travelers often have a higher probability of infection, the transmission risk of males and youth could be a promising focus for COVID-19 prevention.

In the recent Omicron waves, imported infections represent the majority of cases in China, and most positive patients had a travel history to high-risk areas, e.g. Shanghai [33]. As pre-symptomatic and asymptomatic pathogen carriers can travel to a foreign country and initiate the spread of COVID-19 even when there is no community transmission, human migration behaviors are promising candidates to incorporate into epidemiological models. Our findings emphasize that focusing on the representativeness of mobility data is essential for more sophisticated modeling approaches to capture key mechanisms of epidemic propagation. In future work, we tend to further explore how to accurately measure and quantify the inherent biases related to data non-representativeness for accurate epidemiological surveillance and forecasting.

Acknowledgments

The authors are grateful for the generous work the operator made in response to the COVID-19 pandemic, and their help in de-identifying this data and providing it for scientific investigation. This work was supported by the National Nature Science Foundation of China (71790615, 72025405, 91846301, 72088101), the Shenzhen Basic Research Project for Development of Science and Technology (JCYJ20200109141218676, 202008291726500001), and the Innovation Team Project of Colleges in Guangdong Province (2020KCXTD040).

Author Contributions

XL designed the research; CL and XL analyzed the data; CL, PH, SL, XL, and BCO contributed to the interpretation of the results and wrote the manuscript.

Conflicts of Interest

The authors declare that they have no competing interests.

Data and materials availability

De-identified data and code used in the analysis are available for reasonable requests from the corresponding author.

Multimedia Appendix 1

Materials and Methods

Figs. S1 to S8

References

1. Barbosa H, Barthelemy M, Ghoshal G, et al. Human mobility: Models and applications. *Physics Reports*, 2018, 734: 1-74.
2. Tan S, Lai S, Fang F, et al. Mobility in China, 2020: A tale of four phases. *National Science Review*, 2021, 8(11): nwab148.
3. Hou X, Gao S, Li Q, et al. Intracounty modeling of COVID-19 infection with human mobility: Assessing spatial heterogeneity with business traffic, age, and race. *Proceedings of the National Academy of Sciences*, 2021, 118(24).
4. Schlosser F, Maier B F, Jack O, et al. COVID-19 lockdown induces disease-mitigating structural changes in mobility networks. *Proceedings of the National Academy of Sciences*, 2020, 117(52): 32883-32890.
5. Lu X, Tan J, Cao Z, et al. Mobile phone-based population flow data for the COVID-19 outbreak in mainland China. *Health Data Science*, 2021.
6. Xiong C, Hu S, Yang M, et al. Mobile device data reveal the dynamics in a positive relationship between human mobility and COVID-19 infections. *Proceedings of the National Academy of*

Sciences, 2020, 117(44): 27087-27089.

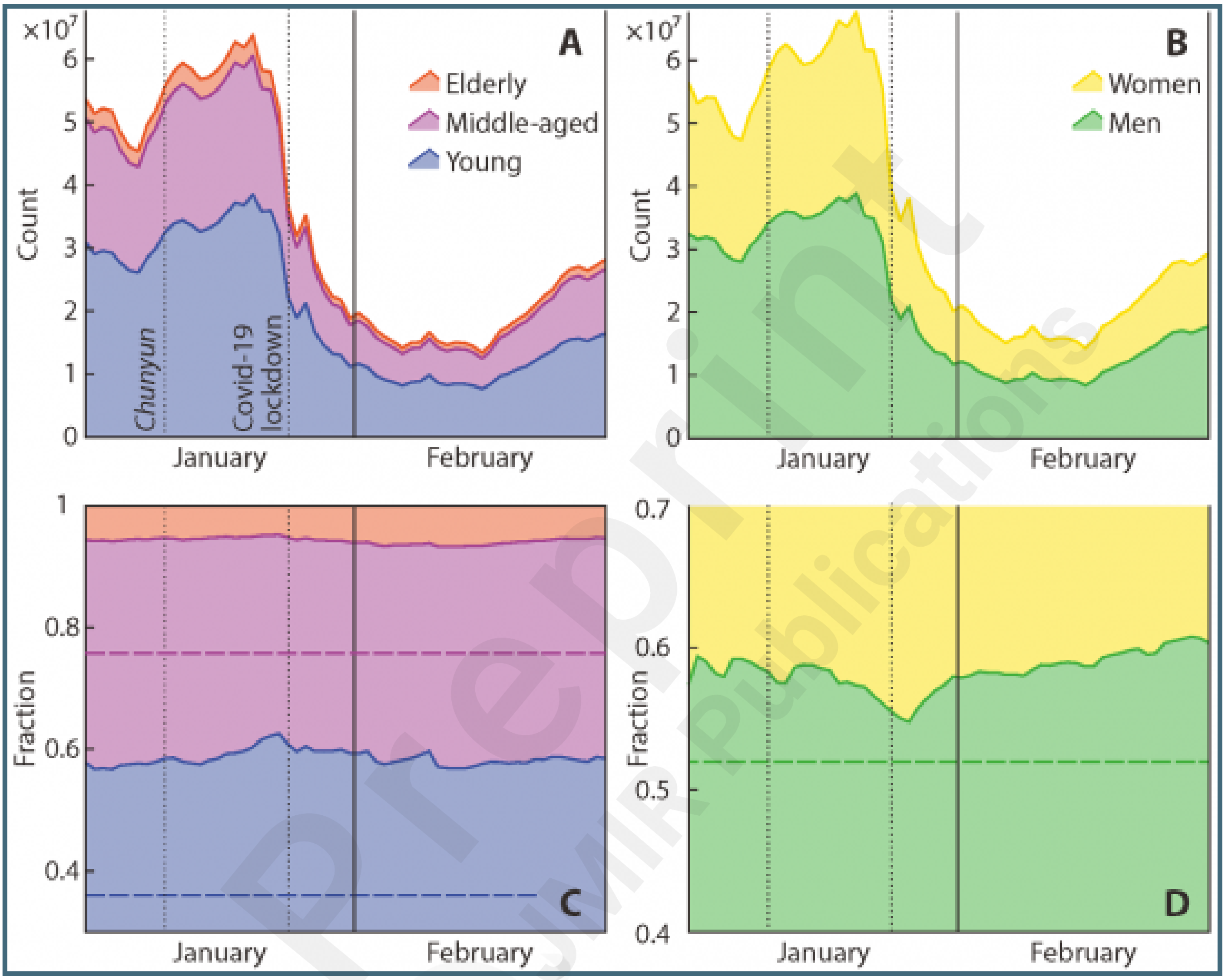
7. Jia J S, Lu X, Yuan Y, et al. Population flow drives spatiotemporal distribution of COVID-19 in China. *Nature*, 2020, 582(7812): 389-394.
8. Kraemer M U G, Yang C H, Gutierrez B, et al. The effect of human mobility and control measures on the COVID-19 epidemic in China. *Science*, 2020, 368(6490): 493-497.
9. Chen P, Liu R, Aihara K, et al. Autoreservoir computing for multistep ahead prediction based on the spatiotemporal information transformation[J]. *Nature communications*, 2020, 11(1): 4568.
10. Liu R, Zhong J, Hong R, et al. Predicting local COVID-19 outbreaks and infectious disease epidemics based on landscape network entropy[J]. *Science Bulletin*, 2021, 66(22): 2265-2270.
11. Oliver N, Lepri B, Sterly H, et al. Mobile phone data for informing public health actions across the COVID-19 pandemic life cycle. *Science Advances*, 2020, 6(23): eabc0764.
12. Lai S, Ruktanonchai N W, Zhou L, et al. Effect of non-pharmaceutical interventions to contain COVID-19 in China. *Nature*, 2020, 585(7825): 410-413.
13. Xia J, Yin K, Yue Y, et al. Impact of human mobility on COVID-19 transmission according to mobility distance, location, and demographic factors in the greater bay area of China: population-based study[J]. *JMIR Public Health and Surveillance*, 2023, 9(1): e39588.
14. Li Z, Li X, Porter D, et al. Monitoring the spatial spread of COVID-19 and effectiveness of control measures through human movement data: proposal for a predictive model using big data analytics[J]. *JMIR Research Protocols*, 2020, 9(12): e24432.
15. Adhikari S, Pantaleo N P, Feldman J M, et al. Assessment of community-level disparities in coronavirus disease 2019 (COVID-19) infections and deaths in large US metropolitan areas. *JAMA Network Open*, 2020, 3(7): e2016938-e2016938.
16. Bavel J J V, Baicker K, Boggio P S, et al. Using social and behavioural science to support COVID-19 pandemic response. *Nature Human Behaviour*, 2020, 4(5): 460-471.
17. Weill J A, Stigler M, Deschenes O, et al. Social distancing responses to COVID-19 emergency declarations strongly differentiated by income. *Proceedings of the National Academy of Sciences*, 2020, 117(33): 19658-19660.
18. Grantz K H, Meredith H R, Cummings D A T, et al. The use of mobile phone data to inform analysis of COVID-19 pandemic epidemiology. *Nature Communications*, 2020, 11(1): 1-8.
19. Sinha I, Sayeed A A, Uddin D, et al. Mapping the travel patterns of people with malaria in Bangladesh. *BMC medicine*, 2020, 18(1): 1-17.
20. Chang S, Pierson E, Koh P W, et al. Mobility network models of COVID-19 explain inequities and inform reopening. *Nature*, 2021, 589(7840): 82-87.
21. Davies N G, Klepac P, Liu Y, et al. Age-dependent effects in the transmission and control of COVID-19 epidemics. *Nature Medicine*, 2020, 26(8): 1205-1211.
22. Pareek M, Bangash M N, Pareek N, et al. Ethnicity and COVID-19: an urgent public health research priority. *The Lancet*, 2020, 395(10234): 1421-1422.
23. Chowkwanyun M, Reed Jr A L. Racial health disparities and Covid-19—caution and context. *New England Journal of Medicine*, 2020, 383(3): 201-203.
24. Buckee C, Noor A, Sattenspiel L. Thinking clearly about social aspects of infectious disease transmission. *Nature*, 2021, 595(7866): 205-213.
25. Buckee C O, Balsari S, Chan J, et al. Aggregated mobility data could help fight COVID-19. *Science*, 2020, 368(6487): 145-146.
26. Mossong J, Hens N, Jit M, et al. Social contacts and mixing patterns relevant to the spread of infectious diseases. *PLoS Medicine*, 2008, 5(3): e74.
27. Karaca-Mandic P, Georgiou A, Sen S. Assessment of COVID-19 hospitalizations by race/ethnicity in 12 states. *JAMA Internal Medicine*, 2021, 181(1): 131-134.
28. Rubin D, Huang J, Fisher B T, et al. Association of social distancing, population density, and temperature with the instantaneous reproduction number of SARS-CoV-2 in counties across the United States. *JAMA Network Open*, 2020, 3(7): e2016099-e2016099.

29. Number of mobile cell phone subscriptions in China from December 2020 to March 2022. China: mobile phone subscriptions by month 2020-2022, https://www.statista.com/statistics/278204/china-mobile-users-by-month/, last accessed 2022/06/02.
30. The seventh national census data from national bureau of statistics in China, http://www.stats.gov.cn/tjsj/tjgb/rkpcgb/qgrkpcgb/202106/t20210628_1818823.html, last accessed 2022/05/12.
31. Prem K, Liu Y, Russell T W, et al. The effect of control strategies to reduce social mixing on outcomes of the COVID-19 epidemic in Wuhan, China: a modelling study. *The Lancet Public Health*, 2020, 5(5): e261-e270.
32. Zhang J, Litvinova M, Liang Y, et al. Changes in contact patterns shape the dynamics of the COVID-19 outbreak in China. *Science*, 2020, 368(6498): 1481-1486.
33. Zhang J, Tan S, Peng C, et al. Heterogeneous changes in mobility in response to the SARS-CoV-2 Omicron BA. 2 outbreak in Shanghai[J]. *Proceedings of the National Academy of Sciences*, 2023, 120(42): e2306710120.

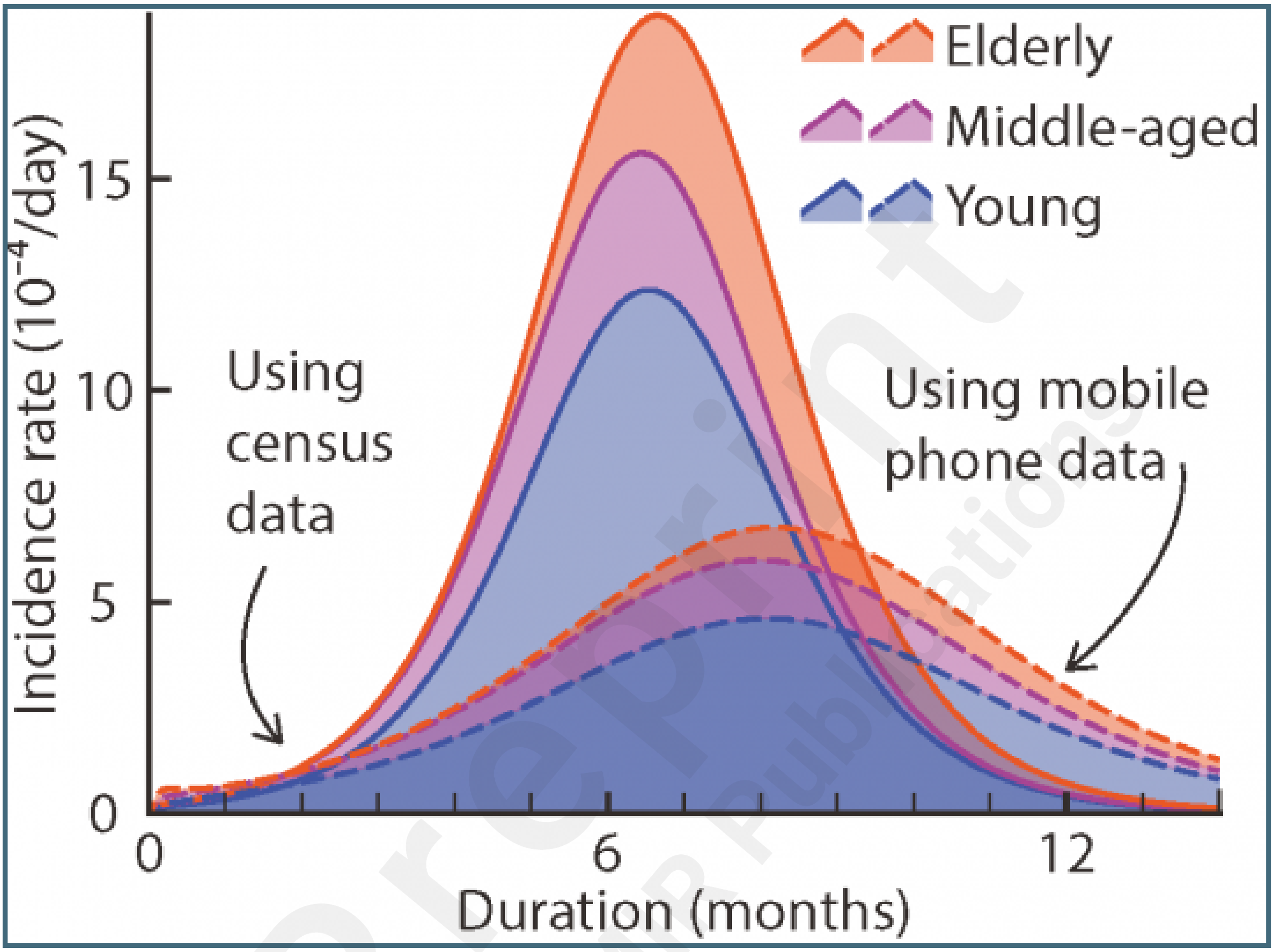
Supplementary Files

Figures

Profiles of the cross-city movements extracted from mobile phone data during Jan 1 to Feb 29, 2020 in China. (A) and (B) shows the daily travel number for different age and gender groups; (C) and (D) shows the respective ratio. Dashed horizontal lines denote the composition of respective groups in the latest 7th census. As children generally do not have mobile phones, the proportions of young, middle-aged, and old people add up to 100%. Young: 20s-30s; middle-aged: 40s-50s; elderly: ?60s.



Dynamics of the incidence rates of different groups predicted by the age-structured model. Solid lines indicate the incidence rates of different age groups from census data, and dashed lines indicate the incidence rates by using traveling data from mobile phones.



Multimedia Appendixes

under-represented in the population flow_Appendix.
URL: http://asset.jmir.pub/assets/b767e92c4bdf9b2d881384191c6b8d83.docx