Bring Your Own Location Data: Use of Google Smartphone Location History Data for Environmental Health Research

Perry Hystad, Ofer Amram, Funso Oje, Andrew Larkin, Kwadwo Boukye, Ally Avery, Assefaw Gebremedhin, and Glen Duncan

1College of Public Health and Human Sciences, Oregon State University, Corvallis, Oregon, USA
2Department of Nutrition and Exercise Physiology, Elson S. Floyd College of Medicine, Washington State University (WSU), Spokane, Washington, USA
3Paul G. Allen School for Global Animal Health, WSU, Pullman, Washington, USA
4School of Electrical Engineering and Computer Science, WSU, Pullman, Washington, USA

BACKGROUND: Environmental exposures are commonly estimated using spatial methods, with most epidemiological studies relying on home addresses. Passively collected smartphone location data, like Google Location History (GLH) data, may present an opportunity to integrate existing long-term time–activity data.

OBJECTIVES: We aimed to evaluate the potential use of GLH data for capturing long-term retrospective time–activity data for environmental health research.

METHODS: We included 378 individuals who participated in previous Global Positioning System (GPS) studies within the Washington State Twin Registry. GLH data consists of location information that has been routinely collected since 2010 when location sharing was enabled within android operating systems or Google apps. We created instructions for participants to download their GLH data and provide it through secure data transfer. We summarized the GLH data provided, compared it to available GPS data, and conducted an exposure assessment for nitrogen dioxide (NO2) air pollution.

RESULTS: Of 378 individuals contacted, we received GLH data from 61 individuals (16.1%) and 53 (14.0%) indicated interest but did not have historical GLH data available. The provided GLH data spanned 2010–2021 and included 34 million locations, capturing 66,677 participant days. The median number of days with GLH data per participant was 752, capturing 442 unique locations. When we compared GLH data to 2-wk GPS data (~1.8 million points), 95% of GPS time–activity points were within 100 m of GLH locations. We observed important differences between NO2 exposures assigned at home locations compared with GLH locations, highlighting the importance of GLH data to environmental exposure assessment.

DISCUSSION: We believe collecting GLH data is a feasible and cost-effective method for capturing retrospective time–activity patterns for large populations that presents new opportunities for environmental epidemiology. Cohort studies should consider adding GLH data collection to capture historical time–activity patterns of participants, employing a “bring-your-own-location-data” citizen science approach. Privacy remains a concern that needs to be carefully managed when using GLH data. https://doi.org/10.1289/EHP10829

Introduction

Environmental exposures, such as air pollution, noise, walkability, and green space, are commonly estimated using spatial methods, with most epidemiological studies relying on home addresses. There are limited well-established biomarkers for most of these exposures, and personal measurements have logistical and cost constraints, especially for the large sample sizes required to power environmental epidemiological studies. The reliance on home addresses to estimate personal exposures likely results in exposure misclassification and potential biases, given the self-selection of individuals into residential locations by sociodemographic characteristics and other unmeasured factors. Recently, the concept of the exposome, defined as “the totality of environmental exposures from conception onwards,” has catalyzed exposure scientists to develop new measurement and modeling methods, but little progress has been made to incorporate nonresidential environmental exposures into environmental epidemiology.

Studies have used Global Positioning System (GPS)–based monitoring to measure location and integrate daily time–activity patterns into environmental exposure assessments; however, these studies are typically composed of small samples and conducted over short time periods. Nevertheless, these studies have shown that residential location alone does not capture how individuals interact with their environments (i.e., their activity space) nor their important environmental exposure circumstances, such as commuting. The largest environmental GPS study to date leveraged GPS data collected from a smartphone travel survey for 5,452 adults (~15/d-participant) in Montreal, Canada, to compare residential- to mobility-based air pollution exposures. Important differences in air pollution exposures were observed, which varied geographically in the city and by participant sociodemographic characteristics. This study did not examine health outcomes, consistent with other GPS studies that have typically focused on environmental exposures or physical activity. For example, a recent review identified 129 studies that used GPS in physical activity research, of which most studies measured location for <7 d in small samples. In addition, the requirement of having participants carry an accelerometer and GPS device results in clear trends of decreasing compliance with longer measurement periods. This highlights a major obstacle to having participants actively collect GPS data to characterize long-term activity patterns and environmental exposures.

Other methods besides GPS are available to capture time–activity patterns, but these have significant limitations in the context of environmental epidemiology. Methods include agent-based modeling and simulation, web-based mapping, qualitative methods, activity space questionnaires, or smartphone mobility surveys. These methods are mostly used for prospective data collection, similar to existing GPS data collection. For example, data...
collected using diaries or mobility surveys require individuals to record their daily time–activity patterns or locations where they spend a large amount of time, which is a subjective and time-intensive task known to have reporting errors and is difficult to implement for long time periods. Agent-based modeling and other simulation methods can predict population-level time–activity patterns well, but they do not capture the individual time–activity patterns required to estimate personal exposures. Given these challenges, and the fact that residential addresses are easy to collect or available with administrative health data, residential addresses remain the predominant spatial location used to estimate individual environmental exposures in health research.

Smartphones and passively collected location data provide a new data source to capture long-term, individual time–activity patterns. In particular, Google Location History (GLH) data are collected on both Android smartphones and Apple iPhones (with the Google Maps app installed) for individuals who enable location services. GLH data are owned and managed by individuals and can be visualized and explored using Google Maps Timeline (https://timeline.google.com). Approximately 85% of Americans own a smartphone,\(^{17}\) and most smartphone users (71%) continuously carry and sleep within an arm’s reach of their smartphones.\(^{18}\) Thus, smartphone-derived time–activity patterns are likely to be highly representative of personal behavior. One study evaluated GLH data against GPS tracker data for 25 Android users in the United Kingdom and found that GLH data were spatially equivalent to GPS tracker data (within 100 m) and that GLH data did not suffer from the compliance issues seen with GPS trackers and the biases in self-reported travel diaries.\(^{19}\) Another study evaluated GLH data and air pollution measurement error (for one individual) and found that the GLH-derived air pollution estimates aligned closely with GPS data logger–derived estimates.\(^{20}\) Another study evaluated GLH data as a tool for travel diary data collection and prospectively collected GLH data from five different individuals’ prescribed itineraries over 12 d to evaluate whether GLH could capture locations and trips.\(^{21}\) Only 51% of locations and 32% of trips were correctly matched to the predetermined travel diary data (with missing locations and trips being predominantly those of shorter duration), suggesting GLH data may not be an accurate method for travel diary data collection. To date, no study has collected retrospective GLH data to assess individual long-term time–activity patterns, which hold the most promise for refining environmental exposure estimates for health research.

In this study, we evaluated the potential use of GLH data for capturing retrospective time–activity patterns for environmental health studies. We targeted GLH data collection to 378 individuals who participated in previous GPS studies within the Washington State Twin Registry (WSTR). We created instructions for participants to download their GLH data and provide it to WSTR researchers through secure data transfer. We report response rates, participant demographics, data coverage, and spatial accuracy of GLH data compared with GPS data. We also report on a nitrogen dioxide (\(\text{NO}_2\)) air pollution exposure assessment conducted to demonstrate the utility of the GLH data. Finally, we discuss the opportunities and limitations of collecting and integrating GLH data within existing cohort studies.

**Methods**

**GLH Data**

GLH data consists of information that has been routinely collected from smartphone devices with location services enabled since 2010.\(^{22}\) Collected data are tied to a user’s Google account, and not to specific phones. As a result, it is continuously archived in the Google account even if the user changes phones over time. Locations and various characteristics about a user’s movements over time are recorded and transmitted to Google. These data are collected passively (i.e., individuals do not need to do anything), eliminating biases in terms of participant compliance. Both Android and Apple devices require a user to opt in to data collection. Location services use GPS, cell phone towers, and Wi-Fi devices to record locations.

GLH data records raw location data elements, as well as additional semantic information created by Google (Table S1). The raw data includes the longitude and latitude of the location, time stamps, device information, address information, and accuracy of the measurement. The semantic data provides more detailed information about specific places (e.g., locations visited, such as stores or parks) and activities (e.g., movement from one place to another). The semantic data are broken down into two categories: (a) “place visits,” which record information about the locations, including the duration of the visit; and (b) “activity segments,” which record movements from one place to another, such as driving from home to work.

**GLH Data Collection**

We contacted 378 individuals from the WSTR who had participated in previous GPS and physical activity studies between 2012 and 2019. The first study was the Physical Activity in Twins (PAT) study,\(^{23,24}\) which collected 14 d of GPS and accelerometer data for 288 individual twins residing in the Puget Sound area of Washington State between 2012 and 2015. A follow-up study collected data in 2018 and 2019 from a subset of the PAT twins using accelerometer and GPS, but only 1 wk of data were collected.\(^{25}\) The second study is the ongoing Environmental Exposures in Twins study, which also collected 14 d of GPS and accelerometer data, in addition to air pollution data, from 108 individual twins between 2018 and 2019.

**Participant Recruitment**

Individuals \((n = 378)\) from the activity studies who were still actively participating in the WSTR were sent a study invitation by email, followed by a mailed letter 1 wk later. If we had not received any response by the end of the second week, we sent another reminder email and began follow-up phone calls. A study coordinator was available to assist participants who experienced trouble in downloading or uploading their GLH file. Individuals who uploaded their GLH file were provided a $10 gift card for participation. Once the GLH data had been processed and environmental exposures assigned, participants received a personalized summary of exposure measures. This report included summaries for outcomes such as environmental exposures (e.g., \(\text{NO}_2\)), time spent outdoors in nature, outdoor physical activity levels, and differences in environmental exposures at home, work, and in transit. All participants in our study provided written, informed consent. The research protocol was approved by the Washington State University (WSU) institutional review board (no. 18473).

**Data Collection Methods**

We developed a survey using REDCap electronic data capture tool\(^{26,27}\) hosted at WSU Spokane to enroll participants into the study. Within REDCap, participants were provided a summary of the study, details, and examples of GLH data (including an example map) and information on how to participate. Interested individuals were provided a consent form to sign within REDCap.
Once individuals consented to participate in the study, they were provided step-by-step instructions, both written and as a video, for downloading their GLH data from their Google takeout account and uploading it directly to a secure server hosted at WSU. The general instructions were to sign into Google Takeout (https://takeout.google.com), check the box for Location History, create an exported ZIP file, and upload the zipped file to the secure server. The time this process takes depends on the size of the GLH file created, but for an average 19-MB ZIP file, the time was ~5 min. Uploaded file sizes ranged from 110 KB to 100 MB, depending on the length of time the GLH data were collected for each participant.

Data Processing
We developed a workflow to create a central GLH database for study participants. The GLH data were first imported using a Python (version 3.9.5) script that structurally loaded data into a SQLite (version 3.0) database. The semantic data (place visits and activity segments) were then joined to the corresponding raw location data by date, time, and location identifier. This allowed us to identify movement paths during an activity and record small movements that occur while visiting a particular place (e.g., moving throughout a workplace).

Further processing was done to normalize the data, such as changing time stamps from epoch to regular date format and splitting the activities into a different schema to aid data consumption and reporting. Once all the data were consolidated into the SQLite database, we developed automated R (version 4.0.4; R Development Core Team) scripts to summarize data elements by participant.

Summary Measures of GLH Data
We calculated summary measures of the collected GLH data up to September 2021 (the end of data collection for this study). First, we calculated response rates and tested for socioeconomic and demographic trends between those participants who provided and those who did not provide GLH data using independent t-tests. These individual measures were collected via surveys completed as part of the activity study. Given our small sample size, we had to collapse several sociodemographic categories, such as age, race, and ethnicity. Next, we calculated summary statistics for GLH measures, including the median number of days of GLH data and the number of place visits per day. We also tested for differences in these measures by participant characteristics using independent t-tests to identify potential bias in GLH data collection and coverage. We also examined the classification of trips to determine whether GLH data could provide information on active transportation. All analyses were completed in R (version 4.0.4; R Development Core Team).

Evaluation of GLH and GPS Locations
We examined the accuracy of GLH locations compared with the GPS-based locations by overlaying corresponding time periods of GLH and GPS data per participant. We intersected GLH and GPS points in ESRI ArcGIS using different buffer distances applied to GLH points, including 10, 100, 200, 500 meters. We then summarized the number and percentage of GPS locations captured by the GLH locations. We conducted this analysis for all GPS points, as well as for nonhome and work location (by removing all GPS points within 100 m of home and work locations), to specifically examine how GLH data captures daily time–activity patterns. All analyses were completed in ESRI ArcGIS Pro (version 10.6.1) and R (version 4.0.4; R Development Core Team).

NO₂ Air Pollution Exposure Assessment
We conducted an exposure assessment for NO₂ air pollution using the GLH data and a global spatial–temporal NO₂ land-use regression model (LUR). The annual model explained 63% of global annual NO₂ variation from 8,250 monitor locations. Owing to processing time restrictions, we applied only the 2019 annual average NO₂ concentrations to GLH point locations. We first removed GLH locations that were classified in a plane and then extracted the corresponding raster NO₂ value to each GLH point in ArcGIS. For each participant, we calculated NO₂ concentrations at the home address, as well as at the work address (classified from GLH data). We then calculated the mean NO₂ for GLH data classified as other locations [this includes all places classified in the GLH data besides home and work (e.g., retail stores or parks), GLH points classified as within a vehicle, and for GLH data points weighted by the time in each location]. We calculated the mean ± standard deviation (SD) of the NO₂ concentration for each of these categories and examined correlations between each individual’s exposure measures based on each classification.

Results
GLH Data Response Rates
Table 1 summarizes the response rates as of 1 September 2021 for the 378 participants who were contacted about providing GLH data. We collected historical GLH files for 61 individuals, a participation rate of 16.1%. We exhausted contact attempts for 118 individuals, meaning that we did not get a response after the initial letter, three follow-up emails, and five phone call attempts at various times of day. Fifty-three (14%) individuals were interested in participating in the study but did not have historical GLH data available. Of the 146 individuals who were not interested in participating in the study, 34 individuals (23.3%) stated that concerns about privacy were their main reason for not participating.

Characteristics of Individuals Providing GLH Data
Characteristics of those individuals who provided, and of those who did not provide, GLH data are summarized in Table 2. Individuals providing GLH data were significantly younger than those not providing data (mean difference of 6.1 y, p < 0.001). Although there were no other statistically significant differences, individuals providing data were more likely to be male, to not be married, and to have a bachelor’s degree or higher.

Spatial and Temporal Coverage of GLH Data
The collected GLH data for the 61 WSTR participants included a total of 34 million data points over 66,677 d, capturing daily time–activity patterns from 27 May 2010 to 10 September 2021. Most participants (54%) did not have GLH data until 2015.

| Table 1. Response summary for GLH data collected in a sample of 378 twins enrolled in the Washington State Twin Registry. |
|---------------------------------------------------------------|
| Responses | Participants [n (%)] |
| GLH data received | 61 (16.1) |
| Interested but no historical GLH data available | 53 (14.0) |
| Not interested in participating | 146 (38.6) |
| Privacy concerns | 34 (23.3) |
| No time/too busy | 15 (10.3) |
| Too complicated | 15 (3.4) |
| No GLH | 19 (13.0) |
| No reason given | 73 (50.0) |
| Exhausted contact attempts | 118 (31.2) |

Note: GLH, Google Location History.
Overall, 76.7% of GLH locations were classified with semantic information (either place visits or trips) in the GLH data provided by Google. Figure 1 illustrates the locations in the GLH data along with annual NO2 concentrations for 2019. Given the large number of locations (34 million), as well as the temporal coverage (2010–2021), it is difficult to visualize GLH data. For the 61 participants, data were present from all continents of the world given that GLH data can be collected during trips as long as a smartphone has connectivity (e.g., cellular or Wi-Fi).

The median number of days per participant with location data was 752. Figure 2 illustrates the temporal coverage of the GLH data collected by week, month, and year. Figure S1 summarizes the distribution of the total number of trips classified per participant as well as the length of trips. Classified trip categories included passenger vehicle (59.8%), walking (20.3%), in vehicle (10.5%), in bus (2.0%), cycling (1.9%), flying (0.4%), in subway (0.4%), and running (0.2%). In terms of outdoor physical activity, GLH captured 21,579 h for walking, 2,401 h for cycling, and 440 h for running for all participants.

### Differences in GLH Data Coverage

We explored differences in provided GLH data coverage based on participant characteristics (Table S2). We did not observe any statistically significant differences, but we had a small sample size to examine these differences. Overall, days of GLH coverage were greater in older individuals, males, unmarried individuals, and lower-income households. Places visited per day were greater for younger individuals, females, unmarried individuals, and individuals with college or higher education. Most individuals did not have device information data available, and we were unable to compare coverage by device type owing to small sample numbers.

### Comparison of GLH Location Accuracy to GPS Locations

To assess GLH accuracy in capturing locational data, we compared 25 study participants who had overlapping GLH and GPS data time periods (Figure 3). First, we examined all GPS locations, which were heavily influenced by home and work locations. We then excluded GLH and GPS points within 100 m from the participants home and work (when available) locations. The

---

Table 2. Descriptive statistics of individuals who provided and those who did not provide GLH data in a sample of 378 twins enrolled in the Washington State Twin Registry.

| Characteristics                                  | GLH data received | GLH not received | Difference (p-value)* |
|--------------------------------------------------|-------------------|------------------|-----------------------|
| Individuals                                      | 61                | 317              | —                     |
| Age (mean ± SD)                                  | 44.6 ± 10.8       | 50.7 ± 13.1      | <0.001                |
| Sex [n (%)]                                      |                   |                  |                       |
| Female                                           | 39 (63.9)         | 233 (73.5)       | 0.128                 |
| Male                                             | 22 (36.1)         | 84 (26.5)        |                       |
| Ethnicity [n (%)]                                |                   |                  |                       |
| Hispanic/Latino                                  | 2 (3.3)           | 13 (4.1)         | 0.763                 |
| Non-Hispanic/Latino                              | 59 (96.7)         | 304 (95.9)       |                       |
| Race [n (%)]                                     |                   |                  |                       |
| White                                            | 53 (86.9)         | 271 (85.5)       | 0.552                 |
| Black                                            | 0 (0.0)           | 7 (2.2)          |                       |
| Asian                                            | 4 (6.6)           | 14 (4.4)         |                       |
| Other                                            | 0 (0.0)           | 6 (1.9)          |                       |
| Two or more races                                 | 4 (6.6)           | 19 (6.0)         |                       |
| Marital status [n (%)]                           |                   |                  |                       |
| Single, never married                            | 15 (25.0)         | 55 (17.5)        | 0.212                 |
| Married or living with partner                   | 40 (66.7)         | 219 (69.5)       |                       |
| Widowed                                          | 3 (5.0)           | 10 (3.2)         |                       |
| Divorced or separated                            | 2 (3.3)           | 31 (9.8)         |                       |
| Education [n (%)]                                |                   |                  |                       |
| High school or less                              | 3 (5.0)           | 22 (7.0)         | 0.763                 |
| Some college                                     | 6 (10.0)          | 42 (13.4)        |                       |
| Associate’s degree                              | 2 (3.3)           | 17 (5.4)         |                       |
| Technical or vocational degree                  | 2 (3.3)           | 14 (4.5)         |                       |
| Bachelor’s degree or higher                     | 47 (78.3)         | 219 (69.7)       |                       |
| Household income per year [n (%)]                |                   |                  |                       |
| <$50,000                                         | 7 (11.9)          | 52 (17.0)        | 0.358                 |
| $50,000–$99,999                                  | 24 (40.7)         | 137 (44.8)       |                       |
| ≥$100,000                                       | 28 (47.5)         | 117 (38.2)       |                       |

Note: Sociodemographic variables were collected by survey at time of enrollment into the WSTR physical activity study. GLH, Google Location History; SD, standard deviation; WSTR, Washington State Twin Registry.

*Independent t-test for linear variables and chi-square test for categorical variables for individuals who provided and did not provide GLH data.
GPS used in previous WSTR physical activity studies was calibrated to record location data every 10 s. A total of 1,832,267 GPS points were examined, including 724,260 time–activity (nonhome and work locations) GPS points, with an average of 9.6 GPS points for every GLH point. Table 3 summarizes the number of GPS points that intersected with the buffers of GLH points calculated per participant. GLH data captured 79.5% and 47.5% of total and time–activity GPS locations in 10-m buffers and 97.1% and 90.6% in 200-m buffers, respectively.

### NO2 Air Pollution Exposure Assessment

Figure 1 illustrates GLH points with annual average NO2 concentrations (in parts per billion) for 2019. Based on time-weighted GLH locations the mean ± SD NO2 exposure level was 12.46 ± 2.23 ppb, compared with 11.69 ± 3.25 ppb at home locations, 14.95 ± 4.13 ppb at work locations, 13.68 ± 1.95 ppb when in a vehicle driving, and 12.06 ± 2.05 ppb at all other locations classified in GLH data (Table S3). We observed large differences in the correlations between NO2 exposures assigned at home locations compared with GLH locations for work ($r = -0.06$), when driving ($r = 0.27$), at other locations ($r = 0.51$), and when using all time-weighted GLH data ($r = 0.74$) (Figure S3).

### Discussion

In the present study, we explored the use of GLH data to collect long-term retrospective time–activity data for environmental health research. Of the 378 individuals contacted, we received historical GLH data from 61 individuals (16.1%), whereas 53 (14.0%) indicated interest but did not have existing GLH data. Collected GLH data included 34 million data points, including 66,677 d of participant time–activity data from May 2010 to September 2021. The median number of days with time–activity data per person was 752 d, capturing a median of 442 unique locations. Most of all trips (76.7%) were classified with semantic information about the places visited (home, work, business types) and the mode of transportation for trips (vehicle, bus, cycling, walking). We observed high spatial correspondence between GLH and GPS locations for 25 individuals with overlapping periods of data. Finally, we demonstrated the application of GLH data for environmental exposure assessment using NO2 air pollution, highlighting the new information captured with GLH data.
Taken together, these findings demonstrate that collecting and analyzing GLH data is a feasible and cost-effective method to capture retrospective long-term time-activity patterns for the large samples needed in environmental health research.

**Feasibility of Collecting GLH Data**

Applying GLH data to environmental health studies has the potential to dramatically improve personal long-term environmental exposure assessment while requiring minimal effort from study participants. It takes participants $\sim 5$ min to download and provide their GLH data. On average, participants provided over 2 y of time-activity data in this study. The response rate was similar to that of survey-based research: 61 individuals (16.1%) provided data, although an additional 53 individuals (14.0%) indicated interest but did not have historical GLH data. We are currently expanding GLH data collection in the WSTR and now ask participants to turn on GLH collection for prospective time-activity collection, which would provide an expected response rate of $\sim 30\%$ for prospective data collection. For comparison, the response rate of the original WSTR PAT study was 27.4% and the GPS and air pollution study was 24.8%. Rarely are response rates provided in GPS studies, which tend to recruit until a desired sample size is reached. Based on the existing literature of GPS and active research studies, our response rate for GLH collection is on par with active participant GPS data collection. The use of GLH data also avoids the compliance issues seen with GPS applications and the bias and measurement errors seen in self-reported travel diaries or surveys. The GLH data that is downloaded is also relatively clean (i.e., it has been formatted
and standardized by Google) with >70% of the locations and trips already classified, reducing the time involved for data cleaning and formatting that is typically required with traditional GPS data.

Moving toward Personal Exposure Measures

The use of GLH data will require a movement toward personalized environmental exposures assessment approaches and big data analytics. The application of NO₂ air pollution presented here is a good example. We applied an annual global LUR model at a 50-m resolution and demonstrated important differences between home NO₂ concentrations (which are used by most epidemiological studies) and work, transit, and other GLH locations. However, this is a relatively simplistic exposure assessment that does not fully capture the power of GLH information. To fully leverage the GLH data, we would need to apply NO₂ concentrations to GLH locations at an hourly temporal resolution to capture the large within-day variations in NO₂. This is feasible by, for example, linking continuous ground NO₂ monitoring data to daily LUR estimates to adjust for diurnal NO₂ patterns. In combination with time–activity patterns, GLH data can provide important information on modification factors for air pollution exposure. For example, infiltration of air pollution may differ substantially based on microenvironments identified using GLH (e.g., home, work, retail locations); travel modes based on driving, biking, walking, or public transit will have different exposure profiles; and inhalation rates change with different behaviors, such as biking, walking, and driving. These types of personalized exposure assessment approaches will require extremely large computational resources. For the 61 participants who provided GLH data here, we had 34 million data points with locations in every continent of the world, and it took 6 d of processing to extract 1 y of NO₂ data. We are currently building data cleaning and analysis pipelines to improve the computational efficiencies of exposure assessments using GLH data.

Opportunities to Collect GLH Data in Existing Health Studies

A major research opportunity is to collect GLH data for existing large cohort studies to refine personal environmental exposure estimates. Given that GLH data have been collected passively since 2010 (although most participants will not have data until 2016–2017 owing to location sharing services), they can be linked to existing cohorts to assess retrospective environmental exposures and behaviors over many years. We are currently expanding our data collection efforts to 3,000 individuals in the WSTR who completed a COVID-19 survey to examine how behaviors changed before, during, and after local COVID-19 lockdown measures. It is important that GLH data be conceptualized and framed to participants as a citizen science approach to environmental health research: Individuals own their GLH data and must consent to provide their location data to a study (thus the phrase, “bring your own location data”). This approach has been successful for other types of personal data, such as Fitbit data collection within the All of Us Research Program. If 16% of participants (the response rate observed in our study) in the All of Us study \( n \approx 345,000 \) provided GLH data, this would establish a substudy of ~55,000 individuals with years of daily time–activity data linked to existing surveys, biomarkers, and electronic medical records, creating an entirely new resource to examine personal environmental exposures, human–environment interactions, and health effects. Such a linkage would facilitate novel research questions, such as a) personal environmental exposures (e.g., air pollution, green space) and health outcomes partitioned by important exposure locations (home, work, school, transit, recreation), b) exposure approaches to capture total personal environmental exposures and health impacts, c) representativeness of home-based exposures for capturing personal exposures and characterization of different types of exposure measurement error, d) selective mobility bias using long-term time–activity patterns and residential mobility, e) GLH-measured physical activity and location-based built environment characteristics, and f) individual food environment use and associations with health outcomes. These are only a sample of new research questions that could be addressed by linking GLH data to existing cohort information.

Limitations of GLH Data

GLH provides a unique opportunity to move the field of environmental epidemiology forward, but limitations need to be acknowledged. Most importantly, GLH data are invasive: They contain all locations a person visits for years (to potentially decades in the future) and, if used incorrectly, could lead to breaches of participant privacy. It is important to highlight that privacy issues for GLH data are different from GPS data because GLH data are collected over long time periods where individuals may not have

Table 3. Summary of the spatial correspondence between GLH data and GPS data for 25 study participants with 2-wk GPS data collection.

| Spatial correspondence (m) | All GPS points captured by GLH [n (%)] | Time–activity* GPS points captured by GLH [n (%)] |
|---------------------------|----------------------------------------|-----------------------------------------------|
| 10                        | 1,456,807 (79.5)                       | 343,859 (47.5)                                |
| 100                       | 1,736,532 (94.8)                       | 612,212 (84.5)                                |
| 200                       | 1,779,120 (97.1)                       | 656,011 (90.6)                                |
| 500                       | 1,805,189 (98.5)                       | 693,404 (95.7)                                |

Note: Total GPS points for the 25 individuals was 1,832,267 and nonresidential GPS points was 724,260. GLH, Google Location History; GPS, Global Positioning System.*Includes all GPS points outside of home and work locations.
been aware that location data was being collected, within a device they carry on a daily basis for other purposes (i.e., texting, phone calls, app usage). In this study, 23.3% of individuals stated that privacy concerns were their major reason for not participating. This number is likely low given that a large percentage of individuals (50%) did not report a reason for not participating. We had originally conducted a survey of a random sample of 1,000 twins in the overall WSTR to gauge whether study participants have GLH data available and if they would consider sharing the data with future studies. Of the 700 individuals who responded, 91% reported that they had GLH data and would provide the data for research. This response rate was much higher than our actual GLH data collection and shows that individuals in this original survey may not have understood exactly what GLH data are and what they comprise. In the consent process of our present study, we showed examples of GLH data to make sure individuals were fully aware of the data they were providing. We are currently developing automated scripts for deriving environmental exposure information from GLH data that would not require downloading the data to a local server (i.e., researchers would never see GLH data). We hypothesized that this may improve response rates, but initial qualitative feedback from individuals showed that participants may not understand the downloading process or view it as an important factor in their decision to provide data.

In addition to privacy concerns, there are other limitations to using GLH data. It is unknown what the response rate will be for collecting GLH data in other studies and populations. Individuals who are willing to provide GLH data to a health study are likely not representative of the general population. Specifically, we observed differences in individuals providing GLH data by age and socioeconomic characteristics. In addition, GLH data are not collected routinely at set time intervals (e.g., every 30 s as in a GPS logger), and some inaccuracies may exist in terms of the completeness of location data. In addition, some location information may have large spatial uncertainties, for example, comparing data from when indoors and signal strength in low or from when cell towers are being used to determine locations. However, because these data are collected over many years, errors should be minimized for long-term exposure measures. GLH data are also collected on different brands of smartphones (e.g., Android and iPhone) and data collection accuracy may be inconsistent across different platforms and/or models. GLH accuracy is also less precise for the earlier years when location services were not as well developed. Another consideration is that Google recently changed its policy regarding the retention time for GLH data.31 Previously, users had to turn this setting on if they did not want their location data stored indefinitely. Now, new users will have a setting to automatically delete location history data after 18 months, which may impact retrospective GLH data collection in the future (although this will only effect new Google users). Finally, it is important to address the potential issue of reverse causality, where a subset of people’s phone use and associated GLH data may be inversely correlated with specific behaviors. For example, individuals with a high affinity for nature and the outdoors may be more likely to pursue nature as an escape from the connected world and therefore be more likely to leave their phones in the car or at home during these activities. Further analysis is required to assess how much of an issue this may be, as well as other personal behavioral factors, and what environmental exposure measures will be most impacted.

Conclusion
We believe that collecting GLH data can provide a feasible and cost-effective method to capture long-term retrospective time-activity patterns for environmental health research. Cohort studies should consider adding GLH data collection to capture historical time-activity patterns of participants, employing a bring-your-own-location-data citizen science approach. Although response rates were relatively low and participants in the present study were not representative of overall study participants, the rates are comparable to GPS and survey studies. Privacy remains a concern that needs to be carefully managed when using GLH data.

Acknowledgments
We thank the Washington State Twin Registry participants for their time and engagement with this study. The research reported in this publication was supported by the National Institutes of Health (NIH)/National Institute of Environmental Health Sciences under awards R21ES031226 (to P.H.) and R33ES024715 (to G.D.) and the National Institute of Aging under award RO1AG042176 (to G.D.). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

References
1. Nuckols JR, Ward MH, Jarup L. 2004. Using geographic information systems for exposure assessment in environmental epidemiology studies. Environ Health Perspect 112(9):1007–1015, PMID: 15198921, https://doi.org/10.1289/ehp.6738.
2. Neveu V, Moussy A, Rouaix H, Wedekind R, Pon A, Knox C, et al. 2017. Exposome-Explorer: a manually-curated database on biomarkers of exposure to dietary and environmental factors. Nucleic Acids Res 45(D1):D979–D984, PMID: 27924041, https://doi.org/10.1093/nar/gkw980.
3. Larkin A, Hystad P. 2017. Towards personal exposures: how technology is changing air pollution and health research. Curr Environ Health Rep 4(4):463–471, PMID: 28983874, https://doi.org/10.1007/s40572-017-0163-y.
4. Setton E, Marshall JD, Brauer M, Lundquist KR, Hystad P, Keller P, et al. 2011. The impact of daily mobility on exposure to traffic-related air pollution and health effects estimates. J Expo Sci Environ Epidemiol 21(1):42–48, PMID: 20583235, https://doi.org/10.1038/jes.2010.14.
5. Guan X, Wang D, Cao X.J. 2020. The role of residential self-selection in land use-travel research: a review of recent findings. Transp Rev 40(3):267–287, https://doi.org/10.1080/01441647.2019.1629265.
6. Wild CP. 2005. Complementing the genome with an “exposome”: the outstanding challenge of environmental exposure measurement in molecular epidemiology. Cancer Epidemiol Biomarkers Prev 14(8):1847–1850, PMID: 16103423, https://doi.org/10.1158/1055-9965.EPI-05-0456.
7. Poom A, Willberg E, Toivonen T. 2021. Environmental exposure during travel: a research review and suggestions forward. Health Place 70:102584, PMID: 34020322, https://doi.org/10.1016/j.healthplace.2021.102584.
8. Ma X, Longley I, Gao J, Salmond J. 2020. Assessing schoolchildren’s exposure to air pollution during the daily commute—a systematic review. Sci Total Environ 737:140389, PMID: 32783874, https://doi.org/10.1016/j.scitotenv.2020.140389.
9. Fallah-Shorshani M, Hatzopoulou M, Ross NA, Patterson Z, Weichenthal S. 2018. Evaluating the impact of neighborhood characteristics on differences between residential and mobility-based exposures to outdoor air pollution. Environ Sci Technol 52(18):10777–10786, PMID: 30119601, https://doi.org/10.1021/acs.est.8b02260.
10. Katapally TR, Bhawra J, Patel P. 2020. A systematic review of the evolution of GPS use in active living research: a state of the evidence for research, policy, and practice. Health Place 66:102453, PMID: 33137684, https://doi.org/10.1016/j.healthplace.2020.102453.
11. Kroen PJ, Titze S, Dja P, Jones A, Onglie D. 2011. Use of global positioning systems to study physical activity and the environment: a systematic review. Am J Prev Med 41(5):508–515, PMID: 22011423, https://doi.org/10.1016/j.amepre.2011.06.046.
12. Garcia LMT, Diez Roux AV, Martins ACR, Yang Y, Florindo AA. 2018. Exploring the emergence and evolution of population patterns of leisure-time physical activity through agent-based modelling. Int J Behav Nutr Phys Act 15(1):112, PMID: 30453997, https://doi.org/10.1186/s12966-018-0750-9.
13. Chais B, Kestens Y, Perchoux C, Karusisi N, Merlo J, Labadi K. 2012. An inter-continental time-use travel research: a review of recent findings. Transp Rev 40(3):267–287, https://doi.org/10.1080/01441647.2019.1629265.
15. Stalvey BT, Dvosley C, Sloane ME, Ball K. 1999. The Life Space Questionnaire: a measure of the extent of mobility of older adults. J Appl Gerontol 18(4):460–478, https://doi.org/10.1177/073346489901800404.

16. Palmer JRB, Espenshade TJ, Bartumeus F, Chung CY, Ozgenci NE, Li K. 2013. New approaches to human mobility: using mobile phones for demographic research. Demography 50(3):1105–1128, PMID: 23192393, https://doi.org/10.1007/s13524-012-0175-z.

17. Pew Research Center. 2021. Mobile Fact Sheet. https://www.pewresearch.org/internet/fact-sheet/mobile/ [accessed 20 December 2021].

18. Bank of America. 2015. Trends in Consumer Mobility Report. https://www.scribd.com/document/269996227/2015-BAC-Trends-in-Consumer-Mobility-Report [accessed 2 November 2022].

19. Ruktanonchai NW, Ruktanonchai CW, Floyd JR, Tatem AJ. 2018. Using Google Location History data to quantify fine-scale human mobility. Int J Health Geogr 17(1):28, PMID: 30049275, https://doi.org/10.1186/s12942-018-0150-z.

20. Yu X, Stuart AL, Liu Y, Ivey CE, Russell AG, Kan H, et al. 2019. On the accuracy and potential of Google Maps Location History data to characterize individual mobility for air pollution health studies. Environ Pollut 252(pt A):924–930, PMID: 31226517, https://doi.org/10.1016/j.envpol.2019.05.081.

21. Cool D, McCallum SC, Rainham D, Taylor N, Patterson Z. 2021. Understanding Google Location History as a tool for travel diary data acquisition. Transp Res Rec 2675(5):238–251. https://doi.org/10.1177/0361198120986169.

22. Google. 2022. Google Account Help. Manage your Location History. https://support.google.com/accounts/answer/3118887?hl=en [accessed 19 May 2022].

23. Boakye KA, Amram O, Schuna JM Jr, Duncan GE, Hystad P. 2021. GPS-based built environment measures associated with adult physical activity. Health Place 70:102602, PMID: 34139613, https://doi.org/10.1016/j.healthplace.2021.102602.

24. Duncan GE, Avery A, Hurvitz PM, Moudon AV, Tsang S, Turkheimer E. 2019. Cohort profile: TWINS study of environment, lifestyle behaviours and health. Int J Epidemiol 48(4):1041–1041h, PMID: 30429093, https://doi.org/10.1093/ije/dyy224.

25. Duncan GE, Hurvitz PM, Moudon AV, Avery AR, Tsang S. 2021. Measurement of neighborhood-based physical activity bouts. Health Place 70:102595, PMID: 34090126, https://doi.org/10.1016/j.healthplace.2021.102595.

26. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. 2009. Research electronic data capture (REDCap)—a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform 42(2):377–381, PMID: 18929686, https://doi.org/10.1016/j.jbi.2008.08.010.

27. Harris PA, Taylor R, Minor BL, Elliott V, Fernandez M, O’Neal L, et al. 2019. The REDCap consortium: building an international community of software platform partners. J Biomed Inform 95:103208, PMID: 31078660, https://doi.org/10.1016/j.jbi.2019.103208.

28. Larkin A, Aarebrot A, Goldberg DL, Mohegh A, Brauer M, Hystad P. 2022. A global spatial-temporal land use regression model for nitrogen dioxide air pollution. Earth Space Sci Preprint posted online 17 May 2022, https://doi.org/10.1002/essoar.10511382.1.

29. U.S. EPA (U.S. Environmental Protection Agency). 2013. EPA ExpoBox (A Toolbox for Exposure Assessors). Updated 15 March 2013. https://www.epa.gov/expobox [accessed 3 June 2019].

30. Ramirez AH, Sulieman L, Schlueter DJ, Halvorson A, Ginn J, Ratsimbazafy F, et al. 2020. The All of Us Research Program: data quality, utility, and diversity. Patterns 3(9):100570, https://doi.org/10.1016/j.jpattern.2022.100570.

31. Google. 2022. Privacy & Terms—how google retains data we collect. https://policies.google.com/technologies/retention?hl=en [accessed 18 May 2022].