Association Between User-Generated Commuting Data and Population-Representative Active Commuting Surveillance Data — Four Cities, 2014–2015

Geoffrey P. Whitfield, PhD[1]; Emily N. Ussery, PhD[1]; Brian Riordan[2]; Arthur M Wendel, MD[3]

Creating environments that support all types of physical activity, including active transportation, is a public health priority (*1*). Public health surveillance that identifies the locations where community members walk and bicycle (i.e., engage in active transportation) can inform such efforts. Traditional population-representative active transportation surveillance incurs a considerable time lag between data collection and dissemination, and often lacks geographic specificity (*2*). Conversely, user-generated active transportation data from Global Positioning System (GPS)-based activity tracking devices and mobile applications can provide near real-time information, but might be subject to self-selection bias among users. CDC analyzed the association between GPS-based commuting data from a company that allows tracking of activity with a mobile application (Strava, Inc., San Francisco, California) and population-representative commuting data from the U.S. Census Bureau's American Community Survey (ACS) (*3*) for four U.S. cities. The level of analysis was the Census block group. The number of GPS-tracked commuters in Strava was associated with the number of ACS active commuters (Spearman's rho = 0.60), suggesting block groups were ranked similarly based on these distinct but related measurements. The correlation was higher in high population density areas. User-generated active transportation data might complement traditional surveillance systems by providing near real-time, location-specific information on where active transportation occurs.

Physical activity, including walking and bicycling for transportation, is a valuable health behavior. Public health surveillance can identify areas with high and low levels of active transportation and guide efficient investments in active transportation programs and infrastructure, as recommended in "Step It Up! The Surgeon General's Call to Action to Promote Walking and Walkable Communities" (*1*).

Historically, active transportation surveillance using surveys or logs has used purposeful sampling to provide population-representative estimates, despite a time lag between the activity and data availability (*2*). User-generated, GPS-tracked active transportation data are available in near real-time but are subject to selection bias because the only persons who contribute data are those who can use the requisite technology and have the interest in doing so. The magnitude of this bias has not been established. If user-generated active transportation data demonstrate some validity in measuring this behavior, they could complement traditional active transportation surveillance. Cities have begun using these data sources to plan infrastructure, so evaluation relative to an established surveillance system is important (*4*). The purpose of this analysis was to determine if the number of GPS-tracked active commuters is associated with the number of workers who walk or bicycle to work at the block group level in four U.S. cities.

Strava is one of several companies that has built an activity-tracking application and repository for GPS-tracked data. It was chosen for this analysis because of its large user base, amassed by offering a free social media platform where users can compare activities with peers. Further, cities have begun using Strava for planning bicycle and pedestrian infrastructure (*4*). Strava analysts identified user-logged commute trips (versus recreational trips) using two methods. First, users could mark or label activities as commutes in the application. Second, a proprietary algorithm was used that analyzed trip origin, destination, and timing to identify trips that were likely commute-oriented. Strava analysts provided CDC with the number of unique application users who started a commute trip in each block group in the study area during May 2014–May 2015 (GPS-tracked commuters). Block groups are subdivisions of census tracts and generally contain 600–3,000 residents. No personally identifiable information was provided to CDC, and the software users agreed to Strava's use of deidentified data at the time of registration.

Comparison data were obtained from ACS. ACS samples approximately 3.5 million addresses each year, and all residents at an address complete the survey. ACS achieves 96%–98% response rates with internet, mail, telephone, and in-person data collection. Employed ACS respondents aged ≥16 years reported the single mode of transportation that accounted for the majority of miles traveled to work during the previous week. The estimated number of commuters per block group who reported bicycling or walking (ACS active commuters) was downloaded from the Census Bureau (*5*). Although GPS-tracked commuters could be counted in any block group where they begin a commute trip, ACS active commuters were only counted in their block group of residence. Five ACS cycles were merged to increase the reliability of block group estimates and maintain respondent anonymity; for this analysis, ACS cycles 2009–2013 were used. Population density, which

is strongly associated with active transportation (*2*), was also obtained from ACS. ACS samples continuously to account for seasonal variation.

Four U.S. cities (Austin, Texas; Denver, Colorado; Nashville, Tennessee; and San Francisco, California) were selected, based on a high number of tracking application users and their geographic diversity across the United States. Because the number of active commuters (both GPS-tracked and ACS) was skewed and had a high prevalence of zero values, this analysis presents medians with interquartile ranges and Spearman's rank correlation coefficients (rho) (*6*). The large number of block groups resulted in uniformly significant correlation coefficients, so interpretation of rho followed Cohen (low = 0.1–0.3; moderate >0.3–0.5; strong >0.5–0.7) (*7*). The number of commuters per block group was analyzed both as a raw count and as a percentage of the block group population. Analyses were stratified by city and by population density tertiles.

Population density within block groups varied across cities; the median ranged from 2,785 persons per square mile in Nashville to 25,567 in San Francisco (Table 1). The median number of GPS-tracked commuters and ACS active commuters per block group was similar within each city and for the sample as a whole, with a maximum difference of five commuters per block group in San Francisco.

Across all block groups in all cities, the number of GPS-tracked commuters was strongly associated with the number of ACS active commuters (rho = 0.60). The correlation differed across cities, ranging from 0.28 in Nashville to 0.58 in San Francisco (Table 1). Analyses examining commuter percentages were similar to the count estimates (Table 1). The correlations were progressively stronger with higher block group population density, reaching rho = 0.61 for both numbers and percentages of active commuters in block groups with at least 10,443 persons per square mile (Table 2).

Discussion

Across block groups in four U.S. cities, the number of GPS-tracked commuters in Strava correlated with the number of ACS active commuters at rho = 0.60, indicating that these distinct but related variables rank block groups similarly regarding the presence of active transportation. This degree of correlation suggests some degree of convergent validity between user-generated, GPS-tracked commuting data and representative data from ACS.

The association between GPS-tracked and ACS commuter variables was stronger in cities and block groups with higher population densities. This finding might be attributable to a higher prevalence of activity tracking application users in more densely populated areas: information given to CDC by the data provider indicated the most densely populated city (San Francisco) also had the most GPS-tracking application users per capita (4.1%). As use of these applications increases within an area, the data produced by these users might more closely approximate the general population's behavior, and better match representative surveys like ACS.

Despite the differences between GPS tracking and ACS in sampling and assessment, the findings from this analysis suggest that user-generated, GPS-based activity tracking can perform similarly to ACS in identifying block groups where active transportation is common. In fact, the magnitude of the overall correlation (rho = 0.60) was larger than that seen in other comparable analyses. For example, when walk and bike commuting from ACS were disaggregated into two separate variables and compared in this same sample of block groups, they

TABLE 1. Correlations between block group level GPS-tracked and ACS active commuting variables, stratified by city — Austin, Denver, Nashville, and San Francisco, 2009–2013* and 2014–2015*

Characteristic	City				Total
	Austin	Denver	Nashville	San Francisco	
No. block groups	527	481	473	581	**2,062**
Population per block group, median (IQR)	1,469 (1,013)	1,134 (653)	1,172 (867)	1,289 (709)	**1,271 (829)**
Population density,† median (IQR)	4,234 (4,114)	7,077 (4,940)	2,785 (2,934)	25,567 (18,024)	**6,214 (13,425)**
Median no. of active commuters					
GPS-tracked, no. (IQR)	19 (30)	16 (27)	2 (6)	54 (89)	**17 (41)**
ACS, weighted, no. (IQR)	16 (43)	18 (52)	0 (13)	59 (118)	**18 (60)**
Spearman's rho§	0.36	0.52	0.28	0.58	**0.60**
Median percentages¶ of active commuters					
GPS-tracked, % (IQR)	1.1 (2.6)	1.5 (2.7)	0.2 (0.7)	4.4 (6.7)	**1.3 (3.5)**
ACS, % (IQR)	0.8 (3.0)	1.6 (4.2)	0 (1.2)	4.7 (9.3)	**1.3 (4.7)**
Spearman's rho§	0.37	0.49	0.27	0.55	**0.59**

Abbreviations: ACS = American Community Survey; GPS = Global Positioning System; IQR = interquartile range.
* ACS from 2009–2013 and GPS-tracked from 2014–2015.
† Persons per square mile of land area.
§ All Spearman's rho have p<0.001.
¶ Within block groups; count divided by total population.

were only moderately correlated at rho = 0.38 (data not shown). Further, previous research has assessed the association between physical activity questionnaires and accelerometer-assessed bodily movement in individual persons. A 2010 review found that only one of 41 questionnaires had a correlation >0.50 with accelerometer data (*8*). The correlation between GPS-tracked and ACS data in these block groups is as strong as or stronger than the correlation between questionnaire and accelerometer-based activity assessment among individual adults.

The findings in this report are subject to at least four limitations. First, although ACS served as a comparison measure, it is not a standard for assessing total population participation in active transportation because it does not capture infrequent and non-work active transportation. Second, user-generated GPS-tracked commuting data only capture trips made by persons who download and use the applications, and this group is likely more active than the general population. Similarly, these results cannot be generalized to all GPS data collection efforts because of potential differences in the user bases across systems. High numbers of users and variation in demographic characteristics and physical activity among users would likely yield more representative systems. Third, the algorithms used in the present study to identify commute-related trips are proprietary, and their actual performance is unknown. Finally, ACS data are self-reported via questionnaire for only the past week and subject to social desirability and recall biases.

Planning and evaluation of interventions to increase active transportation need detailed information about where and when persons engage in active transportation and will therefore benefit from location- and time-specific data. User-generated GPS data from mobile applications can capture this information, but their use could be limited by concerns about the performance of user-generated data in public health surveillance. Surveillance evaluation often includes comparison of a systems' data quality to the quality of existing methods (*9*). These results suggest that user-generated active transportation data might provide valuable information to assist with achieving public health and transportation goals. Additional research into the validity of other information collected from users (e.g., route, heart rate, and speed) might further support their usefulness.

Summary

What is already known about this topic?

City health and transportation officials are increasingly interested in measuring walking and bicycling, and user-generated, Global Positioning System (GPS)-tracked methods are emerging as popular choices. Questions remain about how representative the users of these systems are of the general population.

What is added by this report?

A comparison of user-generated GPS-tracked commuting data with similar data from a representative sample of the general U.S. population suggests that these systems similarly rank census block groups according to the presence of active commuting, and that the similarity might be stronger in areas that have a higher population density.

What are the implications for public health practice?

Public health and transportation officials need information on where and when persons engage in active transportation. User-generated, GPS-tracked data sources might provide critical information regarding active transportation to local health and transportation officials as a complement to traditional active transportation surveillance systems; these data might inform investments in active transportation programs and infrastructure.

TABLE 2. Correlations between block-group level GPS-tracked and ACS active commuting variables, stratified by tertile of population density — Austin, Denver, Nashville, and San Francisco, 2009–2013* and 2014–2015*

Characteristic	Population density tertile I	II	III	Total
No. persons per square mile	0–4,107	4,108–10,442	10,443–175,523	—
Median population density† (IQR)	2,211 (1,866)	6,214 (2,598)	23,254 (17,643)	**6,214 (13,425)**
Median no. of active commuters				
GPS-tracked, no. (IQR)	8 (24)	13 (25)	38 (72)	**17 (41)**
ACS, weighted, no. (IQR)	0 (21)	15 (42)	60 (119)	**18 (60)**
Spearman's rho§	0.40	0.49	0.61	**0.60**
Median percentages¶ of active commuters				
GPS-tracked, % (IQR)	0.7 (2.0)	1.0 (2.3)	2.9 (5.8)	**1.3 (3.5)**
ACS, % (IQR)	0.0 (1.6)	1.2 (3.4)	4.5 (9.3)	**1.3 (4.7)**
Spearman's rho§	0.38	0.50	0.61	**0.59**

Abbreviations: ACS = American Community Survey; GPS = Global Positioning System; IQR = interquartile range.
* ACS from 2009–2013 and GPS-tracked from 2014–2015.
† Persons per square mile of land area.
§ All Spearman's rho have p<0.001.
¶ Within block groups; count divided by total population.

[1]Division of Emergency and Environmental Health Services, National Center for Environmental Health, CDC; [2]Strava, Inc., Hanover, New Hampshire; [3]Division of Community Health Investigations, Agency for Toxic Substances and Disease Registry, Seattle, Washington.

Corresponding author: Geoffrey P. Whitfield, GWhitfield@cdc.gov, 770-488-3976.

References

1. US Department of Health and Human Services. Step it up! The Surgeon General's call to action to promote walking and walkable communities. Washington, DC: US Department of Health and Human Services, Office of the Surgeon General; 2015. http://www.surgeongeneral.gov/library/calls/walking-and-walkable-communities/
2. Whitfield GP, Paul P, Wendel AM. Active transportation surveillance—United States, 1999–2012. MMWR Surveill Summ 2015;64(No. SS-7). http://dx.doi.org/10.15585/mmwr.ss6407a1
3. US Census Bureau. American community survey. Washington, DC: US Bureau of Labor Statistics, US Census Bureau; 2015. https://www.census.gov/programs-surveys/acs/
4. Albergotti R. Strava, popular with cyclists and runners, wants to sell its data to urban planners. Wall Street Journal. May 7, 2014. http://blogs.wsj.com/digits/2014/05/07/strava-popular-with-cyclists-and-runners-wants-to-sell-its-data-to-urban-planners
5. US Census Bureau. American FactFinder. Washington, DC: US Department of Commerce; 2015. http://factfinder.census.gov
6. Huson LW. Performance of some correlation coefficients when applied to zero-clustered data. J Mod Appl Stat Methods 2007;6:530–6.
7. Cohen J. Statistical power analysis for the behavioral sciences. 2nd ed. Mahwah, NJ: Lawrence Erlbaum; 1988.
8. van Poppel MN, Chinapaw MJ, Mokkink LB, van Mechelen W, Terwee CB. Physical activity questionnaires for adults: a systematic review of measurement properties. Sports Med 2010;40:565–600. http://dx.doi.org/10.2165/11531930-000000000-00000
9. German RR, Lee LM, Horan JM, Milstein RL, Pertowski CA, Waller MN; Guidelines Working Group, CDC. Updated guidelines for evaluating public health surveillance systems: recommendations from the Guidelines Working Group. MMWR Recomm Rep 2001;50(No. RR-13).