OPEN

At-risk-measure Sampling in Case–Control Studies with Aggregated Data

Michael D. Garber,[a] *Lauren E. McCullough,*[a] *Stephen J. Mooney,*[b,c] *Michael R. Kramer,*[a] *Kari E. Watkins,*[d] *R.L. Felipe Lobelo,*[e] *and W. Dana Flanders*[a,f]

Abstract: Transient exposures are difficult to measure in epidemiologic studies, especially when both the status of being at risk for an outcome and the exposure change over time and space, as when measuring built-environment risk on transportation injury. Contemporary "big data" generated by mobile sensors can improve measurement of transient exposures. Exposure information generated by these devices typically only samples the experience of the target cohort, so a case-control framework may be useful. However, for anonymity, the data may not be available by individual, precluding a case–crossover approach. We present a method called at-risk-measure sampling. Its goal is to estimate the denominator of an incidence rate ratio (exposed to unexposed measure of the at-risk experience) given an aggregated summary of the at-risk measure from a cohort. Rather than sampling individuals or locations, the method samples the measure of the at-risk experience. Specifically, the method as presented samples person–distance and person–events summarized by location. It is illustrated with data from a mobile app used to record bicycling. The method extends an established case–control sampling principle: sample the at-risk experience of a cohort study such that the sampled exposure distribution approximates that of the cohort. It is distinct from density sampling in that the sample remains in the form of the at-risk measure, which may be continuous, such as person–time or person–distance. This aspect may be both logistically and statistically efficient if such a sample is already available, for example from big-data sources like aggregated mobile-sensor data.

Keywords: Big data; Case–control studies; Epidemiological monitoring; Epidemiologic studies; Location-based studies; Sampling studies

(*Epidemiology* 2021;32: 101–110)

Transient exposures are difficult to measure in epidemiologic studies, especially when both the status of being at risk for an outcome and the exposure change over time. For example, people are at risk of pedestrian fatality only while walking. During this at-risk period, exposures of interest, such as the presence of a sidewalk, may also vary. For etiologic inference, the transience of both at-risk status and exposure should be considered. Measures that account for this transience include person–time (e.g., hours walked), person–distance (kilometers walked), or person–events (intersections crossed on foot). We refer to these, collectively, as measures of the at-risk experience or at-risk measures.

Traditional methods to collect transient exposure measures, such as a survey or a travel log, can be both burdensome and inaccurate.[1] Developments in mobile sensing can make data collection less onerous and more accurate.[2] For example, global positioning system (GPS) measurements from mobile devices can be used to reconstruct a study participant's travel routes, from which time or distance exposed to certain environmental conditions at that place and time can be measured.[3–5] Such exposures might include transportation infrastructure,[6] air pollution,[3,7,8] liquor stores,[9] or areas of malaria risk.[10] In principle, researchers could request that individuals download and share their location history, which is passively tracked by common smartphones.[11,12] However, recruiting willing individuals may be impractical.[13]

Submitted February 28, 2020; accepted September 23, 2020.

[a]Department of Epidemiology, Rollins School of Public Health, Emory University, Atlanta, GA; [b]Department of Epidemiology, School of Public Health, University of Washington, Seattle, WA; [c]Harborview Injury Prevention & Research Center, University of Washington, Seattle, WA; [d]School of Civil and Environmental Engineering, Georgia Institute of Technology, Atlanta, GA; [e]Hubert Department of Global Health, Rollins School of Public Health, Emory University, Atlanta, GA; and [f]Department of Biostatistics and Bioinformatics, Rollins School of Public Health, Emory University, Atlanta, GA.

M.D.G. was supported by the National Heart, Lung, and Blood Institute (F31HL143900) and by the Doctoral Student Research Grant from the American College of Sports Medicine Foundation (18-00663). S.J.M. was supported by the National Library of Medicine (R00LM012868). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or the American College of Sports Medicine Foundation.

W.D.F. owns Epidemiologic Research & Methods LLC which does consulting work for pharmaceutical companies, environmental laboratories, and attorneys. The other authors report no conflicts of interest.

Process of obtaining code and data: Data-use agreements prohibit the authors from sharing the data used in the empirical example. R code following the example with simulated data is available at https://github.com/michaeldgarber/at-risk-measure-sampling.

SDC Supplemental digital content is available through direct URL citations in the HTML and PDF versions of this article (www.epidem.com).

Correspondence: Michael D. Garber, Department of Epidemiology, Claudia Nance Rollins Building, 1518 Clifton Road NE, Atlanta, GA 30322. E-mail: mdgarbe@emory.edu.

ISSN: 1044-3983/21/3201-0101
DOI: 10.1097/EDE.0000000000001268

To avoid individual recruitment, researchers might use existing sources of anonymized mobile-sensor-generated data with activity summarized by location rather than by person. *Strava*, for example, aggregates its users' bicycling data by street segment.[7,14,15] Using this information, the total bicycle–kilometers ridden in an area could then be calculated.

Whether mobile-sensor data are available by individual or are aggregated, the data will often not be available for the entire target population. Case–control sampling strategies should thus be considered. If data are available by person, a case-crossover design could be used.[16] Use of the case-crossover, however, requires data on an individual's exposure for at least one referent period,[17,18] precluding the method if the mobile-sensor sample is aggregated without individual-level follow-up.

If data are instead summarized by geographical location,[6,19] case–control sampling could be done with density sampling,[20,21] wherein entities are sampled with replacement with probability proportional to their total at-risk measure during the follow-up period. The at-risk measure is conventionally person–time at the individual-level,[21] but could also be person–distance summarized by roadway[6] or person–events summarized by intersection. Regardless of the at-risk measure used or the level at which it is summarized, a density sample is tabulated as a discrete frequency.[22] In this article, we argue that if a sample of an at-risk measure is available (e.g., person–distance traveled while cohort members used an app), then applying density sampling to that sample could result in a loss of information. Under plausible assumptions, the sample of the at-risk measure may simply be used *as is* as the control series.

We call this approach *at-risk-measure* sampling. The method's ultimate goal is to estimate the incidence rate ratio (IRR) that would have been observed in a study of the underlying cohort if the full exposure distribution were known. It is distinct from density sampling in that the sample remains in the form of the at-risk measure, which may be continuous like person–time or person–distance. It could also be discrete as in person–events. It is distinct from case-crossover sampling in that it may be used without individual-level information.

We first introduce the concept with a short vignette considering person–time. Second, we describe the method in detail assuming a sample of person–distance aggregated by location is available. Third, we comment on how the method applies to person–events, such as intersection crossings. Fourth, we describe methods to adjust for selection bias. Finally, we apply the method to an example study of the incidence rate of bicycle crashes and include a simulated demonstration informed by the empirical example.

METHODS

The target measure of association is the IRR from a cohort study,

$$\mathrm{IRR} = \frac{\dfrac{\text{Exposed cases}}{\text{Measure of total at}-\text{risk experience exposed}}}{\dfrac{\text{Unexposed cases}}{\text{Measure of total at}-\text{risk experience unexposed}}} = \frac{\dfrac{\text{Exposed cases}}{\text{Unexposed cases}}}{\dfrac{\text{Measure of total at}-\text{risk experience exposed}}{\text{Measure of total at}-\text{risk experience unexposed}}},$$

where the second equality follows by rearranging terms. We assume exposure status of cases at the time of occurrence is known, so, as with density sampling, the specific goal is to estimate the ratio in the denominator:

$$\frac{\text{Measure of total at}-\text{risk experience exposed}}{\text{Measure of total at}-\text{risk experience unexposed}}.$$

Introductory Example with Person–Time

To introduce the sampling concept, we consider a hypothetical cohort study in which 300 at-risk person–hours were exposed, and 800 were unexposed. Some in the cohort intermittently used a mobile sensor to report their person–time at risk. If 50% of the person–hours were sampled (by having sensor on) independently of exposure, then, in expectation, the sample would contain 150 exposed person–hours and 400 unexposed person–hours. The ratio of exposed to unexposed person–hours in the sample (150/400 = 0.375) would, apart from random error, equal that of the cohort (300/800 = 0.375). This vignette captures the method's essence: sampling a continuous measure of the at-risk experience such that the exposed-to-unexposed ratio in the sample estimates that of the cohort.

Notation and Setup: Person–Distance Sampling

We now describe the method assuming a sample of person–distance is available with data summarized by spatially referenced segments. Segments are unique portions of roadways or trails. In applications we consider, they are defined by mapping software like OpenStreetMap as the stretch of roadway or trail between intersections and are about 100 m in length. Distance traveled is often of interest in transportation-related outcomes.[23,24] Apps used to track bicycling,[7,14,15] to hail rides,[25] or to otherwise monitor movement[26,27] may collect distance traveled by users. To protect user privacy, data may be aggregated before being made available for research. A spatial summary might also be preferred by the research question.

The cohort is defined by activity at risk during a time period in a study area, say, a city. People may freely enter and leave the cohort; only their at-risk activity in the study area and timeframe is considered part of the cohort. The goal is to estimate the IRR between a transient exposure and an acute outcome in this cohort. The study area is defined by a set of M segments, m, $m = 1, 2, \ldots, M$, of time-invariant length

L_m, $0 \le L_m < \infty$, classified by a binary exposure condition, $E_m = 1$ or $E_m = 0$. If exposure varies spatially within segment (e.g., a segment is 60% exposed rather than 100% or 0%), we recommend redefining segments with geographic information systems software so that exposure is constant with segments. The number of times any individual travels along segment *m* in either direction while at risk for the outcome is denoted by N_m, $N_m = 0, 1, 2, .., \infty$. Using the vertical bar to denote "given that," the total person–distance in the cohort in exposure category *e* is

$$D_e = \sum_{m=1}^{M} L_m * N_m | E_m = e. \quad (1)$$

Note: the totals in each exposure category can be equivalently defined as the sum of activity of individual people, assuming the same cohort definition is used (eAppendix 1; http://links.lww.com/EDE/B732). Thus, conclusions drawn from this location-based framework can apply to individuals.

Suppose some of the individuals who ever pass through the study area during the timeframe at times use a mobile sensor or smartphone app to record their activity. For anonymity, their activity is summarized by segment before it is made available for research (e.g., https://metro.strava.com/; accessed 3 January 2020). The number of times a person travels along segment *m* in the time period while using the sensor, and thus in the sample, is denoted by $n_m : n_m = 0, 1, 2, \ldots, N_m$, for $m = 1, 2, \ldots, M$. The sample is an element of the sample description space, $\Omega = \left\{ n_1, n_2, \ldots, n_M : n_m = 0, 1, 2, \ldots, N_m ; m = 1, 2, \ldots, M \right\}$, in which every n_m takes one of its possible values. At every segment m, the sampling fraction, f_m, is the ratio of the number of app-users' trips over the segment in the sample (i.e., using the sensor), n_m, to the corresponding total number in the cohort, $N_m : f_m = \frac{n_m}{N_m}; 0 \le f_m \le 1$. If and only if $f_m = 0$, then segment *m* is not sampled. Recall, segments are classified as either exposed ($E_m = 1$) or unexposed ($E_m = 0$). The total sampled person–distance in exposure category *e* is

$$d_e = \sum_{m=1}^{M} L_m \times N_m \times f_m | E_m = e. \quad (2)$$

Finally, the overall proportion of distance sampled in each exposure category is, respectively, $f_{D,1} = \frac{d_1}{D_1}$ and $f_{D,0} = \frac{d_0}{D_0}$.

Note: for simplicity, we consider one time period. If time were a covariate of interest, variables could be indexed by time period, $t, t = 1, 2, \ldots, T$, and total sampled person–distance in exposure category *e* would then be $d_e = \sum_{m=1,t=1}^{M,T} L_m \times N_{m,t} \times f_{m,t} | E_{m,t} = e$.

Condition for Statistical Consistency

We now show a condition sufficient for the ratio of exposed to unexposed sampled person–distance, $\frac{d_1}{d_0}$, to consistently estimate that of the cohort, $\frac{D_1}{D_0}$. The condition is that the ratio of expected values of exposed to unexposed person–distance in the sample equals that of the cohort:

$$\frac{E\left[\sum_{m=1}^{M} L_m \times N_m \times f_m | E_m = 1\right]}{E\left[\sum_{m=1}^{M} L_m \times N_m \times f_m | E_m = 0\right]} = \frac{E\left[\sum_{m=1}^{M} L_{\boldsymbol{m}} \times N_m | E_m = 1\right]}{E\left[\sum_{m=1}^{M} L_m \times N_m | E_m = 0\right]}. \quad (3)$$

Alternatively phrased by re-arranging terms, the expected person–distance exposed in the sample divided by the expected exposed person–distance in the cohort should equal the corresponding fraction for the unexposed:

$$\frac{E\left[\sum_{m=1}^{M} L_m \times N_m \times f_m \mid E_m = 1\right]}{E\left[\sum_{m=1}^{M} L_m \times N_m \mid E_m = 1\right]} = \frac{E\left[\sum_{m=1}^{M} L_m \times N_{\boldsymbol{m}} \times f_m \mid E_m = 0\right]}{E\left[\sum_{m=1}^{M} L_m \times N_m \mid E_m = 0\right]}. \quad (4)$$

That equation 3 is sufficient follows from the Law of Large Numbers. Specifically, $\frac{d_1}{d_0}$ converges in probability to $\frac{E[d_1]}{E[d_0]}$. Once the cohort's activity occurs, D_1 and D_0 are considered constant, so $\frac{E[D_1]}{E[D_0]} = \frac{D_1}{D_0}$. Thus, if $\frac{E[d_1]}{E[d_0]} = \frac{E\left[\sum_{m=1}^{M} L_m \times N_m \times f_m \mid E_m = 1\right]}{E\left[\sum_{m=1}^{M} L_m \times N_m \times f_m \mid E_m = 0\right]} = \frac{E\left[\sum_{m=1}^{M} L_{\boldsymbol{m}} \times N_m \mid E_m = 1\right]}{E\left[\sum_{m=1}^{M} L_m \times N_m \mid E_m = 0\right]} = \frac{E[D_1]}{E[D_0]}$, then $\frac{d_1}{d_0}$ is a consistent estimator of $\frac{D_1}{D_0}$.

The condition is flexible in that it allows dependence between L_m, N_m, and f_m, as depicted in Figure 1. As long as associations are such that equation 3 is satisfied, consistency holds. Nevertheless, the condition may benefit from simplification. One possibility, which is partially verifiable, is to consider whether, in each exposure category, segment length L_m is independent of both the number of times the segment was traveled upon in the cohort, N_m, and in the sample, n_m. If so, then equation 4 simplifies to

$$\frac{E\left[N_m \times f_m | E_m = 1\right]}{E\left[N_m | E_m = 1\right]} = \frac{E\left[N_m \times f_m | E_m = 0\right]}{E\left[N_m | E_m = 0\right]}. \quad (5)$$

Person–Event Sampling

The principles also apply if the at-risk measure were person–events rather than person–distance. Person–events are conceptualized as being countable—discrete occurrences of being at risk for the outcome. For example, Strava counts the

number of times intersections were crossed by app-using bicyclists, and Yelp counts the number of times restaurants were visited by app-using diners.[28] The sampling concept is essentially the same, except the length dimension, L_m, need not be considered (Figure 2). As detailed in eAppendix 2; http://links.lww.com/EDE/B732, the condition for consistency is equation 5, above.

Correcting Selection Bias

The condition, equation 3, may often be plausible. For example, if the mobile sensor is always on and does not require user input, then the sampling mechanism may not be associated with exposure. However, the condition will not always hold. Individuals who use mobile sensors requiring manual input may have distinct behavior patterns from non-users

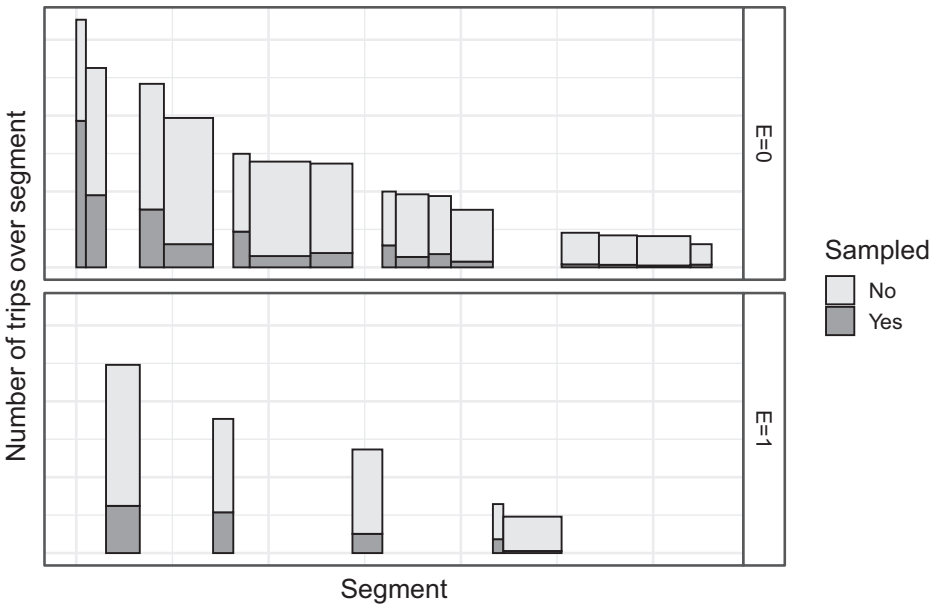

FIGURE 1. A bar chart satisfying the condition, equation 3. Segment length, L_m, is depicted by the width of the bars, and the number of trips over each segment, N_m, is depicted by the height of the bars. Sampled trips, enumerated as n_m, are shaded. Each segment's person–distance, D_m, is the area inside the bar. As defined above, the proportion sampled, f_m, is the proportion of N_m sampled, i.e., $f_m = \frac{n_m}{N_m}$. In this scenario, f_m is positively associated with N_m and negatively associated with L_m. Nevertheless, the ratio of exposed to unexposed person–distance in the sample is about the same as that in the cohort. R code to produce the figure is here: https://github.com/michaeldgarber/at-risk-measure-sampling/blob/master/code/code-for-figures.md.

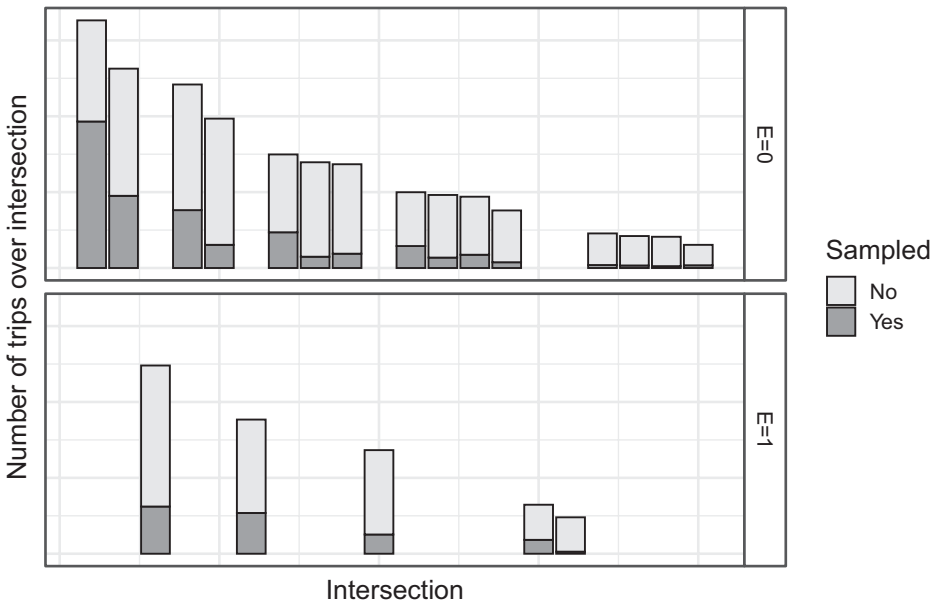

FIGURE 2. An illustration of aggregated person–event sampling at intersections. The notation follows that of Figure 1, except intersections are conceptualized as having no length dimension, reflected by the constant width of the bars. R code to produce the figure is here: https://github.com/michaeldgarber/at-risk-measure-sampling/blob/master/code/code-for-figures.md.

which differ by exposure.[29] For example, app-using bicyclists may tend to prefer certain infrastructure.[30]

To check and correct for selection bias, established principles can be followed.[31–34] Here, we describe inverse-probability-of-selection weighting (IPSW) with internal validation data. In the supplement (eAppendix 3; http://links.lww.com/EDE/B732), we also consider estimating a summary bias "breaker."[33] IPSW uses a subset of S segments, $m = 1, 2, \ldots, S, \ldots, M$, containing measurement of both N_m and the sampled (e.g., app-recorded) version, n_m, such that the true sampling fraction, f_m, is known among the subset. Such data are commonly available in transportation settings, where the total count of motor vehicles,[35] bicyclists,[15,36] or pedestrians[37] may be monitored at some locations. The estimated sampling fraction beyond the subset, $\hat{f}_m$, can then be estimated, possibly with a parametric model.[34] IPSW to estimate N_m then proceeds as follows: $\hat{N}_m = \frac{n_m}{\hat{f}_m}$. Substituting $\hat{N}_m$ for N_m in equation 1, the estimated total person–distance in exposure category e is

$$\hat{D}_e = \sum_{m=1}^{M} L_m \times \frac{n_m}{\hat{f}_m} | E_m = e. \quad (6)$$

$\frac{D_1}{D_0}$ can then be estimated by $\frac{\hat{D}_1}{\hat{D}_0}$. The validity of this approach will depend on the validity of the method used to estimate $\hat{f}_m$. The usual considerations for IPSW apply including the possibility of unstable weights if $\hat{f}_m$ is very small.[38] If internal validation data are not available and bias parameters cannot be drawn from literature, targeted-adjustment sensitivity analyses may be conducted.[32]

EXAMPLE

Data Sources and Exposure Definition

To illustrate the method, we estimate the IRR of police-reported crashes between bicyclists and motor vehicles per bicycle-kilometer traveled comparing roadway types in Atlanta, GA, for 23 months between 1 October 2016 and 31 July 2018. We classified the roadway type as primary, secondary, tertiary, or residential by the OpenStreetMap definition (https://wiki.openstreetmap.org/wiki/Key:highway#Roads; accessed 19 April 2020). We considered residential roadways unexposed and classified the other three types as exposed. The study area is the set of roadways in either condition in the 8.85-km radius around the intersection of Ponce de Leon Ave NE and Monroe Dr NE (Figure 3).

We obtained a sample of bicyclist–distance in the area and timeframe from Strava. Strava outputs the number of times a bicyclist travels along a segment while using the app. Segments, described above, were defined by Strava using the OpenStreetMap basemap (n, segments = 20,276; Table 1). Following equation 2, we calculated sampled bicycle–distance on each segment by multiplying the number of times the segment was ridden in Strava by the segment's length (Figure 3; Table 2). We obtained geolocated police-reported crashes (n = 129) between bicyclists and drivers of motor vehicles from a state registry. Information on Strava use is not available for the crashes. If a crash occurred at an intersection between exposure types (n = 46; e.g., a primary road intersecting a residential road), we classified the case as occurring on the road from which the bicyclist entered the intersection, as described on the police report. We assume that the cases, bicycle crashes reported specifically by police, are accurately classified and are not under-reported.[40] The study was approved by the Emory University Institutional Review Board (IRB00105514). All or a portion of included data was licensed by Strava.

Analysis

We estimate the IRR in four ways. We first report the estimated bicycle–distance from Strava. We then adjust for possible selection bias using IPSW. Third, although causal inference is not necessarily the explicit goal, we adjust for possible confounding. We finally employ multiple bias analysis.

Following equation 2, we report summed bicycle–distance from Strava in each exposure category, yielding an estimated IRR of 3.0 [95% confidence interval (CI): 1.4, 5.9; Table 2]. If the condition, equation 3, holds, then the observed ratio of exposed to unexposed bicycle–distance of 1.4 (95% CI: 1.1, 1.9) estimates that of the cohort, and the $\hat{IRR}$ estimates the *IRR* of the cohort. We checked whether we could simplify the condition per equation 5 but observed a weak association between L_m and n_m (correlation = –0.09), so we did not simplify.

Sampling uncertainty was estimated by hierarchical bootstrapping. Street segments are nested within street, which are also defined by OpenStreetMap. To consider correlation of nested observations, we first resampled streets with any crashes (n = 75) with replacement without weighting. Then, within sampled streets, we resampled segments containing any crashes (n = 115) with replacement, weighted by their number of crashes. Similarly, streets with any bicycle ridership in Strava (n = 2,773) were resampled with replacement without weighting. Then, within sampled streets, segments (n = 20,276) were resampled with replacement weighted by the number of times the segment was ridden. Over 1,000 replications, the 95% CI for each parameter of interest was calculated by the percentile method[41] as the empirical 2.5th and 97.5th percentiles.

Second, to check and adjust for selection bias, we used IPSW. Previous research from Atlanta reported that cyclists who used Strava tended to be more confident cyclists,[30] which may mean that they are more likely than non-app users to ride

TABLE 1. Descriptive Statistics of the Sample of Bicycle Ridership Reported by Strava at the Segment Level Between 1 October 2016 and 31 July 2018 in the 8.85-km Radius Around the Intersection of Ponce de Leon Ave NE and Monroe Dr NE in Atlanta, GA

	Exposed: Primary, Tertiary, or Secondary Roadway[a] (n, Segments = 8,959)		Unexposed: Residential Roadway[a] (n, Segments = 11,317)		Total (n, Segments = 20,276)	
Attribute	**Mean (SD)**	**Median (25th, 75th)**	**Mean (SD)**	**Median (25th, 75th)**	**Mean (SD)**	**Median (25th, 75th)**
Segment length (m), L_m	62 (66)	42 (19, 84)	105 (104)	78 (33, 137)	86 (92)	59 (25, 116)
Number of times traveled in sample, n_m	2,714 (3,924)	1,245 (515; 3,173)	949 (2,047)	200 (35, 840)	1,729 (3,148)	570 (115; 1,855)
Proportion of n_m classified as a commute[b]	0.34 (0.14)	0.34 (0.26, 0.43)	0.24 (0.21)	0.22 (0.00, 0.36)	0.28 (0.19)	0.29 (0.16, 0.40)

25th and 75th refer to percentiles.
[a]Classified by OpenStreetMap: https://wiki.openstreetmap.org/wiki/Key:highway#Roads; accessed 19 April 2020.
[b]Proportion commute is used in the example's logistic-regression model for inverse-probability-of-selection weighting. Rides were classified as commutes by Strava if the ride's start and end are were more than 1 km apart[39] or if the user tagged it as a commute in the app.

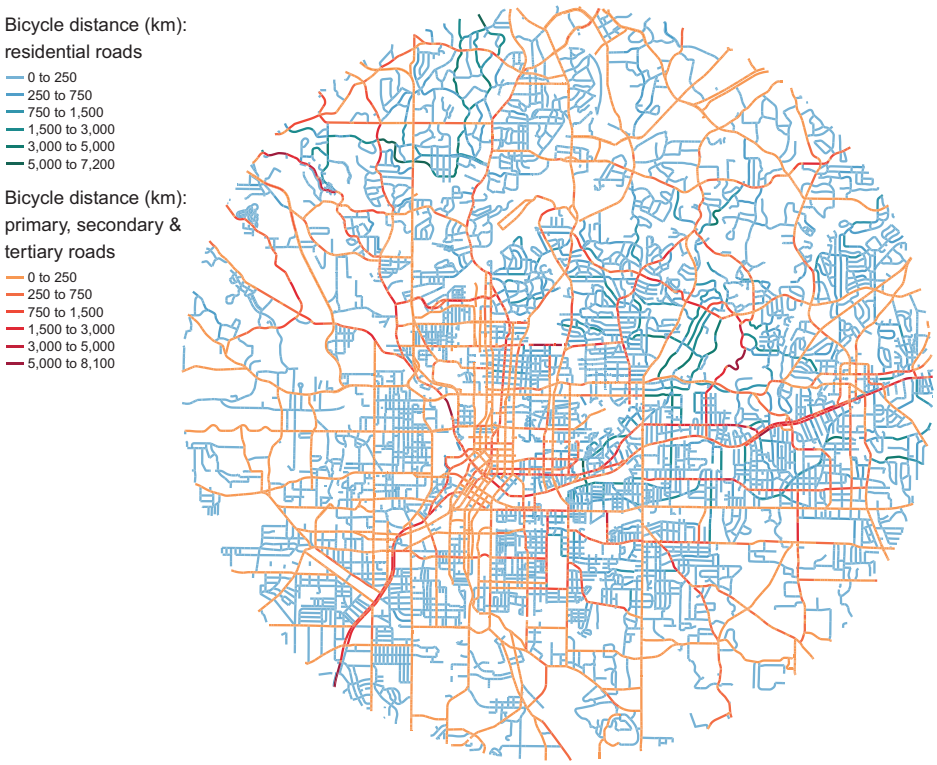

FIGURE 3. Bicycle distance (km) reported by exposure category in Strava between 1 October 2016 and 31 July 2018 in an 8.85-km radius around the intersection of Ponce de Leon Ave NE and Monroe Dr NE in Atlanta, GA.

on nonresidential roadways, possibly inducing selection bias. To implement IPSW, with permission of the City of Atlanta,[42] we gathered internal validation data from nine stationary counters (manufacturer: Eco-Counter Urban ZELT) that monitor bicycle traffic in real time with high accuracy[37,43] located at fixed locations throughout the study area. Data from the nine counters were available, on average, in 17.3 of the 23 study months. We calculated the sampling fraction in the 156 segment-month observations. We then fit an events-trials logistic regression model in those observations to estimate the sampling fraction when and where it was unknown (n, segment-month observations = 466,192). Independent variables and model performance are described in eAppendix 4; http://links.lww.com/EDE/B732. To avoid extreme and implausible values, we truncated[38] the estimated sampling fraction at 0.02 0.5. The total person–distance in each exposure category was then estimated, first summing over month, by equation 6. The resulting estimated exposed-to-unexposed ratio of person–distance was 1.3 (95% CI: 1.1, 1.9; Table 2), suggesting, conditional on the validity of the IPSW model, that the unadjusted exposure ratio of 1.4 (95% CI: 1.1, 1.9) was a slight overestimate and the condition, equation 3, was narrowly violated. Contrary to our hypothesis above, the model suggests Strava-recorded rides were relatively more likely to have occurred on

TABLE 2. An Application of At-risk-measure Sampling in an Example Study Estimating the IRR of Bicycle Crashes Comparing Roadway Types in Atlanta, Georgia Between 1 October 2016 and 31 July 2018

Measure	Scenario	Exposed (95% CI)	Unexposed (95% CI)	Exposure Ratio (95% CI)	IRR (95% CI)
Number of crashes	No bias adjustment	104 (73, 140)	25 (15, 38)	4.2 (2.1, 8.2)	
Bicycle–distance (person–km)	No bias adjustment	1.5×10^{6} (1.2×10^{6}, 1.8×10^{6})	1.1×10^{6} (0.9×10^{6}, 1.2×10^{6})	1.4 (1.1, 1.9)	3.0 (1.4, 5.9)
Bicycle–distance (person–km)	IPSW only	1.3×10^{7} (0.5×10^{7}, 3.0×10^{7})	1.0×10^{7} (0.3×10^{7}, 1.9×10^{7})	1.3 (1.1, 1.9)	3.1 (1.4, 6.0)
Number of crashes	Standardization only			3.4 (1.8, Inf)	
Bicycle–distance (person–km)	Standardization only			1.4 (1.1, 1.9)	
Pseudo-IR[a] (per sampled person–km)	Standardization only	5.6×10^{-5} (3.1×10^{-5}, 7.9×10^{-5})	2.3×10^{-5} (0.0, 3.5×10^{-5})		2.4 (1.2, Inf)
Number of crashes	IPSW, then standardization			3.5 (1.8, Inf)	
Bicycle–distance (person–km)	IPSW, then standardization			1.4 (1.1, 2.0)	
IR (per estimated person–km)	IPSW, then standardization	6.4×10^{-6} (2.2×10^{-6}, 19.5×10^{-6})	2.5×10^{-6} (0.0, 8.0×10^{-6})		2.6 (1.2, Inf)

The 95% CI was estimated by the percentile method using nonparametric hierarchical bootstrapping (N, replications = 1,000). For scenarios including IPSW, the 95% CI also considers the uncertainty due to the parametric sampling-fraction model.

Please see text for exposure definition and additional detail.

[a]The *pseudo-incidence rate* is the number of crashes divided by the sampled person–distance, following the terminology on p. 113 of Modern Epidemiology, 3rd Edition.[21]

residential (unexposed) roadways. To estimate uncertainty, we resampled residuals from the sampling-fraction model in each bootstrapped replicate, so the 95% CI reflects uncertainty from both the sampling and the parametric model.

Third, although the method may be used to estimate associations whether or not causal inference is the goal,[44] we consider confounding adjustment by standardization. A confounder might be time period or another covariate. We considered an area-level socioeconomic status indicator as a possible confounder (defined in eAppendix 5; http://links.lww.com/EDE/B732) and standardized results using a weighted geometric mean. A weighted arithmetic mean[45] could also be used. In the context of case–control studies, the weighted geometric mean has the useful property of preserving the symmetry of the cross-product ratio,

$$IRR_{\text{wgm}} = \frac{\left(\frac{Y_1}{d_1}\right)_{\text{wgm}}}{\left(\frac{Y_0}{d_0}\right)_{\text{wgm}}} = \frac{\left(\frac{Y_1}{Y_0}\right)_{\text{wgm}}}{\left(\frac{d_1}{d_0}\right)_{\text{wgm},}}$$

where Y_e denotes the number of crashes in exposure category e, and the wgm subscript denotes the weighted geometric mean. The resulting IRR_{wgm} was 2.4 [95% CI: 1.2, Infinity (Inf)], with the 95% CI again calculated using hierarchical bootstrapping.

Finally, we conducted multiple bias analysis by first employing IPSW and then standardization, following advice to adjust for possible biases in the expected reverse order in which they may have manifested.[32] The resulting IRR of 2.6 (95% CI: 1.2, Inf) was slightly higher than when adjusting for confounding alone [IRR = 2.4 (95% CI: 1.2, Inf)] but lower than when adjusting for selection bias alone [IRR = 3.1 (95% CI: 1.4, 6.0); Table 2].

Online Demonstration

In an online repository (https://github.com/michaeldgarber/at-risk-measure-sampling), we have posted accompanying R code. One script presents a simple simulation to illustrate the condition, equation 3. Another follows the above empirical analysis with simulated hierarchical data.

DISCUSSION

We have presented a method called at-risk-measure sampling, the goal of which is to estimate the ratio of the exposed to unexposed measure of the experience at risk in a rate-based cohort study. The concept aligns with established case–control principles[20,21,46]: sample the at-risk experience of a cohort study such that the sampled exposure distribution approximates that of the cohort. Its new aspect is that it accommodates sampling a continuous measure of the at-risk experience when aggregated by location. We expect the method to be useful when studying geographically specific exposures and acute outcomes with short induction periods.

Comparison with Existing Methods

The method has similarities with case-base sampling, case-crossover sampling, methods estimating individual-level risk from area-level summaries, and incidence density sampling. Depending on the circumstances, at-risk-measure sampling may have practical and conceptual advantages compared with these established methods. The method is related to the case-base design[47] in that the case-base approach may consider time at risk in the control series when calculating the estimated measure of association.[48] Distinct from the case-base approach, which conventionally samples individuals, the proposed approach samples the possibly more general at-risk experience. To our knowledge, the case-crossover[18,49] and its relatives[50] are the only established methods in the case–control family that have sampled the at-risk experience in continuous

form as the control series.[18,51] In example tables in Maclure's seminal paper on the case-crossover, control data are measured in person–hours.[18] At-risk-measure sampling is similar to generalizations of the case-crossover like the case-time-control[50] in that it may sample a continuous measure of the at-risk experience and may include that of non-cases (see Figure 3 by Schneeweiss et al.[17]). It is distinct from those approaches in that it explicitly accommodates an aggregated summary of the at-risk experience. The example study above might have used a case–crossover,[52] but doing so would have required measurement of exposure in a referent period for each case, information we did not have. The presented method provides another option for estimating epidemiologic measures of association using existing aggregated mobile-sensor data.[14,25,26]

Other methods have considered settings combining individual-level information about cases with area-level or ecological summaries of the source population at risk.[53–56] While similar, the proposed method differs from these approaches in that we assume a sample of the at-risk measure is available in each stratum of the exposure. Although particulars vary, the referenced works have considered the situation wherein only an area-level summary of exposure is available, and within-area variation is not directly accessible.[53–56]

Unmatched incidence density sampling can also be used with data summarized by location.[6,19] Modifying the definition on p. 124[21] to consider locations, under unmatched location-based density-sampling, segments are sampled with replacement with probability proportional to the person–distance at risk occurring on them and then, if exposure varies spatially within the sampled segment, a point along each sampled segment is sampled in a second stage to measure exposure. The study by Aldred and colleagues[6] illustrates this approach. In the proposed approach, rather than sampling places with replacement with probability proportional to their at-risk measure, the aggregated at-risk measure is sampled without replacement directly.

We see three advantages to this distinction. First, it can be practically advantageous if a sample of an at-risk measure is already available in this format, as in our example. Second, the proposed approach is conceptually appealing in that it reinforces the notion that a case–control study is a sample of a hypothetical cohort study.[21,46,57] When the risk ratio is the target parameter, this concept is clear: people are the at-risk measure, and people are sampled.[48,58] In contrast, when the IRR is the target parameter, a density sample is expected to be proportional to person–time but is not itself a direct sample of person–time, depending on the definition.[21] If instead the IRR is estimated by sampling person-moments from the study base,[22,59,60] this concept is perhaps more explicit.[60] The proposed approach might be viewed as extending person-moment sampling[61] to settings with a sample of aggregated, possibly continuous, at-risk data. Third, the proposed approach may be more statistically precise than density sampling. In our example, bicycle–distance was obtained in a single stage (from our perspective) from Strava, and exposure was measured within that sample. Conceivably, we could have applied density sampling to the first-stage sample by sampling segments with replacement with probability proportional to their sampled person–distance. This additional sampling would have lost statistical efficiency without practical benefit. The total sampling variance would have been at least as large as that of the original sample[61] and would have garnered no new information since exposure was known throughout the initial sample.

Scope and Threats to Validity

As with location-based density sampling, we imagine the method being most useful for studies of associations between spatially specific measures of exposure and acute outcomes with brief induction periods. Example questions of interest may be associations between roadway infrastructure and crashes in transportation epidemiology, as illustrated above, or associations between environmental factors and violent assault.[62] As the induction period becomes longer and the exposure less spatially specific, the approach may be less informative. For example, anonymized restaurant check-ins have been used to study foodborne illness.[28,63] Location-level person–event sampling, as described here, could be used to study the IRR between staph infection and exposure to restaurants of concern, but correct specification of the induction period may be a strong assumption.[64,65]

Selection bias may also be a threat to validity. We illustrated IPSW to adjust for possible selection bias. Our IPSW model was limited in that it was fit in only 156 observations and predicted the sampling fraction for 466,192 observations. Although selection bias is possible, an important result is that consistency can hold despite an association between sampling and the at-risk measure (Figure 1). This result is reassuring because people who spend more time at risk may more often sample their activity with a sensor.[29,30]

CONCLUSIONS

In closing, the at-risk-measure sampling method applies a well-established concept: that a sample of the risk experience from a cohort study can be used to estimate the ratio measure of association.[20,21,46,48] Its new aspect is that it accommodates sampling a continuous measure of the at-risk experience for studies aiming to estimate the IRR with aggregated data summarized by location. Compared with the usual density sampling approach, the sample may sometimes be simpler to obtain, like from existing sources of mobile-sensor data.

ACKNOWLEDGMENTS

For assistance with data collection, we thank E. Dave Adams with Georgia Department of Transportation and Haynes Bunn with Strava.

REFERENCES

1. Kang B, Moudon AV, Hurvitz PM, Reichley L, Saelens BE. Walking objectively measured: classifying accelerometer data with GPS and travel diaries. *Med Sci Sports Exerc*. 2013;45:1419–1428.
2. Chaix B. Mobile sensing in environmental health and neighborhood research. *Annu Rev Public Health*. 2018;39:367–384.
3. Breen MS, Long TC, Schultz BD, et al. GPS-based microenvironment tracker (MicroTrac) model to estimate time-location of individuals for air pollution exposure assessments: model evaluation in central North Carolina. *J Expo Sci Environ Epidemiology*. 2014;24:412–420.
4. Lu M, Schmitz O, Vaartjes I, Karssenberg D. Activity-based air pollution exposure assessment: differences between homemakers and cycling commuters. *Health Place*. 2019;60:102233.
5. Morrison CN, Byrnes HF, Miller BA, et al. Assessing individuals" exposure to environmental conditions using residence-based measures, activity location-based measures, and activity path-based measures. *Epidemiology*. 2019;30:166–176.
6. Aldred R, Goodman A, Gulliver J, Woodcock J. Cycling injury risk in London: a case-control study exploring the impact of cycle volumes, motor vehicle volumes, and road characteristics including speed limits. *Accid Anal Prev*. 2018;117:75–84.
7. Lee K, Sener IN. Understanding potential exposure of bicyclists on roadways to traffic-related air pollution: findings from El Paso, Texas, using Strava metro data. *Int J Environ Res Public Health*. 2019;16:371.
8. Bekö G, Kjeldsen BU, Olsen Y, et al. Contribution of various microenvironments to the daily personal exposure to ultrafine particles: Personal monitoring coupled with GPS tracking. *Atmos Environ*. 2015;110:122–129.
9. DiMaggio C, Mooney S, Frangos S, Wall S. Spatial analysis of the association of alcohol outlets and alcohol-related pedestrian/bicyclist injuries in New York City. *Inj Epidemiology*. 2016;3:11.
10. Hast M, Searle KM, Chaponda M, et al; Southern and Central Africa International Centers of Excellence for Malaria Research. The use of GPS data loggers to describe the impact of spatio-temporal movement patterns on malaria control in a high-transmission area of northern Zambia. *Int J Health Geogr*. 2019;18:19.
11. Lindquist M, Galpern P. Crowdsourcing (in)voluntary citizen geospatial data from google android smartphones. *J Digit Landsc Archit*. 2016;1:263–272.
12. Google Maps Timeline - Google Maps Help. Available at: https://support.google.com/maps/answer/6258979?co=GENIE.Platform%3DDesktop&hl=en. Accessed 4 November 2018.
13. Kiriazes RE, Watkins KE, Guin A, Hunter MP. The impact of smartphone applications on trip routing. Paper presented at: 2020 Transportation Research Board Annual Meeting (Abstract 20-04415). 2020.
14. Jestico B, Nelson T, Winters M. Mapping ridership using crowdsourced cycling data. *J Transp Geogr*. 2016;52:90–97.
15. Roy A, Nelson TA, Fotheringham AS, Winters M. Correcting bias in crowdsourced data to map bicycle ridership of all bicyclists. *Urban Sci*. 2019;3:62.
16. Roberts I, Marshall R, Lee-Joe T. The urban traffic environment and the risk of child pedestrian injury: a case-crossover approach. *Epidemiology*. 1995;6:169–171.
17. Schneeweiss S, Stürmer T, Maclure M. Case-crossover and case-time-control designs as alternatives in pharmacoepidemiologic research. *Pharmacoepidemiol Drug Saf*. 1997;6 (Suppl 3):S51–S59.
18. Maclure M. The case-crossover design: a method for studying transient effects on the risk of acute events. *Am J Epidemiology*. 1991;133:144–153.
19. Ranapurwala SI, Mello ER, Ramirez MR. A GIS-based matched case-control study of road characteristics in farm vehicle crashes. *Epidemiol*. 2016;27:827–834.
20. Greenland S, Thomas DC. On the need for the rare disease assumption in case-control studies. *Am J Epidemiology*. 1982;116:547–553.
21. Rothman KJ, Greenland S, Lash TL. Case-Control studies. In: Rothman KJ, Greenland S, Lash TL, eds. *Modern Epidemiology*. 3rd ed. Philadelphia, PA: Wilkins; 2008.
22. Suissa S, Dell"Aniello S, Martinez C. The multitime case-control design for time-varying exposures. *Epidemiology*. 2010;21:876–883.
23. Lusk AC, Furth PG, Morency P, Miranda-Moreno LF, Willett WC, Dennerlein JT. Risk of injury for bicycling on cycle tracks versus in the street. *Inj Prev*. 2011;17:131–135.
24. Vanparijs J, Int Panis L, Meeusen R, de Geus B. Exposure measurement in bicycle safety analysis: a review of the literature. *Accid Anal Prev*. 2015;84:9–19.
25. Zhang B, Chen S, Ma Y, Li T, Tang K. Analysis on spatiotemporal urban mobility based on online car-hailing data. *J Transp Geogr*. 2020;82. doi: 10.1016/j.jtrangeo.2019.102568.
26. Buckee CO, Balsari S, Chan J, et al. Aggregated mobility data could help fight COVID-19. *Science*. 2020;368:145–146.
27. Oliver N, Lepri B, Sterly H, et al. Mobile phone data for informing public health actions across the COVID-19 pandemic life cycle. *Sci Adv*. 2020;6:eabc0764.
28. Oldroyd RA, Morris MA, Birkin M. Identifying methods for monitoring foodborne illness: review of existing public health surveillance techniques. *J Med Internet Res*. 2018;20:e57.
29. Mooney SJ, Garber MD. Sampling and sampling frames in big data epidemiology. *Curr Epidemiol Rep*. 2019;6:14–22.
30. Garber MD, Watkins KE, Kramer MR. Comparing bicyclists who use smartphone apps to record rides with those who do not: Implications for representativeness and selection bias. *J Transp Heal*. 2019;15:100661.
31. Lash TL, Fox MP, Fink AK. Selection bias. In: Lash TL, Fox MP, Fink AK, eds. *Applying Quantitative Bias Analysis to Epidemiologic Data. Statistics for Biology and Health*. New York, NY: Springer New York; 2009.
32. Lash TL, Fox MP, MacLehose RF, Maldonado G, McCandless LC, Greenland S. Good practices for quantitative bias analysis. *Int J Epidemiol*. 2014;43:1969–1985.
33. Geneletti S, Richardson S, Best N. Adjusting for selection bias in retrospective, case-control studies. *Biostatistics*. 2009;10:17–31.
34. Haneuse S, Schildcrout J, Crane P, Sonnen J, Breitner J, Larson E. Adjustment for selection bias in observational studies with application to the analysis of autopsy data. *Neuroepidemiology*. 2009;32:229–239.
35. Chu X; (U.S.) NC for TR. A Guidebook for Using Automatic Passenger Counter Data for National Transit Database (NTD) Reporting Prepared for Florida Department of Transportation Contract Number: BDK85 977-04. National Center for Transit Research (U.S.); 2010. Available at: http://www.nctr.usf.edu. Accessed 8 June 2020.
36. Brum-Bastos V, Ferster CJ, Nelson T, Winters M. Where to put bike counters? Stratifying bicycling patterns in the city using crowdsourced data. *Findings*. 2019. doi: 10.32866/10828.
37. Johnstone D, Nordback K, Lowry M. *Collecting Network-Wide Bicycle and Pedestrian Data : A Guidebook for When and Where to Count*. Portland, OR: State of Washington Department of Transportation; 2017. Available at: https://www.wsdot.wa.gov/NR/rdonlyres/ACBE2F89-6311-4BB0-ABD8-0CFCCD91D927/0/Guidebook_BikePedCounts.pdf. Accessed 2 October 2020.
38. Austin PC, Stuart EA. Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. *Stat Med*. 2015;34:3661–3679.
39. STRAVA METRO Comprehensive User Guide Version 8.0. 2019. Available at: http://metro.strava.com/wp-content/uploads/2019/05/Strava-Metro-Comprehensive-User-Guide-Version-8.0.pdf. Accessed 15 November 2019.
40. Winters M, Branion-Calles M. Cycling safety: quantifying the under reporting of cycling incidents in Vancouver, British Columbia. *J Transp Heal*. 2017;7(Part A):48–53.
41. Efron B, Hastie T. Bootstrap Confidence Intervals. In: *Computer Age Statistical Inference: Algorithms, Evidence, and Data Science*. Cambridge University Press; 2016:181–2014.
42. Lance Bottoms K, Moore FA, Smith C, et al. 2017. City of Atlanta 2017 Annual Bicycle Report. Atlanta. 2017. Available at: https://www.atlantaga.gov/home/showdocument?id=34089. Accessed 2 October 2020.
43. Tin Tin S, Woodward A, Robinson E, Ameratunga S. Temporal, seasonal and weather effects on cycle volume: an ecological study. *Environ Heal A Glob Access Sci Source*. 2012;11:12.
44. Hernán MA. The C-word: Scientific euphemisms do not improve causal inference from observational data. *Am J Public Health*. 2018;108:616–619.
45. Greenland S, Rothman KJ. Measures of occurrence. In: Rothman KJ, Greenland S, Lash TL, eds. *Modern Epidemiology*. 3rd ed. Philadelphia, PA: Wilkins; 2008:33–58.
46. Miettinen O. Estimability and estimation in case-referent studies. *Am J Epidemiol*. 1976. 103:226–235.

47. Kupper LL, McMichael AJ, Spirtas R. A hybrid epidemiologic study design useful in estimating relative risk. *J Am Stat Assoc*. 1975;70:524–528
48. Flanders WD, Dersimonian R, Rhodes P. Estimation of risk ratios in case-base studies with competing risks. *Stat Med*. 1990;9:423–435.
49. Maclure M, Mittleman MA. Should we use a case-crossover design? *Annu Rev Public Health*. 2000;21:193–221.
50. Suissa S. The case-time-control design. *Epidemiology*. 1995;6:248–253.
51. Gullette EC, Blumenthal JA, Babyak M, et al. Effects of mental stress on myocardial ischemia during daily life. *JAMA*. 1997;277:1521–1526.
52. Teschke K, Harris MA, Reynolds CC, et al. Route infrastructure and the risk of injuries to bicyclists: a case-crossover study. *Am J Public Health*. 2012;102:2336–2343.
53. Diggle PJ, Guan Y, Hart AC, Paize F, Stanton M. Estimating individual-level risk in spatial epidemiology using spatially aggregated information on the population at risk. *J Am Stat Assoc*. 2010;105:1394–1402.
54. Chang X, Waagepetersen R, Yu H, et al. Disease risk estimation by combining case-control data with aggregated information on the population at risk. *Biometrics*. 2015;71:114–121.
55. Haneuse S, Wakefield J. Geographic-based ecological correlation studies using supplemental case-control data. *Stat Med*. 2008;27:864–887. doi:10.1002/sim.2979
56. Haneuse SJPA, Wakefield JC. The combination of ecological and case-control data. *J R Stat Soc Ser B Stat Methodol*. 2008;70:73–93.
57. Wacholder S. The case-control study as data missing by design: estimating risk differences. *Epidemiology*. 1996;7:144–150.
58. Sato T. Risk ratio estimation in case-cohort studies. *Environ Health Perspect*. 1994;102 (Suppl 8):53–56.
59. Miettinen OS. Etiologic research: needed revisions of concepts and principles. *Scand J Work Environ Health*. 1999;25:484–490.
60. Suissa S. The Quasi-cohort approach in pharmacoepidemiology: upgrading the nested case-control. *Epidemiology*. 2015;26:242–246.
61. Cochran WG. Subsampling with units of equal size. In: *Sampling Techniques*. Third. New York: John Wiley & Sons; 1977:274–291.
62. Wiebe DJ, Richmond TS, Guo W, et al. Mapping Activity Patterns to Quantify Risk of Violent Assault in Urban Environments. *Epidemiology.* 2016;27:32–41.
63. Yelp Dataset. Yelp. Available at: https://www.yelp.com/dataset. Accessed 9 June 2020.
64. Strömberg U. Does induction time have any bearing on definition of study base? *Epidemiology*. 1994;5:356–359.
65. Salvan A, Stayner L, Steenland K, Smith R. Selecting an exposure lag period. *Epidemiology*. 1995;6:387–390.