MS# EDE20-0132

Appendices to accompany *At-risk-measure sampling in case-control studies with aggregated data*

Michael D. Garber,[1] Lauren E. McCullough,[1] Stephen J. Mooney,[2,3] Michael R. Kramer,[1] Kari E. Watkins,[4] R.L. Felipe Lobelo,[5] W. Dana Flanders[1,6]

[1]Department of Epidemiology, Rollins School of Public Health, Emory University, Atlanta, GA
[2]Department of Epidemiology, School of Public Health, University of Washington, Seattle, WA
[3]Harborview Injury Prevention & Research Center, University of Washington, Seattle, WA
[4]School of Civil and Environmental Engineering, Georgia Institute of Technology, Atlanta, GA
[5]Hubert Department of Global Health, Rollins School of Public Health, Emory University, Atlanta, GA
[6]Department of Biostatistics and Bioinformatics, Rollins School of Public Health, Emory University, Atlanta, GA

eAppendix 1: Relating a segment-level definition of total person-distance with a person-level definition

In the main text, we state that, assuming the same cohort definition is used, the total amount of exposed and unexposed person-distance is the same whether the total is defined by summing over segments without considering individual people—as in Equation (1) in the main text—or by summing over people. To illustrate this relation, we consider a group of people, $i = 1, \ldots, I$, who traveled any distance while at risk for the outcome in the study area and time period. The study area is again defined by the set of M segments, $m = 1,2, \ldots, M$, of time-invariant length L_m, $0 \leq L_m < \infty$ of a binary exposure condition, $E_m = 1$ or $E_m = 0$. The number of times individual *i* travels over segment *m* in either direction in the study period is denoted as $N_{i,m}$. Individual *i*'s person-distance in exposure category e is thus $D_{i,e} = \sum_{m=1}^{M} L_m * N_{i,m} | E_m = e$. Summing over individuals yields the total person-distance in the cohort in exposure category *e*:

$$D_e = \sum_{i=1}^{i=I} \sum_{i=i,m=1}^{i=i,m=M} L_m * N_{i,m} | E_m = e. \quad \text{(S1.1)}$$

Summation is commutative (the order does not matter), so Equation (S1.1) can be alternatively expressed as the sum of the times segment *m* was ridden by these *I* individuals, summed over segments:

$$D_e = \sum_{m=1}^{M} \sum_{i=1,m=m}^{i=I,m=m} L_m * N_{i,m} | E_m = e. \quad \text{(S1.2)}$$

Segment length, L_m, is a constant in the inner sum in Equation (S1.2) so can be drawn out by the constant-multiple rule:

$$D_e = \sum_{m=1}^{M} \left[(L_m | E_m = e) * \sum_{i=1,m=m}^{i=I,m=m} N_{i,m} | E_m = e \right]. \quad \text{(S1.3)}$$

Finally, the quantity $\sum_{i=1,m=m}^{i=I,m=m} N_{i,m}$ is the total number of times any individual travels over segment *m* while in the cohort, as denoted by N_m in the main text. Substituting N_m for

$\sum_{i=1,m=m}^{i=I,m=m} N_{i,m}$, Equation (S1.3) becomes $[D_e = \sum_{m=1}^{M} L_m * N_m | E_m = e]$, which is Equation (1) from the main text.

eAppendix 2: Person-event sampling

As noted in the main text, the at-risk-measure sampling technique may also be used to sample person-events, for example those occurring at intersections[1] between segments. The goal is to estimate the incidence rate ratio (IRR) in this cohort between a transient exposure and an acute outcome. The general form of the IRR is $IRR = \frac{\frac{exposed\ cases}{unexposed\ cases}}{\frac{measure\ of\ total\ at-risk\ experience\ exposed}{measure\ of\ total\ at-risk\ experience\ unexposed}}$.

Here, the measure of the at-risk experience is person-events, so the IRR is defined as $\frac{\frac{exposed\ outcome\ events}{unexposed\ outcome\ events}}{\frac{exposed\ person-events\ at\ risk}{unexposed\ person-events\ at\ risk}}$. We again focus on estimating the denominator, $\frac{exposed\ person-events\ at\ risk}{unexposed\ person-events\ at\ risk}$, assuming all outcome events, i.e., cases, are sampled. The notation closely parallels that of the main text. The main difference is that there is no length dimension, L_m, to consider.

We define *the* cohort by person-events at risk during a time period in a study area. For concreteness, suppose person-events at risk are intersection crossings. People may freely enter and leave the cohort; only their events at risk while in the study area and timeframe is considered part of the cohort. The study area is defined by a set of M dimensionless intersections, m, $m = 1,2,\ldots,M$, classified by a binary exposure condition, $E_m = 1$ or $E_m = 0$. The number of times any individual crosses intersection *m* while at risk for the outcome is denoted by N_m, $N_m = 0,1,2,..,\infty$. The total person-events (intersection crossings) in the cohort in exposure category *e* is

$$N_e = \sum_{m=1}^{M} N_m | E_m = e, \qquad \text{(S2.1)}$$

Suppose, again, some people who ever pass through the study area during the timeframe sometimes use a mobile sensor or smartphone app to record their activity as they travel over the segments and cross the intersections. For anonymity, their activity is summarized by intersection before it is made available for research (e.g., https://metro.strava.com/; accessed January 3rd, 2020). The number of times any individual crosses intersection *m* in the time period while using the sensor, and thus in the sample, is denoted by n_m: $n_m = 0,1,2,\dots,N_m$, for $m = 1,2,\dots,M$.

The sample is an element of the sample description space, $\Omega = \{n_1, n_2, \dots, n_M : n_m = 0,1,2,\dots,N_m; m = 1,2,\dots,M\}$, in which every n_m takes one of its possible values. At every intersection m, the sampling fraction, f_m, is the ratio of the number of times the intersection is crossed in the sample, n_m, to the corresponding total in the cohort, N_m: $f_m = \frac{n_m}{N_m}; 0 \leq f_m \leq 1$. If and only if $f_m = 0$, then segment *m* is not sampled. The total sampled person-events in exposure category *e* is

$$n_e = \sum_{m=1}^{M} N_m * f_m \,|E_m = e. \quad \text{(S2.2)}$$

We now show a condition sufficient for the ratio of exposed to unexposed sampled person-events, $\frac{n_1}{n_0}$, to consistently estimate that of the cohort, $\frac{N_1}{N_0}$. The condition is that the ratio of expected values of exposed to unexposed person-distance in the sample equals that of the cohort:

$$\frac{E\left[\sum_{m=1}^{M} N_m * f_m | E_m = 1\right]}{E\left[\sum_{m=1}^{M} N_m * f_m | E_m = 0\right]} = \frac{E\left[\sum_{m=1}^{M} N_m | E_m = 1\right]}{E\left[\sum_{m=1}^{M} N_m | E_m = 0\right]}. \quad \text{(S2.3)}$$

As noted in the main text, this condition, Equation (S2.3), can be re-arranged to be Equation (5) from the main text:

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

In words, this condition states that the ratio of the expected number of sampled person-events to the expected total person-events in the exposed category is the same as that of the unexposed category.

eAppendix 3. Correcting for inverse-probability-of-selection weighting using a summary bias breaker

Here, we consider an alternative strategy to correct for selection bias in addition to inverse-probability-of-selection weighting (IPSW). The alternative uses a summary 'bias breaker.'[2] A practical advantage of estimating a summary bias parameter compared with IPSW is that it obviates the possibly data-intensive task of estimating segment-specific sampling fractions. Rather than estimating each value of D_e in an absolute sense, a summary bias breaker, $\frac{f_{D,S,0}}{f_{D,S,1}}$, can be estimated to satisfy the equation:

$$\left(\frac{d_1}{d_0}\right)_{adj} = \frac{d_1}{d_0} * \frac{f_{D,S,0}}{f_{D,S,1}}, \quad \text{(S3.1)}$$

where $\left(\frac{d_1}{d_0}\right)_{adj}$ is the bias-adjusted ratio of exposed to unexposed person-distance, $f_{D,S,1} = \frac{\sum_{m=1}^{S} L_m * N_m * f_m | E_m=1}{\sum_{m=1}^{S} L_m * N_m | E_m=1}$, and $f_{D,S,0} = \frac{\sum_{m=1}^{S} L_m * N_m * f_m | E_m=0}{\sum_{m=1}^{S} L_m * N_m | E_m=0}$. By definition, $\frac{d_1}{d_0} = \frac{f_{D,1} * D_1}{f_{D,0} * D_0} = \frac{f_{D,1}}{f_{D,0}} * \frac{D_1}{D_0}$, so $\frac{D_1}{D_0} = \frac{d_1}{d_0} * \frac{f_{D,0}}{f_{D,1}}$, and if $\frac{f_{D,S,0}}{f_{D,S,1}} \approx \frac{f_{D,0}}{f_{D,1}}$, then $\left(\frac{d_1}{d_0}\right)_{adj} \approx \frac{D_1}{D_0}$. The key assumption is that $\frac{f_{D,S,0}}{f_{D,S,1}}$, obtained from the validation subset, generalizes to that of the cohort, $\frac{f_{D,0}}{f_{D,1}}$. To check or correct violations of this assumption, $\frac{f_{D,S,0}}{f_{D,S,1}}$ could be standardized with respect to factors whose distribution differs between the subset and the full cohort and which may influence differences between $\frac{f_{D,S,0}}{f_{D,S,1}}$ and $\frac{f_{D,0}}{f_{D,1}}$. The summary bias-breaker approach may not always be feasible. In our empirical example described in the main text, the validation subset was entirely within the exposed stratum of non-residential roadways, so a summary sampling fraction in the unexposed,

$f_{D,S,0}$, could not have been calculated, precluding this approach. If data to adjust for selection bias are not available in a validation subset, another option would be to draw the bias breaker in Equation (S3.1) from literature.[3]

eAppendix 4. Details of the event-trial logistic-regression model used to estimate the sampling fraction

The dependent variable of the event-trial logistic regression model is $\frac{n_{m,t}}{N_{m,t}}$, the number of rides on segment *m* in month *t* in Strava divided by the corresponding value measured by the Eco-Counter®, described in the main text. On average, data from the nine counters were available in 17.3 of the 23 study months to comprise a total of 156 segment-month observations.

The independent variables are:

$$n\_3cat_{m,t} + prop\_commute\_3cat_{m,t} + year_t + season_t +$$

$$n\_3cat_{m,t} * prop\_commute\_3cat_{m,t} +$$

$$n\_3cat_{m,t} * year_t +$$

$$prop\_commute\_3cat_{m,t} * year_t +$$

$$prop\_commute\_3cat_{m,t} * season_t +$$

$$year_t * season_t$$

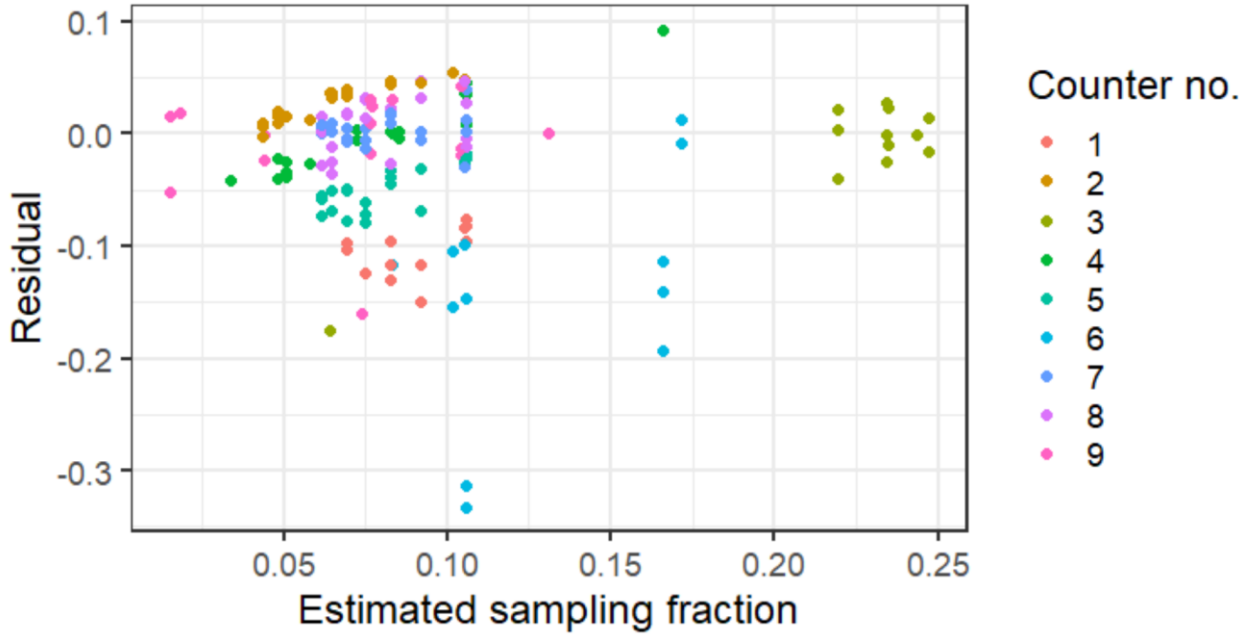
Variable definitions:

- $n\_3cat_{i,j}$ denotes the number of rides reported in Strava on segment *i* in month *j*, categorized into 3 groups: [0, 40), [40, 250), and [250, 3960].
- $prop\_commute\_3cat_{m,t}$ denotes the proportion of rides reported to be commutes in Strava on segment *i* in month *j*, categorized into 3 groups: 0, (0, 0.5), and [0.5, 0.1]. Rides were classified as commutes by Strava if the ride's start and end are were more than 1 kilometer apart[4] or if the individual tagged it as a commute in the app. The distribution of this variable is described in **Table 1** of the main text.

- $year_t$ is either 2017 or 2018, as counter data were not available during the 2016 months under study.
- $season_t$ classifies the month into 4 possible categories of 3 each. Winter is December, January, and February. The remaining seasons follow sequentially.

Model performance:

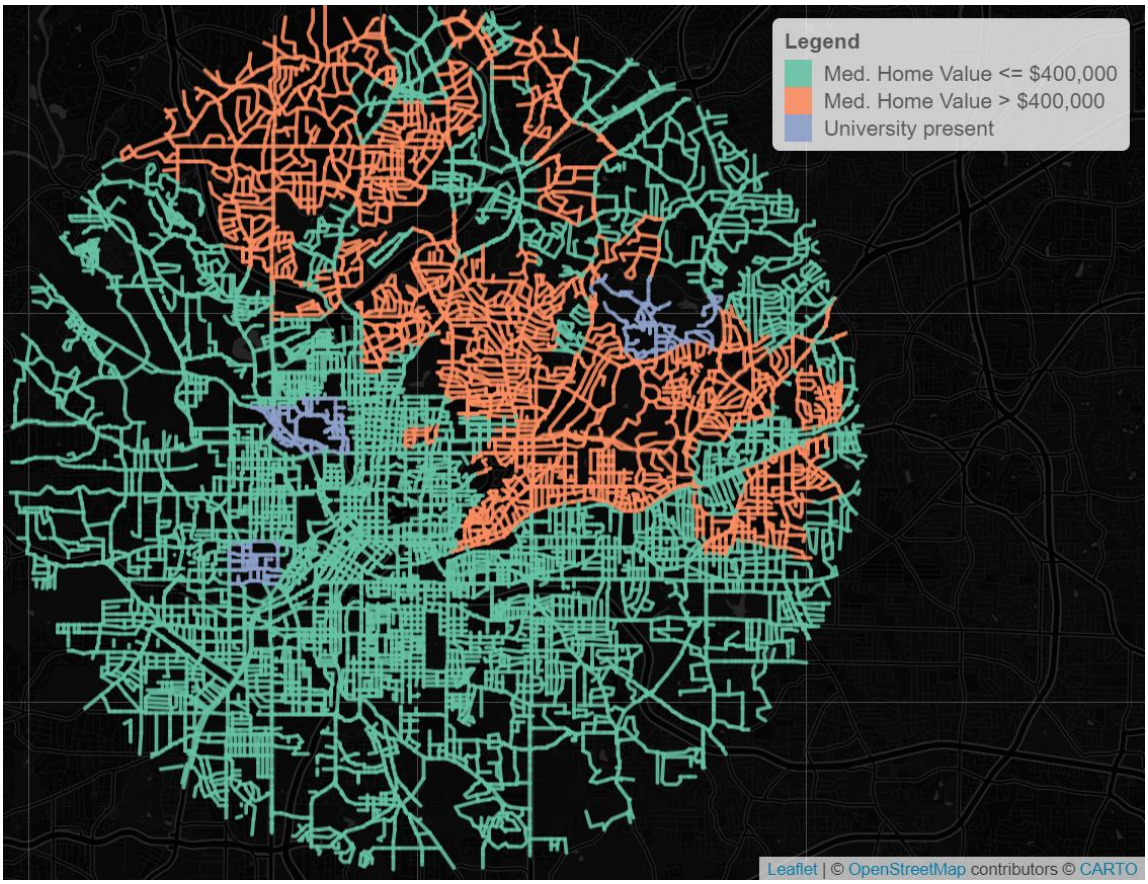
The additive residuals ($predicted\ value - true\ value$) are plotted against the estimated sampling fraction by counter number in **eFigure 1**. Most residuals are near zero but have a small negative bias (median = -0.003, mean = -0.02), on average.



eFigure 1. The additive residuals ($predicted\ value - true\ value$) plotted against the estimated sampling fraction by counter number (n, observations=156).

eAppendix 5. Definition of area-level socioeconomic status: a three-level variable considering the presence of a university and the median home value

In the example in the main text, we consider confounding by an area-level indicator of socioeconomic status. We define that variable here. We observed a relatively high *pseudo*-incidence rate (term from p. 113[5]) of police-reported bicycle crashes near universities and, to a lesser extent, in areas of lower median home value, so we examined whether the estimated *IRR* between exposure and incidence of crashes may be confounded by these factors. We stratified crashes and bicycle-distance by a three-level variable (abbreviated MHV) indicating if a university was present in the census block-group and, if not, indicating whether the block-group's median home value was above or below $400,000 per the 2017 5-year American Community Survey. We then standardized the exposure ratio and the IRR with respect to the marginal distribution of MHV in the unweighted bicycle-distance using a weighted geometric mean. The spatial distribution of MHV is depicted in the map below (**eFigure 2**):



eFigure 2. The spatial distribution of the potential confounding variable indicating whether a university is present (purple) and, if not, whether the median home value is, according to the 2017 5-year American Community Survey, above (orange) or below (turquoise) $400,000.

References

1. Harris MA, Reynolds CCO, Winters M, et al. Comparing the effects of infrastructure on bicycling injury at intersections and non-intersections using a case-crossover design. *Inj Prev*. 2013;19(5):303-310. doi:10.1136/injuryprev-2012-040561
2. Geneletti S, Richardson S, Best N. Adjusting for selection bias in retrospective, case-control studies. *Biostatistics*. 2009;10(1):17-31. doi:10.1093/biostatistics/kxn010
3. Lash TL, Fox MP, Maclehose RF, Maldonado G, Mccandless LC, Greenland S. Good practices for quantitative bias analysis. *Int J Epidemiol*. 2014;43(6):1969-1985. doi:10.1093/ije/dyu149
4. *STRAVA METRO Comprehensive User Guide Version 8.0*.; 2019. http://metro.strava.com/wp-content/uploads/2019/05/Strava-Metro-Comprehensive-User-Guide-Version-8.0.pdf. Accessed November 15, 2019.
5. Rothman KJ, Greenland S, Lash TL. Case-Control studies. In: Rothman KJ, Greenland S, Lash TL, eds. *Modern Epidemiology*. 3rd ed. Wilkins; 2008.