SUPPLEMENTAL MATERIAL

Supplemental Methods

Geocoding

Participant residential locations were geocoded using SAS (SAS Institute, Cary, NC), which uses the North American Datum 1983 (NAD83) as the geographic coordinate system. We projected the participant locations to the USA Contiguous Albers Equal Area Conic projected coordinate system in meters, in order to ease interpretation of the distances.

Preliminary test for spatial clustering

We tested for spatial variation in the prevalence of hypertension, diabetes, smoking, and dyslipidemia. Participants were considered to have dyslipidemia if: (1) they met the criteria for recommended treatment by lipid-lowering medications according to the Adult Treatment Panel III (APT III) guidelines;[1] or (2) they self-reported taking lipid lowering medications. In order to test for spatial variation in disease prevalence, we used the difference in Ripley’s functions test,[2,3] which is the most commonly used test of disease clustering among epidemiologists when locations of diseased and nondiseased participants are available.[4] This method tests whether the risk for the disease is constant across the region of interest.

In order to account for the overlapping points in the REGARDS dataset due to geocoding error, we excluded overlapping points. There is little precedent for “best practices” with a percentage of overlapping points this large. We tested for clustering of each risk factor in turn, using 1,166 equidistant ranges, which spanned from 0 to approximately 725 km. The maximum range (725 km) was chosen as one quarter the height of the enclosing rectangle for the continental US, and the interval between ranges was the median distance to the nearest neighbor (approximately 0.5 km). The polygonal window used for this study was the 1:20,000,000 resolution boundary shapefile of the US, which we modified to remove polygons not representing the continental US. We performed 200 Monte Carlo simulations and tested for clustering at an alpha level of 0.05. The null hypothesis was that the difference between Ripley's K function for the cases and Ripley's K function for the controls was 0 (i.e., $D(h) = K_{cases}(h) – K_{controls}(h) = 0$). Simultaneous critical regions for D(h) over all ranges tested (i.e., all values of h) were constructed in order to prevent inflation of the family-wise error rate (FEW) caused by performing over 1,000 hypothesis tests.[5] The null hypothesis was rejected if was ever outside the simultaneous critical regions, leading to the conclusion of evidence of clustering for that particular risk factor. The test was performed using the spatstat package (v. 1.43-0)[6] in the R statistical environment (v. 3.2.3).[7]

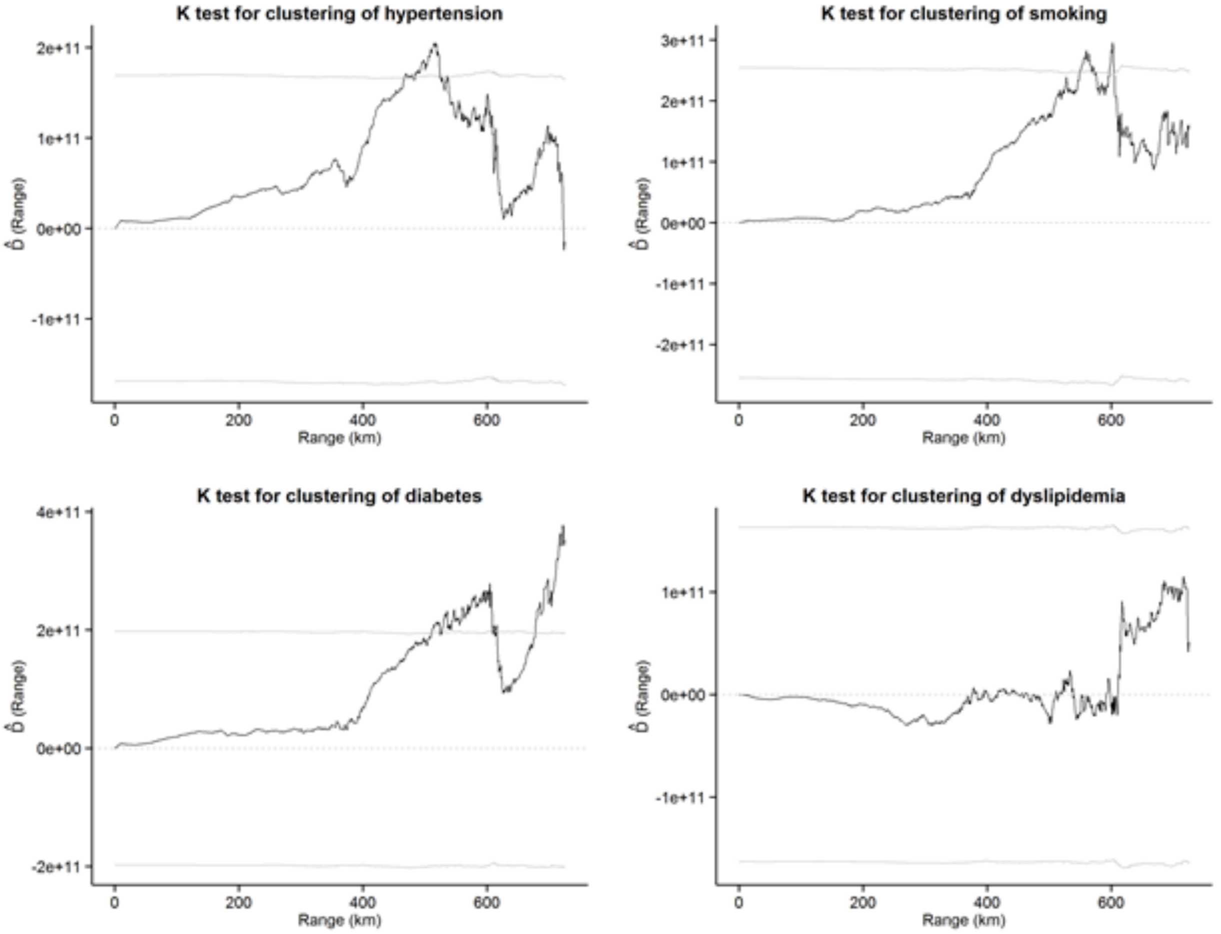
Results

Seven participants were excluded from each test of disease clustering, given that their locations lay outside polygonal window used for the study. The exclusion of these participants, we well as participants that were missing dyslipidemia status, resulted in a sample size of 27,780 for the tests of disease clustering. The value of D(h) for all values of h, as well as the 95% simultaneous critical envelopes for all h, for hypertension, diabetes, smoking, and dyslipidemia are presented in Supplemental Figure 1.

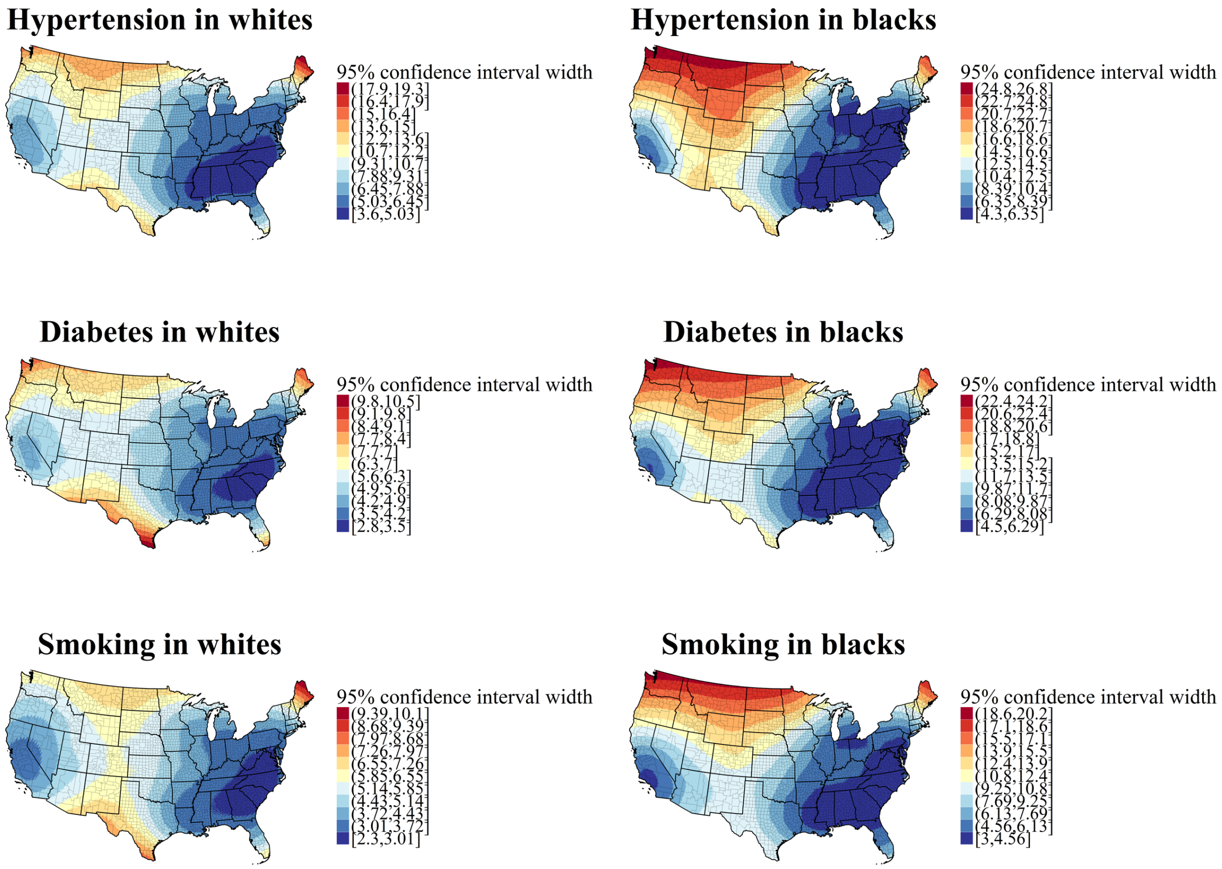
As shown in Supplemental Figure 1, there is evidence of clustering of hypertension up to a range of approximately 500 km, clustering of diabetes up to 600 km and 700 km, and clustering of current smoking up to 550 – 600 km. In other words, within these distances you would find more people around someone with the risk factor of interest, who also had the risk factor, than you would expect by chance. For comparison, the driving distance from Boston, MA to Philadelphia, PA is about 500 km. On the other hand, we found no evidence of clustering of dyslipidemia despite a large sample size (black curve never exited the grey critical envelopes).

Supplemental Figure 2 shows the widths of the 95% confidence intervals of the predicted prevalences across the US, which conveys the uncertainty in our predictions. Mean prevalences whose confidence intervals do not overlap can be considered statistically significantly different. However, no adjustment was made to the interval widths for multiple comparisons, so such results should be interpreted with caution. The widest confidence intervals among blacks tended to be twice as wide as the widest confidence intervals in whites, reflecting both the fewer number of blacks and the low proportion of blacks in large sections of the continental US. Supplemental Figure 3 shows the estimated prevalences when only participants living in the eastern US were used in the prediction models.

Supplemental Figures



Supplemental Figure 1. Estimated difference in K functions, D(h), as a function of range for each risk factor of interest. The grey lines are the critical values for each corresponding range, the black curve is D(h), and the grey dashed line is the estimated mean value for D(h) under the null hypothesis of constant risk. Hypertension, diabetes, and smoking show evidence of clustering, while dyslipidemia does not.



Hypertension in whites

Hypertension in blacks

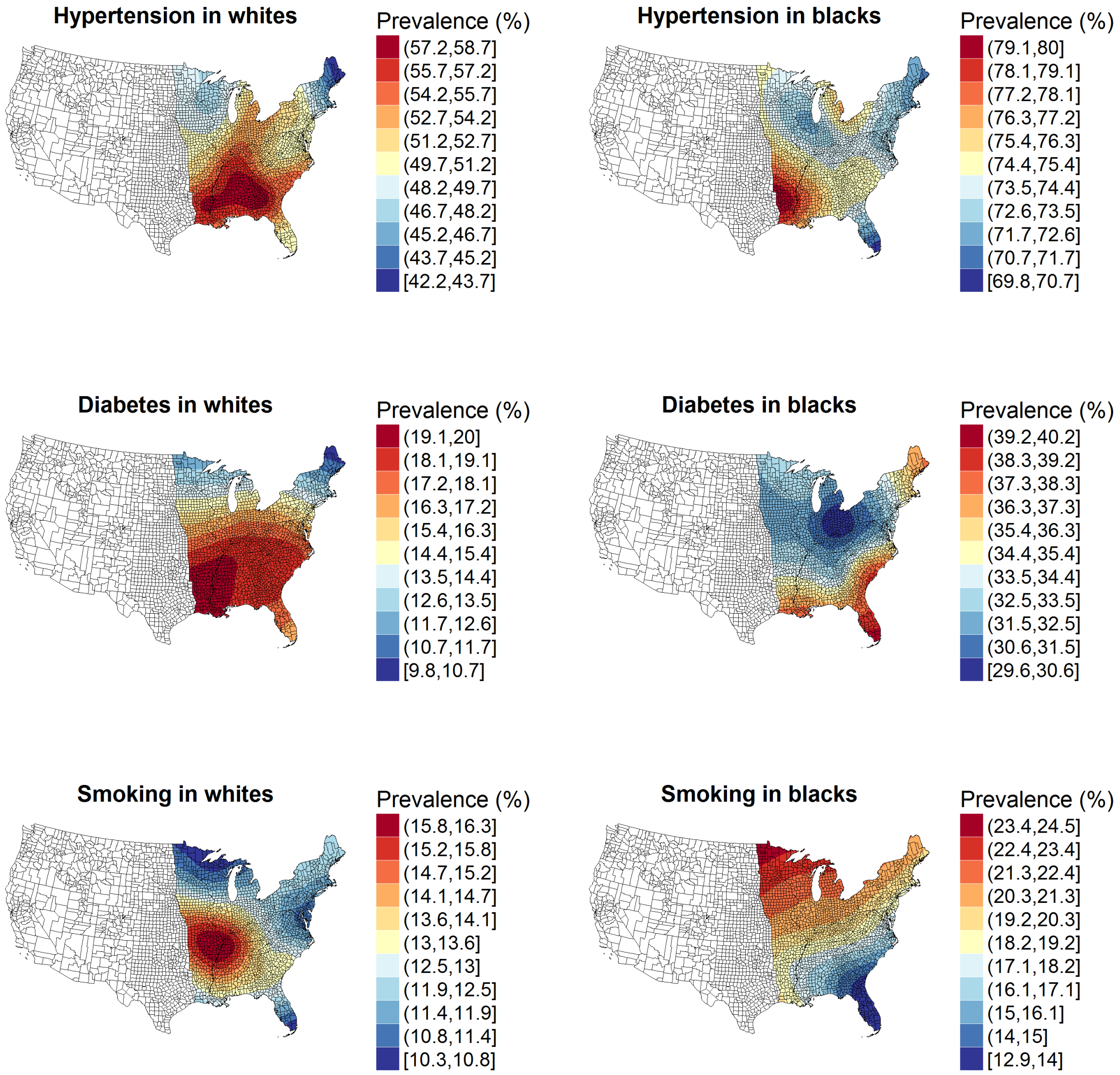
Diabetes in whites

Diabetes in blacks

Smoking in whites

Smoking in blacks

Supplemental Figure 2. Maps of widths for 95% confidence intervals for predicted mean hypertension, diabetes, and smoking prevalence among whites and blacks, adjusted for age and gender. The widest intervals (most uncertainty) are indicated by red, while the shortest intervals (least uncertainty) are indicated by blue.



Hypertension in whites

Hypertension in blacks

Diabetes in whites

Diabetes in blacks

Smoking in whites

Smoking in blacks

Supplemental Figure 3. Maps of estimated hypertension, diabetes, and current smoking prevalence among whites and blacks, adjusted for age and gender, using only REGARDS participants living in the east half of the US. High prevalence is indicated by red, while low prevalence is indicated by blue. Predicted prevalences assumed a population with the same proportion of women for each race and the same age as the mean age of each race. Thus, the prevalences reflect the gender and age composition of REGARDS participants of each race.

Supplemental references

1. Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults. Executive Summary of The Third Report of The National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, And Treatment of High Blood Cholesterol In Adults (Adult Treatment Panel III). *JAMA*. 2001;285:2486–2497.

2. Ripley BD. The Second-Order Analysis of Stationary Point Processes. *J Appl Probab*. 1976;13:255–266.

3. Diggle PJ, Chetwynd AG. Second-order analysis of spatial clustering for inhomogeneous populations. *Biometrics*. 1991;47:1155–1163.

4. Fritz CE, Schuurman N, Robertson C, Lear S. A scoping review of spatial cluster analysis techniques for point-event data. *Geospat Health*. 2013;7:183–198.

5. Loop MS, McClure LA. Testing for clustering at many ranges inflates family-wise error rate (FWE). *Int J Health Geogr*. 2015;14:4.

6. Baddeley A, Turner R. spatstat: An R Package for Analyzing Spatial Point Patterns. *J Stat Softw*. 2005;12:1–42.

7. R Core Team. R: A language environment for statistical computing. 2015.