**Electronic Appendix**

*Table of Contents*

Data sources, indicators and outcome metrics 1

Data mining analysis 2

Additional results 4

Important caveats and limitations 6

Diagnostic plots 8

Summary statistics on indicator variables and outcome metrics 8

References 9

Web figure 1: 11

Web figure 2: 12

Web figure 3 16

Web figure 4 25

Web table 1: 27

Web table 3: 34

Web table 4: 36

In this study, we detected and characterized which indicators of socioeconomic, demographic, behavioral and environmental conditions may best assist public health program managers to objectively identify the most vulnerable or resilient counties across the United States. Here, we “open source” a highly efficient data mining technique, which does not become potentially biased in the presence of multiple collinear variables, to test all complex combinations of interactions among indicators and identify a logical sequence of just a few indicators that are associated with higher or lower premature mortality rates.

## Data sources, indicators and outcome metrics

Web Table 1 summarizes the 50 key indicators of socioeconomic, demographic, behavioral and environmental conditions evaluated in our study, along with calendar years and data links to the 20 geocoded publically available surveys through which these data were collected. As noted in the Table, we preferentially chose metrics that are summarized from CDC and Census Bureau surveys into the 2013 County Health Rankings Database, as this online data repository provides free access to geocoded data with annual updates and extensive documentation of each indicator.1 The primary outcome variable being correlated to these indicators was age-adjusted years of potential life lost before age 75, computed by the National Center for Health Statistics.2 This outcome was chosen because of its importance to the CDC as a principal target for reducing geographical disparities across the country;3 however, we also investigated other outcomes which are major indicators of population health status and for which geographic disparities have been well established in the literature, as itemized below. The three-year mean value for age-adjusted years of potential life lost was employed to provide robust estimates less subject to demographic shift or single-year survey biases; the metric was most recently available from 2008-2010.2

As shown in Web Table 1, we clustered indicators into socioeconomic, demographic, behavioral and environmental metrics by county. All US counties were included. Socioeconomic indicators included both traditional measures such as poverty, income inequality and unemployment as well as novel indicators of social capital and family structure as shown in the Table; demographic variables focused on sex and age structure as well as race/ethnic composition; behavioral variables included tobacco and alcohol use metrics as well as self-reported diet, physical activity and sexually-transmitted infection rates; and environmental metrics included both built environment indicators such as fast food density and environmental quality indicators such as air pollution indicators. Although our primary focus was on the major determinants of population health,4 we additionally included control variables for healthcare access including uninsurance rates, self-reported inability to pay for medical services, and metrics of healthcare quality such as the frequency of indicated disease screening (Web Table 1). As shown in Web Table 1, all indicators were population-size corrected (expressed as a per capita indicator or equivalent rate per population size) and chosen from the most recent available calendar years of data, provided in the table. Analyses were repeated both on the entire dataset and on the subset of counties with at least 10,000 population to evaluate outcome stability. Missing data constituted 5.2% of the dataset, and we performed our analysis on both the subset of counties with complete primary data (*N*=2,653) as well as after imputation of a complete dataset using a standard impurity index and surrogate variable imputation algorithm,5 finding that both analyses resulted in the same results.

## Data mining analysis

While there are numerous data mining algorithms in existence, we chose a regression tree/random forest mining approach because it is: (i) widely accepted as a valid approach, as opposed to some newer methods that remain experimental;6 (ii) easily, rapidly and freely implemented without specialized programming knowledge (i.e., user-friendly); (iii) visually straightforward to interpret results from, in that the approach assembles clusters of more vulnerable or more resilient counties; and (iv) capable of handling large volumes of alternative indicators and being easily extended to alternative outcome metrics.

We specifically employed validated regression tree algorithms,5 which can be generated through the free rpartpackage in the software program *R*, which is also freely available (http://www.r-project.org/). Detailed information on the package is provided in an accompanying manual.5 Splits are chosen using the standard Gini index of deviance, calculating a complexity parameter that prevents overfitting by finding a parsimonious set of correlates to avoid an overly-complex tree and maintain adequate sample size for inference at each branch point.7 Cross-validation using repeated out-of-sample prediction was performed; to improve predictive accuracy, we performed a random forest analysis, which generates a large number of bootstrapped trees using random samples of the variables. The algorithm then combines results across all of these trees through a classification system among the “forest” of these trees.8 Below, we provide parsimonious and efficient code that can be generalized and extended to other datasets and indicators of interest. To facilitate replication of our results, and extension by other researchers, we have provided an organized version of our dataset as a comma-separated values (.csv) file on our website (<http://www.stanford.edu/~basus/code.html>).

After downloading the dataset from the above website, researchers wishing to replicate or extend our analysis must first download the free statistical program *R* (available at: <http://cran.us.r-project.org/>) and paste the following statistical code on to the command line after installation. Lines designated with “#” indicate instructions to the user and labels for code segments.

# install necessary packages

install.packages("rpart")

getwd()

# set the following to the working directory containing the data:

setwd("~/[insert data folder extension here]")

mydata <- read.web table("countydata.csv", header=TRUE,sep=",")

attach(mydata)

# if you wish to visualize the data, type:

# View(mydata)

# summary statistics on the dataset

library(pastecs)

stat.desc(mydata,basic=F)

library(rpart)

# analyze all counties, no exclusions, using all indicators - ypll

fita <-rpart(ypll~smoke+obese+inactive+drink+mva+sti+teenbirth+uninsured+avoidhosp+scrdiab+scrmammo+highschool+college+unemp+poverty+support+singleparent+crime+pollution+water+recfacil+healthyfood+fastfood+young+old+black+indian+asian+otherrace+hispanic+white+nonenglish+female+rural+diabetic+hiv+healthcarecost+uninsuredadult+uninsuredchild+cantaccess+income+housingcost+freelunch+homicides+drivealone+park, method="anova",data=mydata)

# if additional indicators are desired, add these to the above list

# print initial results

printcp(fita)

# visualize complexity parameter web table

plotcp(fita)

# detailed summary statistics on tree selection

summary(fita)

# visualize cross-validation results

par(mfrow=c(1,2))

rsq.rpart(fita)

# plot base tree structure

par(mfrow=c(1,1))

plot(fita, uniform=TRUE,main="Regression Tree for YPLL")

# tree labels

text(fita, use.n=TRUE, all=TRUE, cex=.8)

# prune tree to indicators that pass complexity parameter criteria

pfita<-prune(fita, cp=fita$cpweb table[which.min(fita$cpweb table[,"xerror"]),"CP"])

# plot the pruned tree

par(mfrow=c(1,1))

# save results as png file

png('tree.png')

plot(pfita,uniform=TRUE,compress=TRUE,branch=0,main="Regression Tree for YPLL")

text(pfita,use.n=TRUE,all=TRUE,cex=0.8)

dev.off()

# repeating analysis on only large counties

newdata <- mydata[which(mydata$popsize>= 10000),]

attach(newdata)

fitm <-rpart(ypll~smoke+obese+inactive+drink+mva+sti+teenbirth+uninsured+avoidhosp+scrdiab+scrmammo+highschool+college+unemp+poverty+support+singleparent+crime+pollution+water+recfacil+healthyfood+fastfood+young+old+black+indian+asian+otherrace+hispanic+white+nonenglish+female+rural+diabetic+hiv+healthcarecost+uninsuredadult+uninsuredchild+cantaccess+income+housingcost+freelunch+homicides+drivealone+park, method="anova",data=newdata)

printcp(fitm)

plotcp(fitm)

summary(fitm)

par(mfrow=c(1,2))

rsq.rpart(fitm)

par(mfrow=c(1,1))

plot(fitm, uniform=TRUE,main="Regression Tree for YPLL, large counties")

text(fitm, use.n=TRUE, all=TRUE, cex=.8)

pfitm<-prune(fitm, cp=fitm$cpweb table[which.min(fitm$cpweb table[,"xerror"]),"CP"])

par(mfrow=c(1,1))

png(‘treelarge.png')

plot(pfitm,uniform=TRUE,compress=TRUE,branch=0,main="Regression Tree for YPLL, large counties")

text(pfitm,use.n=TRUE,all=TRUE,cex=0.8)

dev.off()

save.image("~/results.RData")

## Additional results

Web figure 1 provides the distribution of premature mortality rates across US counties. As shown, the distribution has a median value of 7,760 years of potential life lost before age 75 per 100,000 people per year, and is right-skewed with a mean of 7,986 years lost and a large standard deviation of 2,393 years lost, reflecting the large geographical disparities. The highest rate of premature mortality was in Sioux County, North Dakota, which experienced 24,670 years lost per 100,000 population, as compared to the lowest rate of premature mortality in Polk County, Nebraska, which lost only 2,950 years per 100,000. Both of these counties contain less than 10,000 people, however, so in the subsample of counties with at least 10,000 population size, the highest number of years of life lost was 23,850 per 100,000 per year (Shannon County, South Dakota, which is within the Pine Ridge Indian Reservation) and the lowest was 3,290 per 100,000 per year (Loudoun County, Virginia, a suburb of Washington D.C.).

Web figure 2 illustrates disparities among key socioeconomic, demographic, behavioral and environmental indicators across all US counties, along with their correlations to each other and to premature mortality. As illustrated in Web figure 2, most of the indicators of demographic and behavioral status were strongly correlated to rates of premature mortality, as were some of the socioeconomic variables, but few of the environmental indicators. The indicators most significantly correlated (p < 0.001) to premature mortality rates were: child poverty rate (**= 0.70), teen birth rate (** = 0.68), homicide rate (**= 0.68), motor vehicle accident rate (**= 0.67), physical inactivity prevalence (**= 0.61), adults reporting inadequate social support (**= 0.51), percent of population below age 18 (**= 0.13), percent of population not English proficient (**= -0.16), and rate of mammographic screening among women (**= -0.45). Web table 3 further provides summary statistics of all indicators, and Web table 4 provides standard multivariate regressions revealing how the indicators correlated to the studied health outcomes.

Web figure 3 presents results of the complete data mining analysis. The analysis among all counties, using premature mortality as the key outcome metric, is provided in main text figure 1. Given that the main text regression tree analysis included small populations among whom mortality rates may be extreme values due to small population size denominators, we repeated the analysis among the subsample of counties having at least 10,000 individuals. Running the data mining algorithm on this subset of counties (*N* = 2,193), we found that teen birth rate continued to be selected as the first branching point, as shown in Web figure 3I. As shown in the Web figure, the second branch became diabetes prevalence, which did not appear among the full county list, as the counties with large Native American populations were excluded by the population size criteria. As illustrated in Web figure 3I, those counties having low teen birth rate (<46.5 per 1,000), lowest diabetes prevalence rates (<9.5% of adults), and lowest motor vehicle accident rates (<14.5 per 100,000) had the lowest number of potential years of life lost (5,456 per 100,000, *N*=380; Group 1 in Web figure 3I). Conversely, a high teen birth rate (>46.5 per 1,000), diabetes prevalence rate (>11.5%), and high sexually-transmitted infection rate (>1,392 chlamydia cases per 100,000) had the highest premature mortality rates (15,688 years lost per 100,000, *N*=10; Group 10 in Web figure 3I).

While these regression tree analyses offered insights into what indicators may serve as correlates for complex, multi-level social processes that relate to the CDC’s preferred metric of premature mortality, we also conducted further experiments to examine how the most useful indicators may vary if the CDC and other agencies convert to focus on other targets of geographic health disparities. As illustrated in Web figure 4, we performed our data mining approach on age-adjusted overall premature mortality rate (CDC, 2008-2010), infant mortality rate (CDC, 2006-2010), child mortality rate (CDC, 2007-2010), and self-reported poor or fair health (percent of adults sampled in the Behavioral Risk Factor Surveillance System, 2005-2011). As with the primary regression tree analysis displayed in main text Figure 1, all of the regression trees included all of the indicator variables listed in Web table 1 as candidate regressors. We found that the order and selection of indicators chosen by the data mining algorithm did vary depending on the choice of outcome metric (see Web figure 3). However, we found notable consistencies among which indicators were chosen as key branching points, with teen birth, household income, diabetes prevalence, motor vehicle accident rates, and (when including smaller counties) Native American prevalence appearing consistently among top predictors.

## Important caveats and limitations

As with any survey-based statistical study, the results of this analysis have several notable limitations. First, we focused the current analysis on geographic disparities that occur at the county level, to focus on federal disparities-reduction goals.3 Aggregate county-level statistics, however, can mask important within-county variations (e.g., within New York County, which contains Manhattan), and more disaggregated data on local community characteristics are needed to further understand the implications of within-county segregation and related features common to some communities. The analysis performed here can be easily extended to such disaggregated analysis once more fine-grained data are available. Individual counties have begun assembling such data, which may be promoted through federal survey programs.9

Second, our goal is not to perform causal inferences about the association between community-level factors and individual health. Numerous advanced methods have been presented to perform multi-level causal inferences on the social determinants of health;10–13 here, we wished to focus on the correlates of population health metrics and, in particular, on the needs of health program managers for whom finding just a few key indicators to separate highly vulnerable from highly resilient geographical spaces is an important first step towards programmatic need identification, requiring efficient and simple approaches to the analysis of common datasets. Related to this issue, our analyses cannot provide an indication of the reason(s) that these predictors are particularly significant for determining population health. That requires further investigation of the ‘vicious and virtuous cycles’ mentioned earlier.

Third, our analysis is subject to the limitations of the available public use datasets. The datasets include not only directly-observed metrics of mortality, but also self-reported measures that—particularly for behavioral factors such as alcohol drinking rates—are subject to recall and misreporting biases, despite their widespread use.

Finally, the metrics used here involve cross-sectional studies. More extensive data mining can also be performed in the future on longitudinal panel datasets in order to get a better sense of lagged relationships among indicators and outcomes, but this also falls into the realm of causal inference and may be useful when data mining is integrated with novel large-scale time-series causal inference techniques such as convergent cross-mapping.14

Despite these limitations, our study approach offers novel opportunities to investigate pathways of interaction between numerous alternative indicators of geographically-relevant health-altering conditions and actual health outcomes. As a plethora of data become available in the future, further research can work to further automate and streamline the understanding of which data are most important to pay attention to among large datasets, in order to effectively direct resources when many competing claims are made in the literature correlating single indicators to single outcomes. Future research should test multiple alternative strategies for generating composite decisions to direct federal aid towards local municipalities, and given our results, it appears important to understand which factors may lead to differential selection of indicators—such as the choice of whether to include smaller counties, where mortality rates may be skewed, or omit them, which may systematically discriminate or bias results against some of the most vulnerable populations, as illustrated here with the indicator of Native American population prevalence. How to best use data mining approaches in public health practices also requires future research, as field-based strategies for rapidly gathering social determinants of health data are generating new types of indicators, such as through citizens groups who offer real-time feedback through mobile devices capturing both text and picture commentaries of barriers to health in their neighborhood (e.g., poor transportation to access healthy foods, unsafe walking paths, etc.). Incorporating this “patient-centered” data may require further ethnographic analysis and understanding of how to integrate such data with more formal data collection conducted through national surveys, which is not a simple extension of the framework we present here.

## Diagnostic plots

Visualization of cross-validation results is provided in Appendix Web figure 4 for both the full set of counties and the subset of large counties having population of at least 10,000 people. As shown, the chosen regression trees are considered optimal choices based on the complexity parameter, which balances the added complexity of each additional indicator versus the additional explanatory power of the indicator to find a parsimonious tree.

## Summary statistics on indicator variables and outcome metrics

Detailed summary statistics on the values of the indicator variables and outcome metrics are provided in Web table 3. Results of a standard multivariate regression of the outcome metrics on the indicator variables is provided in Web table 4.

## References

1. Robert Wood Johnson Foundaiton. *County Health Rankings Database*.; 2013. Available at: http://www.countyhealthrankings.org/.

2. National Center for Health Statistics. Health, United States, 2012. 2013. Available at: /nchs/hus.htm. Accessed December 24, 2013.

3. U.S. Department of Health and Human Services. *Healthy People 2020: General Health Status Topics and Objectives*. Washington D.C.; 2011.

4. Evans RG, Barer ML, Marmor TR. *Why are some people healthy and others not?: The determinants of the health of populations*. Transaction Books; 1994. Available at: http://books.google.com/books?hl=en&lr=&id=nuvrg2AWCT0C&oi=fnd&pg=PR11&dq=social+policy+and+population+health&ots=vr6Em1SjDd&sig=EKtHvfXV08Pq3oVdSsiKnZjovW0. Accessed December 18, 2013.

5. Therneau TM, Atkinson EJ. *An introduction to recursive partitioning using the RPART routines*. Technical Report 61, Section of Biostatistics, Mayo Clinic, Rochester. URL http://www. mayo. edu/hsr/techrpt/61. pdf; 1997.

6. Ateniese G, Felici G, Mancini LV, Spognardi A, Villani A, Vitali D. Hacking Smart Machines with Smarter Ones: How to Extract Meaningful Data from Machine Learning Classifiers. *ArXiv Prepr ArXiv13064447*. 2013. Available at: http://arxiv.org/abs/1306.4447. Accessed December 24, 2013.

7. Lewis RJ. An introduction to classification and regression tree (CART) analysis. In: *Annual Meeting of the Society for Academic Emergency Medicine in San Francisco, California*. Citeseer; 2000:1–14.

8. Breiman L. Random forests. *Mach Learn*. 2001;45(1):5–32.

9. Hardt NS, Muhamed S, Das R, Estrella R, Roth J. Neighborhood-Level Hot Spot Maps to Inform Delivery of Primary Care and Allocation of Social Resources. *Perm J*. 2013;17(1):4–9. doi:10.7812/TPP/12-090.

10. Snowden JM, Rose S, Mortimer KM. Implementation of G-Computation on a Simulated Data Set: Demonstration of a Causal Inference Technique. *Am J Epidemiol*. 2011;173(7):731–738. doi:10.1093/aje/kwq472.

11. Gangl M. Causal inference in sociological research. *Annu Rev Sociol*. 2010;36:21–47.

12. Glass TA, Goodman SN, Hernán MA, Samet JM. Causal inference in public health. *Annu Rev Public Health*. 2013;34:61–75.

13. Galea S, Riddle M, Kaplan GA. Causal thinking and complex system approaches in epidemiology. *Int J Epidemiol*. 2010;39(1):97–106. doi:10.1093/ije/dyp296.

14. Sugihara G, May R, Ye H, et al. Detecting causality in complex ecosystems. *Science*. 2012;338(6106):496–500.

15. U.S. Census Bureau. *DataFerrett*. Washington D.C.; 2013.

16. National Center for Education Statistics. *Common Core of Data (CCD)*. Washington D.C.; 2013.

17. Bureau of Labor Statistics. Databases, Tables & Calculators by Subject. 2013. Available at: http://data.bls.gov/timeseries/LNS14000000. Accessed December 24, 2013.

18. Federal Bureau of Investigation. *Uniform crime reporting*. Washington D.C.; 2013.

19. CDC. *Environmental Public Health Tracking Network*. Atlanta; 2013.

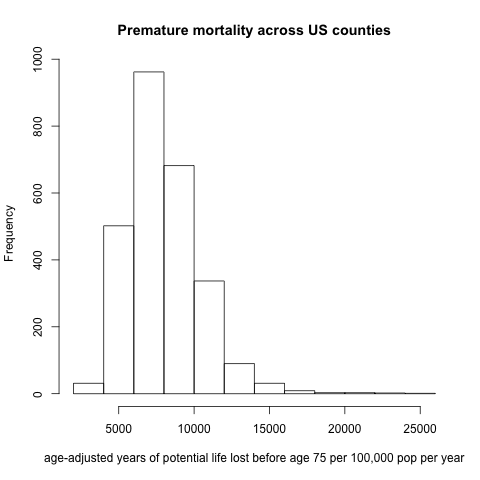
20. Environmental Protection Agency. *Safe Drinking Water Information System*. Washington D.C.; 2013.

21. U.S. Department of Agriculture. *Food Environment Atlas*. Washington D.C.; 2013.

22. Dartmouth University. *Dartmouth Atlas of Healthcare*. Hanover; 2013.

## Web figure 1:

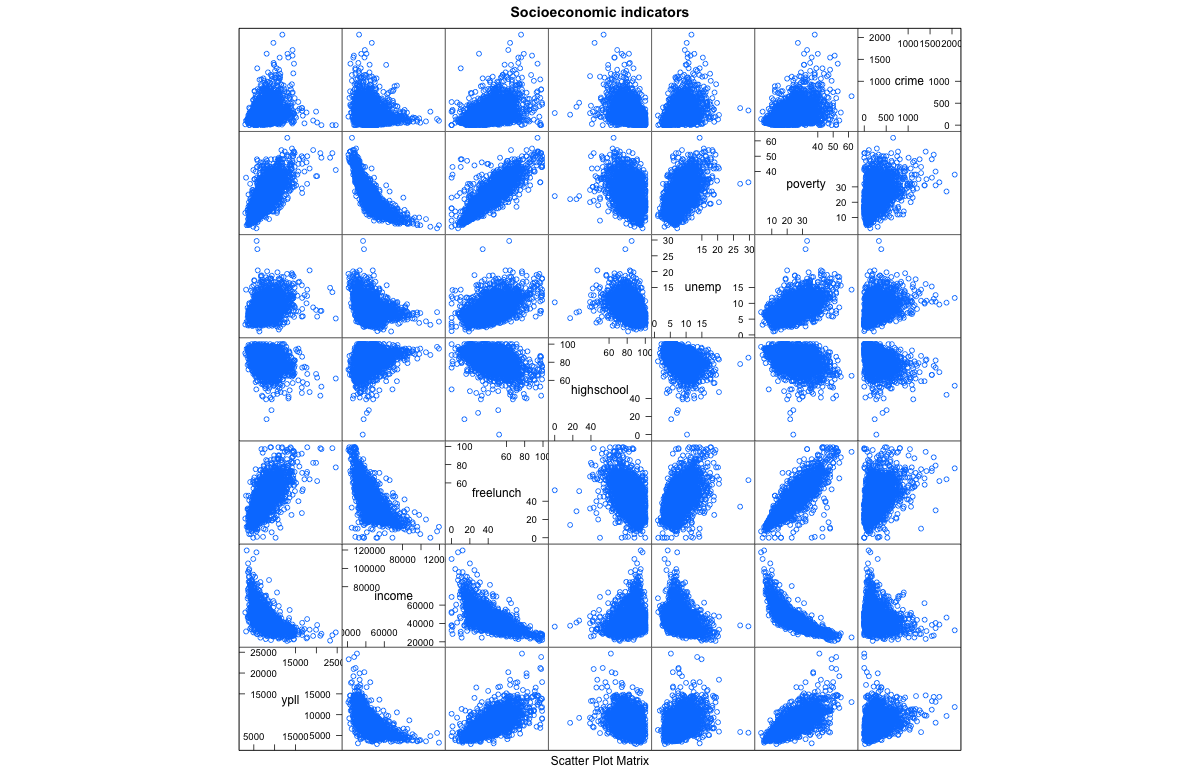
Distribution of premature mortality across US counties, expressed as mean annual age-adjusted potential years of life lost before age 75 per 100,000 population, 2008-2010.



## Web figure 2:

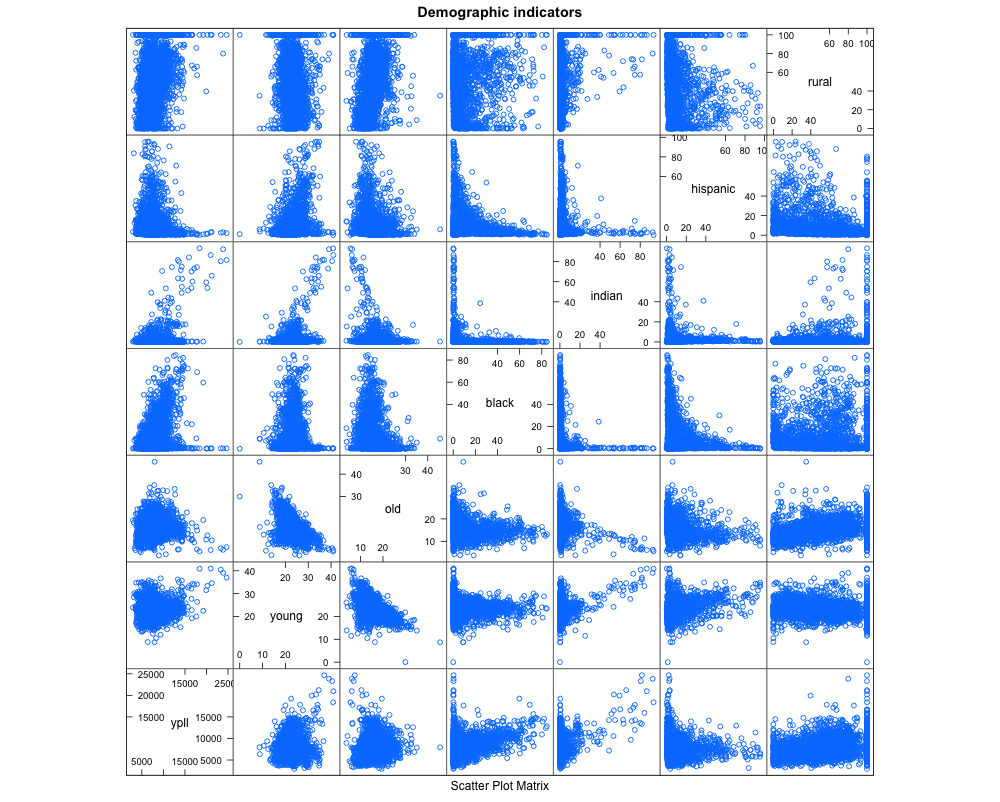
Disparities and correlations among (A) socioeconomica, (B) demographicb, (c) behavioralc and (d) environmentald indicators across all US counties. In all graphs, years of potential life lost are shown in the bottom left box.

(A)



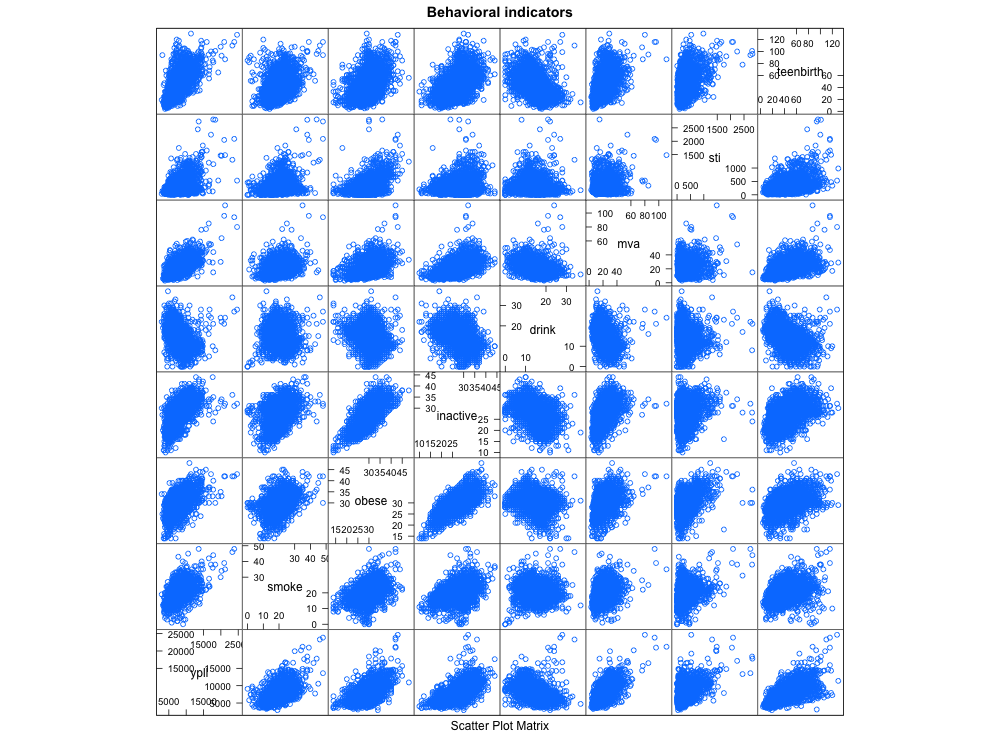
a ypll = years of potential life lost; income = median household income; freelunch = percent of children eligible for federal free lunch program; highschool = high school graduation rate; unemp = unemployment rate; poverty = child poverty rate; crime = violent crime rate. All indicators are defined in Web table 1.

(B)



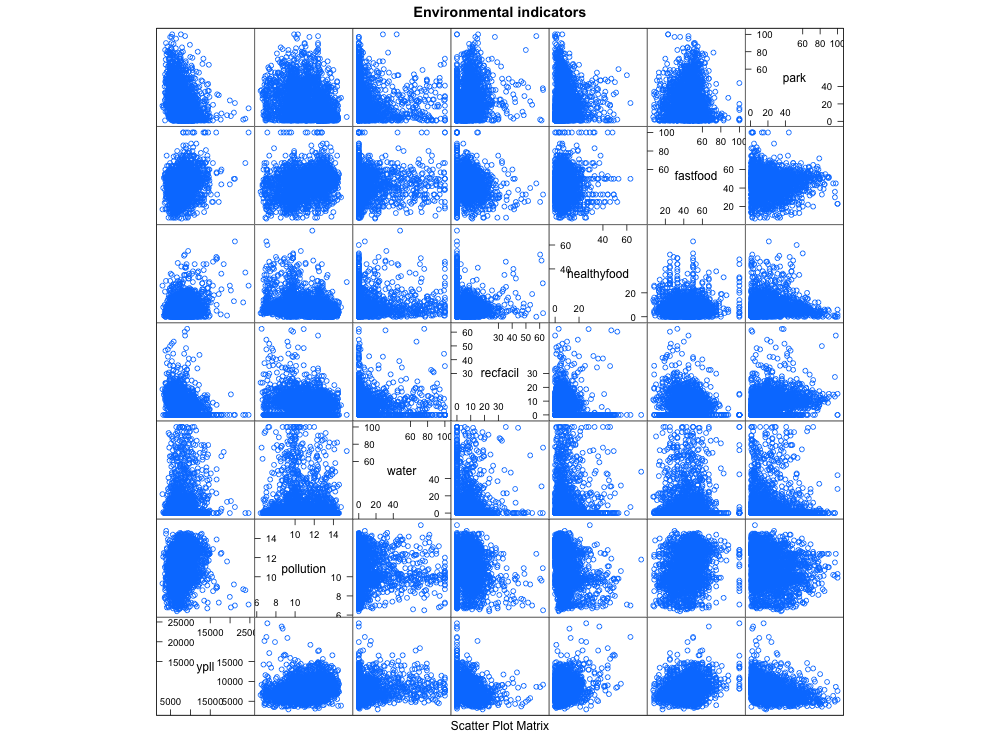
b ypll = years of potential life lost; young = % below 18 years of age; old = % 65 and older; black = % non-Hispanic African American; indian = % American Indian and Alaskan Native; Hispanic = % Hispanic; rural = % rural. All indicators are defined in Web table 1.

(C)



c ypll = years of potential life lost; smoke = % adults who smoke tobacco; obese = % adults obese; inactive = % adults with no leisure time physical activity; drink = % adults engaging in heavy or binge drinking; mva = motor vehicle accident death rate; sti = sexually transmitted infection rate (chlamydia); teenbirth = births per 1,000 females aged 15-19. All indicators are defined in Web table 1.

(D)



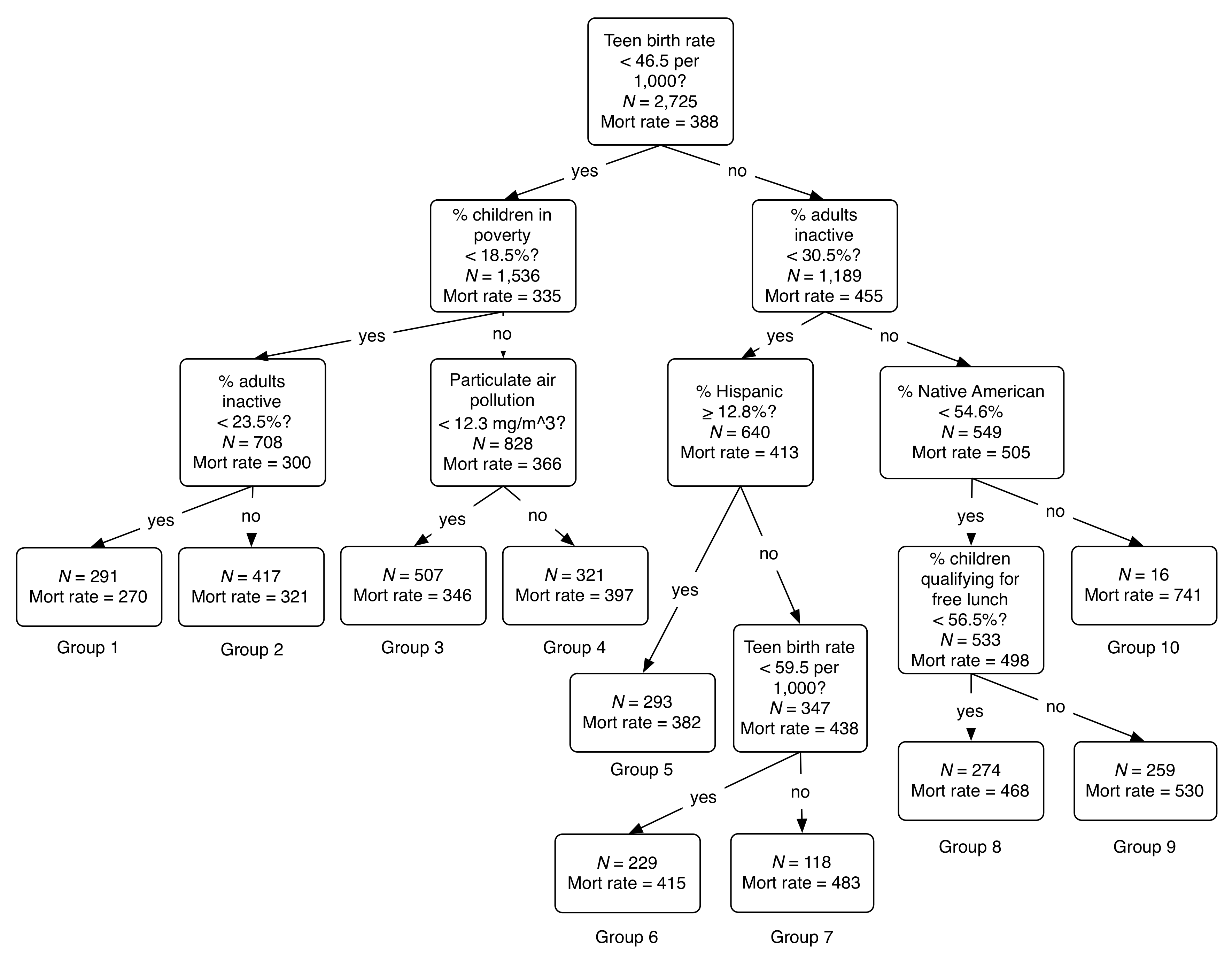
d ypll = years of potential life lost; pollution: daily fine particulate matter; water = drinking water safety; recfacil = access to recreational facilities; healthyfood = % of people more than 10 miles from grocery store; fastfood = % of restaurants that are fast food; park = % of population with access to parks. All indicators are defined in Web table 1.

# 

## Web figure 3

Regression trees among alternative outcome metrics, including: (A) age-adjusted overall premature mortality rate (deaths prior to age 75 per 100,000 population) among all counties and (B) among counties with at least 10,000 population; (C) infant mortality rate among all counties and (D) among counties with at least 10,000 population; (E) child mortality rate among all counties and (F) among counties with at least 10,000 population; and (G) self-reported poor or fair health (percent of adults) among all counties and (H) among counties with at least 10,000 population. For contrast with main text figure 1, we also present (I) disparities in premature mortalitymeasured as potential years of life lost per 100,000 population before age 75, among only U.S. counties with at least 10,000 population.

(A)



# (B)

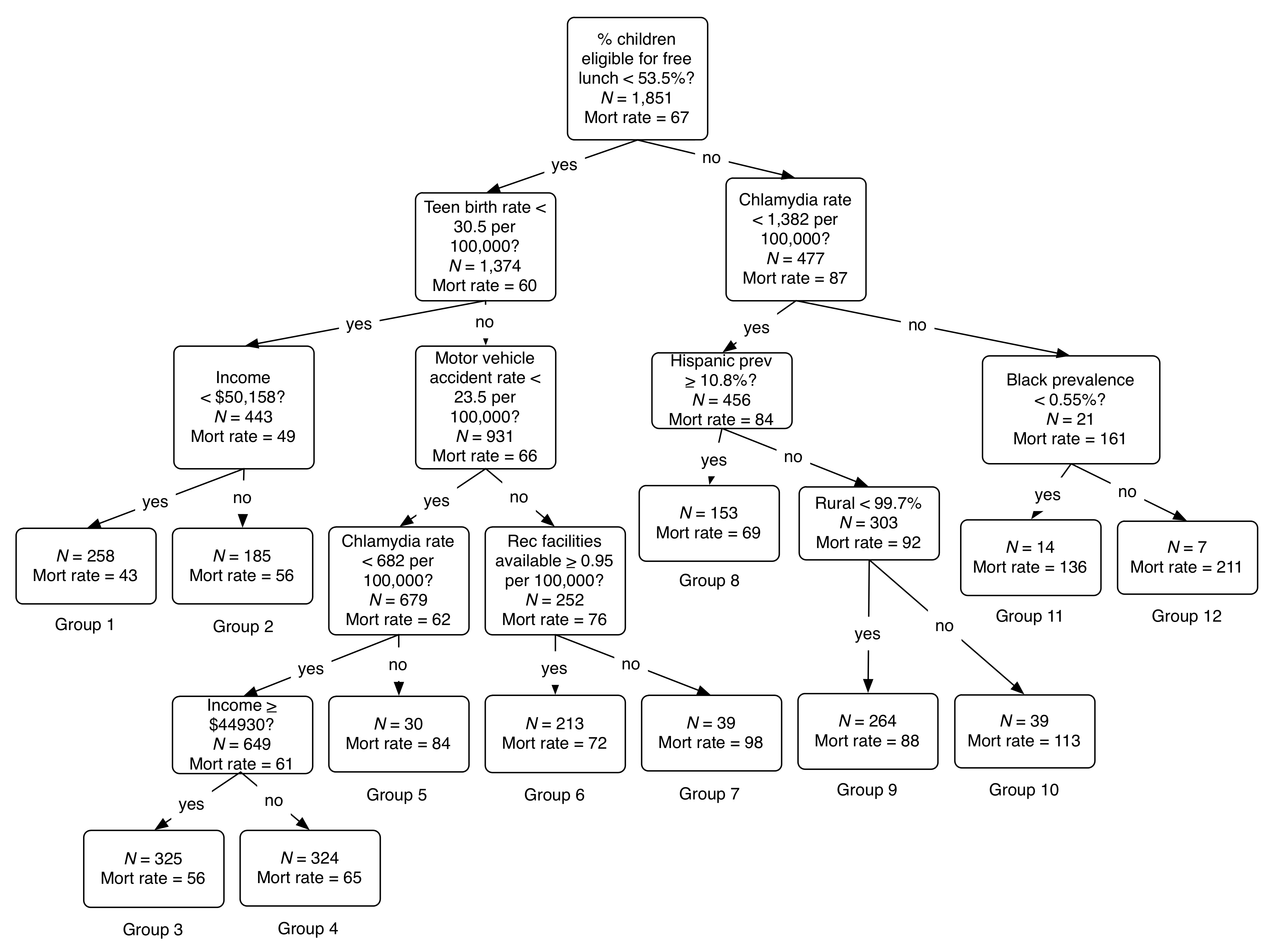
# Macintosh HD:Users:sbasu:Data:County Health Rankings:Web Fig2B.tiff

# (C)

# Macintosh HD:Users:sbasu:Data:County Health Rankings:Web Fig3A.tiff

# (D)

# Macintosh HD:Users:sbasu:Data:County Health Rankings:Web Fig2D.tiff(E)



# (F)

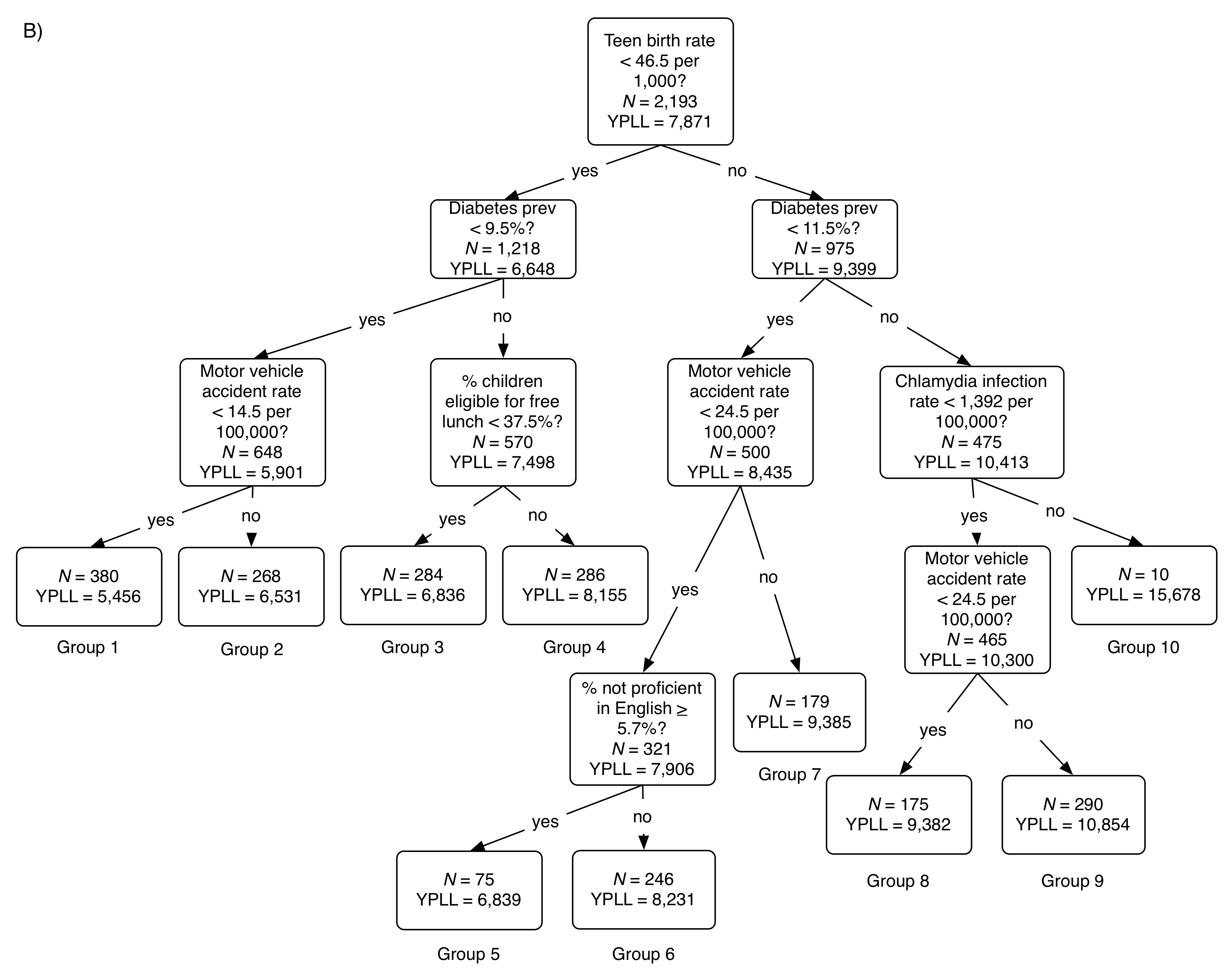
# Macintosh HD:Users:sbasu:Data:County Health Rankings:Web Fig2F.tiff (G)

# Macintosh HD:Users:sbasu:Data:County Health Rankings:Web Fig2G.tiff

# (H)

# Macintosh HD:Users:sbasu:Data:County Health Rankings:Web Fig2G.tiff

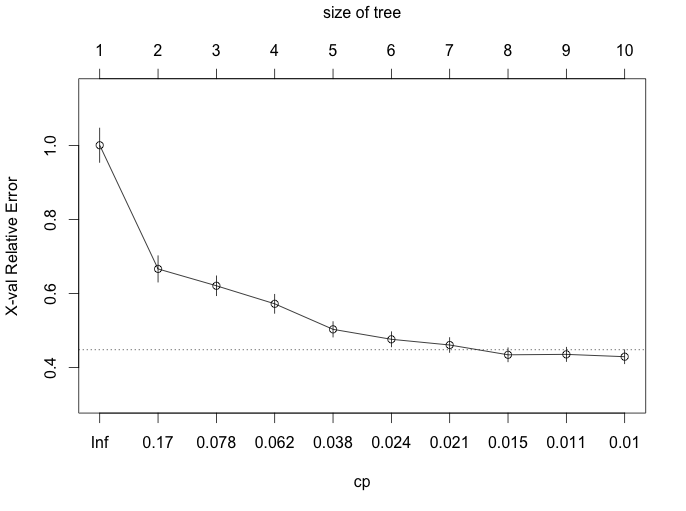
# (I)



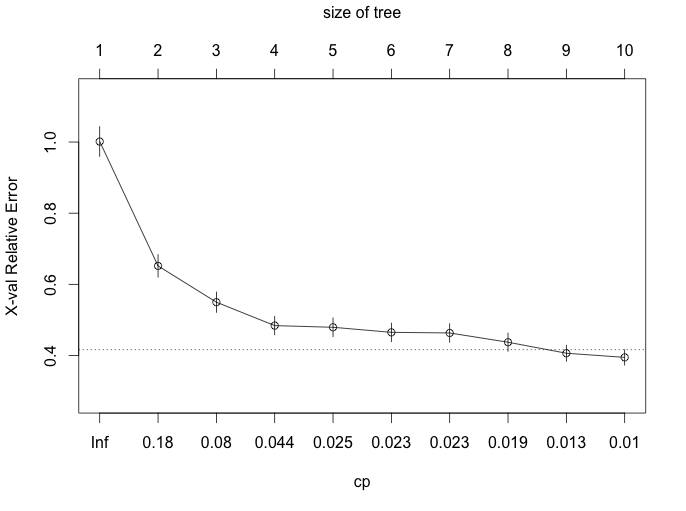
## Web figure 4

Cross-validation results. The web figures display the complexity parameter on the x axis, which increases with each additional indicator included in the tree. The y-axis plots the mean and standard deviation of the errors during cross-validation at each stage of tree construction. An accepted metric for “pruning” the tree to the most parsimonious tree with accepweb table level of error is the leftmost value for which the mean relative error lies below the horizontal dashed line, which is drawn one standard error above the minimum of the curve. Here we display cross-validation for (A) the tree predicting years of potential life lost among all counties and (B) the tree restricted to the subsample of counties with at least 10,000 population. Similar complexity parameter-based pruning was performed on all other regression trees, show in Web figure 4. Abbreviations: cp = complexity parameter.

(A)



# (B)



## Web table 1:

List of key indicators of socioeconomic, demographic, behavioral, environmental and healthcare conditions by county, evaluated in the study.a

|  |  |  |  |
| --- | --- | --- | --- |
| Category | Indicator | Source | Years |
| Socioeconomic | Median household income | Small Area Income and Poverty Estimates15 | 2011 |
|  | % households who qualify for federal designation of ‘high’ housing costs | American Community Survey15 | 2007-2011 |
|  | % children eligible for free lunch | National Center for Education Statistics16 | 2011 |
|  | Homicide rate (murders per 100,000 population per year) | National Center for Health Statistics2 | 2004-2010 |
|  | High school graduation rate (percent of ninth grade cohort that graduates in 4 years) | National Center for Education Statistics16 | 2008-2012 |
|  | Some college (Percent of adults aged 25-44 years with some post-secondary education) | American Community Survey15 | 2007-2011 |
|  | Unemployment rate (percent of population age 16+ unemployed) | Bureau of Labor Statistics17 | 2011 |
|  | Children in poverty (percent of children under age 18 in poverty) | Small Area Income and Poverty Estimates15 | 2011 |
|  | Inadequate social support (percent of adults without social/emotional support) | Behavioral Risk Factor Surveillance System2 | 2005-2010 |
|  | Percent of children that live in single-parent household | American Community Survey15 | 2007-2011 |
|  | Violent crime rate (per 100,000 population per year) | Federal Bureau of Investigation18 | 2008-2010 |
| Demographic | % below 18 years of age | US Census Bureau15 | 2011 |
|  | % 65 and older | US Census Bureau15 | 2011 |
|  | % Non-Hispanic African American | US Census Bureau15 | 2011 |
|  | % Native American Indian or Alaskan Native | US Census Bureau15 | 2011 |
|  | % Asian | US Census Bureau15 | 2011 |
|  | % Native Hawaiian/Other Pacific Islander | US Census Bureau15 | 2011 |
|  | % Hispanic | US Census Bureau15 | 2011 |
|  | % Non-Hispanic white | US Census Bureau15 | 2011 |
|  | % not proficient in English | US Census Bureau15 | 2007-2011 |
|  | % Females | US Census Bureau15 | 2011 |
|  | % Rural | US Census Bureau15 | 2010 |
| Behavioral | Adult smoking (percent of adults that smoke) | Behavioral Risk Factor Surveillance System2 | 2005-2011 |
|  | Adult obesity (percent of adults that report a BMI >= 30) | National Center for Chronic Disease Prevention and Health Promotion2 | 2009 |
|  | Physical inactivity (percent of adults that report no leisure time physical activity) | National Center for Chronic Disease Prevention and Health Promotion2 | 2009 |
|  | Excessive drinking (percent of adults who report heavy or binge drinking) | Behavioral Risk Factor Surveillance System2 | 2005-2011 |
|  | Motor vehicle crash deaths (per 100,000 population per year) | National Center for Health Statistics2 | 2004-2010 |
|  | Sexually transmitted infections (chlamydia rate per 100,000 population per year) | National Center for HIV/AIDS, Viral Hepatitis, STD, and TB Prevention2 | 2010 |
|  | Teen birth rate (per 1,000 females ages 15-19 per year) | National Center for Health Statistics2 | 2004-2010 |
| Environmental | Access to parks (% of population with neighborhood access per federal definition) | CDC19 | 2010 |
|  | Daily fine particulate matter (average daily measure in micrograms per cubic meter) | CDC WONDER2 | 2008 |
|  | Drinking water safety (percent of population exposed to water exceeding a violation limit in the past year) | Environmental Protection Agency20 | 2012 |
|  | Access to recreational facilities (rate per 100,000 population) | County Business Patterns15 | 2010 |
|  | Limited access to healthy foods (percent of population who lives in poverty and more than 10 miles from a grocery store) | US Department of Agriculture21 | 2012 |
|  | Fast food restaurants (percent of all restaurants that are fast food) | County Business Patterns15 | 2010 |
|  | % commuters driving alone | American Community Survey15 | 2007-2011 |
| Healthcare | Health care cost per capita (among Medicare population) | Dartmouth Atlas of Health Care | 2009 |
|  | Uninsured adults (% over age 18) | Small Area Health Insurance Estimates15 | 2010 |
|  | Uninsured children (% under age 18) | Small Area Health Insurance Estimates15 | 2010 |
|  | Could not see doctor due to cost (self-reported, % of adults) | Behavioral Risk Factor Surveillance System2 | 2005-2011 |
|  | Uninsured (% of population < age 65 without health insurance) | Small Area Health Insurance Estimates15 | 2010 |
|  | Prevenweb table hospital stays (rate per 1,000 Medicare enrollees) | Dartmouth Atlas of Health Care22 | 2010 |
|  | Diabetic screening (% of diabetics that receive HbA1c screening) | Dartmouth Atlas of Health Care22 | 2010 |
|  | Mammography screening (% of females that receive screening) | Dartmouth Atlas of Health Care22 | 2010 |

a All listed indicators have not been uploaded to the 2013 County Health Indicators Database for public data retrieval.1

Web table 2: Summary statistics among county clusters identified by the data mining algorithm.a

|  |  |  |  |
| --- | --- | --- | --- |
| Tree A Groupsb (see Figure 1) | *N* | YPLLa | 95% CI |
| Group 1 (healthiest) | 503 | 5,623 | 5,542, 5,717 |
| Group 2 | 535 | 6,852 | 6,782, 6,968 |
| Group 3 | 382 | 7,649 | 7,491, 7,778 |
| Group 4 | 67 | 9,338 | 8,929, 9,661 |
| Group 5 | 324 | 7,898 | 7,640, 7,963 |
| Group 6 | 399 | 9,297 | 9,109, 9,416 |
| Group 7 | 295 | 10,120 | 9,939, 10,259 |
| Group 8 | 122 | 11,947 | 11,654, 12,272 |
| Group 9 | 12 | 13,771 | 8,550, 8,945 |
| Group 10 (least healthy) | 14 | 19,102 | 9,581, 9,840 |
| Observations | 2,653 | | |
| *R*2 | 0.64 | | |
| Tree B Groupsc (see Web figure 3) |  |  |  |
| Group 1 (healthiest) | 380 | 5,456 | 5,364, 5,538 |
| Group 2 | 268 | 6,531 | 6,377, 6,647 |
| Group 3 | 284 | 6,836 | 5,622, 5,807 |
| Group 4 | 286 | 8,155 | 7,982, 8,261 |
| Group 5 | 75 | 6,839 | 6,610, 7,109 |
| Group 6 | 246 | 8,231 | 8,082, 8,366 |
| Group 7 | 179 | 9,385 | 9,178, 9,636 |
| Group 8 | 175 | 9,382 | 9,208, 9,567 |
| Group 9 | 290 | 10,854 | 10,700, 11,056 |
| Group 10 (least healthy) | 10 | 15,678 | 12,431, 16,899 |
| Observations | 2,193 | | |
| *R*2 | 0.67 | | |

Abbreviations: YPLL = years of potential life lost per 100,000 population per year, mean of 2008-2010

a Clusters were identified based on the outcome variable of years of potential life lost per 100,000 population per year. All Group numbers correspond to the labels shown in Web figure 3.

b Including all counties, see main text figure 1.

c Including only counties with population size >10,000, see Web figure 3.

## Web table 3:

Summary statistics on indicators and outcome metrics.

|  |  |  |  |
| --- | --- | --- | --- |
| Indicator | Median | Mean | Standard deviation |
| % < 18 years old | 22.9 | 23.1 | 3.5 |
| % adults reporting no social-emotional support | 19 | 19.5 | 5.5 |
| % adults uninsured | 22 | 22.4 | 6.8 |
| % Asian | 0.6 | 1.3 | 2.7 |
| % Black | 2.1 | 8.9 | 14.4 |
| % can’t access doctor due to cost | 14 | 13.8 | 5.2 |
| % children eligible for free lunch | 41 | 42.5 | 16.3 |
| % children living with single parent | 30 | 31.1 | 10.3 |
| % children uninsured | 9 | 9.6 | 4.8 |
| % college grad | 54.3 | 54.3 | 12 |
| % commuters driving alone | 79 | 77.6 | 8.1 |
| % diabetic | 10 | 10.3 | 2.2 |
| % diabetics screened with hemoglobin A1c | 85 | 83.6 | 7 |
| % drinking excessively or binging | 15 | 14.7 | 5.7 |
| % elderly | 15.9 | 16.2 | 4.3 |
| % Female | 50.4 | 50 | 2.3 |
| % high school graduate | 85 | 82.8 | 10.1 |
| % Hispanic | 3.6 | 9 | 13.8 |
| % inactive | 28 | 27.6 | 5.1 |
| % Native American | 0.6 | 2.4 | 8 |
| % not English proficient | 0.8 | 1.9 | 3 |
| % obese | 30 | 30.1 | 4.2 |
| % other race | 0 | 0.1 | 1 |
| % paying high housing cost | 28 | 28.1 | 7.3 |
| % population with limited access to healthy food | 6 | 8.4 | 8.2 |
| % population with unsafe water | 0 | 9.2 | 19.5 |
| % rural | 58.9 | 58 | 31.5 |
| % tobacco smoking | 20 | 20.3 | 6 |
| % unemployed | 8.4 | 8.7 | 3 |
| % uninsured | 18 | 18.7 | 5.8 |
| % white | 85 | 77.1 | 20.2 |
| % with access to park | 17 | 22 | 18.8 |
| % women screened with mammogram | 63.6 | 63.2 | 8.3 |
| Age-adjusted mortality rate | 376.7 | 387.8 | 99.1 |
| Average healthcare cost for Medicare | 9068 | 9156.8 | 1538.9 |
| Child mortality rate | 62.3 | 67.3 | 26.2 |
| Child poverty rate | 24 | 24.7 | 8.9 |
| Chlamydia rate per 1,000 | 233 | 311.7 | 274.6 |
| Fast food at % of restaurants | 47 | 45.1 | 13.4 |
| HIV prevalence | 103 | 177.9 | 230.2 |
| Homicide rate | 5 | 6.2 | 4.5 |
| Infant mortality rate | 690.1 | 757.4 | 286.9 |
| Median household income | 42105 | 44044 | 11101.9 |
| Motor vehicle accident death rate | 20 | 21.7 | 10.5 |
| Particulate pollution microg/m3 | 11.2 | 11.1 | 1.8 |
| Rate of prevenweb table hospital stays | 72 | 75.9 | 27.6 |
| Rec facilities per 100,000 | 6.8 | 7.6 | 7.7 |
| Self-rated fair or poor health | 16 | 16.4 | 5.4 |
| Teen birth rate per 1,000 females 15-19 | 44 | 45.6 | 20.5 |
| Violent crime rate | 214 | 272.5 | 224.6 |
| Years of potential life lost before age 75 per 100,000 | 7670 | 7985.6 | 2392.5 |

## Web table 4:

Ordinary least squares multivariate regression of indicators against health outcome variables. (A) years of potential life lost before age 75 per 100,000; (B) age-adjusted overall premature mortality rate (deaths prior to age 75 per 100,000 population); (C) infant mortality rate among all counties; (D) child mortality rate; and (E) self-reported poor or fair health (percent of adults).a,b

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model: | A | B | C | D | E |
| % < 18 years old | -74.1\*\*\* [-113.0,-35.1] | -3.88\*\*\* [-5.43,-2.33] | -5.13 [-14.0,3.72] | -0.41 [-1.07,0.24] | -0.23\*\*\* [-0.35,-0.11] |
| % >65 years old | -9.51 [-40.0,21.0] | -3.07\*\*\* [-4.30,-1.85] | -5.19 [-12.1,1.67] | -0.49\* [-0.97,-0.014] | -0.10\* [-0.19,-0.016] |
| % adults inactive | 28.6\*\* [7.06,50.1] | 1.86\*\*\* [0.96,2.76] | 1.34 [-3.51,6.19] | -0.035 [-0.39,0.32] | 0.14\*\*\* [0.072,0.21] |
| % adults obese | -13.4 [-37.4,10.6] | -0.056 [-1.02,0.90] | -3.08 [-9.11,2.95] | 0.026 [-0.36,0.41] | -0.010 [-0.088,0.067] |
| % adults uninsured | -72.3 [-155.4,10.9] | -1.99 [-5.57,1.59] | -6.79 [-25.5,11.9] | -1.59\* [-3.02,-0.16] | -0.030 [-0.31,0.25] |
| % Asian | -83.7\*\* [-146.7,-20.7] | -4.91\*\*\* [-7.63,-2.19] | -5.14 [-19.6,9.32] | 0.025 [-1.03,1.08] | -0.053 [-0.25,0.14] |
| % Black | -74.3\* [-133.8,-14.8] | -4.78\*\*\* [-7.38,-2.18] | 3.05 [-11.0,17.1] | 0.34 [-0.66,1.34] | -0.11 [-0.28,0.068] |
| % can't access doctor due to cost | 13.0 [-6.54,32.5] | 0.56 [-0.25,1.37] | -5.31\* [-10.3,-0.31] | -0.47\* [-0.84,-0.091] | 0.26\*\*\* [0.19,0.33] |
| % children eligible for free lunch | 2.94 [-2.82,8.70] | 0.22 [-0.039,0.48] | 0.40 [-1.75,2.55] | -0.052 [-0.19,0.082] | 0.020 [-0.0039,0.045] |
| % children in single parent household | -1.53 [-18.0,15.0] | 0.059 [-0.64,0.75] | -0.28 [-4.23,3.67] | -0.11 [-0.41,0.19] | -0.053 [-0.11,0.00023] |
| % children uninsured | -53.9\*\* [-93.5,-14.3] | -2.01\* [-3.74,-0.27] | -0.29 [-10.5,9.87] | -0.12 [-0.87,0.64] | -0.24\*\*\* [-0.37,-0.10] |
| % college grad | 0.52 [-10.0,11.1] | -0.36 [-0.80,0.084] | 0.81 [-1.79,3.41] | 0.027 [-0.17,0.23] | -0.0070 [-0.044,0.030] |
| % commuters driving alone | 18.5\*\* [7.47,29.5] | 0.66\*\* [0.17,1.15] | 2.80\* [0.021,5.57] | -0.076 [-0.28,0.13] | -0.029 [-0.070,0.013] |
| % diabetes | 74.1\*\* [18.1,130.0] | 2.81\* [0.62,5.00] | 10.3 [-2.69,23.2] | -0.30 [-1.19,0.58] | 0.30\*\* [0.11,0.48] |
| % diabetics screened with hemoglobin A1c | -18.1\* [-34.3,-1.96] | -0.76\* [-1.38,-0.14] | 2.61 [-1.37,6.59] | 0.0057 [-0.26,0.27] | -0.036 [-0.079,0.0064] |
| % excessive or binge drinking | -22.2\*\* [-38.4,-6.04] | -0.91\*\* [-1.55,-0.27] | -1.04 [-4.63,2.55] | -0.18 [-0.43,0.068] | -0.13\*\*\* [-0.18,-0.074] |
| % female | 37.8 [-13.8,89.4] | 1.58 [-0.76,3.92] | 0.90 [-12.5,14.3] | 0.17 [-0.81,1.15] | -0.038 [-0.22,0.14] |
| % high housing cost | 6.45 [-6.49,19.4] | 0.46 [-0.098,1.02] | 1.90 [-1.05,4.85] | 0.030 [-0.17,0.23] | -0.0082 [-0.052,0.036] |
| % high school grad | -5.19 [-12.7,2.36] | -0.12 [-0.44,0.20] | -0.45 [-2.30,1.39] | -0.082 [-0.20,0.039] | 0.0022 [-0.020,0.024] |
| % Hispanic | -78.2\*\* [-135.9,-20.5] | -5.07\*\*\* [-7.57,-2.56] | -2.70 [-16.0,10.6] | -0.12 [-1.08,0.85] | -0.019 [-0.18,0.15] |
| % inadequate social support | 34.9\*\*\* [16.6,53.2] | 1.59\*\*\* [0.84,2.34] | 0.13 [-4.25,4.51] | -0.15 [-0.46,0.16] | 0.14\*\*\* [0.071,0.21] |
| % Native American | -70.4\* [-134.6,-6.17] | -5.58\*\*\* [-8.29,-2.88] | 1.09 [-13.9,16.1] | 0.28 [-0.81,1.36] | -0.11 [-0.29,0.069] |
| % non English fluent | -90.9\*\*\* [-129.4,-52.4] | -4.28\*\*\* [-5.95,-2.61] | 2.78 [-5.96,11.5] | 0.17 [-0.49,0.82] | 0.14 [-0.011,0.29] |
| % Other race | 102.0 [-142.4,346.3] | 2.19 [-8.03,12.4] | 45.5 [-31.3,122.4] | 3.81 [-1.43,9.05] | 0.41 [-0.69,1.52] |
| % rural | -6.00\*\* [-10.5,-1.48] | -0.23\* [-0.43,-0.039] | 0.44 [-0.71,1.59] | -0.032 [-0.11,0.049] | 0.0021 [-0.014,0.018] |
| % tobacco smoking | 21.8\*\* [5.35,38.3] | 0.82\* [0.089,1.56] | 6.16\*\* [2.33,10.00] | 0.55\*\*\* [0.27,0.83] | 0.11\*\*\* [0.044,0.17] |
| % unemployed | 0.44 [-27.2,28.1] | 0.46 [-0.65,1.57] | -3.37 [-9.54,2.80] | -0.68\*\* [-1.16,-0.20] | 0.13\* [0.022,0.23] |
| % uninsured | 105.0 [-5.06,215.1] | 3.16 [-1.70,8.02] | 2.55 [-23.5,28.6] | 1.79 [-0.14,3.72] | 0.0024 [-0.37,0.37] |
| % White | -67.1\* [-126.4,-7.82] | -4.41\*\*\* [-6.97,-1.85] | -2.05 [-16.0,11.9] | 0.051 [-0.95,1.05] | -0.089 [-0.26,0.081] |
| % with acces to park | 5.00\*\* [1.76,8.24] | 0.21\*\* [0.068,0.35] | 0.39 [-0.44,1.21] | 0.015 [-0.050,0.081] | -0.0071 [-0.019,0.0050] |
| % with healthy food access | -6.16 [-21.4,9.06] | -0.32 [-0.98,0.35] | -1.05 [-5.06,2.96] | -0.069 [-0.36,0.22] | -0.056\* [-0.10,-0.0084] |
| % with unsafe water | -0.61 [-3.81,2.59] | -0.026 [-0.16,0.10] | 0.13 [-0.67,0.93] | -0.0020 [-0.067,0.063] | -0.0045 [-0.017,0.0080] |
| % women screened with mammogram | -26.7\*\*\* [-38.7,-14.8] | -1.36\*\*\* [-1.86,-0.86] | -1.31 [-3.93,1.32] | -0.063 [-0.27,0.14] | 0.0034 [-0.037,0.044] |
| child poverty rate | 37.6\*\*\* [20.9,54.3] | 0.93\*\* [0.25,1.60] | 2.71 [-1.37,6.80] | 0.46\*\* [0.17,0.74] | 0.055 [-0.0023,0.11] |
| chlamydia infection rate per 100,000 | 0.069 [-0.30,0.44] | 0.0024 [-0.013,0.018] | 0.078 [-0.030,0.19] | 0.012\*\* [0.0030,0.020] | -0.00065 [-0.0019,0.00061] |
| fastfood as % of restuarants | 5.39 [-3.31,14.1] | 0.16 [-0.19,0.51] | 0.22 [-1.83,2.27] | 0.023 [-0.13,0.18] | 0.048\*\* [0.018,0.077] |
| Healthcare cost per Medicare | 0.030 [-0.029,0.090] | 0.00025 [-0.0022,0.0027] | 0.017\* [0.0026,0.031] | 0.0011\* [0.000012,0.0021] | 0.00011 [-0.00010,0.00031] |
| HIV prevalence | -0.18 [-0.58,0.21] | -0.0098 [-0.027,0.0078] | 0.012 [-0.081,0.10] | -0.00038 [-0.0071,0.0063] | -0.0012\* [-0.0025,-0.0000020] |
| Homicide rate | 92.8\*\*\* [61.5,124.2] | 2.88\*\*\* [1.58,4.17] | 4.25 [-2.94,11.4] | 0.62\*\* [0.22,1.02] | -0.0066 [-0.076,0.063] |
| Median household income | -0.0026 [-0.013,0.0077] | -0.00048\* [-0.00091,-0.000057] | -0.0028\* [-0.0053,-0.00033] | -0.00021\* [-0.00039,-0.000031] | -0.000027 [-0.000061,0.0000069] |
| motor vehicle accident mortality rate | 77.7\*\*\* [63.3,92.0] | 2.38\*\*\* [1.80,2.97] | 1.38 [-2.07,4.83] | 0.53\*\*\* [0.27,0.79] | 0.042\* [0.00011,0.085] |
| particulate pollution microgram/m^3 | -13.6 [-50.8,23.6] | 0.50 [-1.04,2.04] | -0.21 [-9.01,8.60] | -0.21 [-0.80,0.38] | -0.043 [-0.16,0.075] |
| rate of prevenweb table hospital stays | 2.01 [-1.87,5.89] | 0.23\*\* [0.078,0.39] | -0.41 [-1.24,0.43] | -0.030 [-0.10,0.045] | 0.0029 [-0.010,0.016] |
| rec facilities per 100,000 | -0.21 [-14.7,14.3] | 0.062 [-0.56,0.68] | 3.16 [-0.49,6.81] | 0.22 [-0.063,0.50] | -0.015 [-0.064,0.034] |
| teen birth rate per 1,000 females 15-19 | 25.3\*\*\* [17.6,33.0] | 1.29\*\*\* [0.97,1.61] | 3.00\*\*\* [1.34,4.66] | 0.38\*\*\* [0.25,0.51] | 0.035\*\* [0.012,0.058] |
| violent crime rate | -0.14 [-0.44,0.17] | -0.0046 [-0.018,0.0086] | -0.047 [-0.13,0.040] | -0.0042 [-0.0100,0.0015] | -0.00017 [-0.0012,0.00084] |
| Constant | 10571.8\*\* [4091.8,17051.8] | 742.2\*\*\* [459.4,1025.0] | 303.9 [-1281.0,1888.7] | 55.4 [-53.9,164.6] | 23.9\* [3.03,44.8] |
| Observations | 969 | 969 | 935 | 961 | 966 |
| R2 | 0.912 | 0.914 | 0.56 | 0.687 | 0.803 |

a \* *P <* 0.05, \*\* *P <* 0.01, \*\*\* *P <* 0.001

b 95% confidence intervals in brackets

# 