Paper 882-2017

The Development and Application of a Composite Score for Social Determinants of Health

Paul A. LaBrec and Ryan Butterfield, 3M Corporation

ABSTRACT

Socioeconomic status (SES) is a major contributor to health disparities in the United States. Research suggests that those with a low SES versus a high SES are more likely to have lower life expectancy; participate in unhealthy behaviors such as smoking and alcohol consumption; experience higher rates of depression, childhood obesity, and ADHD; and experience problems accessing appropriate health care. Interpreting SES can be difficult due to the complexity of data, multiple data sources, and the large number of socioeconomic and demographic measures available. When SES is expanded to include additional social determinants of health (SDOH) such as language barriers and transportation barriers to care; access to employment and affordable housing; adequate nutrition, family support and social cohesion; health literacy; crime and violence; quality of housing; and other environmental conditions, the ability to measure and interpret the concept becomes even more difficult. This paper presents an approach to measuring SES and SDOH using publicly available data. Various statistical modeling techniques are used to define state-specific composite SES scores at local areas-ZIP Code and Census Tract (CT). Once developed, the SES/SDOH models are applied to healthcare claims data to evaluate the relationship between health services utilization, cost, and social factors. The analysis includes a discussion of the potential impact of social factors on population risk adjustment.

INTRODUCTION

Socioeconomic status (SES) is a major contributor to health disparities existing in the United States (Winkleby et al., 1992). In a meta-analysis of the public health literature between the late 1970s and early 1990s, McGinnis and Foege observed that “although deaths are attributed to a specific disease (e.g., heart disease or cancer), the actual causes of death reside in the factors that determine whether and when an individual develops and succumbs to disease” (McGinnis and, 1993). Underlying social and behavioral factors were found to include tobacco, alcohol, and illicit drug use, diet and activity patterns, microbial and toxic agents, firearms, risky sexual behaviors, and motor vehicle accidents (ibid). Galea and colleagues estimated that social determinants of health (SDoH) such as poverty, income inequality, and racial segregation accounted for more than 800,000 deaths in 2000, which is "comparable to the number attributed to pathophysiological and behavioral causes” (Adler and Prather, 2015). A recent study of the relationship between poverty and cancer incidence in the US, found a greater incidence of cancers in ‘high-poverty’ versus ‘low-poverty’ areas. In addition to a higher cancer incidence, ‘high-poverty’ areas also experienced a later stage at diagnosis than ‘low-poverty’ areas for the majority of cancers studied (Boscoe et al., 2015). Furthermore, the chronic stress experienced by socially disadvantaged persons with limited access to economic and health care resources exacerbates existing health disparities (Adler and Prather, 2015).

Healthy People 2020 identifies five key environment-related determinants of health: Economic Stability, Education, Social and Community Context, Health and Health Care, and Neighborhood and Built Environment (Healthy People 2020). Social determinants of health are defined to be the social, environmental, and economic conditions that are shown to contribute to health outcome inequalities (Fig. 1). SDoH are not only defined under domains from areas such as social, environmental and behavioral factors, but genetics and health care as well. Taylor et al. estimate that social, environmental and behavioral factors account for 60% of the determinants of health, with genetics and health care each accounting for 20% (2015). These socioeconomic related inequalities are a major contributor to the existence of health disparities in the United States.

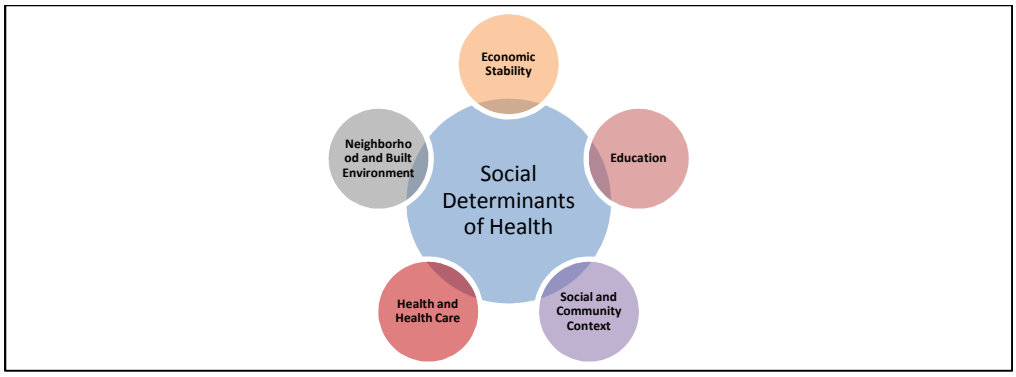


Figure 1. Relationship between SDoH and SES Index Factors

Interpreting and measuring SDoH can be difficult due to the complexity of data, multiple data sources, variation in measurements, and the large number of socioeconomic, environmental, and demographic measures available. In this paper we present one approach to identifying and selecting measures of social determinants of health, building a composite score of SDoH by small geographic area, and linking that score to administrative health care claims data for analysis of the relationship between social and clinical factors in assessing a sample of health care outcomes and costs. Finally, we present examples of mapping geocoded data on social determinants of health.

METHODS

DATA SOURCES

The model used to calculate the composite score for SDoH presented in this paper was developed using publically-available data from multiple sources. Primary sources include the American Community Survey conducted by the US Census Bureau (US Census Bureau, 2017) and the Food Access Research Atlas (Economic Research Service, 2016).

The American Community Survey (ACS) is a household survey of a representative sample of the United States population. In conducting the ACS, the Census Bureau mails survey invitations to approximately 295,000 addresses each month across the US. Surveys are generally completed online. The series of monthly samples produces annual estimates for the same small areas (census tracts and block groups) formerly surveyed via the decennial census long-form sample. Content areas surveyed include but are not limited to housing infrastructure and cost, demographics, language, employment, income, education, transportation (US Census, 2013). For this analysis we used 2015, 5-year estimate data from the ACS.

The Food Access Research Atlas (FARA) maps food access indicators for census tracts using ½-mile and 1-mile demarcations to the nearest supermarket for urban areas, 10-mile and 20-mile demarcations to the nearest supermarket for rural areas, and vehicle availability for all tracts. The Atlas assesses both food availability (including proximity of sources and availability of transportation) and household income when assessing food access. Through a combination of variables, the Economic Research Service can document areas with both low income and low access as “food deserts” (Economic Research Service, 2017). For this analysis we used 2015 data from the FARA.

Primary databases for the analysis were downloaded from the US Census Bureau for ACS data (US Census Bureau, 2017) or Economic Research Service for FARA data (Economic Research Service, 2016). Source databases were converted from their native format to SAS® v 9.4 datasets prior to analysis.

MODEL DEVELOPMENT

Define Domains

We began our work on a composite score for social determinants of health by using an approach similar to the SES model developed by Nancy Krieger and colleagues (2003, 2007). We organized our selected variables using the five domains of SDoH identified in Healthy People 2020. For this proof-of-concept

exercise, we looked for metrics from the easily accessible public data sources that would fit the domains defined in Healthy People 2020.

The 24 variables in our model can be grouped into the five domains as follows:

Education

- Low Education (% persons in CT with < HS degree)
- High Education (% persons in CT with Associate's or Bachelor's Degree)
- Current Education_PreElem (% persons in CT who are enrolled in Pre-Elementary)
- Current Education Elementary (% persons in CT who are enrolled in Gr 1-8)
- Current Education High School (% persons in CT who are enrolled in Gr 9-12)
- Current Education Undergraduate School (% persons in College, Undergraduate)
- Current Education Graduate School (% persons in Graduate or Professional School)
- Current Education Some College or Graduate School (% persons enrolled in College or Graduate School)

Economic Stability

- Persons Employed (% persons in CT who are employed)
- Persons Unemployed (% persons in CT who are unemployed)
- Households Below Poverty (% households below federal poverty level)
- Households Government Assistance (% households receiving government assistance)
- Household Income Low (% households with income <= $24,999)
- Household Income Mid (% households with income between $25,000 and $49,999)
- Household Income Mid-High (% households with income between $50,000 and $99,999)
- Household Income High (% households with income between >= $100,000)

Health and Health Care

- Insured (% persons in CT who are insured)
- Uninsured (% persons in CT who are uninsured)

Neighborhood and Built Environment

- Homes with Gas (% of homes with primary fuel source = Gas)
- Homes with Electricity (% of homes with primary fuel source = Electricity)
- Homes with Solar (% of homes with primary fuel source = Solar)
- Homes with NoFuel (% of homes with primary fuel source = No Fuel)
- Homes with Other Fuel (% of homes with primary fuel source = Other)

Social and Community Context

- Rural Urban (CT in urban versus rural area)

Prepare the Environmental Data

We refer to the public data used to create the composite score by geographic area as 'environmental data' or 'geographically-based data.' In preparing these data for use in our modeling, we collapsed some income and education ranges. All variables selected from the source data were formatted as proportions, except the urban/rural distinction which was binary. The variables in these formats were used in the PCA model.

Prepare the Claims Data

The administrative claims database used for this analysis was derived from beneficiary information and health care claims for an insured population in Iowa for the year 2015. Administrative claims include information on patient demographics, utilization of inpatient, outpatient, professional, and pharmacy services, and allowed charges—the amounts paid by the payer—for these services. Claim-level and claim-line-level detail were aggregated as necessary to produce person-level summaries for this analysis.

Linking SDoH to Claims Data

Geocoding

Health care claims data were geocoded using Pitney Bowes' MapMarker software. MapMarker returns Census Block IDs or partial Census information as output, depending on the quality of the match between the input addresses and the MapInfo reference data. The complete Census Block ID is a 15-character code (Pitney Bowes, 2016). In order to link the geographically-based SDoH data to insured members' claims data, we first geocoded individual member street addresses using the MapMarker software. The software was able to successfully geocode 90% of beneficiaries, across all claim years. Among the calendar year 2015 study period, 97% of eligible members had a geocoded census tract. Initially, 676,364 members were eligible for having their street addresses geocoded in 2015. The MapMarker software generated Census Block IDs for 658,831 or 97% of the eligible geocoded members (658,831/676,364 = 97%).

The census tract level SDoH Score was matched to individual members found in the claims database of insured enrollees for calendar year 2015 (n=571,598, 87% of geocoded patients) by geocoding addresses and linking them to the census tract of residence. This index score was then used as a covariate to explore the relationship between SDoH and costs and utilization outcomes determined from claims data.

Statistical Methods

Descriptive analysis of all variables in the analysis, including mean, median, standard deviation, first and third quartiles, was produced using PROC MEANS and PROC FREQ. The statistical methods outlined here for conducting the Principal Components Analysis are based on recommendations by O'Rourke and Hatcher (2013).

PCA

Principal Components Analysis (PCA) is a clustering technique specific for quantitative variables. SAS® has several options for this analysis and either the FACTOR or PRINCOMP procedures may be used (SAS Institute, 2016a; 2016b). This technique provides a way to identify those variables accounting for the most variation in the dataset. Additionally, PCA allows for dimensional reduction in the data set and calculation of standardized scoring coefficients. Principal components analysis also creates an optimal fit through a multivariate data "cloud" using vectors of linear combinations, each of which are either at orthogonal (uncorrelated) or oblique (correlated) angles to each other depending on the rotation used. The rotational options in SAS® 9.4 include orthomax, varimax, quartimax, equamax, and promax (the only oblique rotation). Once the orientation is settled upon, the principal components are used to either reduce the number of elements needed in the data analysis or to produce standardized scoring values. These scoring values form the basis of the scoring algorithm which generates the 3M SDoH Index Score.

Stage 1 consists of data visualizations and a principal components analysis (PCA). Major outcomes and variables of importance as identified by Stage 1 were summarized descriptively and reported according to good statistical practice. The specified outcomes were reported similarly. PCA is used for identifying those variables from the variable set which are responsible for the most variation present in the data. Those principal components which collectively comprise greater than 80% of the cumulative variance are considered for inclusion. Skree plots and scatter plots were used to identify the number of components. Communality was calculated using the formula from O'Rourke and Hatcher (2013) which is the sum of the

squared factor loadings for a given variable across factors. Using these tools, PCA(s) were run until a reduced variable set was identified.

A confirmatory factor analysis (CFA) was used to verify findings from Stage 1. This step was used to produce the factor scores which were then used to create the SDoH algorithm and score. Theoretically, the SDoH acts as a latent variable where it is hypothetically, the thematic interpretation of the underlying commonality shared by the variables in this analysis (and potentially others). The outputs from this step were used as inputs into PROC SCORE where the information from the data and from the CFA were combined. This linear combination of data elements and factor scores is what actually "creates" the score which represents SDoH for a census tract. This development of a score is our attempt at creating an SDoH measurement that is reliable and repeatable for all states at the census tract geographic level.

This index score was then used as a covariate in the multivariable regression stage of the analysis for determining the balance of influence on variation found in different outcomes comparing SDoH score and clinical risk. Clinical risk was represented by a weight and calculated using a combination of 3M's Clinical Risk Group™ (CRG), age group, and sex of each patient (Averill et al., 1999). The following utilization and cost outcome measures were modeled:

- Outpatient emergency department (ED) visits
- Outpatient emergency services
- Outpatient emergency services per visit
- Outpatient emergency department visit allowed dollars

Multivariable general linear regression was used to model the outcomes of outpatient emergency services, outpatient emergency services per visit, and outpatient ED allowed dollars. Ordinal regression was used in modeling the emergency department visits metric. SDoH and CRG/age/sex weight were the covariates included in the modeling, both as continuous effects.

Using PROC GLIMMIX we created a multivariable linear regression model, modeling outpatient ED Total Allowed ($) by a CRG/age/sex weight variable, and our newly created composite SDoH scoring variable. Some example SAS code on this model:

```
proc glimmix data=test;
  model OP_ED_Allowed = score person_Weight /
    solution distribution=normal;

run;
```

RESULTS

PCA Model

The development of an SDoH index follows the traditional Principal Components Analysis to Confirmatory Factor Analysis process, with the end goal not being dimension reduction but instead being a composite score based on all available information. We first look for which linear combination of variables (e.g. Principle Components or Factors) explain the majority of variation in the data. This is done using eigenvalues (Table 1), scree plots (Fig. 2), and cumulative variance explained by each factor (Table 2). Once the principal components analysis is conducted, a confirmatory factor analysis is performed. This process includes results based on the use of a rotation method; in this analysis the varimax and promax rotations were evaluated with the promax being selected for use, as there was a strong indication of correlation between the variables. Results from the initial PCA and the CFA are presented here and are used for the creation of the SDoH Score.

Eigenvalues represent the amount of variance captured in a given factor. Once 70-80% of the variance is accounted for and the other tools align, then a decision on the number of factors can be reached (Goldberg, 1997). As discussed by O'Rourke and Hatcher (2013), eigenvalues can also be used as a decision criterion, where those factors or components with eigenvalues greater than 1.00 are retained. SAS can be used for automating this by using the MINEIGEN=1 option in PROC FACTOR; alternatively, this can done through manual inspection by looking at those factors beyond the eigenvalue equals 1.00

mark. An additional tool is the graphical visualization of the eigenvalues and variation known as a scree plot.

Eigenvalues of the Correlation Matrix: Total = 26; Average = 1				
	Eigenvalue	Difference	Proportion	Cumulative
1	7.45027822	2.65725650	0.2865	0.2865
2	4.79302172	2.38740396	0.1843	0.4709
3	2.40561776	0.79611320	0.0925	0.5634
4	1.60950456	0.41973864	0.0619	0.6253
5	1.18976592	0.06261889	0.0458	0.6711
6	1.12714703	0.13714045	0.0434	0.7144
7	0.99000658	0.09189484	0.0381	0.7525
8	0.89811173	0.07578365	0.0345	0.7871
9	0.82232808	0.03995672	0.0316	0.8187
10	0.78237136	0.03590386	0.0301	0.8488

Table 1. Eigenvalues

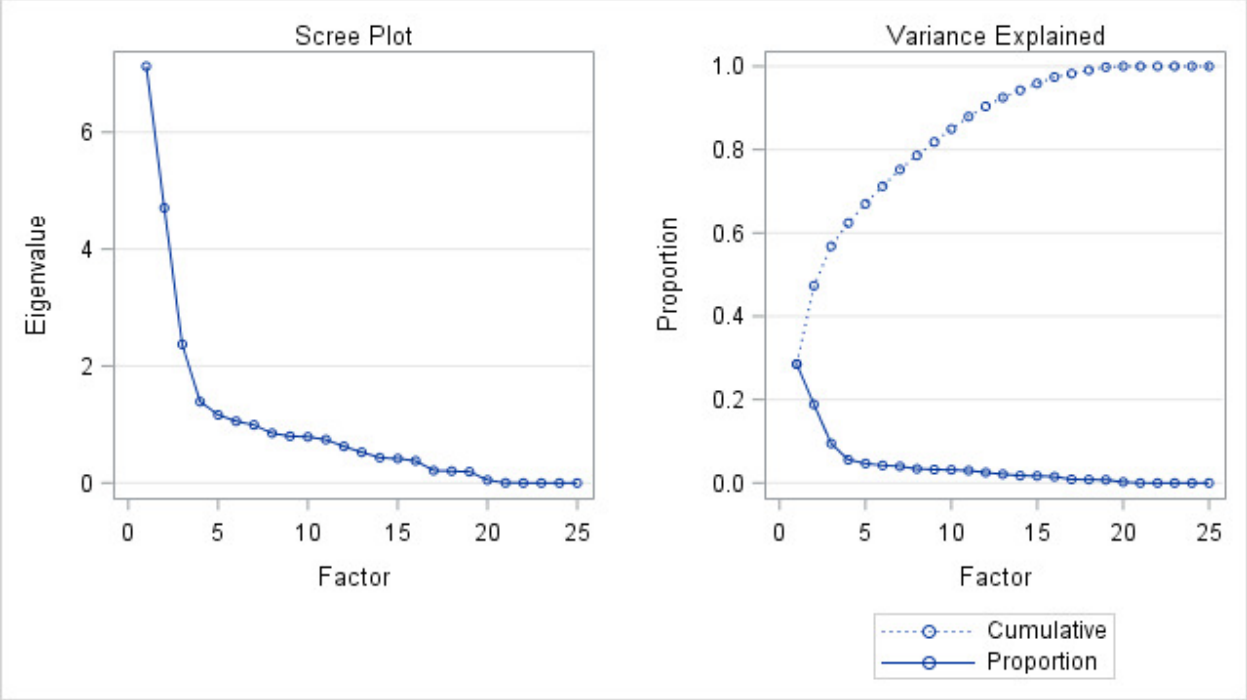


Figure 2. Scree Plots

A scree plot is useful for generating an idea as to how many factors should be retained, the heuristic commonly used is to look for the “elbow” of the graph. This indicates how many factors should be retained and is almost always in line with those indications found from the eigenvalues. An additional technique used commonly in these methods is to look at the variance explained by each factor, as the number of factors increases the variation accounted for should also increase.

Variance Explained by Each Factor (PCA-Non Rotated)					
Factor1	Factor2	Factor3	Factor4	Factor5	Factor6
7.4502782	4.7930217	2.4056178	1.6095046	1.1897659	1.1271470

Table 2. Variance by Factor

With looking at the variance explained (Table 2), clearly the first two factors account for a high amount of variation; however, given the chain of evidence, it is reasonable to include the identified six factors in the scoring model. In this analysis we decided that six factors was adequate (although a case could have been made for seven). A confirmatory factor analysis was then run to optimize the fit and make sure the number of factors selected was indeed appropriate.

To verify this, we look at the same tools as before and also use scatterplot comparisons of the factors. There are two scatterplots that we will use, one can be output directly from PROC FACTOR and another can be generated in PROC SGPLOT. These plots are presented in Figure 3. We find that using PROC SGPLOT gives more data related information, while the PROC FACTORS scatterplots display the correlation between the factors.

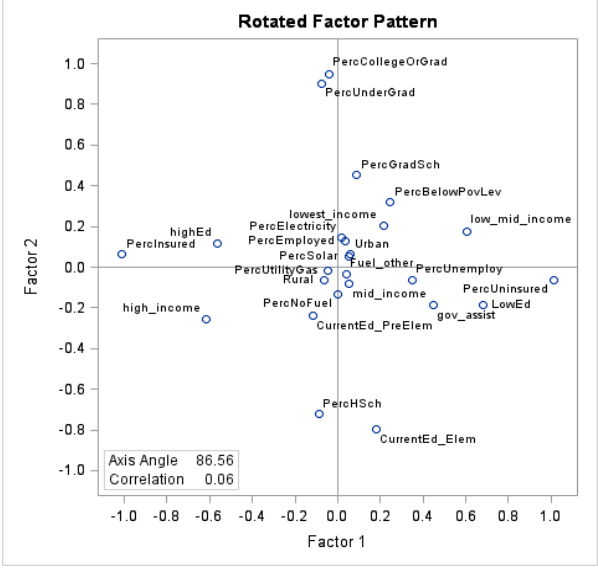
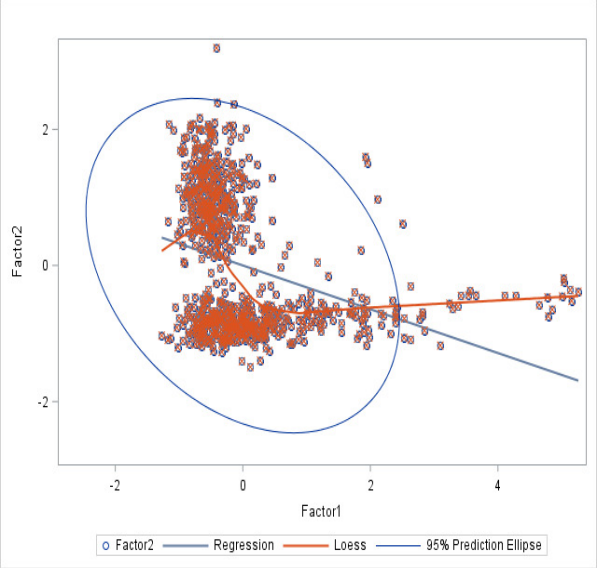
From PROC FACTOR	From PROC SGPLOT
	

Figure 3. PROC FACTOR and PROC SGPLOT Output

For example, we can use both tools to look at Factor 1 by Factor 2, which were generated using a promax rotation. In the SGPLOT figure, we include 95% predictive ellipses, a linear regression line, and a LOESS line, just for examples of the kinds of tools that can be used. The utility of using both graphics is

apparent as we identify two clear clusters of data in the SGPLOT, and we see multiple clusters of variables in the variable scatterplot from FACTOR. Looking at the pattern of factors via their permutations graphically can be overwhelming, fortunately PROC FACTOR produces a factor structure table (Table 3) based on correlations, where any correlation above 0.2 is flagged by an asterisk (*). This flag-value can be altered to suit any specification.

In attempting to assign thematic interpretation to the factors, we recall how Healthy People 2020 identifies domains that are inherent to social determinants of health. These domains were previously identified but as a reminder they are: economic stability, education, social and community context, health and health care, and neighborhood and built environment. We use the factor structure to attempt to align the metrics and clusters to domains. In doing so, we attempt to get closer and closer to a complete model of SDoH. In the table below we identify possible thematic interpretations of each factor. Factor 1 indicates that there is a cluster of census tracts positively correlated to variables related to low income, low education, and a general low SES, and there is also indication that these census tracts are negatively correlated to High SES which involves mid to high income and higher education, which is found in Factor 2. Interestingly, as the two Factors are compared side by side, both are positively correlated to living in an urban environment and living below the poverty line; however, Factor 2 has a lower correlation to unemployment. This may indicate that there are individuals with higher SES but also a group who have higher education but lower income. These dichotomies create the spectrum of SDoH and serve to actually improve the realistic interpretation of this scoring system. Factor 3 is similar to Factor 2 but some differences are seen in the urban variable. Factor 4 seems to align to the idea of the wealthy rural areas, maybe successful farmers; whereas, Factor 5 is split between rural and urban but indicates a middle class that has some education and some income. Factor 6 indicates a mid-income level but inversely correlates to education.

Closer study of these results is needed, in part to determine whether or not expanding the model to include further variables would possibly be a closer alignment to the known and published domains of SDoH.

Building upon these findings we use PROC SCORE to generate standardized scoring values. This leads us to the final scoring distribution, which will be used as a covariate to model Emergency Department Allowed amounts in an Outpatient population.

Factor Structure (Correlations) – part 1												
	Factor1		**Factor2**		**Factor3**		**Factor4**		**Factor5**		**Factor6**	
PercUninsured	92	*	0		15		-35	*	18		-7	
LowEd	80	*	-8		14		-53	*	12		17	
low_mid_income	63	*	8		2		-36	*	-24	*	54	*
gov_assist	74	*	5		37	*	-71	*	20		14	
highEd	-72	*	14		19		61	*	18		-39	*
high_income	-76	*	-23	*	-6		63	*	10		-42	*
PercInsured	-92	*	0		-15		35	*	-18		7	
PercCollegeOrGrad	-1		97	*	35	*	-16		17		-36	*
PercUnderGrad	1		92	*	30	*	-27	*	6		-30	*
PercGradSch	-7		46	*	28	*	25	*	38	*	-31	*
PercHSch	-8		-69	*	-34	*	6		-12		-5	
CurrentEd_Elem	10		-80	*	-26	*	15		-20		19	

Table 3a. Factor Structure

Factor Structure (Correlations) – part 2												
	Factor1		Factor2		Factor3		Factor4		Factor5		Factor6	
PercUtilityGas	8		27	*	89	*	-9		-13		-9	
Urban	22	*	39	*	93	*	-20		17		-15	
PercSolar	0		-2		-18		2		-8		-3	
Rural	-22	*	-39	*	-93	*	20		-17		15	
Fuel_other	-14		-35	*	-93	*	16		-17		7	
PercEmployed	-38	*	-5		0		81	*	3		-7	
mid_income	-32	*	-34	*	-34	*	65	*	-47	*	11	
PercUnemploy	58	*	14		31	*	-60	*	11		-4	
PercBelowPovLev	57	*	52	*	34	*	-76	*	35	*	-9	
lowest_income	60	*	41	*	28	*	-85	*	36	*	2	
PercElectricity	13		20		6		-15		70	*	6	
PercNoFuel	11		-5		-4		-15		59	*	-4	
CurrentEd_PreElem	-4		-33	*	-6		9		5		75	*
Printed values are multiplied by 100 and rounded to the nearest integer. Values greater than 0.2 are flagged by an '*'.												

Table 3b. Factor Structure

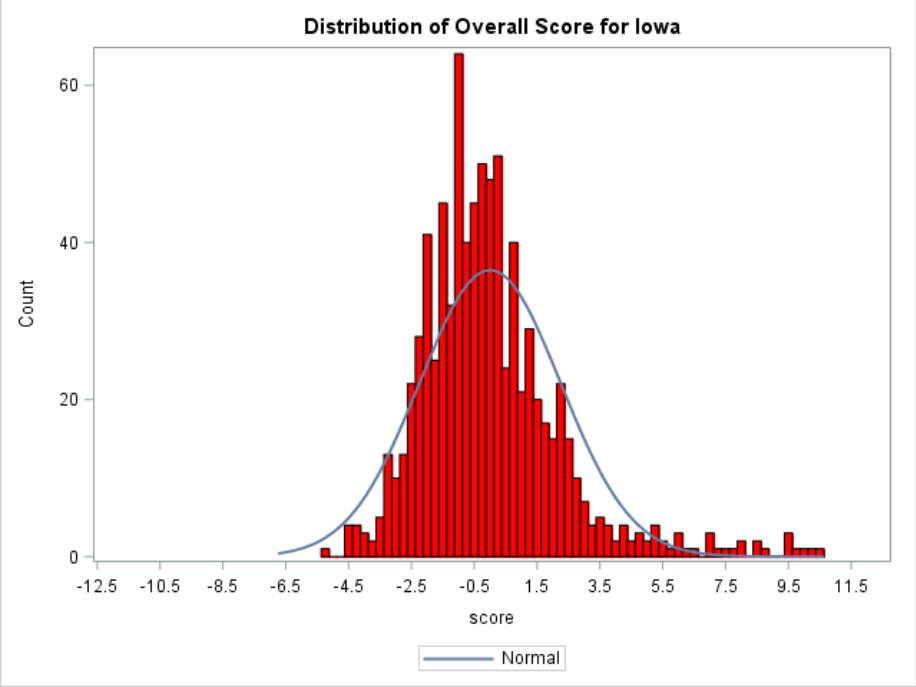


Figure 4. Distribution of Census Tract Scores for SDoH Composite Measure

Centered almost at zero, the SDoH Score shows the state of Iowa's scores at the census tract level (Fig. 4). There were n=822 census tracts used in this analysis and each one received a single score. We assume a similarity in those persons living within a census tract which may not be wholly accurate, but for our purposes, is a useful conclusion (we could go down to the block level but finding SDoH data at this level is extremely rare). Creating a composite index allows for widely diverse applications, from risk

adjustment to statistical modeling, to geographical assessments. We provide an example looking at a subset of subjects who are outpatients with ED usage.

Claims Data and Multivariate Models

Descriptive statistics for the study population attributes and outcomes of this analysis are presented in Tables 4a and 4b.

ED Visits (Ordinal)	Frequency	Percent			
0 Visits	430,263	75.27			
1 Visit	79,317	13.88			
2+ Visits	62,018	10.85			
Location					
Rural	170,154	29.77			
Urban	401,444	70.23			
Chronic Risk Segmentation					
At Risk	64,614	11.30			
Complex Chronic	36,578	6.40			
Critical	2,822	0.49			
Healthy	214,774	37.57			
Non User	170,002	29.74			
Simple Chronic	57,384	10.04			
Stable	25,424	4.45			
Gender					
F	314,653	55.05			
M	256,945	44.95			
Variable	**Mean**	**Std Dev**	**Median**	**Minimum**	**Maximum**
Age	24.5	18.5	21.0	0.0	115.0
Score	0.2	2.4	-0.3	-5.2	10.5
test_rate	6.5	6.3	4.3	1.0	99.0

Table 4a. Study Population Attributes and Outcomes

ED Visits (Continuous)	Mean	Std Dev	Median	Minimum	Maximum
OP_ED_Visits	0.5	1.3	0.0	0.0	59.0
OP_ED_Services	3.4	12.1	0.0	0.0	1218.0
OP_ED_Allowed	$215	$819	$0	$0	$61,460
Person_Weight (CRG)	0.9	2.9	0.3	0.0	248.3

Table 4b. Study Population Attributes and Outcomes

Results from the linear regression models are presented in Table 5. Interpreting this linear regression model, we estimate that for every unit increase in SDoH Score, there is a 3.3445 increase in Total ED Allowed, and for every unit increase in clinical risk there was a $64.1354 increase in Total ED Allowed.

	Slope Estimates		Standard Error	p-values
Total ED Allowed				
Intercept	156.12		1.1108	<.0001
SDoH Score	3.3445		0.4404	<.0001
Clinical Person Weight	64.1354		0.3699	<.0001
Scale	636686		1190.96	.
Count Of ED Services				
Intercept	2.4474		0.01641	<.0001
SDoH Score	0.07587		0.006506	<.0001
Clinical Person Weight	1.0008		0.005465	<.0001
Scale	138.96		0.2599	.
Ratio Of Services Per Visit				
Intercept	5.9325		0.01790	<.0001
SDoH Score	0.003358		0.006695	0.6160
Clinical Person Weight	0.3084		0.004095	<.0001
Scale	38.0862		0.1433	.
Ordinal Representation Of ED Visits				
Effect	OP ED ordinal	Estimate	Standard Error	Pr > |t|
Intercept	0 vs 2	1.2267	0.003330	<.0001
Intercept	1 vs 2	2.2408	0.004539	<.0001
Score		-0.01502	0.001263	<.0001
Person_Weight		-0.1158	0.001228	<.0001

Table 5. Linear Regression Model Results

These results indicate statistically significant relationships between SDoH Score, clinical person-weight, and the outcome of Total ED Allowed Amount ($).

Both SDoH Score and clinical person-weight are significantly associated to the number of emergency department services used, with the slopes indicating that for every unit increase in SDoH there is a slight increase in services, while with the clinical risk weight, for every unit increase there is a corresponding increase of 1 service.

When we examine the ordinal representation of emergency department visits (0, 1, 2+) we see again that both SDoH Score and clinical person-weight are statistically significant but have an inverse association with visits. Both variables indicate a slight decrease in visits.

Combining the two previous outcomes to form the ratio of services per visit, we find that there is a significant relationship with the clinical risk weight only, which indicates an increase in the ratio of services for every unit increase in weight. SDoH did not have a statistical relationship with the ratio of services per visit.

In addition to describing and analyzing the population using statistical techniques, we can use mapping techniques to view the distribution of SDoH Scores and their association with various outcomes of interest. Figure 5 present a map of SDoH Score by Iowa census tracts. For this map SDoH Scores are categorized by quintiles and output in a heat map, with darker areas representing the lower SDoH Scores.

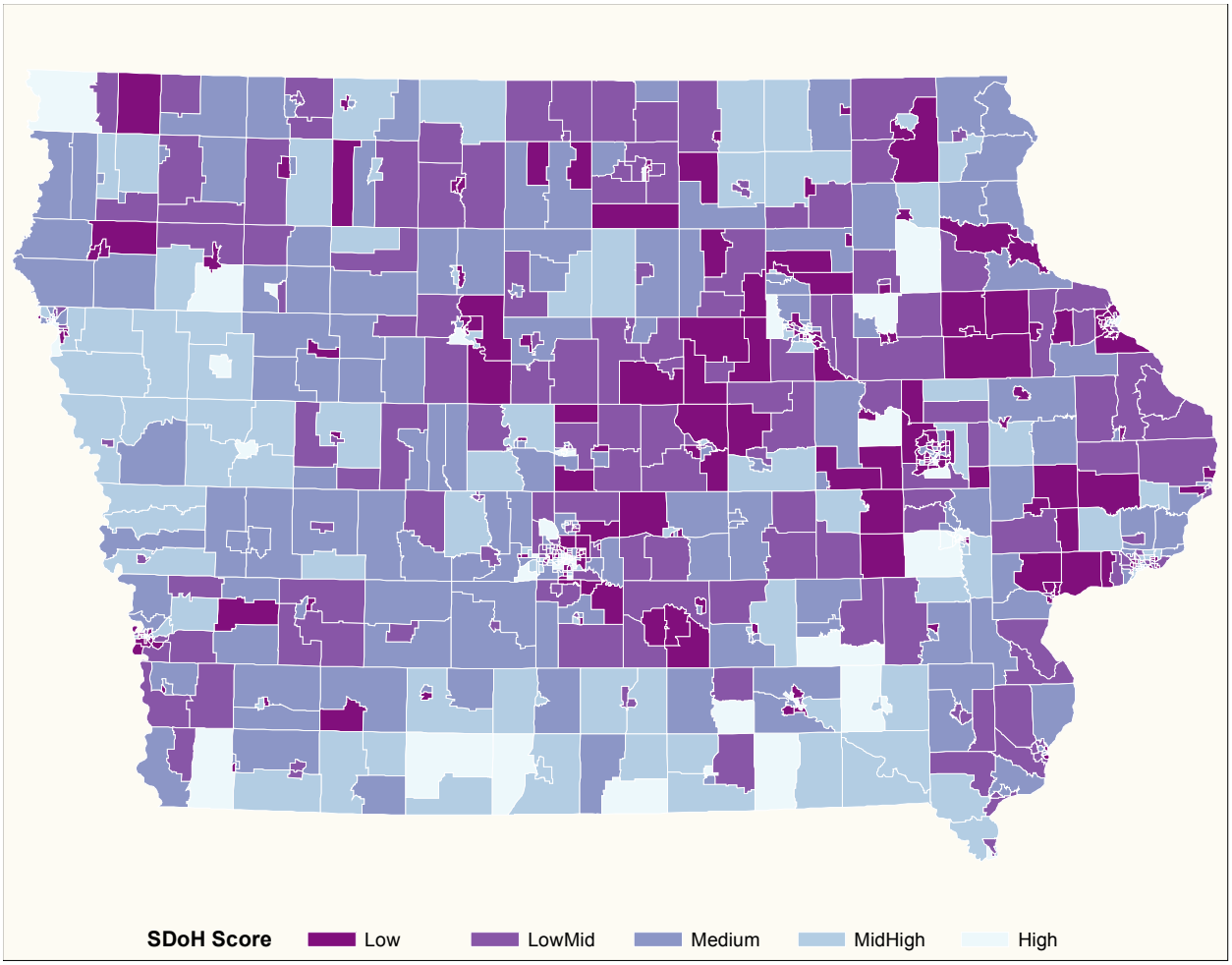


Figure 5. Heat Map of SDoH Scores by Census Tract - Iowa

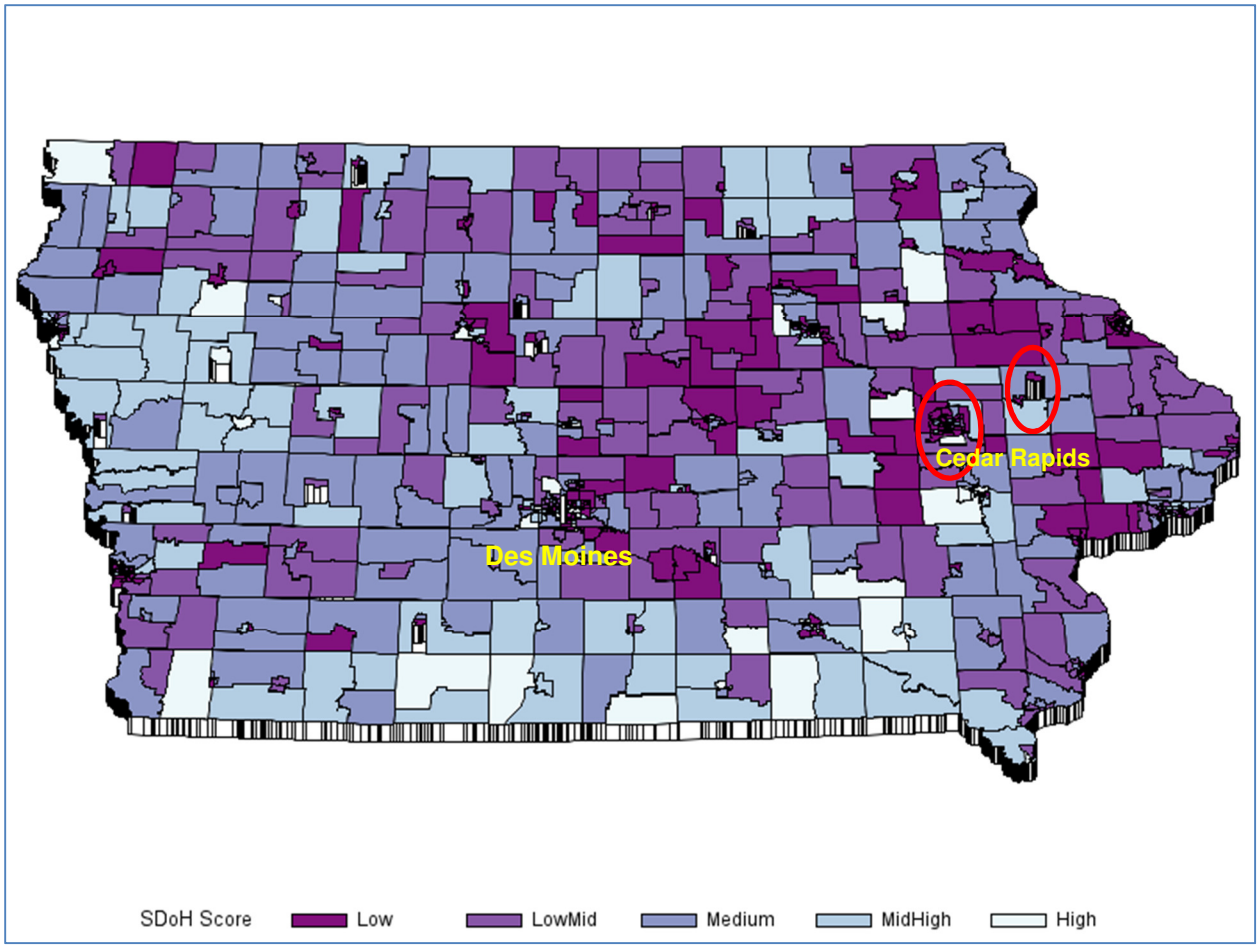


Figure 6. SDoH Heat Map Overlaid with Average OP ED Services per Visit by Census Tract – Iowa

To produce Figure 6 we began with the SDoH heat map and added the average number of services per emergency department visit using the PRISM command in PROC GMAP to create raised areas for higher services per visit. By combining these two variables we can see areas with both a low SDoH Score and a high number of services per ED visit—two examples are circled in red. This may be an indicator of neighborhoods with persons needing more regular intervention so they don't wind up in the ED requiring multiple services.

DISCUSSION

The gathering and organizing of large amounts of information is still the central tenant to statistical analysis in the current era of large data. This collection and organization is one of the greatest challenges in this analysis—which variables measure what construct and to which domain should they belong? There were over 15 data sources searched for over 400 different variables evaluated. Due to the fact that not all data is collected and stored equally, we arrived at a subset of 220 variables at the census tract level. From this we sorted variables into domains as identified by Healthy People 2020 and attempted to narrow down the variable list using PCA. There were issues of sample size and power, as the unit of measure is the census tract, so there were too many variables to be included in a single PCA. To meet the goal of this paper we restricted the variable list qualitatively based on the literature and identified at least a few variables under each domain which we included here.

While this paper reflects a small subset of variables that represent SDoH, the reality is that an initial run of the PCA process was done with approximately 220 variables, some were measuring similar constructs but in their own unique way contributing to the developing field of knowledge that is social determinants of health. In choosing the limited set of variables (we decided upon a final 24 variables) for this analysis, we

perhaps have some misclassification within this score, but overall we found it to match up to metadata we identified from online sources (www.datausa.io).

There are additional steps that are planned for this development aside from adding additional data sources, metrics, and domains. These steps include further statistical development, the use of more advanced machine learning techniques to verify clustering results, and more complex modeling that incorporates additional information such as correlation or covariance. We imagine a statistical process that incorporates all available public information and uses it to gauge the effect that these determinants have on a person's health, wealth, and well-being. This type of process would allow for repeatable and reliable estimation of SDoH, and its applications would be widespread.

Next steps in the application of this analysis include: a) an investigation on the potential role of an SDoH Score in supplementing clinical risk adjustment; and b) an assessment of the relationship between the geographically defined SDoH Score based on a patient's residence and indications of SDoH factors reflected in a range of Z Codes, which are now available with the advent of ICD-10 as supplemental diagnosis codes on healthcare claims.

The National Quality Forum recently convened an expert panel to issue recommendations for supplementing clinical risk adjustment (or case-mix adjustment) with risk adjustment for socioeconomic and other sociodemographic factors (NQF, 2014). In its recommendations, the panel gave a cautious endorsement of the use of sociodemographic factors when risk-adjusting clinical performance measures. The recommendations outlined not a blanket endorsement, but specific criteria for the use of sociodemographic factors in risk-adjustment for clinical performance measures (p. vii) including:

- Clinical/conceptual relationship with the outcome of interest
- Empirical association with the outcome of interest
- Variation in prevalence of the factor across the measured entities
- Present at the start of care
- Is not an indicator or characteristic of the care provided (e.g., treatments, expertise of staff)
- Resistant to manipulation or gaming
- Accurate data that can be reliably and feasibly captured
- Contribution of unique variation in the outcome (i.e., not redundant)
- Potentially, improvement of the risk model (e.g., risk model metrics of discrimination, calibration)
- Potentially, face validity and acceptability

Following similar guidelines we are beginning research into the potential of supplementing clinical risk assessment with adjustment for social determinants of health.

> With the advent of the International Classification of Diseases-Tenth Revision (ICD-10), coders of health care records now have the ability to add supplemental codes (Z Codes) providing information on additional factors affecting a patient's health status. A section of these codes (Z55-Z65) indicates the presence of one or more "potential health hazards related to socioeconomic and psychosocial circumstances" (ICD-10 Data.com, 2016). These codes include items such as the following:
>
> Z59 - Problems related to housing and economic circumstances
>
> - Z59.0 - Homelessness
> - Z59.1 - Inadequate housing
> - Z59.4 - Lack of adequate food and safe drinking water
> - Z59.5 - Extreme poverty
> - Z59.6 - Low income
> - Z59.7 - Insufficient social insurance and welfare support

The presence of codes such as these on medical records indicate a social determinant that is specific to a patient, not just an attribute of their neighborhood. While our initial investigations indicate the

prevalence of these codes on healthcare claims is fairly low, the codes present a unique source of data for the validation of geographically-based SDoH models, as well as measuring the association between these codes and various clinical outcomes.

In addition to potential refinement of risk adjustment, we believe a geographically-based SDoH Score can have utility for care management programs in healthcare delivery. Having knowledge about the likelihood that patients from a particular neighborhood will have adverse circumstances affecting the success of their treatment or care plans, can allow care managers to augment their care management strategies to help address these circumstances.

CONCLUSION

The development of a standardized score for social determinants of health based on an individual's place of residence can be an enterprise undertaken primarily by scientists. Determining where the healthcare system would benefit most from such a score, on the other hand, is a conversation for a much broader group of participants. The application of such a social score and the potential actions resulting from the score occur in a social context involving researchers, patients, healthcare payers and providers, and public and private institutions of various types. With this type of statistical insight comes a great responsibility - that as citizen scientists we handle the application of these predictive analytics with care.

REFERENCES

Adler, N and Prather, A. Determinants of Health and Longevity. Content last reviewed July 2015. Agency for Healthcare Research and Quality, Rockville, MD. http://www.ahrq.gov/professionals/education/curriculum-tools/population-health/adler.html

Averill, R et al. "Development and Evaluation of Clinical Risk Groups (CRGs)." 3M HIS Research Report 9-99, 1999. Accessed March 6, 2017. http://solutions.3m.com/3MContentRetrievalAPI/BlobServlet?lmd=1225920653000&assetId=1180606514454&assetType=MMM_Image&blobAttribute=ImageFile

Data USA: Iowa. Accessed March 6, 2017. https://datausa.io/profile/geo/iowa/#health

Economic Research Service (ERS), U.S. Department of Agriculture (USDA). Food Access Research Atlas, Accessed November 9, 2016 at https://www.ers.usda.gov/data-products/food-access-research-atlas/

Economic Research Service (ERS), U.S. Department of Agriculture (USDA). Food Access Research Atlas, Updated January 17, 2017. https://www.ers.usda.gov/data-products/food-access-research-atlas/

Goldberg, R. "PROC FACTOR: How to Interpret the Output of a Real-World Example." Proceedings of the 22nd International SAS Users Group International Conference. San Diego, CA, March 16-19, 1997.

Healthy People 2020 [Internet]. Washington, DC: U.S. Department of Health and Human Services, Office of Disease Prevention and Health Promotion [cited February 28, 2017]. Available from: http://www.healthypeople.gov/2020/topics-objectives/topic/social-determinants-health

ICD-10 Data.com. "Factors influencing health status and contact with health services Z00-Z99." Accessed November 15, 2016. http://www.icd10data.com/ICD10CM/Codes/Z00-Z99

Krieger N, Chen JT, Waterman PD, Soobader MJ, Subramanian SV, Carson R. Choosing area based socioeconomic measures to monitor social inequalities in low birth weight and childhood lead poisoning: The Public Health Disparities Geocoding Project (US). *Journal of Epidemiology & Community Health.* 2003 Mar; 57(3): 186-199

Krieger N, Williams DR, Moss NE. Measuring Social Class in US Public Health Research: Concepts, Methodologies, and Guidelines. *Annual Reviews in Public Health* vol. 10 no. 16 (March 2007): 341-78.

National Quality Forum. Risk Adjustment for Socioeconomic Status or Other Sociodemographic Factors: Technical Report. August 15, 2014. Accessed June 9, 2016. http://www.qualityforum.org/Risk_Adjustment_SES.aspx

O'Rourke, N. and Hatcher, L. A Step-by-Step Approach to Using SAS for Factor Analysis and Structural Equation Modeling. Cary, NC: SAS Institute, 2013.

SAS Institute. 2016a. "Introduction to Structural Equation Modeling with Latent Variables." STAT® 14.2 User's Guide. Cary, NC. Accessed March 10, 2017. http://documentation.sas.com/?docsetId=statug&docsetVersion=14.2&docsetTarget=statug_introcalis_sect001.htm&locale=en

SAS Institute. 2016b. "The PRINCOMP Procdure." STAT® 14.2 User's Guide. Cary, NC. Accessed March 10, 2017. http://documentation.sas.com/?docsetId=statug&docsetVersion=14.2&docsetTarget=statug_princomp_toc.htm&locale=en

Taylor L, Coyle, C, Ndumele, C, Rogan, E, Canavan, M, Curry, L, and Bradley, E. "Leveraging the Social Determinants of Health: What Works?" Global Social Service Workforce Alliance. 1 June 2015. Web. 2 Dec. 2015. http://www.socialserviceworkforce.org/resources/leveraging-social-determinants-health-what-works

United States Census Bureau. American Community Survey Information Guide, Issued April 2013. Accessed February 28, 2017 at https://www.census.gov/content/dam/Census/programs-surveys/acs/about/ACS_Information_Guide.pdf

United States Census Bureau. American FactFinder. *2011 – 2015 American Community Survey*. U.S. Census Bureau's American Community Survey Office, 2015. Web. 1 January 2017 http://factfinder2.census.gov

Winkleby, M., Jatulis, D., Frank, E., Fortmann, S. (1992). Socioeconomic Status and Health: How Education, Income, and Occupation Contribute to Risk Factors for Cardiovascular Disease. *AJPH,* 82(6): 816-820.

ACKNOWLEDGMENTS

The authors acknowledge Melissa Gottschalk for her work in literature review, data collection, map generation, and editing. We could not have met our deadline without her help!

RECOMMENDED READING

- *STAT® 14.2 User's Guide*
- *Healthy People 2020*

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Paul A, LaBrec
3M Health Information Systems, Inc.
518.426.4315
plabrec@mmm.com
www.3m.com