An Analysis of Independent Causes Leading to Medical Conditions Treatable by Implantable Medical Devices

by

John T Giles

Dr. Piotr Szczurek, Advisor

Masters project submitted in partial fulfillment of the

requirements for the Master of Science in Data Science degree

in the College of Arts and Sciences of

Lewis University

***Abstract*—This article presents graphical and statistical results produced during the discovery and analysis of data collected from various sources. The datasets are related to the causes of medical conditions which may lead to a need for specific implantable medical devices. Data collected on a yearly basis by the Center for Disease Control (CDC) provides the basis for the independent medical condition causes. A publicly available website containing information about implantable medical devices offered by the author's current employer is the source of the data about the conditions leading to a possible need for the devices. The geographic and demographic data comes from the US Census Bureau. The paper strives to determine the human behaviors and ailments suspected of causing health conditions known to be improved or eliminated by specific implantable medical devices. The primary goal is to geographically locate the human behaviors and ailments in an effort to suggest opportunities for companies operating or interested in operating in the implantable medical device spaces.**

Keywords—Medical device; Implantable medical device, Human behavior, Medical condition, Geographic location data, Demographic data;

I. Introduction

There are medical device companies who manufacture products designed to improve the human condition by alleviating or eliminating certain medical conditions. One of these companies is W.L. Gore & Associates (Gore). Gore employ's the author. While at Gore the author has thought about methods for finding individuals who may benefit from Gore's medical devices. This project was chosen as means to determine if this is possible.

Companies are always looking for the physical location of individuals who may benefit from the company's product lines or products. It is also important to understand what motivates individuals to purchase a product. A person's health is a great motivator. The information gathered from this project may be useful for finding concentrations of individual's whose biological features, genetics, and behaviors may benefit from implantable medical devices at some point in the future. An alternative use for this information is to educate people in these geographic areas about how some behavior changes may improve their health and decrease their chances developing medical conditions treatable by implantable medical devices.

II. Project Idea

The project idea is to discover through publicly available datasets the geographic location for high concentrations of individuals who have biological and genetic features, and exhibit behaviors known to influence medical conditions treatable by certain implantable medical devices. Medical devices designed to alleviate or eliminate those medical conditions. A statistical discovery of the data is to be conducted. The discovery process is focused on determining the best subset of factors to use for predicting the number of individuals who may develop the treatable medical conditions.

This information would be useful to any implantable medical device company such as Gore for marketing and product development. One marketing use for this information is to focus a medical device image management campaign on the medical personnel who have the skill to implant the medical device in the geographic areas discovered. The information could also be used for social media marketing campaigns targeted at the geographic areas discovered.

A product development use could be discovering which medical conditions have the highest potential benefit based on the highest potential number of individuals who may develop the condition. The medical conditions with the highest potential number of uses would be a good place to invest the largest amount of Research and Development (R&D) dollars.

III. Objectives

A. Primary Objective

The primary objective is to provide a graphical representation of the geographic location of groups of individuals whose biological and genetic features along with their behavior are suspected of or known to cause medical conditions which may lead to a need for implantable medical devices such as those manufactured by Gore.

B. Secondary Objective

The secondary objective is to provide a visual representation of the underlying analytical and statistical support for the primary objective. An exploratory analysis will be conducted to discover correlations between the variables.

IV. Research methodology

The research methodology was to define the datasets necessary to meet the primary objective. Datasets were collected or created based on the defined dataset definitions. A physical examination of the defined datasets was conducted. The physical examination provided the ability to determine the geographic layout of the main dataset. A statistical analysis of the main dataset was performed to further examine the data and discover information not uncovered during the physical examination of the datasets. Information discovered during the statistical examination of the data were applied to the geographic layout to assist with meeting the secondary objective.

A. Dataset Definintion

The data requirements for this project are loosely defined. It was determined three datasets were required. A medical history dataset, geographic dataset, and a medical device to medical condition dataset. One dataset needs to contain a large sampling of accurate medical history data from individuals in the United States (US). Each individual providing their medical history data must freely consent to give the information. The medical history data cannot be traceable to any individual in any way. However, it must be possible to locate the geographic area from where the data was gathered for each individual. Their exact location, such as address, is not necessary. The city, county, state, and other defined geographic areas are sufficient. The dataset must also be publicly available and free to use for data analysis.

The geographic dataset must come from a reputable source. It must be capable of being joined to the medical history dataset on predefined key values. Each record from the medical history dataset should be easily located on a common digital geographic map of the US. The geographic dataset must also be publicly available and free to use for data analysis.

A third dataset is necessary to map the implantable medical devices to medical conditions. This dataset also needs to map to responses from the medical history dataset. The data for this dataset must be publicly available and free to use as well.

B. Data Collection

An ideal long-term solution for this project is to provide real-time data analysis of the factors suspected of causing ailments known to be precursors to the defined medical conditions. The method for collecting real-time data directly from the population is typically social media. Data found on social media is extremely random and often cryptic. An understanding of the factors which influence the conditions is useful for deciphering the random and cryptic data from social media. A starting point for understanding the factors suspected of causing the medical conditions is to look to experts in the human behavioral risk arena.

The CDC has performed behavioral risk analysis since 1984. Their data, analysis results, and analysis tools are currently made available to the public through the online Behavioral Risk Factor Surveillance System (BRFSS) [2]. The BRFSS uses questionnaires to collect medical and behavioral information about individuals throughout the United States. The CDC works with US States and some US Territories to develop the questionnaires. Questionnaires are administered over the phone by personnel from the CDC, US States and Territories [3].

The main dataset created each year by the CDC is the Combined Landline Telephone and Cellular Telephone (LLCP) dataset. The data is collected, analyzed, enhanced, and published within a years' time. The data for 2016 is the latest available and was published in late 2017 [25]. The enhancements made to the data are based on statistics the CDC discovers during their analysis. The enhancements are called calculated fields and are developed from responses of multiple questions [6]. Statistics collected include information such as the frequency, percentage, and weighted percentage of each response for each question from each individual.

The Selected Metropolitan/Micropolitan Area Risk Trends (SMART) dataset is a subset of the 2016 LLCP dataset. It is produced to provide some local area estimates. The SMART dataset improves upon the LLCP dataset by including more complex calculated fields, including a weighted value for each response for each individual estimating the number of individuals they represent in the full population. The weight value is calculated using the BRFSS raking methodology defined in [19] and [25].

A few drawbacks exist with the SMART dataset. It does not contain all the fields from the LLCP dataset and has approximately half the number of records. The SMART dataset was used for this project based on the availability of the extra calculated fields.

The CDC suggests using geographic delineation data files from the US Census Bureau. The delineation data files contain geographic data but no demographic data, such as population. The population data is necessary to add meaning to the SMART data by providing a visual sense of proportion. Fortunately, the Census Bureau provides demographic datasets compatible with the CDC datasets [23].

The medical devices manufactured by Gore are designed to help treat specific medical conditions. The probability of an individual developing one of these conditions is influenced by biological and genetic features, and human behaviors. Gore makes available an online list of implantable medical devices [7] they offer. They also provide the primary medical conditions [8] leading to the use of their medical devices.

Gore began manufacturing medical devices over 40 years ago. They offer a wide range of implantable medical devices to treat a variety of conditions. These conditions include: Abdominal Aortic Aneurysm (AAA), Abdominal Wall Reconstruction, Atrial Septal Defects (ASD), Biliary Disease, Carotid Artery Stenosis, Colon Disease, Dialysis, Heart Defect, Hernia, Peripheral Disease (PAD) (PVD), Portal Hypertension and Liver Disease, Thoracic Aortic Aneurysm (TAA), and Weight Loss Surgery [8].

The causes to the medical conditions treated by Gore's implantable medical devices were researched through publicly available websites. The method for determining the validity of the medical condition was an intuitive discovery process. All the information gathered about the known or suspected causes of the medical conditions came from Stanford Healthcare and Mayo Clinic websites.

Stanford Healthcare is associated with the Stanford academic medical institution. Stanford Healthcare is known to be a leading medical institution whose reputation could be damaged by providing inaccurate information about medical conditions. They have incentive to provide accurate information. The Mayo Clinic is also known to be a leading medical institution whose reputation could be damaged by providing inaccurate information about medical conditions. The author regards the information from these institutions as reliable.

C. Dataset Examination

The Metropolitan Micropolitan Statistical Area (MMSA) SMART dataset for 2016 contains 249011 records with 147 fields and each record being 325 Bytes wide. The Census Bureau Census Statistical Area (CSA) 2017 estimate dataset contains 1896 records with 80 fields. The MMSA key field within the SMART dataset maps directly to the Core Based Statistical Area (CBSA) [24] key field within US Census Bureau datasets. The SMART MMSA key field also maps directly to the CBSA key field in the Census Bureau dataset.

There are three classification fields within the SMART dataset: MMSA key field, MMSA weight field, and MMSA name field. The MMSA key field is described above. The MMSA weight field contains the estimated number of individuals within the general population represented by the individual respondent answering the question. The weight is based on a formula designed by the CDC called raking. The MMSA name field contains the primary cities and the 2-character US state or territory abbreviation in which the respondent resides.

The CDC raking methodology is described in their BRFSS overview document [3] in the section titled Weighting the Data. At a high level, "the design weight considers the probability of selection, adjusts for nonresponse bias and non-coverage errors. Design weights are calculated using the weight of each geographic stratum (_STRWT), the number of phones within a household (NUMPHON2), and the number of adults aged 18 years and older in the respondent's household (NUMADULT). For cellphone respondents, both NUMPHON2 and NUMADULT are set to 1. The formula for the design weight is

Design Weight = _STRWT * (1/NUMPHON2) * NUMADULT". [3] There is formula for the geographic stratum weight as well. There are "16 raking margins used in the following order—county by gender, county by age, county by race or ethnicity, county, region by race or ethnicity, region by gender, region by age, region, telephone service (landline, cellular telephone or dual user), age by race or ethnicity, gender by race or ethnicity, tenure (rent or own), marital status, education, race or ethnicity, and gender by age." [25]

Several dataset were created of the Gore medical device conditions mapped to possible causes of those conditions. A final dataset was created from the many iterations mapping each medical condition to its own set of causes and the closest matching SMART dataset field name. There are 14 medical conditions mapped to 183 causes. 112 of the causes map to 18 unique SMART fields. The remaining 71 causes have no apparent equivalent in the SMART dataset. There are medical condition causes duplicated across multiple medical conditions. The significance of the duplicates will not be addressed in this project. Appendix A shows the entire table.

The CDC and Census Bureau datasets were loaded into Tableau and Hadoop. Hadoop was used to perform basic data analysis of the data, mainly record counts and different ways of grouping the data. Tableau was used to join the datasets together for creating the geographic and initial statistical analysis of the data.

D. Examine dataset geographic layout

Tableau was used to create the visual representation for the geographic layout of the SMART dataset. The MMSA weight field provides the respondent count for MMSA per filter application. The SMART and Census Bureau datasets contain estimated data for multiple geographic area designators. In the Census Bureau dataset there the following geographic area designators: CENSUS2010POP, County or equivalent, Metropolitan Division, Metropolitan Statistical Area, and Micropolitan Statistical Area. In the SMART dataset the following geographic area designators exist: Metropolitan Division, Metropolitan Statistical Area, and Micropolitan Statistical Area. The geographic layout of data will use the Metropolitan Statistical Area.

A view of all respondents in a geographic layout helps provide a midlevel perspective of all respondents for each Metropolitan Statistical Area as seen in Fig. 1. A test of the weighted respondent count in each Metropolitan Statistical Area in relation to the estimated population for the same area uncovered an apparent error in the weighted respondent count. The weighted respondent count should equal the population for the area, and it does not. The range for the error is between 17% and 34%, which is significant. An example of the error is found in Table 1. The weight % of Population row should at or near 100%. A value lower than 95% indicates an error.

Table 1
Example of Weight % Error

Name	Riverside-San Bernardino-Ontario, CA
Weight % of Population	0.743376647
Population. Estimate 2016	4,523,653
Weight Count	3,362,778.00
Number of Respondents	1,119

The cause of the error has not yet been identified. Possible causes for the error include wrong interpretation was made about the value contained in the MMSA weight field. Another possible cause for the error may be value placed in the MMSA weight field may have used different population estimates in the equation than used in this project.

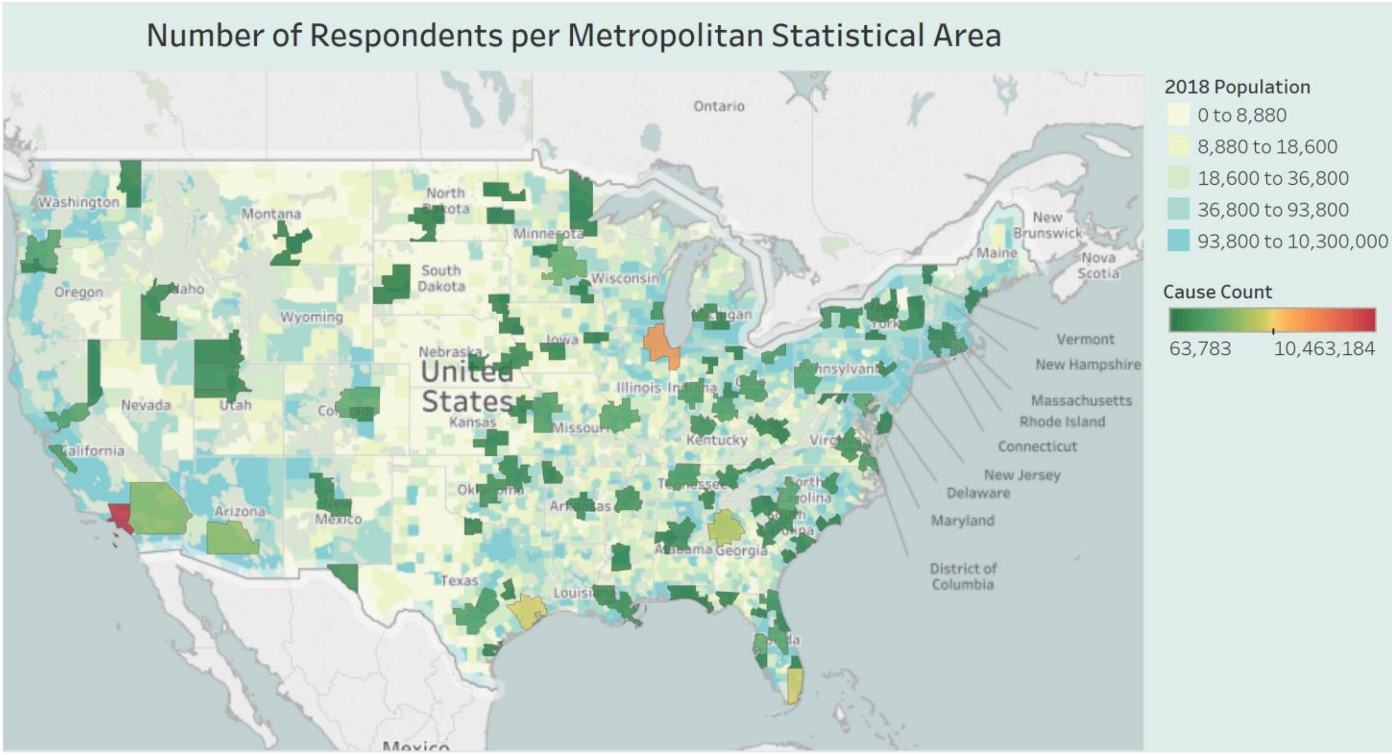


Fig. 1. Number of Respondents per Metropolitan Statistical Area. The legend on the right shows the 2018 population estimate provided by Tableau. The cause count heat map bar provides a means to visually locate large numbers of respondents. The redder an area is, the higher number of respondents. The greener an area is, the lower number of respondents. Note the data used for the heat map used 2016 population estimates.

Drilling down into each Metropolitan Statistical Area in Tableau shows how filtering on a combination of factors highlights the different opportunities available to medical device companies and medical personnel. Each medical condition and related causes produced different results, as anticipated. A graphic representation of each medical condition is found in the Appendix B-L.

E. Statistically analyze datasets

The physical examination of the data and the visual geographic exploration of the data is not enough to fully understand how the data interacts with each other. The best combination of medical condition causes is exposed through statistical analysis of the data. The CDC SMART dataset was loaded into SAS for the bulk of the statistical analysis. SAS Enterprise Guide version 7.15 HF2 (7.100.5.6112) (64-bit) ODA was used. SAS is a very powerful statistics application. The student version was the only version available to the author. The results provided by the tool are assumed to be as accurate as the full version. There were issues with dataset longevity in the application. It was necessary to continually reload the data after a short amount of time. The initial dataset never changed. Reloading the data required subordinate tasks to be rerun on a regular basis, including random samples.

Attempts were made to perform tasks on the entire dataset. These attempts were abandon after one job was still running after 12 hours. Random samples of 50%, 10% and 5% were generated to improve the processing time. A random sample of 5% of the full dataset provided enough useful information for this project. Factor analysis, Principal Component analysis, Correlation analysis, and Cluster analysis was performed on the 18 SMART variables associated with medical conditions found in Appendix A. The author chose to highlight the Factor analysis process and results.

Table 2 shows there were 12451 records in the random sample with only 10671 records actually used. 1780 records were omitted due to missing values. Table 3 shows the Factor analysis correlation matrix between the 18 SMART variables. Mitochondrial heart disease (c_michd) and cardiovascular infarction (cvdinfr4) have a high correlation. Alcohol use (drnkany5) and alcohol abuse (c_rfdrhv5) have an average correlation. All other variables have unacceptable correlation. Table 4 shows the partial correlations. Mitochondrial heart disease and cardiovascular infarction still have a middling correlation. Cardiovascular heart disease (cvdchrhd) and mitochondrial heart disease move into a miserable correlation range. Alcohol use and alcohol abuse correlation value went down slightly. Table 5 shows an adequate Measure of Sampling Adequacy of 0.54792545. The correlation coefficient values are low except for those few already mentioned.

Table 2
Factor Analysis Data Sample Numbers

Input Data Type	Raw Data
Number of Records Read	12451
Number of Records Used	10671
N for Significance Tests	10671

Table 6 shows the unrotated factor pattern matrix. There are variables with high loadings in more than one factor. These are candidates for extraction. Table 7 shows the orthogonally rotated factor pattern matrix. The fewer variable with high loadings in more than one factor. Tables 8 and 9 shows the first seven variables with Eigen values greater than one. These seven variables are candidates for extraction. Fig. 3 and 4 show the graphical representation of Tables 8 and 9.

The recommendation is to extract the following variables: c_ageg5yr, chcocncr, havarth3, c_smoker3, usenow3, and c_totinda. The factor analysis was executed again with orthogonal rotation. Looking only at the rotated factor matrix shows one more variable candidate for extraction, c_prace1. The third factor analysis execution with orthogonal rotation produced a good set of factors with the remaining 11 variables.

F. Apply statistical discovery to the geographic results

The 18 SMART variables provide geographic results for 13 of the 14 medical conditions. Using all 18 variables should produce individuals who are highly likely to develop all 13 medical conditions. This should be a very short list. Tableau was used to apply all 18 variables to discover if there were any individuals who might develop all 13 medical conditions. Fig. 4 shows the results. There is one individual in the Phoenix-Mesa-Scottsdale, AZ Metropolitan Statistical Area who meets the criteria. Applying the Factor analysis variable extraction results should not change the Tableau filter results. Fig. 5 shows the exact same results as Fig. 4. This is evidence the Factor analysis produced valid results.

Table 3

Correlations

	c_ageg5yr	c_bmi5cat	c_impsex	c_michd	c_prace1	c_rfdrhv5	c_smoker3	c_totinda	chckidny	chcocncr	cvdcrhd4	cvdinfr4	cvdstrk3	diabete3	drnkany5	havarth3	hivrisk4	usenow3
c_ageg5yr	1.0000																	
c_bmi5cat	0.0415	1.0000																
c_impsex	0.0824	-0.0896	1.0000															
c_michd	-0.2331	-0.0538	0.0672	1.0000														
c_prace1	-0.0227	0.0006	-0.0148	0.0018	1.0000													
c_rfdrhv5	-0.0315	-0.0326	-0.0168	0.0015	0.0098	1.0000												
c_smoker3	0.0170	0.0090	0.0576	0.0538	0.0131	-0.0279	1.0000											
c_totinda	0.1215	0.1022	0.0401	-0.1053	0.0297	0.0117	-0.0859	1.0000										
chckidny	-0.0204	-0.0276	-0.0135	0.0245	-0.0129	0.0238	0.0032	-0.0155	1.0000									
chcocncr	-0.1797	0.0005	-0.0346	0.0504	-0.0001	0.0296	0.0033	-0.0412	0.0027	1.0000								
cvdcrhd4	-0.1098	-0.0240	0.0381	0.4734	-0.0057	-0.0002	0.0174	-0.0210	0.0360	0.0292	1.0000							
cvdinfr4	-0.1616	-0.0444	0.0602	0.7046	0.0055	-0.0052	0.0550	-0.0747	0.0033	0.0293	0.1459	1.0000						
cvdstrk3	-0.0520	-0.0178	-0.0206	0.0453	0.0114	0.0171	0.0123	-0.0475	0.0311	0.0231	0.0233	0.0499	1.0000					
diabete3	-0.1860	-0.1899	0.0186	0.1779	-0.0038	0.0212	0.0103	-0.1132	0.0612	0.0446	0.0869	0.1332	0.0345	1.0000				
drnkany5	0.0759	0.0169	0.0654	-0.0534	0.0219	0.5523	0.0420	0.0953	0.0026	0.0090	-0.0237	-0.0413	-0.0114	-0.0722	1.0000			
havarth3	-0.3084	-0.1099	-0.0942	0.1463	0.0281	0.0188	0.0328	-0.1030	0.0715	0.1047	0.0719	0.0978	0.0445	0.1220	-0.0625	1.0000		
hivrisk4	0.0909	-0.0016	0.0304	-0.0273	0.0174	0.1292	0.1070	0.0187	0.0124	-0.0156	-0.0163	-0.0174	0.0039	-0.0132	0.1988	-0.0191	1.0000	
usenow3	0.0517	-0.0274	0.0766	-0.0120	0.0309	0.0832	0.1557	-0.0005	0.0024	-0.0120	-0.0009	-0.0141	-0.0053	0.0006	0.1164	-0.0143	0.2066	1.0000

Table 4

Partial Correlations Controlling all other Variables

	c_ageg5yr	c_bmi5cat	c_impsex	c_michd	c_prace1	c_rfdrhv5	c_smoker3	c_totinda	chckidny	chcocncr	cvdcrhd4	cvdinfr4	cvdstrk3	diabete3	drnkany5	havarth3	hivrisk4	usenow
_ageg5yr	1.0000																	
_bmi5cat	-0.0202	1.0000																
_impsex	0.0585	-0.0994	1.0000															
_michd	-0.1090	0.0047	0.0422	1.0000														
_prace1	-0.0213	-0.0015	-0.0162	-0.0004	1.0000													
_rfdrhv5	-0.0539	-0.0343	-0.0582	0.0128	-0.0055	1.0000												
_smoker3	0.0253	0.0289	0.0437	0.0158	0.0097	-0.0696	1.0000											
_totinda	0.0610	0.0810	0.0401	-0.0556	0.0336	-0.0319	-0.0903	1.0000										
hckidny	0.0111	-0.0103	-0.0102	0.0050	-0.0148	0.0174	0.0004	-0.0017	1.0000									
hcocncr	-0.1454	0.0151	-0.0157	0.0049	-0.0056	0.0105	0.0039	-0.0171	-0.0062	1.0000								
vdcrhd4	0.0005	-0.0012	0.0118	0.5193	-0.0068	-0.0039	-0.0002	0.0353	0.0225	0.0047	1.0000							
vdinfr4	0.0011	-0.0074	0.0188	0.7152	0.0036	-0.0055	0.0234	0.0128	-0.0126	-0.0062	-0.3003	1.0000						
vdstrk3	-0.0245	-0.0066	-0.0162	-0.0042	0.0113	0.0189	0.0086	-0.0335	0.0264	0.0117	0.0111	0.0283	1.0000					
iabete3	-0.1195	-0.1655	0.0182	0.0651	-0.0047	0.0455	0.0028	-0.0586	0.0487	0.0101	0.0077	0.0130	0.0125	1.0000				
rnkany5	0.0470	0.0188	0.0739	-0.0165	0.0170	0.5507	0.0538	0.0828	-0.0020	0.0184	-0.0001	-0.0024	-0.0126	-0.0668	1.0000			
avarth3	-0.2493	-0.0930	-0.0839	0.0461	0.0245	0.0249	0.0380	-0.0402	0.0609	0.0508	0.0022	-0.0057	0.0185	0.0324	-0.0404	1.0000		
ivrisk4	0.0698	0.0008	-0.0006	-0.0041	0.0092	0.0300	0.0751	0.0030	0.0113	-0.0044	-0.0043	0.0016	0.0085	0.0095	0.1280	0.0116	1.0000	
senow3	0.0235	-0.0236	0.0586	-0.0016	0.0274	0.0325	0.1347	-0.0025	-0.0004	-0.0045	0.0038	-0.0113	-0.0054	0.0060	0.0415	-0.0018	0.1717	1.0000

Table 5

Kaiser's Measure of Sampling Adequacy

Overall MSA = 0.54792545

c_ageg5yr	c_bmi5cat	c_impsex	c_michd	c_prace1	c_rfdrhv5	c_smoker3	c_totinda	chckidny	chcocncr	cvdcrhd4	cvdinfr4	cvdstrk3	diabete3	drnkany5	havarth3	hivrisk4	usenow3
0.6984	0.5780	0.5671	0.5160	0.5353	0.5094	0.5564	0.6907	0.6264	0.6815	0.4330	0.4945	0.7599	0.7366	0.5323	0.6835	0.6791	0.6374

Table 6

Unrotated Factor Pattern

	Factor1	Factor2	Factor3	Factor4	Factor5	Factor6	Factor7
c_michd	0.8210	0.1777	0.3575	-0.1787	0.0401	0.0238	-0.0056
cvdinfr4	0.6912	0.1563	0.3498	-0.1404	0.0435	-0.0192	-0.0215
cvdcrhd4	0.4995	0.1252	0.2773	-0.1879	0.0126	0.1219	0.0508
diabete3	0.4156	0.0326	-0.2198	0.2636	-0.3676	-0.0535	0.0740
c_ageg5yr	-0.5245	0.0512	0.4502	0.0926	-0.1482	0.1558	0.0175
drnkany5	-0.2211	0.7726	-0.1411	-0.2709	-0.0937	-0.0214	-0.0605
c_rfdrhv5	-0.0802	0.7218	-0.3126	-0.3125	-0.1670	0.0060	-0.0739
hivrisk4	-0.1329	0.4996	0.0533	0.2808	0.2102	0.0927	0.0297
c_impsex	0.0146	0.1901	0.4056	0.2280	-0.3654	-0.2358	-0.0024
chcocncr	0.1903	0.0218	-0.3653	-0.0971	0.2609	-0.2740	-0.2359
havarth3	0.4160	-0.0156	-0.4863	0.0970	0.1484	-0.0204	0.1058
c_smoker3	0.0719	0.2203	0.1514	0.5026	0.4179	0.0217	-0.1501
usenow3	-0.0753	0.4213	0.1062	0.4584	0.2427	-0.0516	0.0563
c_totinda	-0.2854	0.0583	0.2228	-0.3703	-0.0157	-0.0578	0.3310
c_bmi5cat	-0.2065	-0.0769	0.1814	-0.3978	0.5934	0.1714	-0.1127
chckidny	0.0903	0.0486	-0.1577	0.0882	-0.1371	0.7028	0.2388
cvdstrk3	0.1312	0.0276	-0.1291	0.0481	0.0835	0.4608	0.1400
c_prace1	0.0051	0.0635	-0.0525	0.0042	0.2310	-0.3104	0.8365

Table 7

Orthogonally Rotated Factor Pattern

	Factor1	Factor2	Factor3	Factor4	Factor5	Factor6	Factor7
c_michd	0.8210	0.1777	0.3575	-0.1787	0.0401	0.0238	-0.0056
cvdinfr4	0.6912	0.1563	0.3498	-0.1404	0.0435	-0.0192	-0.0215
cvdcrhd4	0.4995	0.1252	0.2773	-0.1879	0.0126	0.1219	0.0508
diabete3	0.4156	0.0326	-0.2198	0.2636	-0.3676	-0.0535	0.0740
c_ageg5yr	-0.5245	0.0512	0.4502	0.0926	-0.1482	0.1558	0.0175
drnkany5	-0.2211	0.7726	-0.1411	-0.2709	-0.0937	-0.0214	-0.0605
c_rfdrhv5	-0.0802	0.7218	-0.3126	-0.3125	-0.1670	0.0060	-0.0739
hivrisk4	-0.1329	0.4996	0.0533	0.2808	0.2102	0.0927	0.0297
c_impsex	0.0146	0.1901	0.4056	0.2280	-0.3654	-0.2358	-0.0024
chcocncr	0.1903	0.0218	-0.3653	-0.0971	0.2609	-0.2740	-0.2359
havarth3	0.4160	-0.0156	-0.4863	0.0970	0.1484	-0.0204	0.1058
c_smoker3	0.0719	0.2203	0.1514	0.5026	0.4179	0.0217	-0.1501
usenow3	-0.0753	0.4213	0.1062	0.4584	0.2427	-0.0516	0.0563
c_totinda	-0.2854	0.0583	0.2228	-0.3703	-0.0157	-0.0578	0.3310
c_bmi5cat	-0.2065	-0.0769	0.1814	-0.3978	0.5934	0.1714	-0.1127
chckidny	0.0903	0.0486	-0.1577	0.0882	-0.1371	0.7028	0.2388
cvdstrk3	0.1312	0.0276	-0.1291	0.0481	0.0835	0.4608	0.1400
c_prace1	0.0051	0.0635	-0.0525	0.0042	0.2310	-0.3104	0.8365

Table 9

Unrotated Eigenvalues of the Correlation Matrix: Total = 18 Average = 1

	Eigenvalue	Difference	Proportion	Cumulative
1	2.2918	0.5696	0.1273	0.1273
2	1.7222	0.3282	0.0957	0.2230
3	1.3940	0.1401	0.0774	0.3004
4	1.2539	0.1236	0.0697	0.3701
5	1.1303	0.1088	0.0628	0.4329
6	1.0215	0.0114	0.0567	0.4896
7	1.0101	0.0229	0.0561	0.5458
8	0.9872	0.0328	0.0548	0.6006
9	0.9544	0.0601	0.0530	0.6536
10	0.8943	0.0393	0.0497	0.7033
11	0.8550	0.0047	0.0475	0.7508
12	0.8503	0.0375	0.0472	0.7980
13	0.8128	0.0250	0.0452	0.8432
14	0.7878	0.0182	0.0438	0.8870
15	0.7696	0.1350	0.0428	0.9297
16	0.6346	0.2148	0.0353	0.9650
17	0.4198	0.2091	0.0233	0.9883
18	0.2107		0.0117	1.0000

Table 8

Orthogonally Rotated Eigenvalues of the Correlation Matrix: Total= 18 Average = 1

	Eigenvalue	Difference	Proportion	Cumulative
1	2.2918	0.5696	0.1273	0.1273
2	1.7222	0.3282	0.0957	0.2230
3	1.3940	0.1401	0.0774	0.3004
4	1.2539	0.1236	0.0697	0.3701
5	1.1303	0.1088	0.0628	0.4329
6	1.0215	0.0114	0.0567	0.4896
7	1.0101	0.0229	0.0561	0.5458
8	0.9872	0.0328	0.0548	0.6006
9	0.9544	0.0601	0.0530	0.6536
10	0.8943	0.0393	0.0497	0.7033
11	0.8550	0.0047	0.0475	0.7508
12	0.8503	0.0375	0.0472	0.7980
13	0.8128	0.0250	0.0452	0.8432
14	0.7878	0.0182	0.0438	0.8870
15	0.7696	0.1350	0.0428	0.9297
16	0.6346	0.2148	0.0353	0.9650
17	0.4198	0.2091	0.0233	0.9883
18	0.2107		0.0117	1.0000

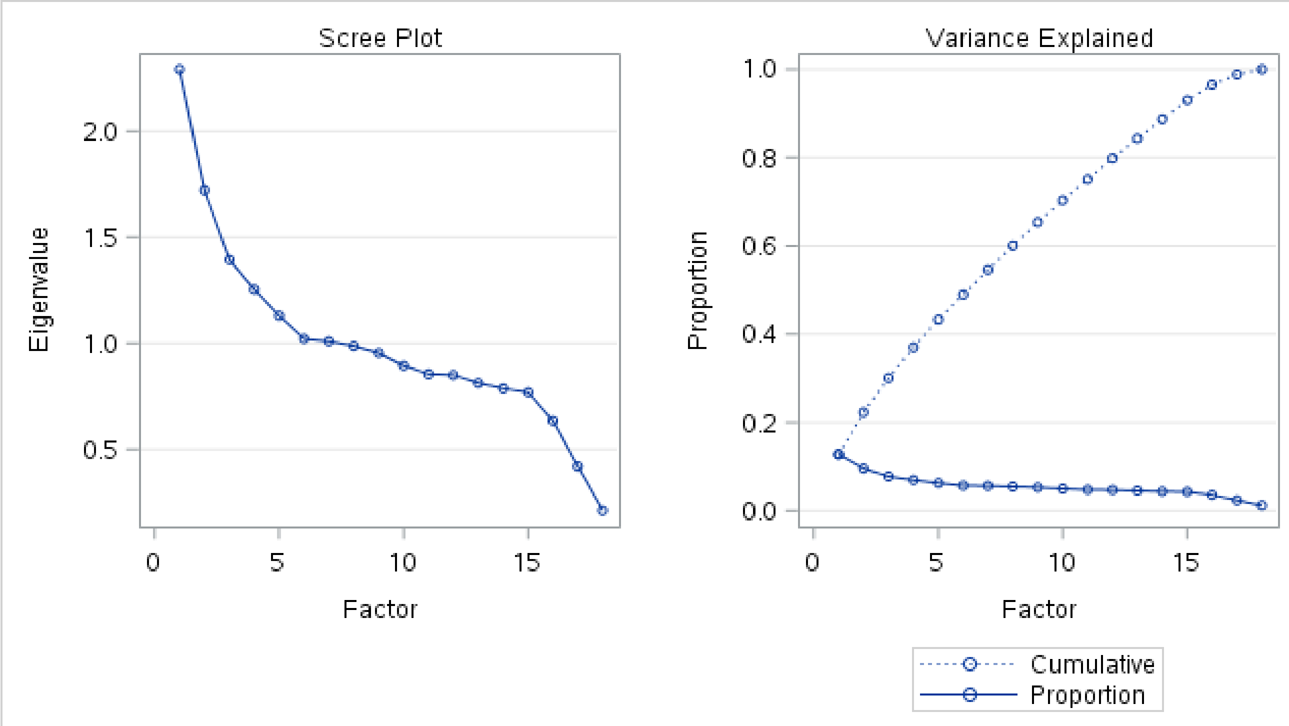


Fig. 2. Unrotated factor analysis scree plot and variance explanation. A graphical representation of Table 8 data.

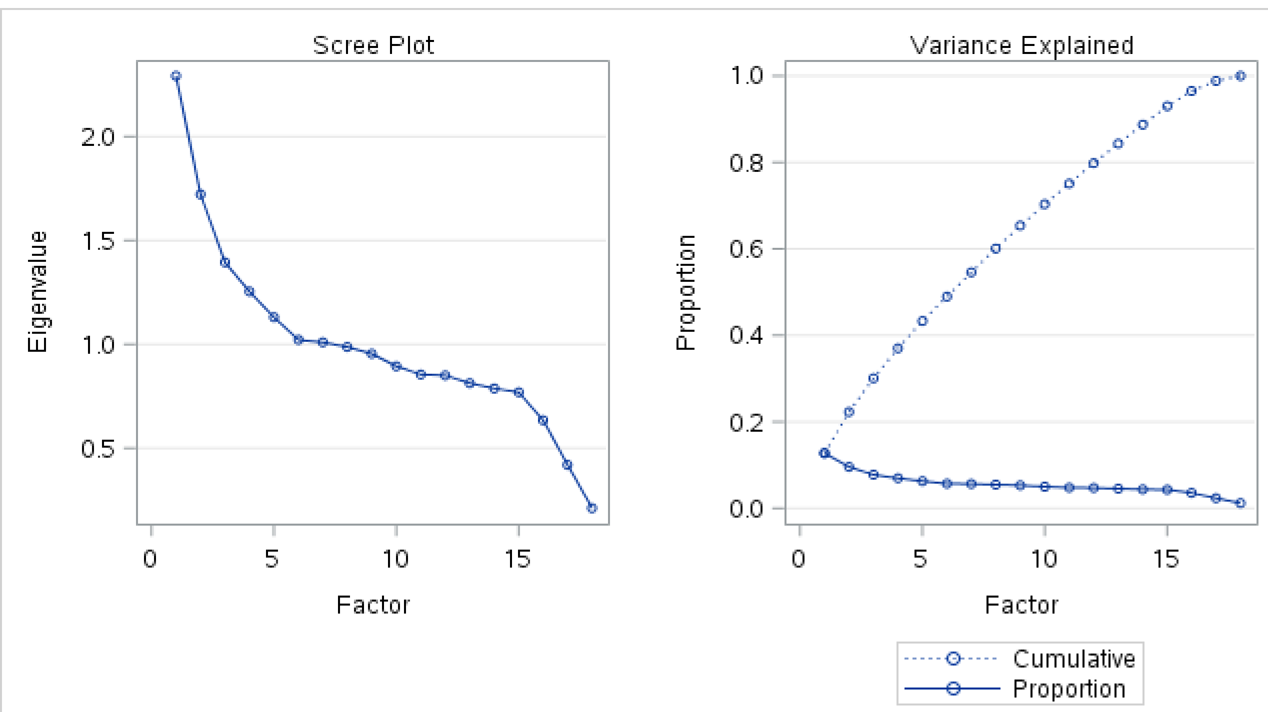


Fig. 3. Rotated factor analysis scree plot and variance explanation. A graphical representation of the Table 9 data.

Table 11

Factor Pattern Second Execution

	Factor1	Factor2	Factor3	Factor4	Factor5
c_michd	0.88476	0.05986	0.20391	-0.08341	-0.02835
cvdinfr4	0.74224	0.05845	0.16053	-0.10228	-0.23194
cvdcrhd4	0.53397	0.04093	0.23742	-0.03561	0.28447
drnkany5	-0.13293	0.84523	0.05157	-0.06226	0.01798
c_rfdrhv5	-0.03524	0.83153	-0.01777	0.09304	-0.01134
hivrisk4	-0.02638	0.46608	-0.00193	-0.10177	-0.02134
c_bmi5cat	-0.19800	-0.03593	0.76569	0.08119	-0.21545
diabete3	0.38053	0.00891	-0.56463	0.35699	-0.04749
c_prace1	-0.00191	0.12485	0.01452	0.34520	0.29646
c_impsex	0.12486	0.05539	-0.29170	-0.73774	0.09712
cvdstrk3	0.22572	0.02049	0.05009	0.35355	0.55065
chckidny	0.17264	0.08080	-0.16540	0.31420	-0.65842

Table 10

Factor Pattern Third Execution

	Factor1	Factor2	Factor3	Factor4	Factor5
c_michd	0.88478	0.05951	0.20421	-0.08640	0.05236
cvdinfr4	0.74226	0.05866	0.16101	-0.09300	0.28116
cvdcrhd4	0.53398	0.04145	0.23769	-0.04600	-0.31880
drnkany5	-0.13261	0.84642	0.05260	-0.05259	-0.01768
c_rfdrhv5	-0.03494	0.83374	-0.01706	0.11829	-0.02506
hivrisk4	-0.02624	0.46933	-0.00043	-0.07091	-0.00321
c_bmi5cat	-0.19801	-0.03625	0.76561	0.11188	0.20179
diabete3	0.38055	0.00671	-0.56619	0.36581	-0.00954
c_impsex	0.12483	0.06000	-0.28861	-0.77009	0.03532
chckidny	0.17266	0.08181	-0.16545	0.41202	0.58766
cvdstrk3	0.22575	0.01778	0.04829	0.32458	-0.66610

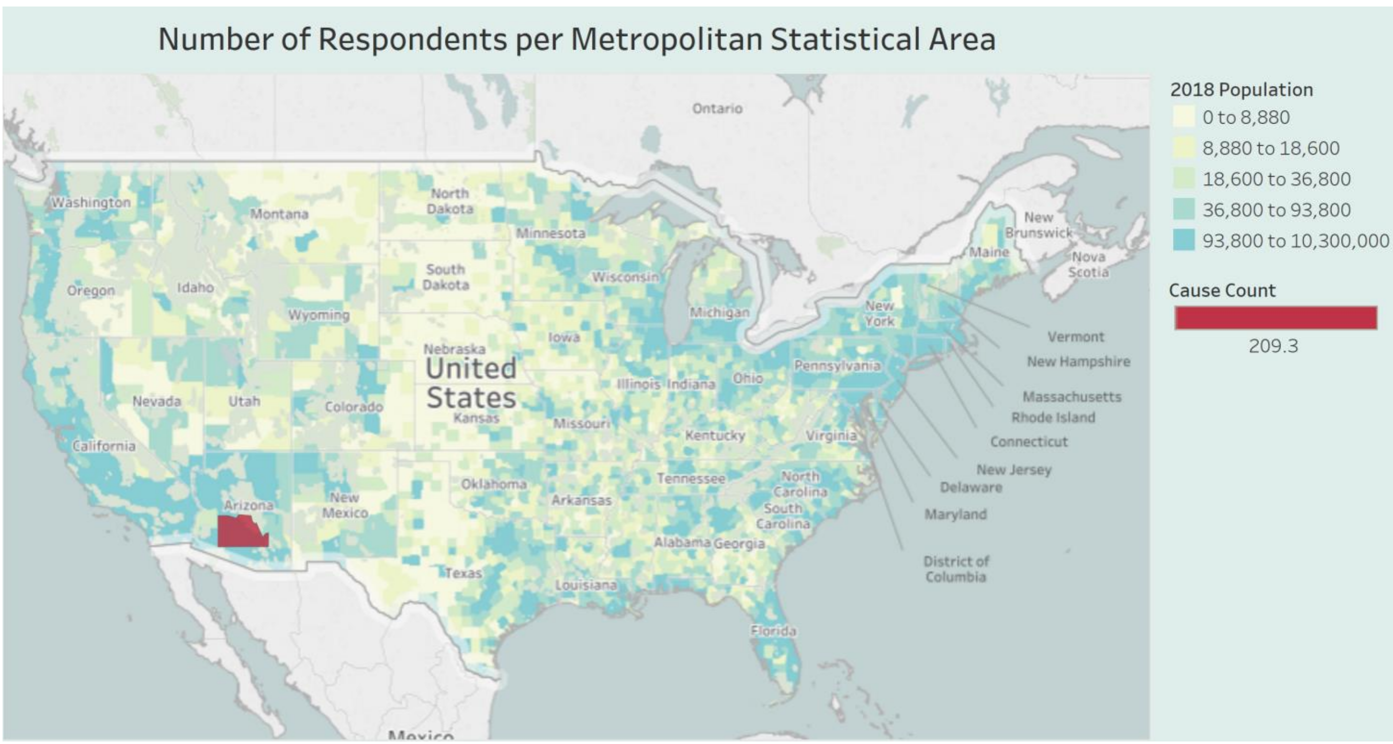


Fig. 4. Number of respondents per Metropolitan Statistical Area who meet all 18 medical condition causes.

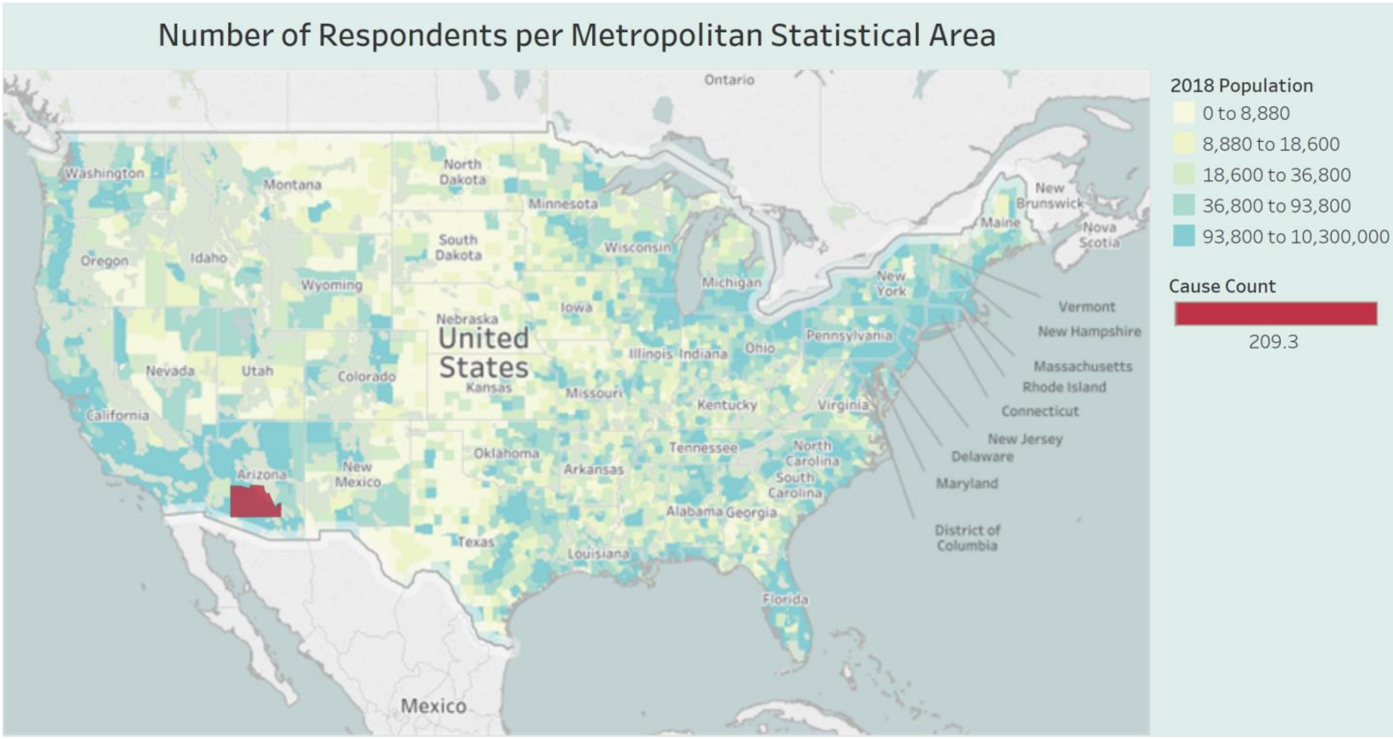


Fig. 5. **Number of respondents per Metropolitan Statistical Area who meet the 11 medical condition causes suggested by the Factor analysis.**

V. Related Work

The author discovered two other projects related to the prediction of medical conditions. Tyler H. McCormick, Cynthia Rudin, and David Madigan developed the Hierarchical Associate Rule Model (HARM). HARM uses medical history records to produce a list of ranked association rules for predicting an individual's next medical condition. The following is synopsis of how it works.

- "Given data from a sequence of past encounters, predict the next condition that a patient will report.
- Given basic demographic information, predict the first condition that a patient will report.
- Given partial data from an encounter (and possibly prior encounters) predict the next condition." [26]

A group of researchers at the University of Nottingham, UK developed a machine learning algorithm to predict heart attacks and strokes based on medical history records. Using a Neural network, the algorithm processed 283.692 medical records. The algorithm focused on hypertension, high cholesterol, age, smoking, and diabetes. It was 72.8% accurate. 4,998 heart attacks were accurately predicted, 335 more predicted heart attacks than real doctors. [27]

VI. Conclusion

It is possible to locate individuals within a geographic area whose biological and genetic features and behaviors coupled together may lead them to develop medical conditions treatable by implantable medical devices using only freely available data. The 18 variables found in the CDC dataset matching medical condition causes can be reduced to 11 and produce the same results.

Others have developed new statistical algorithms for accurately predicting certain types of medical conditions. There projects were based on medical data collected from doctors who got the data from the patients they treat. The data is assumed to be more comprehensive and accurate than the data used for this project.

The next steps for this project would be to use better medical history data. It is assumed better medical history data will more comprehensively match the medical condition causes. A variable correlation process would still need to be performed to eliminate unnecessary variables from the analysis. An additional step would be to track the data over time to show trends.

VII. Appendix

A. Medical Condition to Cause to SMART Field Table

The table contains each medical condition from Gore's website. Some of the condition causes are duplicated across two or more medical conditions. The duplicates across conditions are not an issue. Notice there are blank values in the SMART Field Name per cause per condition column. Not all condition causes were a match or near match to a SMART field. For example, the CRGVPRB2 SMART field is found only the LLCP dataset and not found in the SMART dataset. Also note there are some duplicate SMART Field Name per cause per condition values within some medical conditions. These duplicates are due to multiple CDC responses meeting the matching criteria. The multiple SMART Field values for the multiple CDC responses for a single cause improves the accuracy of the results by ensuring all possible individuals meeting a condition cause criteria are included in the counts. Unfortunately, the unmatched condition causes reduce the accuracy of the results by not including individuals who might meet the criteria from the counts.

Medical Condition	**Condition Cause**	**SMART Field Name per cause per condition**	**SMART Field value**
Abdominal Aortic Aneurysm (AAA)	tobacco use	USENOW3	1
Abdominal Aortic Aneurysm (AAA)	tobacco use	USENOW3	2
Abdominal Aortic Aneurysm (AAA)	atherosclerosis		
Abdominal Aortic Aneurysm (AAA)	hypertension	CRGVPRB2	8
Abdominal Aortic Aneurysm (AAA)	aorta blood vessel disease		
Abdominal Aortic Aneurysm (AAA)	aorta infection		
Abdominal Aortic Aneurysm (AAA)	trauma		
Abdominal Aortic Aneurysm (AAA)	genetics		
Atrial Septal Defects (ASD)	diabetes	CRGVPRB2	7
Atrial Septal Defects (ASD)	drug abuse	CRGVPRB2	12
Atrial Septal Defects (ASD)	rubella infection		
Atrial Septal Defects (ASD)	phenylketonuria (PKU)		
Atrial Septal Defects (ASD)	tobacco use	USENOW3	1
Atrial Septal Defects (ASD)	tobacco use	USENOW3	2
Atrial Septal Defects (ASD)	lupus	HAVARTH3	1
Atrial Septal Defects (ASD)	alcohol use	DRNKANY5	1
Atrial Septal Defects (ASD)	obesity	_BMI5CAT	4
Atrial Septal Defects (ASD)	exposure to certain substances		
Atrial Septal Defects (ASD)	phenylketonuria (PKU)		
Abdominal Wall Reconstruction			
Biliary Disease	female	_IMPSEX	2
Biliary Disease	age 30-60	_AGEG5YR	3
Biliary Disease	age 30-60	_AGEG5YR	4
Biliary Disease	age 30-60	_AGEG5YR	5
Biliary Disease	age 30-60	_AGEG5YR	6
Biliary Disease	age 30-60	_AGEG5YR	7
Biliary Disease	age 30-60	_AGEG5YR	8
Biliary Disease	genetics		
Biliary Disease	Europe		
Biliary Disease	North America		
Biliary Disease	Infections caused by bacteria		
Biliary Disease	Infections caused by fungi		
Biliary Disease	Infections caused by parasites		
Biliary Disease	smoking	_SMOKER3	1
Biliary Disease	smoking	_SMOKER3	2
Biliary Disease	smoking	_SMOKER3	3
Biliary Disease	toxic chemicals		
Carotid Artery Stenosis	hypertension	CRGVPRB2	8
Carotid Artery Stenosis	tobacco use	USENOW3	1
Carotid Artery Stenosis	tobacco use	USENOW3	2
Carotid Artery Stenosis	diabetes	CRGVPRB2	7
Carotid Artery Stenosis	high blood-fat levels		
Carotid Artery Stenosis	genetics		

Medical Condition	Condition Cause	SMART Field Name per cause per condition	SMART Field value
Carotid Artery Stenosis	age 65+	_AGE65YR	2
Carotid Artery Stenosis	obesity	_BMI5CAT	4
Carotid Artery Stenosis	sleep apnea		
Carotid Artery Stenosis	inactivity	_TOTINDA	2
Colon Disease	familial adenomatous polyposis (FAP)		
Colon Disease	age 50+	_AGEG5YR	7
Colon Disease	age 50+	_AGEG5YR	8
Colon Disease	age 50+	_AGEG5YR	9
Colon Disease	age 50+	_AGEG5YR	10
Colon Disease	age 50+	_AGEG5YR	11
Colon Disease	age 50+	_AGEG5YR	12
Colon Disease	age 50+	_AGEG5YR	13
Colon Disease	african-american	_PRACE1	2
Colon Disease	colorectal cancer	CHCOCNCR	1
Colon Disease	colorectal cancer	CRGVPRB2	3
Colon Disease	colorectal cancer	CNCRTYP1	10
Colon Disease	colorectal cancer	CNCRDIFF	1
Colon Disease	colorectal cancer	CNCRDIFF	2
Colon Disease	colorectal cancer	CNCRDIFF	3
Colon Disease	colorectal polyps		
Colon Disease	inflammatory intestinal conditions		
Colon Disease	crohn's disease		
Colon Disease	ulcerative colitis		
Colon Disease	familial adenomatous polyposis		
Colon Disease	lynch syndrome		
Colon Disease	family history of colon cancer		
Colon Disease	low-fiber, high-fat diet		
Colon Disease	inactivity	_TOTINDA	2
Colon Disease	diabetes	CRGVPRB2	7
Colon Disease	obesity	_BMI5CAT	4
Colon Disease	smoking	_SMOKER3	1
Colon Disease	smoking	_SMOKER3	2
Colon Disease	smoking	_SMOKER3	3
Colon Disease	alcohol abuse	_RFDRHV5	2
Colon Disease	radiation therapy	CSRVTRT1	1
Colon Disease	radiation therapy	CSRVTRT1	2
Kidney Failure	diabetes	CRGVPRB2	7
Kidney Failure	hypertension	CRGVPRB2	8
Kidney Failure	lupus	HAVARTH3	1
Kidney Failure	IgA nephropathy		
Kidney Failure	genetic diseases		
Kidney Failure	polycystic kidney disease	CHCKIDNY	1

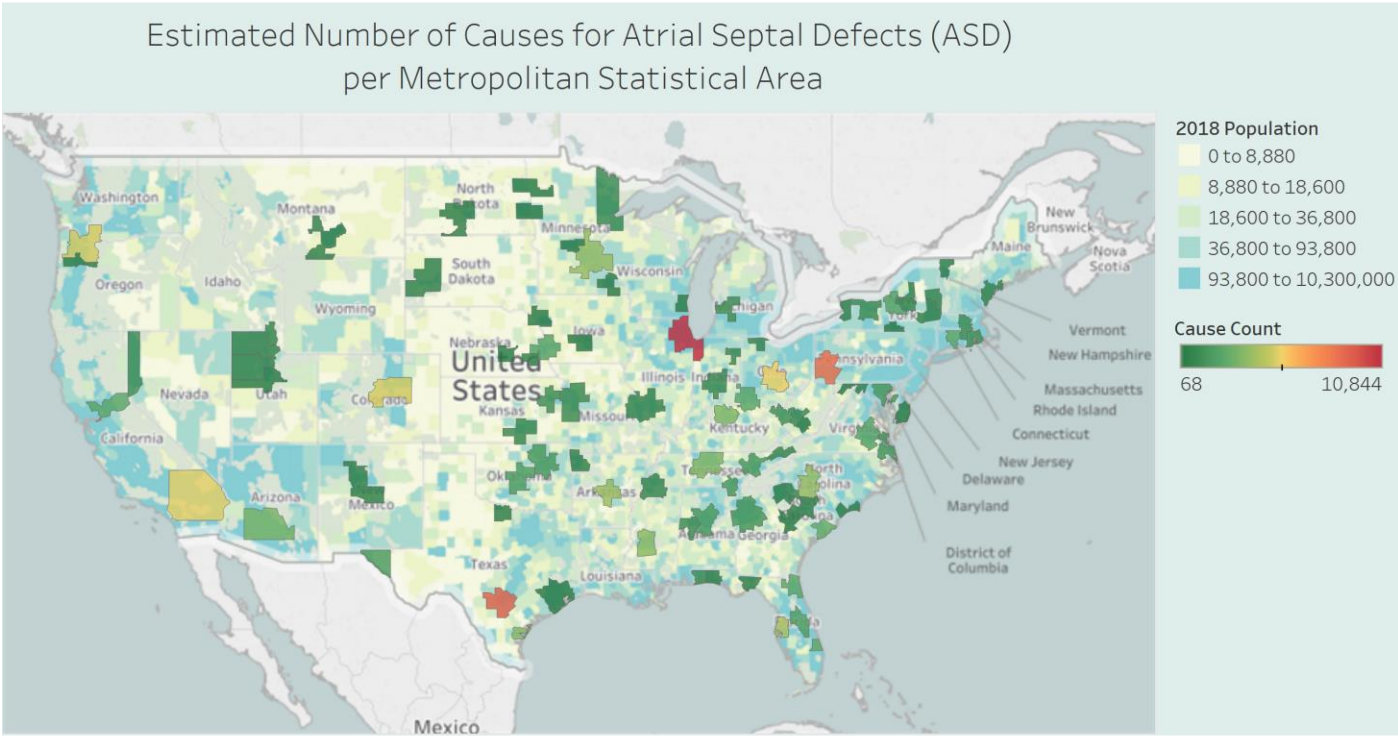
Medical Condition	Condition Cause	SMART Field Name per cause per condition	SMART Field value
Kidney Failure	nephrotic syndrome		
Kidney Failure	urinary tract problems		
Kidney Failure	heart attack	CVDINFR4	1
Kidney Failure	drug abuse	CRGVPRB2	
Kidney Failure	low blood flow to kidneys		
Heart Defect			
Hernia	age 50+	_AGEG5YR	7
Hernia	age 50+	_AGEG5YR	8
Hernia	age 50+	_AGEG5YR	9
Hernia	age 50+	_AGEG5YR	10
Hernia	age 50+	_AGEG5YR	11
Hernia	age 50+	_AGEG5YR	12
Hernia	age 50+	_AGEG5YR	13
Hernia	trauma		
Hernia	large hiatus		
Hernia	Persistent and intense pressure		
Hernia	obesity	_BMI5CAT	4
Peripheral Disease (PAD) (PVD)	smoking	_SMOKER3	1
Peripheral Disease (PAD) (PVD)	smoking	_SMOKER3	2
Peripheral Disease (PAD) (PVD)	smoking	_SMOKER3	3
Peripheral Disease (PAD) (PVD)	diabetes	CRGVPRB2	7
Peripheral Disease (PAD) (PVD)	obesity	_BMI5CAT	4
Peripheral Disease (PAD) (PVD)	hypertension	CRGVPRB2	8
Peripheral Disease (PAD) (PVD)	high cholesterol		
Peripheral Disease (PAD) (PVD)	age 50+	_AGEG5YR	7
Peripheral Disease (PAD) (PVD)	age 50+	_AGEG5YR	8
Peripheral Disease (PAD) (PVD)	age 50+	_AGEG5YR	9
Peripheral Disease (PAD) (PVD)	age 50+	_AGEG5YR	10
Peripheral Disease (PAD) (PVD)	age 50+	_AGEG5YR	11
Peripheral Disease (PAD) (PVD)	age 50+	_AGEG5YR	12
Peripheral Disease (PAD) (PVD)	age 50+	_AGEG5YR	13
Peripheral Disease (PAD) (PVD)	A family history of peripheral artery disease		
Peripheral Disease (PAD) (PVD)	A family history of heart disease	CVDCRHD4	1
Peripheral Disease (PAD) (PVD)	A family history of heart disease	_MICHD	1
Peripheral Disease (PAD) (PVD)	A family history of heart disease	CRGVPRB2	8
Peripheral Disease (PAD) (PVD)	A family history of stroke	CRGVPRB2	8
Peripheral Disease (PAD) (PVD)	A family history of stroke	CVDSTRK3	1
Peripheral Disease (PAD) (PVD)	high levels of homocysteine		
Portal Hypertension	cirrhosis of the liver	CRGVPRB2	11
Portal Hypertension	nonalcoholic fatty liver disease	CRGVPRB2	11
Portal Hypertension	iron buildup		
Portal Hypertension	cystic fibrosis		
Portal Hypertension	poorly developed bile ducts		

Medical Condition	**Condition Cause**	**SMART Field Name per cause per condition**	**SMART Field value**
Portal Hypertension	liver infections	CRGVPRB2	11
Portal Hypertension	methotrexate reaction		
Liver Disease	hepatitis A		
Liver Disease	hepatitis B		
Liver Disease	hepatitis C		
Liver Disease	autoimmune hepatitis		
Liver Disease	primary biliary cirrhosis		
Liver Disease	primary sclerosing cholangitis		
Liver Disease	liver cancer	CRGVPRB2	3
Liver Disease	liver cancer	CNCRTYP1	12
Liver Disease	liver cancer	CNCRDIFF	1
Liver Disease	liver cancer	CNCRDIFF	2
Liver Disease	liver cancer	CNCRDIFF	3
Liver Disease	bile duct cancer	CRGVPRB2	3
Liver Disease	bile duct cancer	CNCRDIFF	1
Liver Disease	bile duct cancer	CNCRDIFF	2
Liver Disease	bile duct cancer	CNCRDIFF	3
Liver Disease	liver adenoma	CRGVPRB2	11
Liver Disease	alcohol abuse	_RFDRHV5	2
Liver Disease	shared needles		
Liver Disease	tattoos		
Liver Disease	body piercings		
Liver Disease	blood transfusion before 1992		
Liver Disease	blood exposure		
Liver Disease	bodily fluid exposure		
Liver Disease	unprotected sex	HIVRISK4	1
Liver Disease	chemical exposure		
Liver Disease	toxin exposure		
Liver Disease	diabetes	CRGVPRB2	7
Liver Disease	obesity	_BMI5CAT	4
Thoracic Aortic Aneurysm (TAA)	cystic medial degeneration (necrosis)		
Thoracic Aortic Aneurysm (TAA)	marfan syndrome		
Thoracic Aortic Aneurysm (TAA)	ehlers-danlos syndrome		
Thoracic Aortic Aneurysm (TAA)	genetics		
Thoracic Aortic Aneurysm (TAA)	atherosclerosis		
Thoracic Aortic Aneurysm (TAA)	syphilis		
Thoracic Aortic Aneurysm (TAA)	takayasu's arteritis		
Thoracic Aortic Aneurysm (TAA)	atherosclerosis		
Thoracic Aortic Aneurysm (TAA)	ascending aortic aneurysm		
Thoracic Aortic Aneurysm (TAA)	descending aortic aneurysm		
Thoracic Aortic Aneurysm (TAA)	age 55+	_AGEG5YR	8
Thoracic Aortic Aneurysm (TAA)	age 55+	_AGEG5YR	9
Thoracic Aortic Aneurysm (TAA)	age 55+	_AGEG5YR	10

Medical Condition	Condition Cause	SMART Field Name per cause per condition	SMART Field value
Thoracic Aortic Aneurysm (TAA)	age 55+	_AGEG5YR	11
Thoracic Aortic Aneurysm (TAA)	age 55+	_AGEG5YR	12
Thoracic Aortic Aneurysm (TAA)	age 55+	_AGEG5YR	13
Thoracic Aortic Aneurysm (TAA)	male	_IMPSEX	1
Thoracic Aortic Aneurysm (TAA)	genetics		
Thoracic Aortic Aneurysm (TAA)	genetics		
Thoracic Aortic Aneurysm (TAA)	hyperlipidemia		
Thoracic Aortic Aneurysm (TAA)	hypertension	CRGVPRB2	8
Thoracic Aortic Aneurysm (TAA)	smoking	_SMOKER3	1
Thoracic Aortic Aneurysm (TAA)	smoking	_SMOKER3	2
Thoracic Aortic Aneurysm (TAA)	smoking	_SMOKER3	3
Thoracic Aortic Aneurysm (TAA)	diabetes	CRGVPRB2	7
Weight Loss Surgery	obesity	_BMI5CAT	4

B. Estimated Number of Causes for Atrial Septal Defects (ASD) per Metropolitan Statistical Area

The medical condition causes found in the SMART dataset for ASD are tobacco use, alcohol use, lupus, and obesity. The medical condition causes not found in the SMART dataset for ASD are atherosclerosis, hypertension, aorta blood vessel disease, aorta infection, trauma, and genetics.



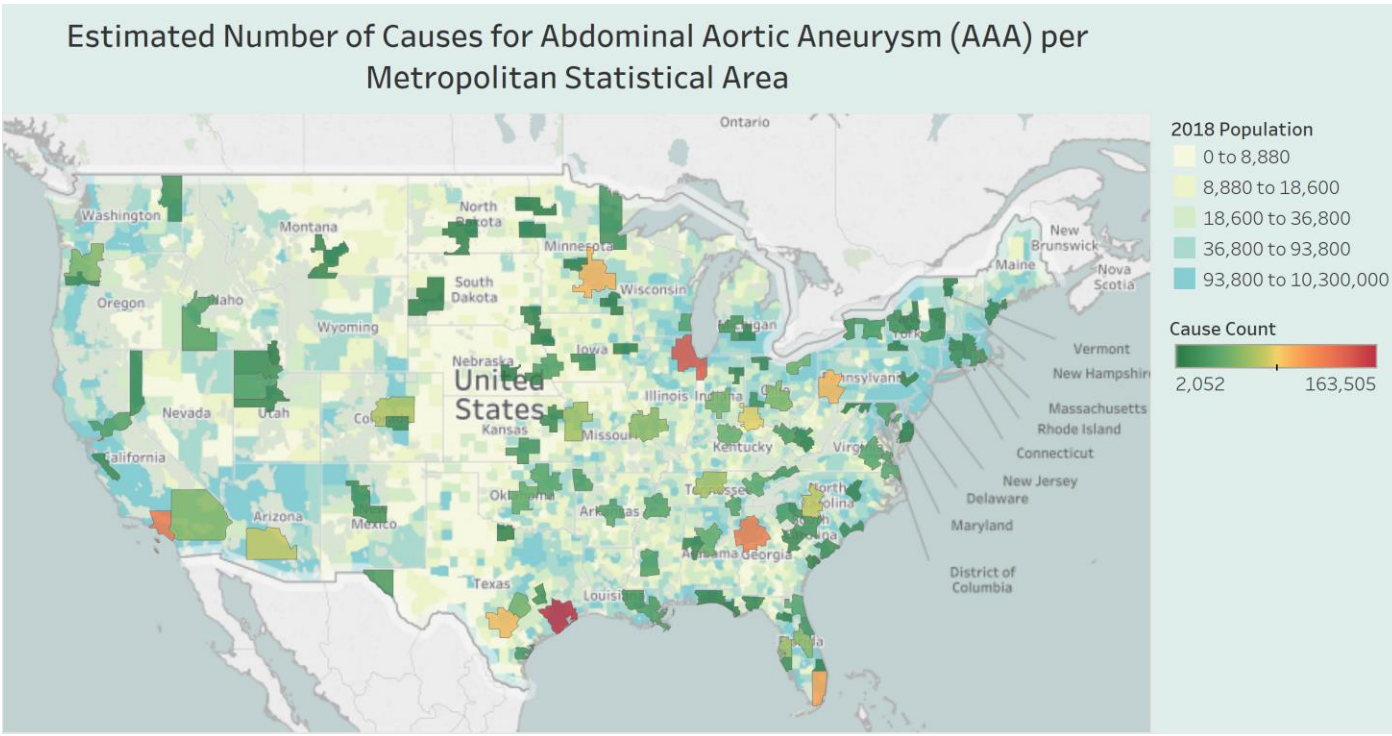
The value in the name column is the MMSA. The 2016 population estimate column contains the population estimate from the Census Bureau for the corresponding MMSA. The weighted count column contains the estimated number of individuals in the MMSA population in 2016 meeting the ASD criteria defined in Appendix A.

MMSA Name	2016 Population Estimate	Weighted Count
Chicago-Naperville-Elgin, IL-IN-WI	9546326	10844.0627
San Antonio-New Braunfels, TX	2426211	9001.97837
Pittsburgh, PA	2341536	8468.717067
Columbus, OH	2046977	5334.004866
Riverside-San Bernardino-Ontario, CA	4523653	5261.8234
Denver-Aurora-Lakewood, CO	2851848	5025.181588
Portland-Vancouver-Hillsboro, OR-WA	2423102	5013.74155
Tampa-St. Petersburg-Clearwater, FL	3036525	4200.487024
Little Rock-North Little Rock-Conway, AR	733461	3828.680758
Minneapolis-St. Paul-Bloomington, MN-WI	3557276	3771.629531
Charlotte-Concord-Gastonia, NC-SC	2475519	3673.225006
Jackson, MS	579558	3577.575296
Corpus Christi, TX	454066	3501.12146
Louisville/Jefferson County, KY-IN	1284848	3291.174889
Nashville-Davidson--Murfreesboro--Franklin, TN	1868855	3250.265209

MMSA Name	2016 Population Estimate	Weighted Count
Phoenix-Mesa-Scottsdale, AZ	4648498	3066.934594
Charleston-North Charleston, SC	761904	2679.639225
Hartford-West Hartford-East Hartford, CT	1210075	2505.871659
Port St. Lucie, FL	464563	2472.188883

C. Estimated Number of Causes for Abdominal Aortic Aneurysm (AAA) per Metropolitan Statistical Area

The medical condition causes found in the SMART dataset for AAA is tobacco use. The medical condition causes not found in the SMART dataset for AAA are diabetes, drug abuse, Rubella infection, Phenylketonuria (PKU), and exposure to certain chemicals.



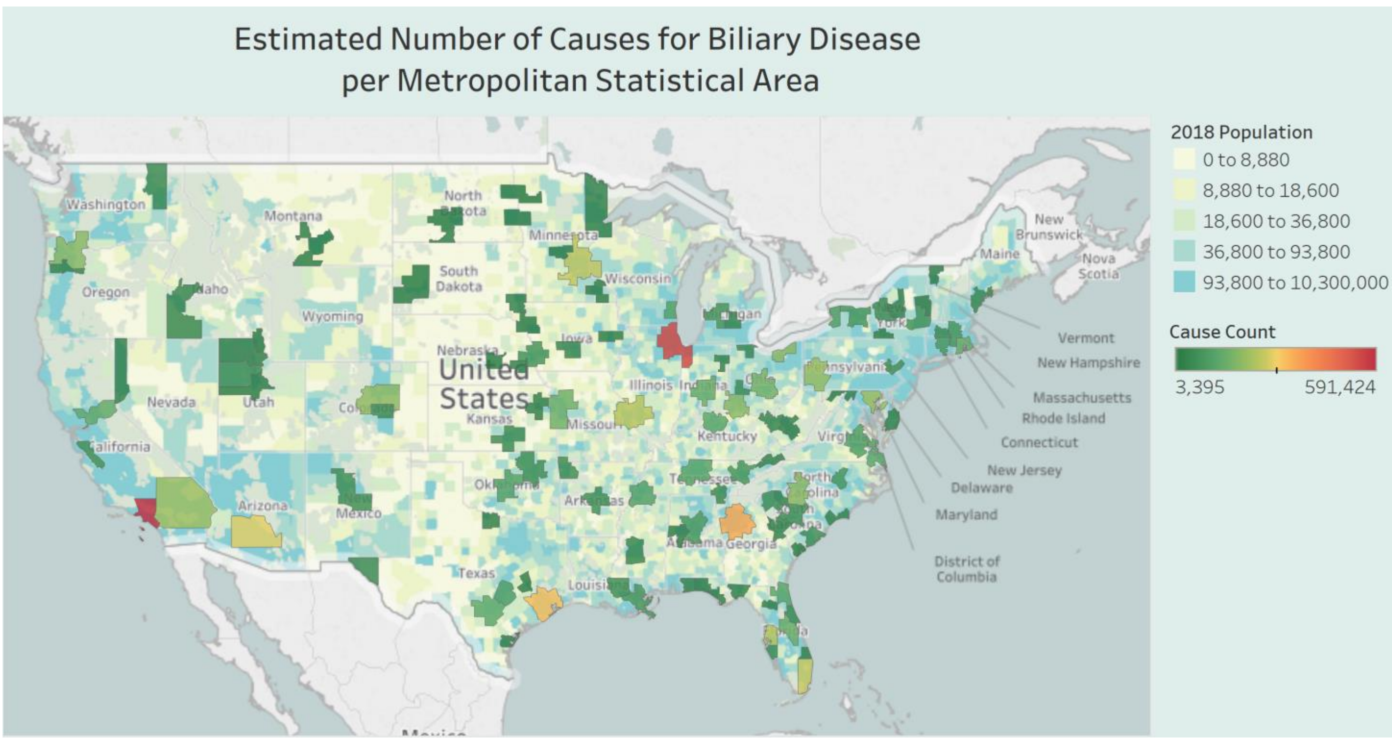
The value in the name column is the MMSA. The 2016 population estimate column contains the population estimate from the Census Bureau for the corresponding MMSA. The weighted count column contains the estimated number of individuals in the MMSA population in 2016 meeting the ASD criteria defined in Appendix A.

MMSA Name	2016 Population Estimate	Weighted Count
Houston-The Woodlands-Sugar Land, TX	6798010	163504.8445
Chicago-Naperville-Elgin, IL-IN-WI	9546326	142170.818
Los Angeles-Long Beach-Anaheim, CA	13328261	123525.8463
Atlanta-Sandy Springs-Roswell, GA	5795723	121230.3126
Miami-Fort Lauderdale-West Palm Beach, FL	6107433	101073.819
Minneapolis-St. Paul-Bloomington, MN-WI	3557276	94040.36685
Pittsburgh, PA	2341536	91423.71876
San Antonio-New Braunfels, TX	2426211	89998.33667
Cincinnati, OH-KY-IN	2166029	79753.72667

MMSA Name	2016 Population Estimate	Weighted Count
Phoenix-Mesa-Scottsdale, AZ	4648498	73193.52583
Charlotte-Concord-Gastonia, NC-SC	2475519	71383.20599
Denver-Aurora-Lakewood, CO	2851848	64827.13733
Nashville-Davidson--Murfreesboro--Franklin, TN	1868855	63199.88877
Kansas City, MO-KS	2106382	60544.13118
Tampa-St. Petersburg-Clearwater, FL	3036525	59367.292
St. Louis, MO-IL	2806782	54225.85479
Orlando-Kissimmee-Sanford, FL	2453333	53289.02907
Riverside-San Bernardino-Ontario, CA	4523653	51889.4579
Portland-Vancouver-Hillsboro, OR-WA	2423102	49760.68056
Columbus, OH	2046977	49344.30618

D. Estimated Number of Causes for Biliary Disease per Metropolitan Statistical Area

The medical condition causes found in the SMART dataset for Biliary Disease is female, age 30 to 60, and smoking. The medical condition causes not found in the SMART dataset for Biliary Disease is Europeans, North Americans, infections caused by bacteria, infections caused by fungi, infections caused by parasites, and toxic chemicals.

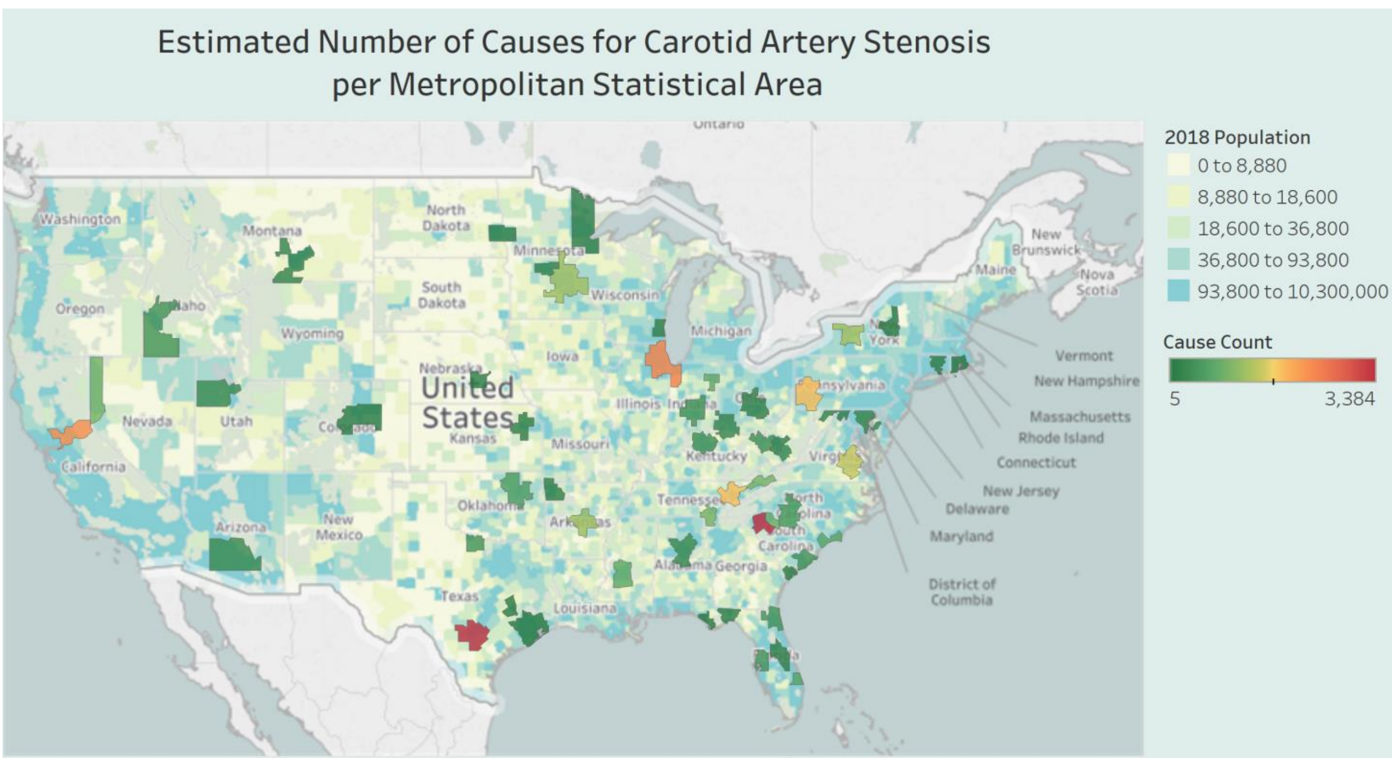


MMSA Name	2016 Population Estimate	Weighted Count
Los Angeles-Long Beach-Anaheim, CA	13328261	591423.5542
Chicago-Naperville-Elgin, IL-IN-WI	9546326	541887.0158
Atlanta-Sandy Springs-Roswell, GA	5795723	356171.168
Houston-The Woodlands-Sugar Land, TX	6798010	320292.6911
Phoenix-Mesa-Scottsdale, AZ	4648498	287041.6141

MMSA Name	2016 Population Estimate	Weighted Count
Miami-Fort Lauderdale-West Palm Beach, FL	6107433	265940.8844
Minneapolis-St. Paul-Bloomington, MN-WI	3557276	256319.0379
St. Louis, MO-IL	2806782	249138.2967
Tampa-St. Petersburg-Clearwater, FL	3036525	239279.4044
Baltimore-Columbia-Towson, MD	2801028	214515.2935
Riverside-San Bernardino-Ontario, CA	4523653	213772.8827
Pittsburgh, PA	2341536	208709.7902
Denver-Aurora-Lakewood, CO	2851848	193032.6265
Cleveland-Elyria, OH	2060065	192151.0688
Cincinnati, OH-KY-IN	2166029	190282.926
Charlotte-Concord-Gastonia, NC-SC	2475519	186053.8634
Portland-Vancouver-Hillsboro, OR-WA	2423102	185145.1531
Columbus, OH	2046977	160916.8446
Kansas City, MO-KS	2106382	155703.515
Orlando-Kissimmee-Sanford, FL	2453333	147554.7084

E. Estimated Number of Causes for Carotid Artery Stenosis per Metropolitan Statistical Area

The medical condition causes found in the SMART dataset for Carotid Artery Stenosis is tobacco use, diabetes, age 65+, obesity, and inactivity. The medical condition causes not found in the SMART dataset for Carotid Artery Stenosis high blood-fat levels, genetics, and sleep apnea.

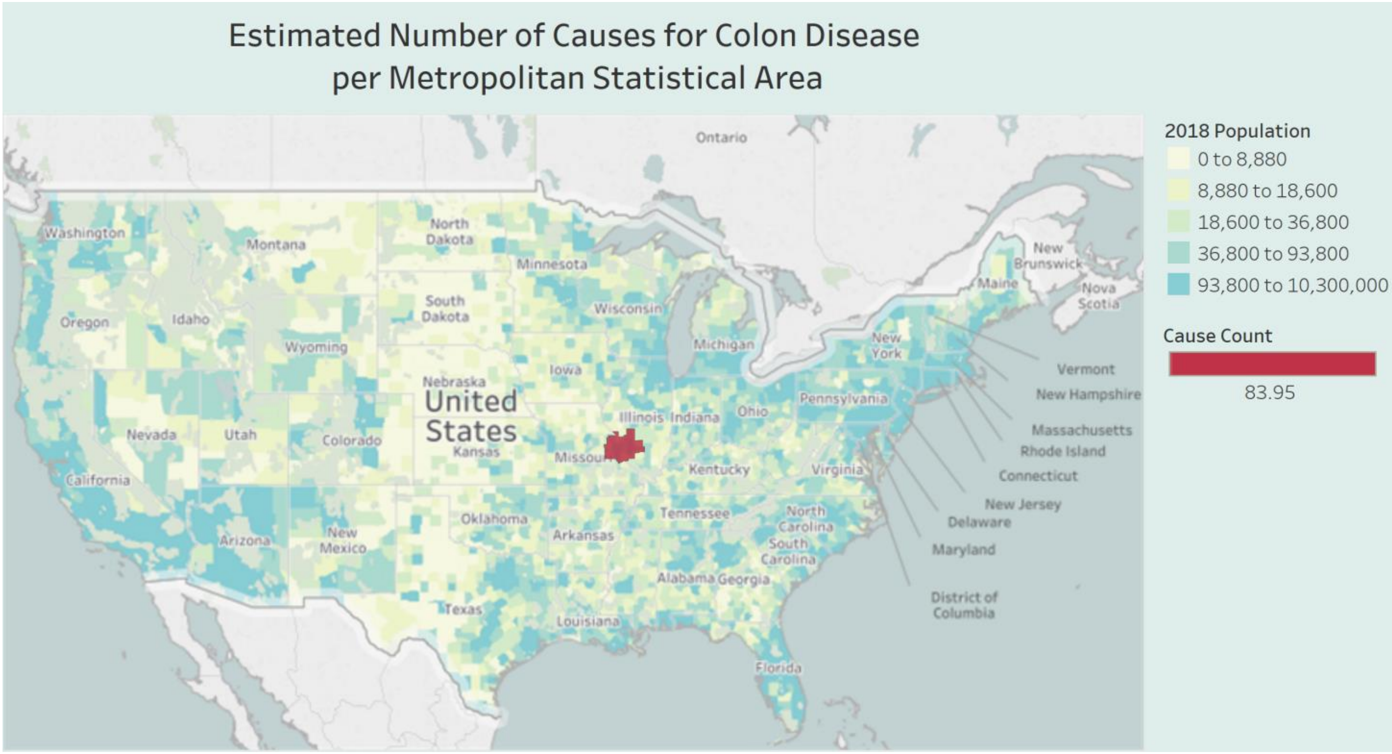


MMSA Name	2016 Population Estimate	Weighted Count

MMSA Name	2016 Population Estimate	Weighted Count
Providence-Warwick, RI-MA	1615878	4.72227134
Houston-The Woodlands-Sugar Land, TX	6798010	7.18965412
Panama City, FL	199092	10.260631
Anchorage, AK	401499	36.5051606
Utica-Rome, NY	293752	47.7355408
College Station-Bryan, TX	254230	50.021949
Grand Island, NE	84968	75.2437774
Denver-Aurora-Lakewood, CO	2851848	83.170297
Topeka, KS	232948	90.4732303
Hilton Head Island-Bluffton-Beaufort, SC	211568	92.4484653
Fargo, ND-MN	237483	99.9841555
Hartford-West Hartford-East Hartford, CT	1210075	101.835231
Duluth, MN-WI	278954	121.4045434
Hagerstown-Martinsburg, MD-WV	263080	122.726031
Orlando-Kissimmee-Sanford, FL	2453333	124.854625
Tallahassee, FL	379047	129.4186229
Fayetteville-Springdale-Rogers, AR-MO	525176	132.009606
Charleston, WV	217735	132.464967
Columbus, OH	2046977	143.544837
Milwaukee-Waukesha-West Allis, WI	1576143	149.552209

F. Estimated Number of Causes for Colon Disease per Metropolitan Statistical Area

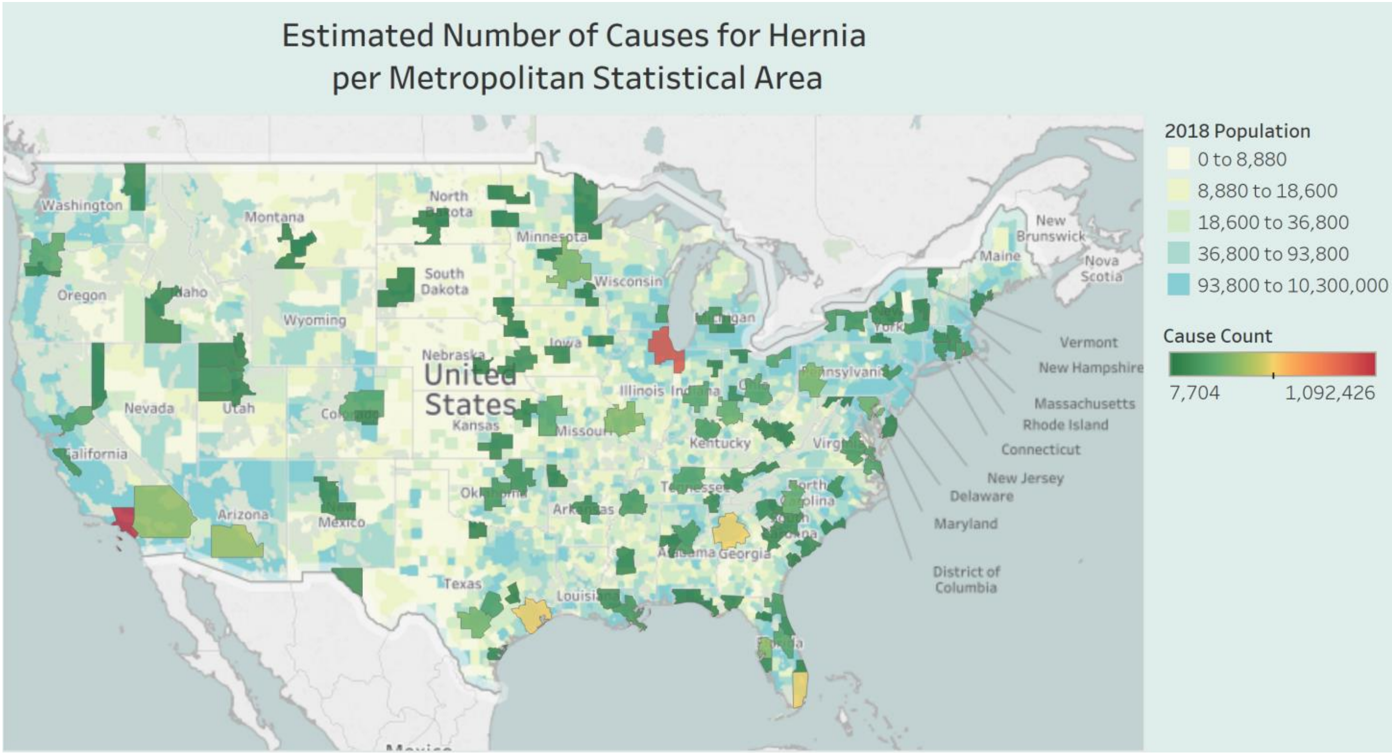
The medical condition causes found in the SMART dataset for Colon Disease is age 50+, African-American race, Colorectal cancer, inactivity, diabetes, obesity, smoking, and alcohol abuse. The medical condition causes not found in the SMART dataset for Colon Disease is familial adenomatous polyposis (FAP), colorectal polyps, inflammatory intestinal conditions, Crohn’s disease, ulcerative colitis, lynch syndrome, family history of colon cancer, and low-fiber and high-fat diet.



MMSA Name	Population Estimate 2016	Weight Count
St. Louis, MO-IL	2806782	193.8225888
Raleigh, NC	1304896	479.144769
Knoxville, TN	867870	234.079747

G. Estimated Number of Causes for Hernia per Metropolitan Statistical Area

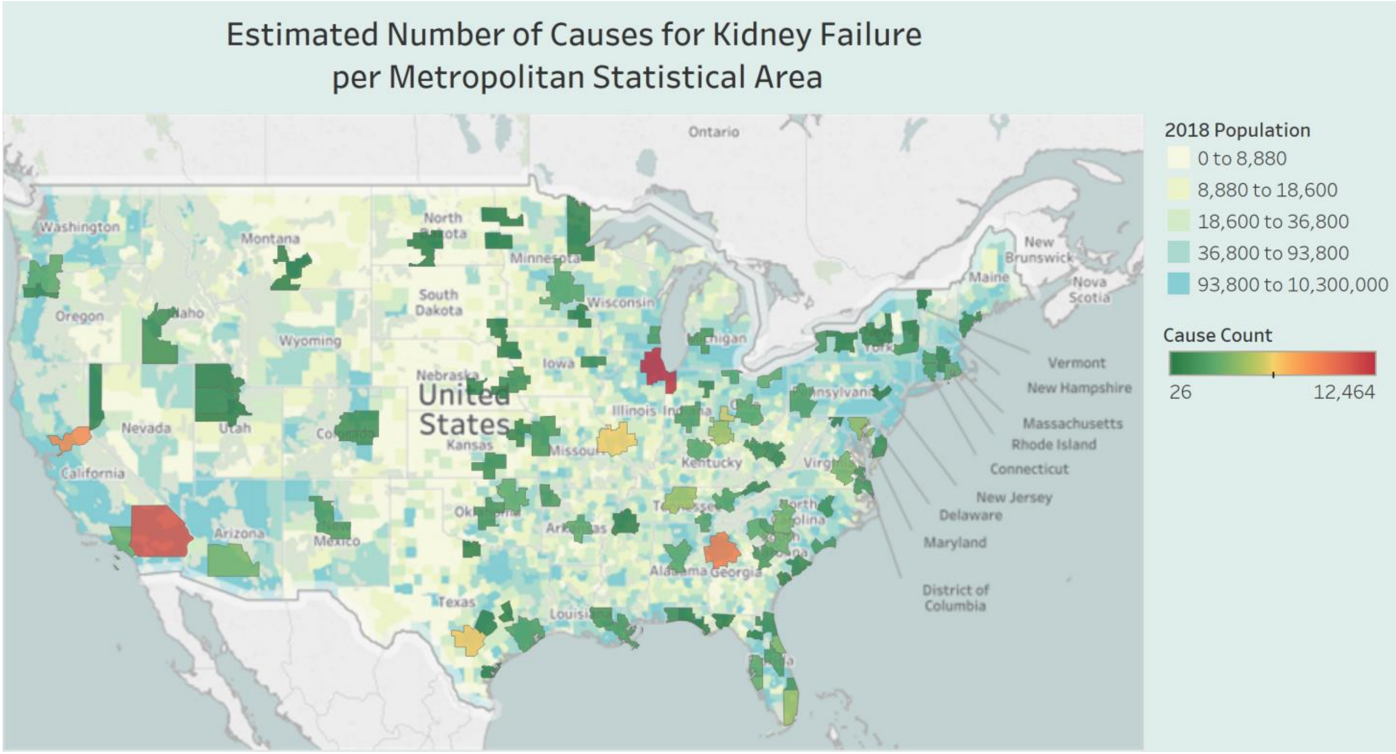
The medical condition causes found in the SMART dataset for Hernia is age 50+ and obesity. The medical condition causes not found in the SMART dataset for Hernia is trauma, large hiatus, and persistent and intense pressure on the surrounding muscles, such as while coughing, vomiting, straining during a bowel movement, exercising or lifting heavy objects.



MMSA Name	Population Estimate 2016	Weight Count
Los Angeles-Long Beach-Anaheim, CA	13328261	1092425.636
Chicago-Naperville-Elgin, IL-IN-WI	9546326	968277.3125
Houston-The Woodlands-Sugar Land, TX	6798010	558188.1647
Atlanta-Sandy Springs-Roswell, GA	5795723	555947.0104
Miami-Fort Lauderdale-West Palm Beach, FL	6107433	548427.721
Phoenix-Mesa-Scottsdale, AZ	4648498	409127.3982
Riverside-San Bernardino-Ontario, CA	4523653	374785.9192
Tampa-St. Petersburg-Clearwater, FL	3036525	349908.5843
St. Louis, MO-IL	2806782	335954.6279
Minneapolis-St. Paul-Bloomington, MN-WI	3557276	330058.4097
Pittsburgh, PA	2341536	326154.4135
San Antonio-New Braunfels, TX	2426211	311621.9485
Baltimore-Columbia-Towson, MD	2801028	302334.4362
Cincinnati, OH-KY-IN	2166029	265690.4216
Charlotte-Concord-Gastonia, NC-SC	2475519	249084.2684
Portland-Vancouver-Hillsboro, OR-WA	2423102	218001.9084
Kansas City, MO-KS	2106382	214151.6497
Cleveland-Elyria, OH	2060065	212136.697
Indianapolis-Carmel-Anderson, IN	2005612	211720.8927
Orlando-Kissimmee-Sanford, FL	2453333	211135.6768

H. Estimated Number of Causes for Kidney Failure per Metropolitan Statistical Area

The medical condition causes found in the SMART dataset for Kidney Failure is age diabetes, lupus, polycystic kidney disease, and heart attack. The medical condition causes not found in the SMART dataset for Kidney Failure is hypertension, lgA nephropathy, genetic diseases, nephrotic syndrome, urinary tract problems, drug abuse, and low blood flow to kidneys.

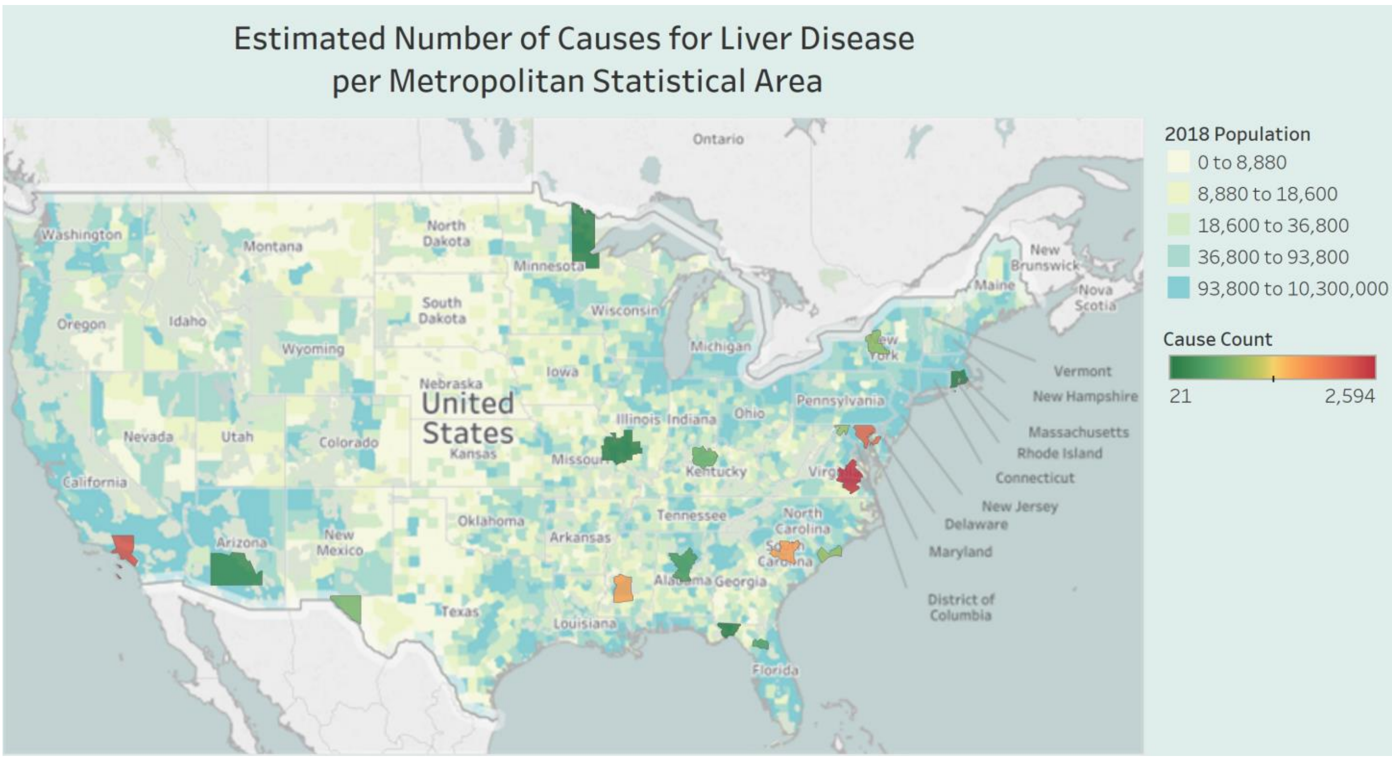


MMSA Name	Population Estimate 2016	Weight Count
Chicago-Naperville-Elgin, IL-IN-WI	9546326	18598.66402
Los Angeles-Long Beach-Anaheim, CA	13328261	17309.8298
Riverside-San Bernardino-Ontario, CA	4523653	15329.92521
Atlanta-Sandy Springs-Roswell, GA	5795723	14163.86034
Sacramento--Roseville--Arden-Arcade, CA	2295233	10370.50235
St. Louis, MO-IL	2806782	10060.26533
Nashville-Davidson--Murfreesboro--Franklin, TN	1868855	8231.53668
Miami-Fort Lauderdale-West Palm Beach, FL	6107433	8217.608229
Phoenix-Mesa-Scottsdale, AZ	4648498	7803.857943
San Antonio-New Braunfels, TX	2426211	7740.985108
Dayton, OH	800886	7294.964394
Louisville/Jefferson County, KY-IN	1284848	6424.274763
Baltimore-Columbia-Towson, MD	2801028	6316.972422
Pittsburgh, PA	2341536	6257.184559
Cincinnati, OH-KY-IN	2166029	5849.996331
Birmingham-Hoover, AL	1146888	5757.200768
Portland-Vancouver-Hillsboro, OR-WA	2423102	5707.762762
Deltona-Daytona Beach-Ormond Beach, FL	636843	5637.913524
Tampa-St. Petersburg-Clearwater, FL	3036525	5140.156382

MMSA Name	Population Estimate 2016	Weight Count
Buffalo-Cheektowaga-Niagara Falls, NY	1134914	4933.070519

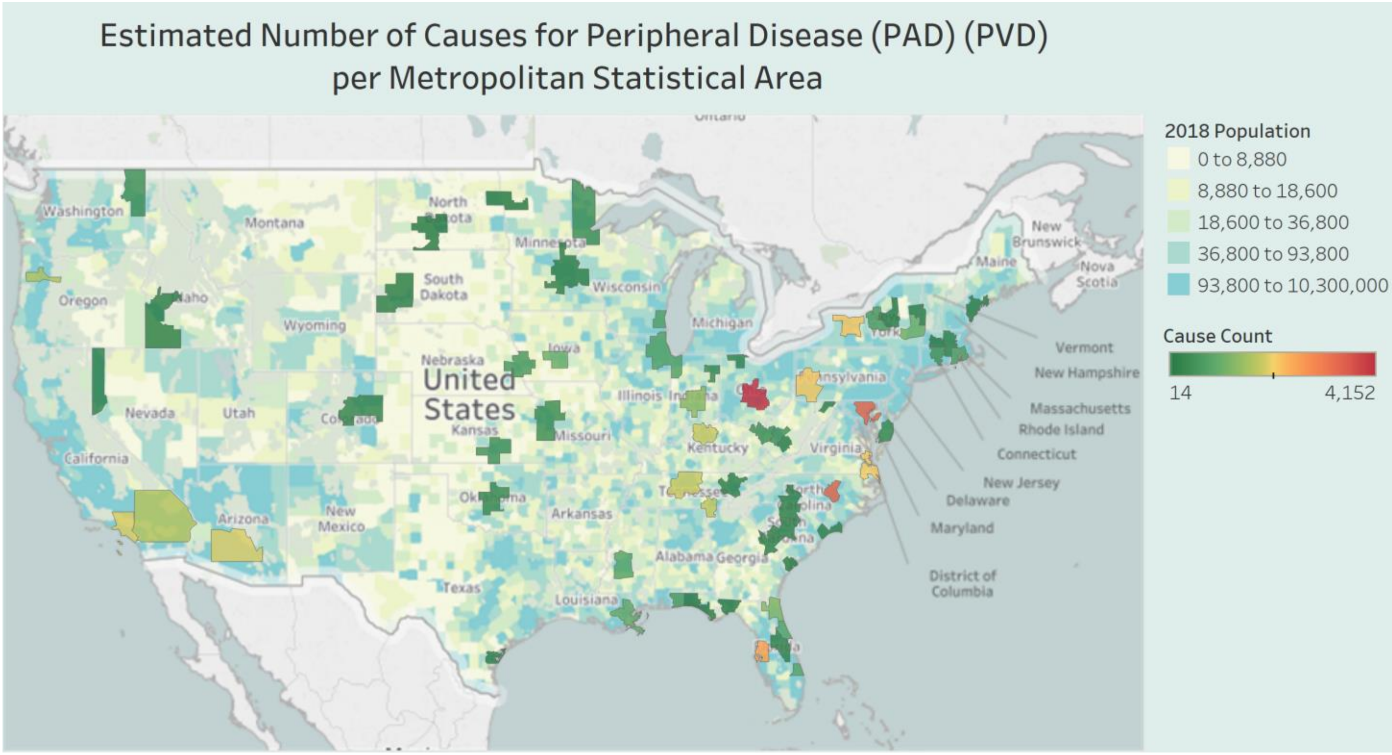
I. Estimated Number of Causes for Liver Disease per Metropolitan Statistical Area

The medical condition causes found in the SMART dataset for Liver Disease is alcohol abuse, unprotected sex, diabetes, and obesity. The medical condition causes not found in the SMART dataset for Liver Disease is hepatitis A, B, and C, autoimmune hepatitis, primary biliary cirrhosis, primary sclerosing cholangitis, liver cancer, bile duct cancer, liver adenoma, alcohol abuse, shared needles, tattoos, body piercings, blood transfusion before 1992, blood exposure, bodily fluid exposure, chemical exposure, and toxin exposure.



J. Estimated Number of Causes for Peripheral Disease (PAD) (PVD) per Metropolitan Statistical Area

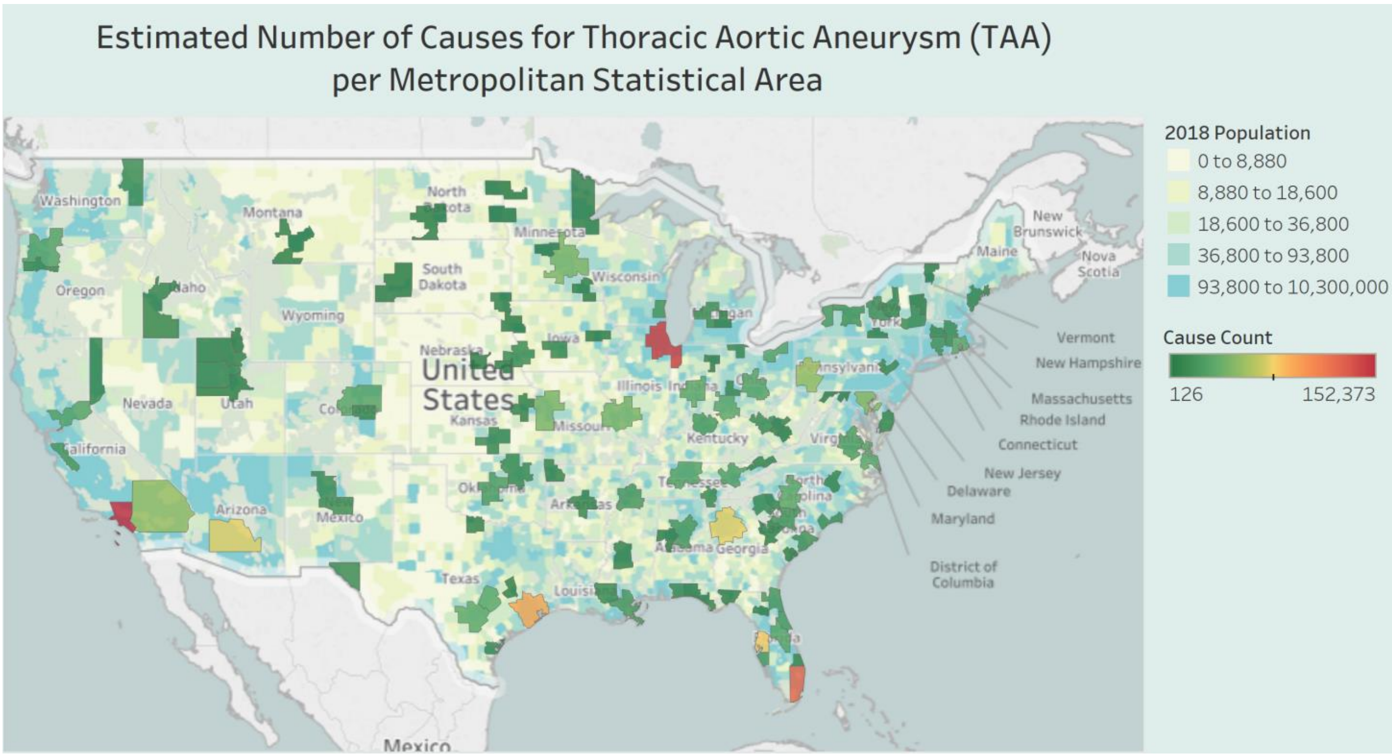
The medical condition causes found in the SMART dataset for PAD/PVD is smoking, diabetes, obesity, age 50+, family history of heart disease, family history of stroke, and family history of heart attack. The medical condition causes not found in the SMART dataset for PAD/PVD is family hypertension, high cholesterol, history of peripheral artery disease, and high levels of homocysteine.



MMSA Name	Population Estimate 2016	Weight Count
Los Angeles-Long Beach-Anaheim, CA	13328261	19885.98619
Miami-Fort Lauderdale-West Palm Beach, FL	6107433	19489.64167
Baltimore-Columbia-Towson, MD	2801028	7932.263054
Atlanta-Sandy Springs-Roswell, GA	5795723	7519.41266
Chicago-Naperville-Elgin, IL-IN-WI	9546326	6673.93963
Houston-The Woodlands-Sugar Land, TX	6798010	6616.55486
Portland-Vancouver-Hillsboro, OR-WA	2423102	5899.623293
Richmond, VA	1282205	5434.026116
Toledo, OH	604591	5427.79986
Milwaukee-Waukesha-West Allis, WI	1576143	5193.39834
Phoenix-Mesa-Scottsdale, AZ	4648498	4816.751028
St. Louis, MO-IL	2806782	4765.692013
Minneapolis-St. Paul-Bloomington, MN-WI	3557276	4504.433279
Worcester, MA-CT	936723	4279.76101
Hartford-West Hartford-East Hartford, CT	1210075	3932.436472
Kansas City, MO-KS	2106382	3714.549179
Charleston-North Charleston, SC	761904	3550.391731
Birmingham-Hoover, AL	1146888	3493.785227
Riverside-San Bernardino-Ontario, CA	4523653	3457.01118
Orlando-Kissimmee-Sanford, FL	2453333	3183.189518

K. Estimated Number of Causes for Thoracic Aortic Aneurysm (TAA) per Metropolitan Statistical Area

The medical condition causes found in the SMART dataset for Peripheral Disease TAA is age 55+, male, smoking, and diabetes. The medical condition causes not found in the SMART dataset for TAA is cystic medial degeneration (necrosis), Marfan syndrome, Ehlers-Danlos syndrome, genetics, atherosclerosis, syphilis, Takayasu's Arteritis, ascending aortic aneurysm, descending aortic aneurysm, genetics, hyperlipidemia, and hypertension.

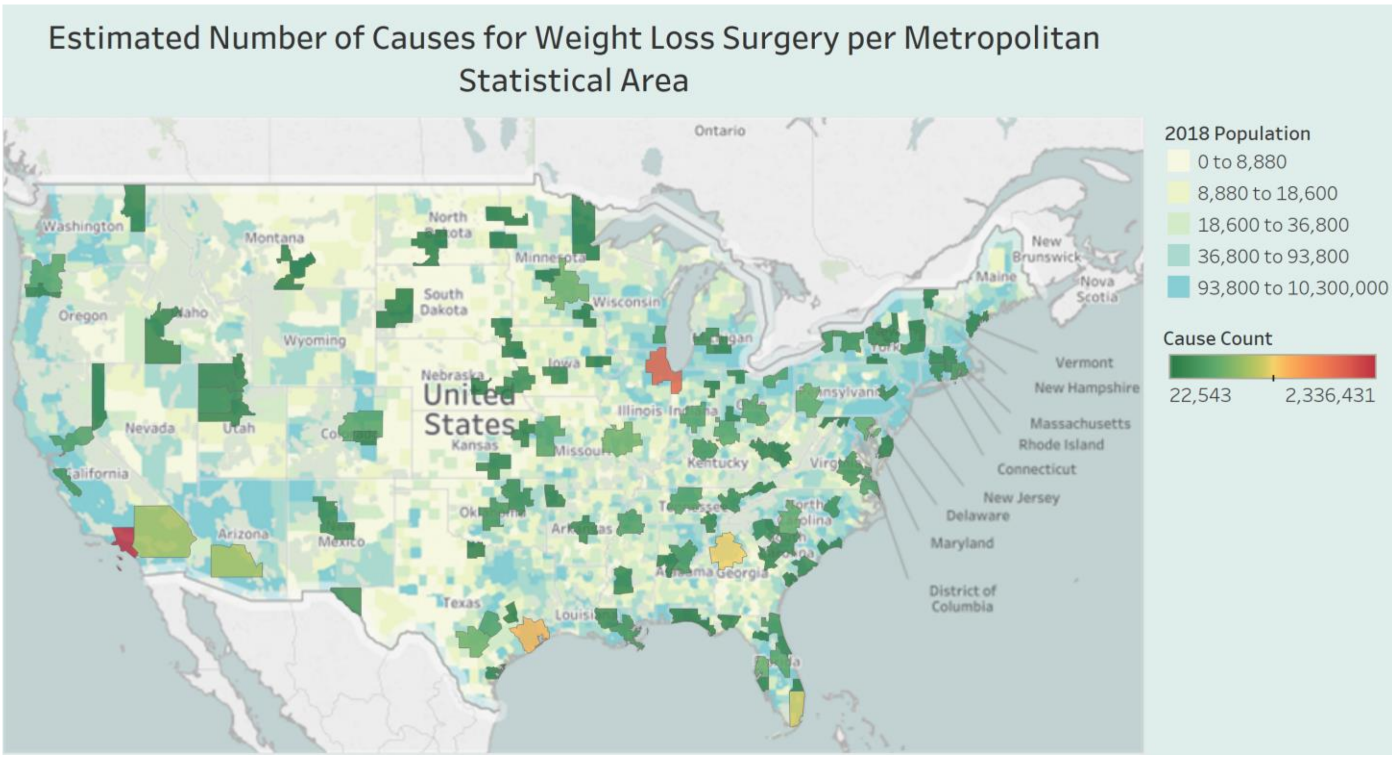


MMSA Name	Population Estimate 2016	Weight Count
Los Angeles-Long Beach-Anaheim, CA	13328261	605553.2212
Chicago-Naperville-Elgin, IL-IN-WI	9546326	594218.8193
Miami-Fort Lauderdale-West Palm Beach, FL	6107433	444305.7363
Houston-The Woodlands-Sugar Land, TX	6798010	362199.7423
Atlanta-Sandy Springs-Roswell, GA	5795723	300964.3972
Phoenix-Mesa-Scottsdale, AZ	4648498	300114.5227
Tampa-St. Petersburg-Clearwater, FL	3036525	281212.5222
Riverside-San Bernardino-Ontario, CA	4523653	257825.8928
Minneapolis-St. Paul-Bloomington, MN-WI	3557276	229167.6147
St. Louis, MO-IL	2806782	224389.2023
Pittsburgh, PA	2341536	218361.0024
Baltimore-Columbia-Towson, MD	2801028	191371.348
Cleveland-Elyria, OH	2060065	183425.6893
Denver-Aurora-Lakewood, CO	2851848	167275.4948
Orlando-Kissimmee-Sanford, FL	2453333	162701.2487
Kansas City, MO-KS	2106382	159348.7842
Portland-Vancouver-Hillsboro, OR-WA	2423102	150442.3095

MMSA Name	Population Estimate 2016	Weight Count
Cincinnati, OH-KY-IN	2166029	148132.3442
Charlotte-Concord-Gastonia, NC-SC	2475519	144159.6498
San Antonio-New Braunfels, TX	2426211	143758.3774

L. Estimated Number of Causes for Weight Loss Surgery per Metropolitan Statistical Area

The only medical condition cause found in the SMART dataset for Weight Loss Surgery is obesity. No missing medical condition causes for Weight Loss Surgery.



MMSA Name	Population Estimate 2016	Weight Count
Los Angeles-Long Beach-Anaheim, CA	13328261	2336430.58
Chicago-Naperville-Elgin, IL-IN-WI	9546326	1919986.067
Houston-The Woodlands-Sugar Land, TX	6798010	1316835.899
Atlanta-Sandy Springs-Roswell, GA	5795723	1186623.291
Miami-Fort Lauderdale-West Palm Beach, FL	6107433	1088812.707
Phoenix-Mesa-Scottsdale, AZ	4648498	914500.1156
Riverside-San Bernardino-Ontario, CA	4523653	899043.5692
San Antonio-New Braunfels, TX	2426211	654940.513
St. Louis, MO-IL	2806782	647022.7779
Minneapolis-St. Paul-Bloomington, MN-WI	3557276	645167.1733
Tampa-St. Petersburg-Clearwater, FL	3036525	601276.6542
Baltimore-Columbia-Towson, MD	2801028	592082.4802
Pittsburgh, PA	2341536	569445.697
Charlotte-Concord-Gastonia, NC-SC	2475519	531794.1312

Cincinnati, OH-KY-IN	2166029	478747.3605
Orlando-Kissimmee-Sanford, FL	2453333	474598.1528
Portland-Vancouver-Hillsboro, OR-WA	2423102	453206.7677
Denver-Aurora-Lakewood, CO	2851848	451786.1146
Columbus, OH	2046977	439510.9758
Cleveland-Elyria, OH	2060065	438373.7339

VIII. References

[1] Granville, V. (2012, August 19). Sample proposal for a data science / big data project. Retrieved May 18, 2018, from https://www.analyticbridge.datasciencecentral.com/profiles/blogs/sample-proposal-for-a-fraud-detection-project

[2] BRFSS Web Enabled Analysis Tool. (n.d.). Retrieved May 18, 2018, from https://nccd.cdc.gov/weat/#/

[3] Behavioral Risk Factor Surveillance System OVERVIEW: BRFSS 2016 [PDF]. (2017, June 29). Center For Disease Control and Prevention. https://www.cdc.gov/brfss/annual_data/2016/pdf/overview_2016.pdf

[4] LLCP 2016 Codebook Report Overall version data weighted with _LLCPWT Behavioral Risk Factor Surveillance System [PDF]. (2017, August 3). Center For Disease Control and Prevention. https://www.cdc.gov/brfss/annual_data/2016/pdf/codebook16_llcp.pdf

[5] U.S. Census Bureau QuickFacts: UNITED STATES. (n.d.). Retrieved May 18, 2018, from https://www.census.gov/quickfacts/fact/table/US/PST045217

[6] Behavioral Risk Factor Surveillance System: Summary Matrix of Calculated Variables (CV) in the 2016 Data File. (2017, August 25). Retrieved from https://www.cdc.gov/brfss/annual_data/2016/Summary_Matrix_16.html

[7] Gore Medical Products. (n.d.). Retrieved May 18, 2018, from https://www.goremedical.com/products

[8] Gore Medical Products Conditions. (n.d.). Retrieved May 25, 2018, from https://www.goremedical.com/conditions

[9] Thoracic Aortic Aneurysm (TAA) Causes. (n.d.). Retrieved May 18, 2018, from https://stanfordhealthcare.org/medical-conditions/blood-heart-circulation/thoracic-aortic-aneurysm/causes.html

[10] US Census Bureau. (2017, November 28). Delineation Files. Retrieved May 25, 2018, from https://www.census.gov/geographies/reference-files/time-series/demo/metro-micro/delineation-files.html

[11] Behavioral Risk Factor Surveillance System. (2017, November 13). Retrieved June 1, 2018, from https://www.cdc.gov/brfss/smart/smart_data.htm

[12] Behavioral Risk Factor Surveillance System. (NA). MMSA Variable Layout. Retrieved June 1, 2018, from https://www.cdc.gov/brfss/smart/2016/mmsa_varlayout_16.pdf

[13] Abdominal aortic aneurysm. (2018, March 13). Retrieved June 1, 2018, from https://www.mayoclinic.org/diseases-conditions/abdominal-aortic-aneurysm/symptoms-causes/syc-20350688

[14] Atrial septal defect (ASD). (2018, January 25). Retrieved June 1, 2018, from https://www.mayoclinic.org/diseases-conditions/atrial-septal-defect/symptoms-causes/syc-20369715

[15] Primary biliary cholangitis. (2018, March 09). Retrieved June 1, 2018, from https://www.mayoclinic.org/diseases-conditions/primary-biliary-cholangitis-pbc/symptoms-causes/syc-20376874

[16] S, V. (2015, December 14). Fixed width files in Hive. Retrieved June 2, 2018, from http://www.hadooptechs.com/hive/fixed-width-files-in-hive

[17] Unable to load hive table with fixed width file using SerDe. (n.d.). Retrieved June 2, 2018, from https://stackoverflow.com/questions/40988493/unable-to-load-hive-table-with-fixed-width-file-using-serde

[18] How to export a Hive table into a CSV file? (n.d.). Retrieved June 3, 2018, from https://stackoverflow.com/questions/17086642/how-to-export-a-hive-table-into-a-csv-file

[19] 2016 SMART BRFSS MMSA Methodology. (n.d.). Retrieved May 18, 2018, from https://www.cdc.gov/brfss/smart/2016/2016_SMART_BRFSS_MMSA_Methodology.pdf

[20] Wicklin, R. (2017, October 02). How to understand weight variables in statistical analyses. Retrieved June 18, 2018, from https://blogs.sas.com/content/iml/2017/10/02/weight-variables-in-statistics-sas.html

[21] Calculated Variables in the 2016 Data File of the Behavioral Risk Factor Surveillance System [PDF]. (2017, June 2). Center For Disease Control and Prevention. Retrieved June 18, 2018 https://www.cdc.gov/brfss/annual_data/2016/pdf/2016_calculated_variables_version4.pdf

[22] Data Access and Dissemination Systems (DADS). (2010, October 05). Results. Retrieved June 22, 2018, from https://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?src=bkmk

[23] CBSA-EST2017-alldata: Annual Resident Population Estimates File Layout [PDF]. (2018, March). Census Bureau. Retrieved June 18, 2018, from https://www2.census.gov/programs-surveys/popest/technical-documentation/file-layouts/2010-2017/cbsa-est2017-alldata.pdf

[24] Behavioral Risk Factor Surveillance System. (2017, December 06). Retrieved June 26, 2018, from https://www.cdc.gov/brfss/annual_data/annual_2016.html

[25] Behavioral Risk Factor Surveillance System [PDF]. (2017, July). Retrieved May 25, 2018, from https://www.cdc.gov/brfss/annual_data/2016/pdf/compare_2016.pdf

[26] McCormick, T. H., Rudin, C., & Madigan, D. (2010). Bayesian Hierarchical Rule Modeling for Predicting Medical Conditions [PDF]. New York: Columbia University. http://www.stat.columbia.edu/~madigan/PAPERS/rec_sys23_ssrn.pdf

[27] Strickland, E. (2017, May 01). AI Predicts Heart Attacks and Strokes More Accurately Than Standard Doctor's Method. Retrieved July 3, 2018, from https://spectrum.ieee.org/the-human-os/biomedical/diagnostics/ai-predicts-heart-attacks-more-accurately-than-standard-doctor-method

John T Giles earned a B.S. B.A. degree in computer information systems from Northern Arizona University in Flagstaff, AZ in 1994, and M.S. B.A. degree in computer information systems from University of Phoenix in Phoenix, AZ in 2004. He is currently working on a M.S. degree in computer science with a concentration in data science from Lewis University in Romeoville, IL.

He has worked in Information Technology for 20+ years. In those 20+ years he developed, configured, implemented, and improved and maintained software solutions both built from scratch and purchased off the shelf. From 1995 to 2001, he worked for the City of Mesa Information Services Division. From 2001 to 2006, he worked for the City of Mesa Police Department. From 2006 to present, he has worked for W.L. Gore and Associates, currently as their MDM Specialist.