How to Identify Health Innovation Gaps? Insights from Data on Diseases' Costs, Mortality, and Funding

Lopez, Claude and Roh, Hyeongyul and Butler, Brittney

Milken Institute, Milken Institute, Milken Institute

January 2021

How to Identify Health Innovation Gaps? Insights from Data on Diseases' Costs, Mortality, and Funding

Claude Lopez, PhD, Hyeongyul Roh, PhD, and Brittney Butler

About the Milken Institute

The Milken Institute is a nonprofit, nonpartisan think tank. We catalyze practical, scalable solutions to global challenges by connecting human, financial, and educational resources to those who need them.

We leverage the expertise and insight gained through research and the convening of top experts, innovators, and influencers from different backgrounds and competing viewpoints to construct programs and policy initiatives. Our goal is to help people build meaningful lives in which they can experience health and well-being, pursue effective education and gainful employment, and access the resources required to create ever-expanding opportunities for themselves and their broader communities.

Contents

Abstract

The Health Innovation Gap Ranking Project focuses on creating a framework to systematically identify priority areas among diseases and conditions that would benefit from research and development (R&D) investment.

This report aims to identify disease categories with the highest economic and social costs and a low level of R&D investment. First, we combine data sets on diseases' medical expenses, patient counts, death rates, and research funding. We then use text mining and machine learning methods to identify gaps between diseases' social and economic costs and research investments in therapeutic areas.

We find that only 25 percent of disease categories causing high economic and social costs received more than 1 percent of National Institutes of Health (NIH) funding over 12 years. In addition, rare diseases imposing high medical costs per patient collected 0.3 percent of research investments on average.

A disease's cost and impact on society are challenging to assess. Our results highlight that the different measures may lead to different conclusions if considered separately: A disease can have a very high cost per patient but a low death rate. They also show that merging information across data sets becomes more complicated when the sources do not focus on diseases specifically.

Our analysis reveals that a formalized procedure to define the correspondence between data sets is needed to successfully develop a metric that allows a systematic assessment of diseases' cost, impact on society, and investment level. Furthermore, the simplification of the large dimensional decision space will only be useful to the questions at hand if there is a clear order of priorities. In our case, the first was the costs and then funding. These priorities dictate how to merge the data sets.

Introduction

As part of the partnership between FasterCures, a Center of the Milken Institute, and the US Department of Health and Human Services (HHS) Accelerate Clinical Innovation (ACI) Initiative, the Milken Institute research department was asked to provide methodological and technical input to the Health Innovation Gap Ranking Project and to contribute to the Biomedical Ecosystem Metrics Initiative. This initiative is part of the 2018-2022 HHS strategic plan that focuses on five goals: "(i) Reforming, Strengthening, and Modernizing the Nation's Healthcare System, (ii) Protecting the Health of Americans Where They Live, Work, Learn, and Play, (iii) Strengthening the Economic and Social Well-Being of Americans Across the Life-Span, (iv) Fostering Sound, Sustained Advances in the Sciences, and (v) Promoting Effective Management and Stewardship."[1]

The Health Innovation Gap Ranking Project focuses on creating a framework to systematically identify priority areas among diseases and conditions that would benefit from research and development (R&D) investment. The prioritization of the diseases and conditions would reflect their impact on public health, their cost to the health-care system, and the absence of recent related biomedical innovation.

This paper contributes to this effort by identifying disease categories with the highest economic and social costs and a low level of R&D investment. We approximate the economic and social costs by a disease's medical expense level, its number of patients, and mortality rate. The level of NIH funding received is a proxy for the R&D investment. Indeed, Packalen and Bhattacharya (2020) show the importance of NIH funding for innovative research, which often leads to biomedicine advancement. NIH funding was $2 trillion over the past 12 years.

Specifically, we suggest a methodology on (1) how to merge information from different sources to assess the economic and social cost per disease so that the match between data sets is close

1. See hhs.gov for more details.

to 100 percent and (2) how to identify clusters of diseases with high medical cost, number of patients, or high mortality rate.

Our results illustrate the benefit of new methods such as text mining and machine learning in merging and sorting information. They also emphasize the necessity of a standardized equivalence procedure between databases' disease and condition categories to reconcile information sources. The reconciliation of information sources is an essential step when designing the framework for any systematic assessment of the cost, funding, and other dimensions of the disease or condition level. Finally, the order in which the data sets are merged impacts the final data set's information, highlighting the importance of prioritizing information when merging data.

The paper first describes the different data sets and how to merge them. Then it explains the methodology before presenting the results and some concluding remarks.

Data

We need to combine the following data sets to obtain information about health-care spending, the number of patients, mortality rates, and research funding per disease.

Health Care Spending: The Blended Account database, from the Bureau of Economic Analysis (BEA) of the US Department of Commerce, estimates the annual health-care expenses and the number of patients per disease type.[2] The Blended Account database relies on three data sources: the Medical Expenditure Panel Survey (MEPS), a patient-level health-care claims database from Truven Health Analytics, and a 5 percent random sample of Medicare beneficiaries. The MEPS collects data on approximately 15,000 families and 35,000 individuals each year. Because of its relatively small sample size compared to the total US population, the MEPS produces volatile estimates across years. The Blended Account database overcomes this issue by including the other two broad claims databases: the Truven Health MarketScan

2. Dunn, Rittmueller, and Whitmire (2015).

Commercial Database, which contains patient-level health-care claims information from employers and health plans, and Medicare data, which consist of claims from a 5 percent random sample of beneficiaries in fee-for-service Medicare. Both the Truven Health and Medicare 5 percent claims data capture information on millions of enrollees and billions of claims. We use estimates of annual spending per patient, the annual patient counts, and yearly total medical expenditures from the data. The Blended Account database has a total of 262 disease categories for the period from 2000 to 2016.

Mortality Rates: The Centers for Disease Control and Prevention's (CDC's) Wide-ranging Online Data for Epidemiologic Research (WONDER) project reports the total number of deaths, based on the death certificates of US residents. The underlying cause of death is selected from the conditions entered by the physician on the death certificate.[3] The WONDER data contain 5,016 causes of deaths for the period from 2000 to 2018.

Funding for Innovative Research: In 2008, NIH implemented a new process to improve consistency and transparency in the reporting of its funded research. The Research, Condition, and Disease Categorization (RCDC) system uses text data mining (categorizing and clustering using words and multiword phrases) in conjunction with NIH-wide definitions to assign the funded research topics to categories. The NIH data report 296 RCDC research categories for the period from 2008 to 2019.

Methodology

Each data set has its definition and number of disease categories. While the cost-related data (i.e., health-care spending, number of patients, and mortality rates) are disease-related, the NIH funding categories are related to research area. Hence, merging these data sets requires several steps.

3. The World Health Organization defines "underlying cause of death" as "the disease or injury which initiated the train of events leading directly to death, or the circumstances of the accident or violence which produced the fatal injury." Specifically, the underlying cause of death is selected from the conditions entered by the physician on the cause of death section of the death certificate. When more than one cause or condition is entered by the physician, the underlying cause is determined by the sequence of conditions on the certificate, provisions of the International Classification of Diseases, and associated selection rules and modifications.

We first identify a set of common disease categories based on the cost-related data. We then use the combined information to sort the disease categories based on their medical expenses, number of patients, and mortality rates. Finally, we match these classes of disease categories with the NIH funding. Below we provide more technical details on this three-step approach.

Step 1: Common Disease Categories

The health-care spending and number of patients data from BEA use 262 Clinical Classification Software (CCS) disease categories. In contrast, the CDC's mortality data follow the *International Classification of Diseases, Tenth Revision, Clinical Modification* (ICD-10-CM). The mortality data include 6,088 unique causes of deaths denoted as ICD-10-CM, from 2000 to 2018. We excluded deaths related to unavoidable accidents, terrorism, and war, which results in 5,016 causes of death considered for this study.[4] Merging both data sets requires the creation of a common set of exclusive disease categories.

We choose the CCS disease categories as a benchmark for the common disease categories because of their simplicity and widespread use in the literature. The metafile in the Clinical Classifications Software Refined (CCSR), an off-the-shelf software product, allows us to match the CCS and ICD-10-CM. We find the CCS codes for 4,601 of 5,016 ICD-10-CM codes. This direct approach matches 92 percent of the data.[5]

Then, we use a text mining method to match the remaining 415 ICD-10-CM codes to CCS disease categories. More specifically, we collect text files describing the disease related to each ICD-10-CM code and each cause of death. We extract key medical terminologies from these descriptions and employ the Edit Distance algorithm to find the most similar pairs between the ICD-10-CM descriptions and causes of death descriptions. The text-analytic method examines the patterns of letters in a word and calculates the minimum number of operations required to

4. We dropped U, V, W, X, and Y codes in ICD-10-CM.
5. The CCSR is one in a family of databases and software tools developed as part of the Healthcare Cost and Utilization Project (https://www.hcup-us.ahrq.gov/overview.jsp), a federal-state-industry partnership sponsored by the Agency for Healthcare Research and Quality. The metafile contains 72,436 ICD-10-CM codes with descriptions and corresponding CCS codes with descriptions.

transform one word into another.[6] For example, "Delusional disorder" needs one operation (the insertion of a letter) to become "Delusional disorders." We select the description pairs with the lowest edit distance—that is, the most similar sentences. Our manual crosschecks of the results find that simple typographical errors in a word, discrepancies in the use of singular or plural forms of medical terminology, or a different order of words in a sentence cause most mismatches created by the metafile in the CCSR software tool. Overall, we can match 99 percent of the data. Only 62 causes of death remain unmatched and are removed from the analysis.[7]

Step 2: Clustering of the Disease Categories

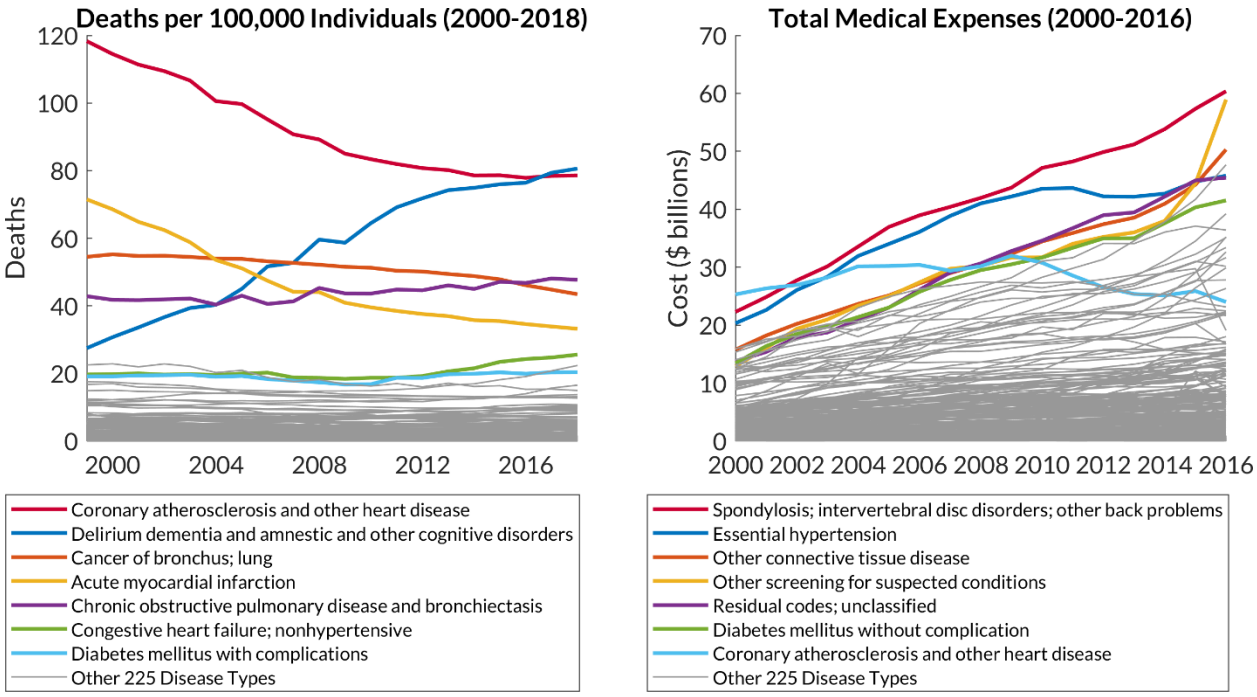
As noted above, our primary goal is to identify the disease categories with relatively high health-care costs or mortality rates and low innovative research funding.

Figure 1 confirms that only a few CCS disease categories have substantially higher death rates and medical expenses than the others. It also shows that high mortality diseases do not correspond to diseases with high medical expenses. For example, only one disease category is among the top seven annual averages for both mortality rates and medical costs: coronary atherosclerosis and other heart diseases. Similarly, there is no benchmark defining relatively and statistically higher death rates and medical costs, except for a few outliers, making conventional statistics inappropriate to identify our groups of interest. Finally, we consider estimates of annual spending per patient to identify costly disease categories and the yearly patient counts as a proxy for the prevalence of diseases, in addition to two variables in Figure 1. Using multiple variables further complicates a conventional statistical methods approach.

6. In computational linguistics and computer science, edit distance is a way to quantify how similar two strings (e.g., words) are to one another by counting the minimum number of operations required to transform one string into the other. Edit distances find applications in natural language processing, where automatic spelling correction can determine candidate corrections for a misspelled word by selecting words from a dictionary that have a low distance to the word in question. In bioinformatics, it can be used to quantify the similarity of DNA sequences, which can be viewed as strings of the letters A, C, G, and T (Wikipedia, n.d.).

7. Combining the mortality data with health-care data from BEA excludes 30 disease categories in the BEA data, which are related to pure accidents or not involved with death.

Figure 1: CCS Disease Categories with High Mortality Rates and High Health-Care Costs



Note: *The left panel shows the total number of deaths per 100,000 individuals caused by specific CCS disease categories from 2000 to 2018. All values are yearly aggregate values. The right panel indicates annual health-care expenses from 2000 to 2016 to treat different diseases in CCS categories. The top seven CCS disease categories in terms of average values are highlighted in different colors and described in the box.*

Source: Authors' calculation using CDC's WONDER and Blended Account database from BEA (2020)

In contrast, machine learning methods automatically partition data into mutually exclusive clusters based on the data's intrinsic structure. Our analysis focuses on above-average behavior, high health-care cost, or mortality rate. As a result, we prefer k-means clustering to other popular algorithms, such as Density-Based Spatial Clustering of Applications with Noise (DBSCAN) and Gaussian Mixture Models, which would combine the outliers in a separate cluster. We also let the data dictate the optimal number of clusters, k.[8] More specifically, we calculate

8. k-means algorithm tries to partition the data set into k pre-determined distinct subgroups. It assigns data points to a cluster such that the sum of the squared distance between the data points and the cluster's centroid is at the minimum. The less variation we have within clusters, the more similar the data points are within the same cluster.

the silhouette value, which measures how close each point in one cluster is to points in the neighboring clusters (Rousseeuw, 1987). The silhouette ranges from -1 to +1, where a high value indicates that the object is well matched to its cluster and poorly matched to neighboring clusters. If most objects have a high value, then the clustering configuration is appropriate. If most objects have a low or negative value, then the clustering configuration may have too many or too few clusters. The best number of clusters, k, maximizes the average silhouette values for all observations.

Before implementing the method, we normalize the mortality and medical expenses due to the different values' scales. In line with any machine learning algorithms, the k-means method involves a numerical minimization problem. To avoid local minima, we repeat the clustering process starting from different randomly selected points for centroids of clusters. We then choose the solution with the lowest total sum of distances among all the replicates.

Our results cluster the CCS disease categories into four classes. The summary statistics reported in Table 1 help describe them as follows:

- Class 1 consists of the 183 CCS disease categories that have neither a high cost nor a high mortality rate. We disregard this class as not relevant to our question.
- Class 2 consists of the 21 CCS disease categories with the most patients and the overall highest yearly medical expenses.
- Class 3 consists of the 23 CCS disease categories with the highest yearly cost per patient.
- Class 4 consists of the 5 CCS disease categories with the highest mortality rates.

Table 1: Clustering Results, Summary Statistics

Class	Deaths per 100,000 Individuals	Number of Patients (thousands)	Yearly Costs per Patient ($)	Total Yearly Expenses ($ in millions)	Number of CCS Disease Categories
1	1.7 (3.5)	3,034 (3,575)	1,915 (1,210)	3,745 (3,679)	183
2	1.9 (2.6)	25,116 (11,929)	1,099 (424)	24,283 (6,946)	21
3	4.6 (4.1)	444 (555)	9,992 (3,643)	4,017 (4,578)	23
4	58.2 (19.5)	5,024 (5,033)	5,647 (4,276)	14,516 (8,906)	5

Note: *This table shows a yearly average of four variables within a class. Deaths per 100,000 individuals are based on the CDC data from 2000 to 2018. The annual patient counts and health-care expenses within a class are averaged from 2000 to 2016. Standard deviations are in parentheses.*

Source: Authors' calculation using CDC's WONDER and Blended Account database from BEA (2020)

Step 3: NIH Funding of the Clusters 2, 3, and 4

We focus on the classes that capture either high medical costs (Classes 2 and 3) or high mortality rates (Class 4) and reconcile the CCS disease categories included in each class with NIH classification.

The categories represented in the RCDC of the NIH funding data differ from the CCS disease categories. Many RCDC categories are research areas (such as genetics or neuroscience), specific populations (such as pediatrics or minority health), or rare diseases (such as Pick's disease). Given RCDC's focus on categorizing research areas rather than classifying specific causes of death and medical expenses, the category definitions and delineations used in the RCDC do not always match those of the CCS. We have to match the 296 different RCDC research categories with the 49 CCS disease categories remaining. We use the extracted key medical terminologies to link RCDC and CCS disease categories.[9] Among the 49 CCS disease categories, we found 32

9. The differing characteristics of the two data sets mean that there are caveats to interpreting the matched data and that there are still judgment calls required in some cases to determine the best fits between the two sources.

CCS matched to the RCDC. The 32 CCS are matched to 92 different RCDCs, which comprise 25 percent ($509 billion) of NIH funding from 2008 to 2019.

We focus on 32 CCS disease categories for our final analysis, which identifies low-invested but high-cost and -mortality diseases. Classes 2 and 3 have 14 CCS disease categories, and Class 4 has 4.

Results

Figures 2 to 6 plot the NIH funding level, death rates, the number of patients, medical expenses per patient, and total medical expenses for the CCS disease categories included in Classes 2, 3, and 4. Table 2 and Appendix 4 provide more details on the CCS disease categories by listing the NIH categories included in each one.

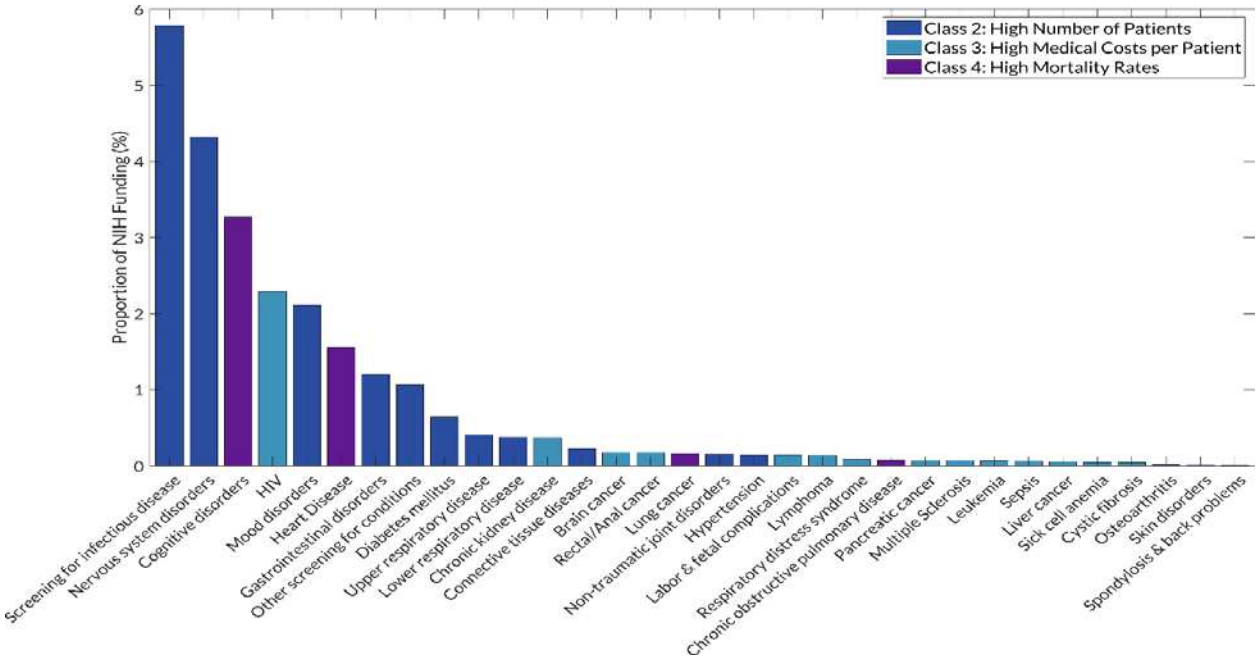
Overall, our results highlight three key findings. First, we do not find any clear pattern between funding allocations and the diseases' social and economic costs. Specifically, as Figure 2 indicates, NIH funding allocated to the top five CCS disease categories in terms of the aggregate funding amount (17.8 percent) almost doubles the funding distributed to all other 27 CCS disease categories with high medical costs or mortality (9.7 percent). Similarly, 12 CCS disease categories in Figure 2 have received less than 1 percent of NIH funding during the past 12 years. Figures 3 to 6 also show that NIH funding is disproportionately allocated to specific CCS disease categories within a class and does not align with the order of the death rates, patient counts, or costs per patient.

Second, Figure 5 shows that most CCS disease categories imposing high medical costs per patient have received little NIH funding. Specifically, all 14 CCS disease categories in Class 3 collected 3.8 percent of NIH funding, but that percentage becomes 1.5 percent if we remove funding focused on HIV infection.

Third, different measures of a disease's cost and social impact lead to different conclusions: Some diseases cause many deaths, while others impose a significant monetary burden on society. For example, a CCS category, immunizations and screening for infectious disease, has

the largest number of patients (46 million) but has the smallest yearly cost per patient ($487). Another CCS category, cystic fibrosis, is the costliest disease ($19,802 for an average annual cost per patient) but rare (less than 50,000 cases nationally), while cancer of bronchus (lung) is one of the deadliest diseases but causes relatively lower medical expenses to society. Our data-centric approach considers all the different measures while simplifying the decision-making process in identifying funding gaps of economic and socially costly diseases.

Figure 2: Proportion of NIH Funding toward CCS Diseases Categories



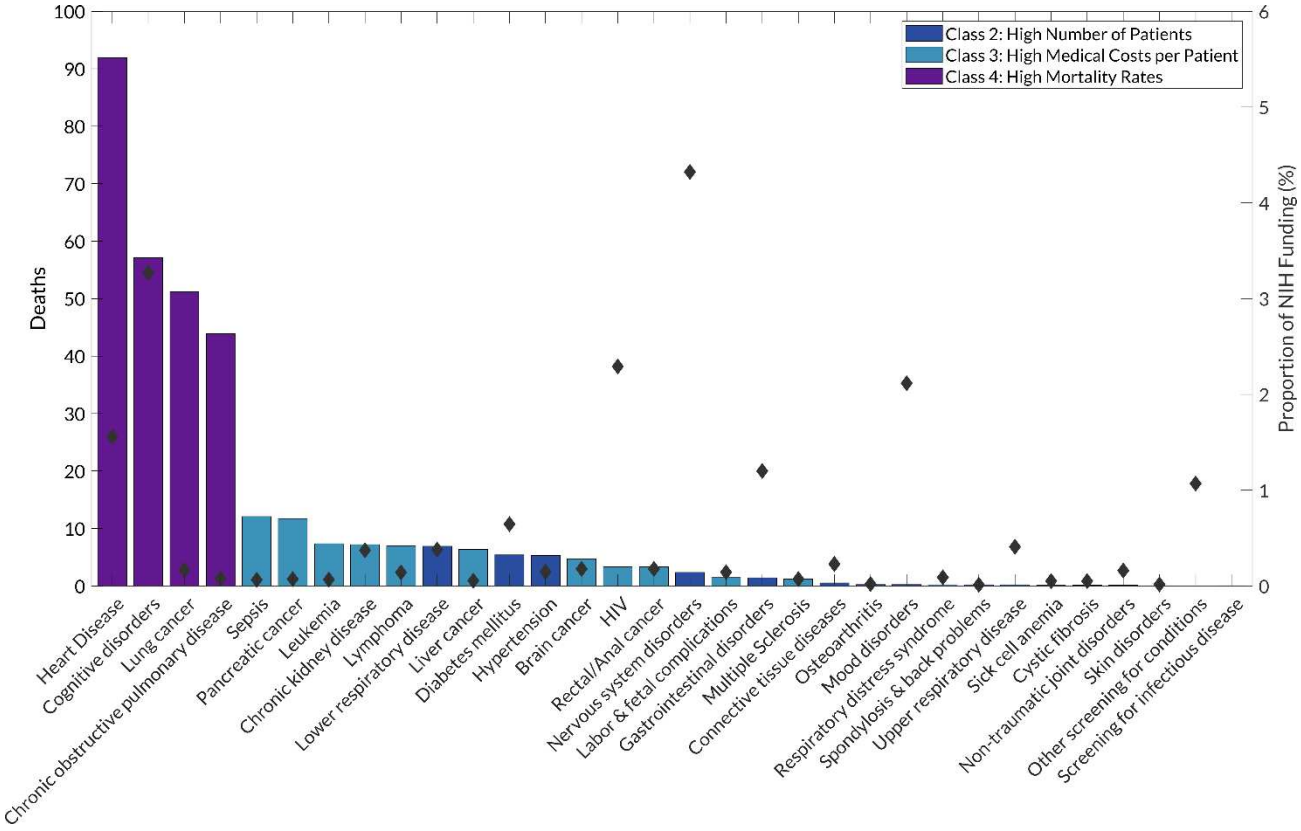
Note: *Proportions of NIH funding onto CCS disease categories are based on an aggregate amount of NIH funding from 2008 to 2019. Some CCS category names are abbreviated for better readability. See footnote for the full version.*[10]

[10] This table shows the full CCS disease category descriptions of the ones abbreviated by authors for better readability in Figures 2 to 6.

Abbreviated CCS Disease Categories	Full CCS Disease Category Descriptions
Brain cancer	Cancer of brain and nervous system
Lung cancer	Cancer of bronchus; lung
Liver cancer	Cancer of liver and intrahepatic bile duct
Rectal/Anal cancer	Cancer of rectum and anus
Chronic obstructive pulmonary disease	Chronic obstructive pulmonary disease and bronchiectasis
Heart Disease	Coronary atherosclerosis and other heart disease
Cognitive disorders	Delirium dementia and amnestic and other cognitive disorders
Diabetes mellitus	Diabetes mellitus without complication
Hypertension	Essential hypertension
Screening for infectious disease	Immunizations and screening for infectious disease
Lymphoma	Non-Hodgkin`s lymphoma
Connective tissue diseases	Other connective tissue disease
Gastrointestinal disorders	Other gastrointestinal disorders
Lower respiratory disease	Other lower respiratory disease
Nervous system disorders	Other nervous system disorders
Non-traumatic joint disorders	Other non-traumatic joint disorders

Source: Authors' calculation using CDC's WONDER, Blended Account database from BEA, and NIH's estimates of funding for RCDC (2020)

Figure 3: Deaths per CCS Disease Category

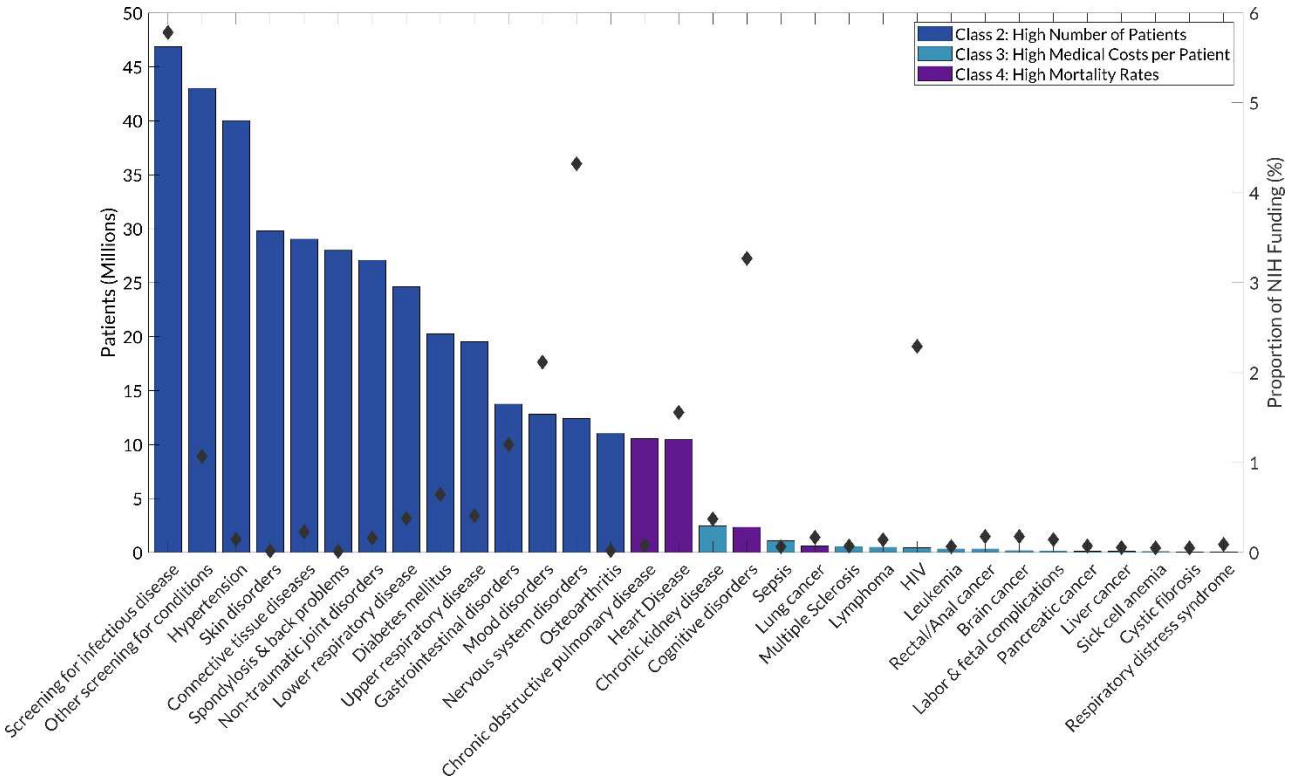


Note: Each bar indicates the average number of annual deaths per 100,000. Diamonds show the proportion of NIH funding allocated to a CCS disease category. Some CCS category names are abbreviated for better readability. See footnote 10 for the full version.

Source: Authors' calculation using CDC's WONDER, Blended Account database from BEA, and NIH's estimates of funding for RCDC (2020)

Other screening for conditions	Other screening for suspected conditions (not mental disorders or infectious disease)
Skin disorders	Other skin disorders
Upper respiratory disease	Other upper respiratory disease
Sepsis	Septicemia (except in labor)
Labor & fetal complications	Short gestation; low birth weight; and fetal growth retardation
Spondylosis & back problems	Spondylosis; intervertebral disc disorders; other back problems

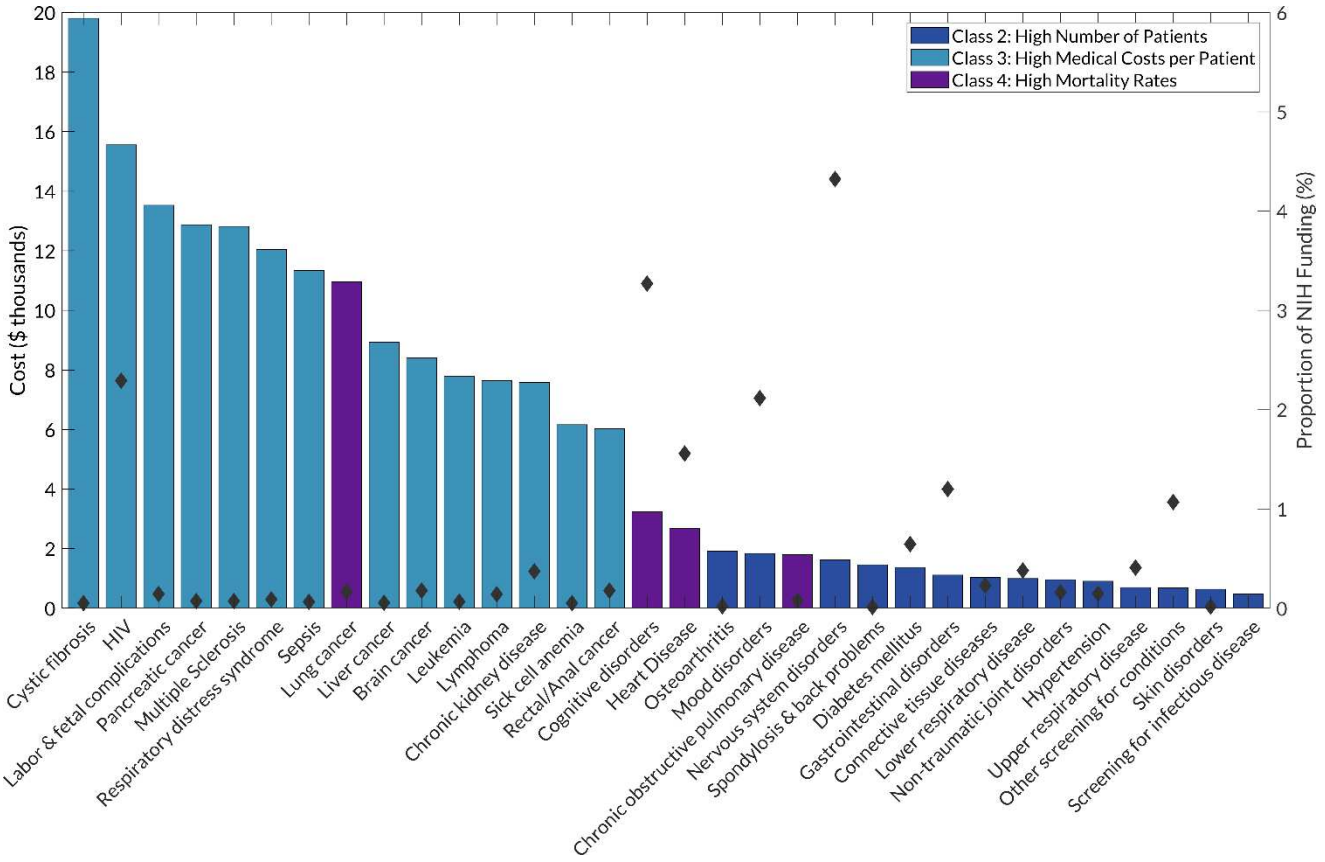
Figure 4: Number of Patients per CCS Disease Category



Note: *Each bar indicates the average number of patients in a year. Diamonds show the proportion of NIH funding allocated to a CCS disease category. Some CCS category names are abbreviated for better readability. See footnote 10 for the full version.*

Source: Authors' calculation using CDC's WONDER, Blended Account database from BEA, and NIH's estimates of funding for RCDC (2020)

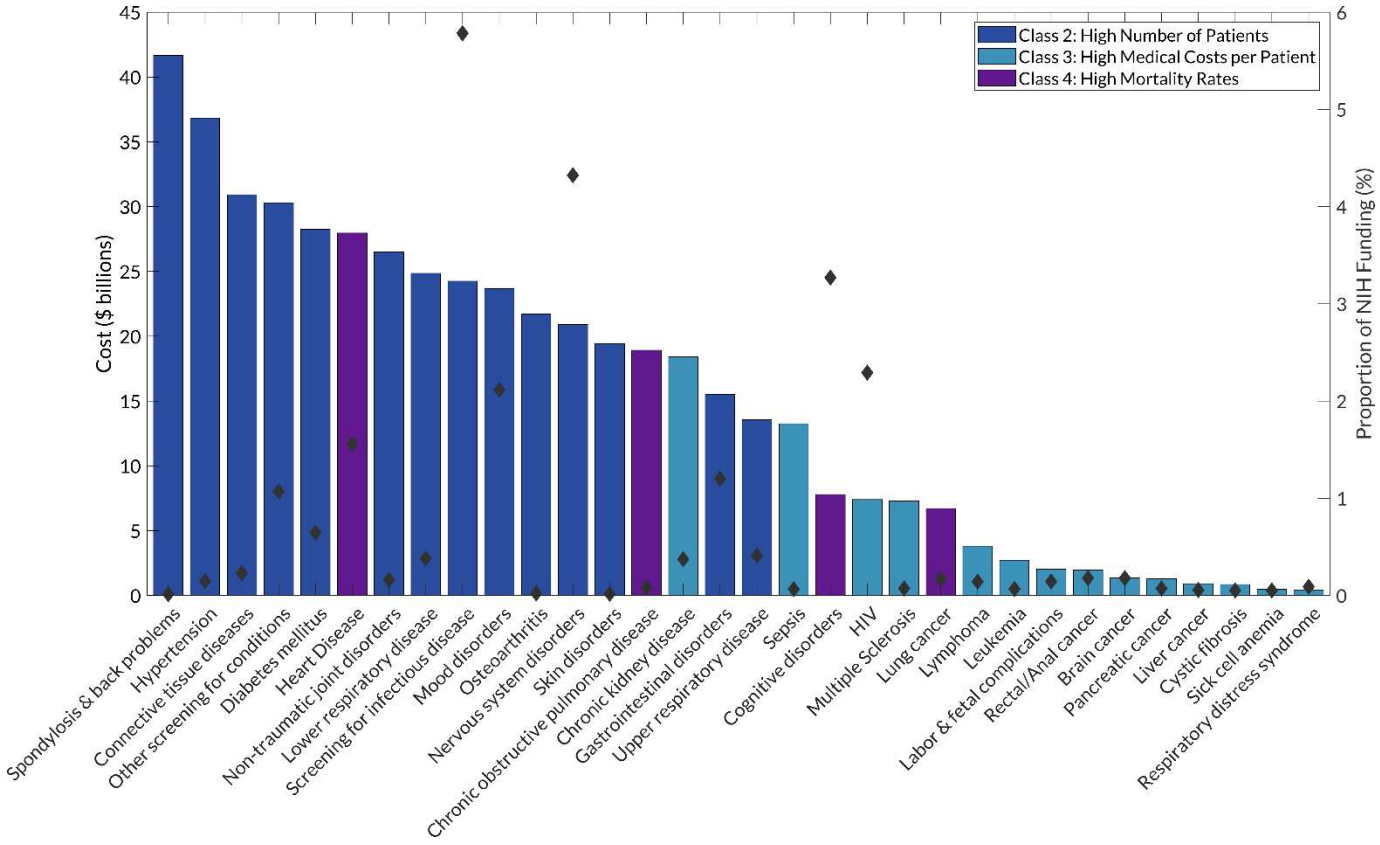
Figure 5: Medical Expenses per Patient per CCS Disease Category



Note: *Each bar indicates the average annual medical spending of patients having diseases associated with a CCS disease category. Diamonds show the proportion of NIH funding allocated to a CCS disease category. Some CCS category names are abbreviated for better readability. See footnote 10 for the full version.*

Source: Authors' calculation using CDC's WONDER, Blended Account database from the BEA, and NIH's estimates of funding for RCDC (2020)

Figure 6: Total Medical Expenses per CCS Disease Category



Note: Each bar indicates the yearly average total medical expenses spent on a CCS disease category. Diamonds show the proportion of NIH funding allocated to a CCS disease category. Some CCS category names are abbreviated for better readability. See footnote 10 for the full version.

Source: Authors' calculation using CDC's WONDER, Blended Account database from the BEA, and NIH's estimates of funding for RCDC (2020)

Table 2: CCS/NIH Disease Categories with the Most NIH Funding

Disease Categories		Funding Allocation	
CCS	NIH	NIH	CCS
Immunizations and screening for infectious disease	Infectious Diseases	2.9%	5.7%
	Emerging Infectious Diseases	1.4%	
	Immunization	1.1%	
	Vector-Borne Diseases	0.3%	
	Malaria Vaccine	<0.1%	
	Tuberculosis Vaccine	<0.1%	
Other nervous system disorders	Brain Disorders Neurodegenerative, Neuroblastoma, Myasthenia Gravis, Fibromyalgia	2.7%	4.3%
	Neurodegenerative	1.2%	
	Epilepsy	<0.1%	
	Peripheral Neuropathy	<0.1%	
	Spinal Cord Injury	<0.1%	
	ALS	<0.1%	
	Transmissible Spongiform Encephalopathy (TSE)	<0.1%	
	Neuroblastoma	<0.1%	
	Neurofibromatosis	<0.1%	
	Rett Syndrome	<0.1%	
	Spina Bifida	<0.1%	
	Charcot-Marie-Tooth Disease Injury	<0.1%	
	Fibromyalgia	<0.1%	
	Ataxia Telangiectasia	<0.1%	
	Myasthenia Gravis	<0.1%	
	Tourette Syndrome	<0.1%	
	Batten Disease	<0.1%	
Delirium, dementia, and amnestic and other cognitive disorders	Neurodegenerative	1.2%	3.3%
	Alzheimer's Disease	0.5%	
	Dementia	0.4%	
	Acquired Cognitive Impairment	0.4%	
	Alzheimer's Disease including Alzheimer's Disease-Related Dementias (AD/ADRD)	0.4%	
	Parkinson's Disease	<0.1%	
	Alzheimer's Disease-Related Dementias (ADRD)	<0.1%	
	Vascular Cognitive Impairment/Dementia	<0.1%	
	Huntington's Disease	<0.1%	

	Frontotemporal Dementia (FTD)	<0.1%
	Aphasia	<0.1%
	Lewy Body Dementia	<0.1%
	Pick's Disease	<0.1%

Disease Categories		Funding Allocation	
CCS	NIH	NIH	CCS
HIV infection	HIV/AIDS	1.8%	2.3%
	Vaccine-related (AIDS)	0.3%	
	Pediatric AIDS	0.1%	
Mood disorders	Mental Health	1.5%	2.1%
	Depression	0.3%	
	Mental Illness	0.2%	
	Serious Mental Illness	0.1%	
	Major Depressive Disorder	<0.1%	
	Bipolar Disorder	<0.1%	
Coronary atherosclerosis and other heart disease	Heart Disease	0.8%	1.6%
	Atherosclerosis	0.3%	
	Heart Disease—Coronary Heart Disease	0.3%	
	Stroke	0.2%	
	Congenital Heart Disease	<0.1%	
	Pediatric Cardiomyopathy	<0.1%	
Other gastrointestinal disorders	Digestive Disease	1%	1.2%
	Inflammatory Bowel Disease	<0.1%	
	Crohn's Disease	<0.1%	
	Digestive Diseases (Peptic Ulcer)	<0.1%	
	Digestive Diseases (Gallbladder)	<0.1%	
Other screening for suspected conditions (not mental disorders or infectious disease)	Vaccine-related	1.1%	1.1%
Diabetes mellitus without complication	Diabetes	0.6%	0.6%
Other upper respiratory diseases	Influenza	0.2%	0.4%
	Asthma	0.2%	
	Acute Respiratory Distress Syndrome	<0.1%	
	Allergic Rhinitis (Hay Fever)	<0.1%	

Note: The colors correspond to the classes: Class 2 is blue, 3 is orange, and 4 is yellow.

Source: Authors' calculation using CDC's WONDER, Blended Account database from the BEA, and NIH's estimates of funding for RCDC (2020)

Conclusion

This paper presents a data set that provides the medical cost, the number of patients, and the number of deaths per disease for 232 disease categories by identifying a set of common disease categories. We then sort the disease categories based on their medical expenses per patient, the total number of patients, total medical expenses, and mortality rates. Finally, we focus on the three groups that perform the worst on the four dimensions and estimate their NIH funding level.

We leverage text mining and machine learning methods to facilitate merging the data and simplify large dimensional decision space. We find that only 25 percent of disease categories causing high economic and social costs received more than 1 percent of research investments over 12 years. In addition, rare diseases imposing high medical costs per patient collected 0.3 percent of research investments on average over 12 years.

Data availability may not be the issue when assessing a disease's economic and social impact, but the ability to combine the existing information and process it, is. Our analysis reveals that a formalized procedure to define the correspondence between data sets is needed to successfully develop a metric that automatizes the assessment of diseases' cost, impact on society, and investment level. It also requires us to define the set of priorities that will guide how the data sets will be merged. In our case, we first focus on the costs (economic and social) to sort the diseases into four categories. Then we match the funding information for the three groups accounting for the costliest diseases.

References

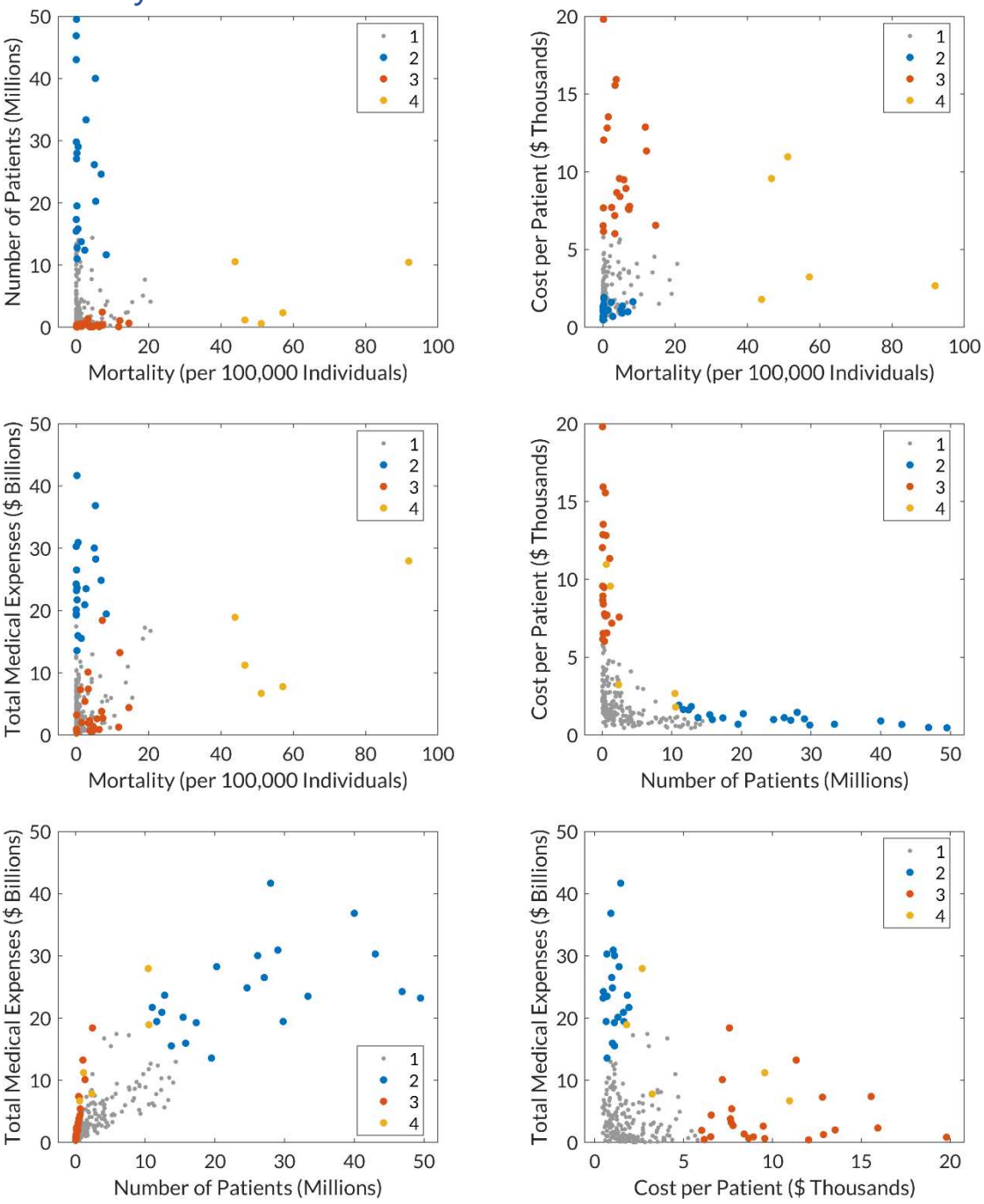
Dunn, Abe, Lindsey Rittmueller, and Bryn Whitmire. "Introducing the New BEA Health Care Satellite Account." *The Bureau of Economic Analysis*. Vol. 95 (2015). https://apps.bea.gov/scb/pdf/2015/01%20January/0115_bea_health_care_satellite_account.pdf

Packalen, Mikko, and Jay Bhattacharya. "NIH Funding and the Pursuit of Edge Science." *Proceedings of the National Academy of Sciences* 117 (22) (2020): 12011 LP–12016. https://doi.org/10.1073/pnas.1910160117.

Rousseeuw, Peter J. "Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis." *Journal of Computational and Applied Mathematics* 20 (C) (1987): 53–65. https://doi.org/10.1016/0377-0427(87)90125-7.

Wikipedia. "Edit Distance." n.d. https://en.wikipedia.org/wiki/Edit_distance.

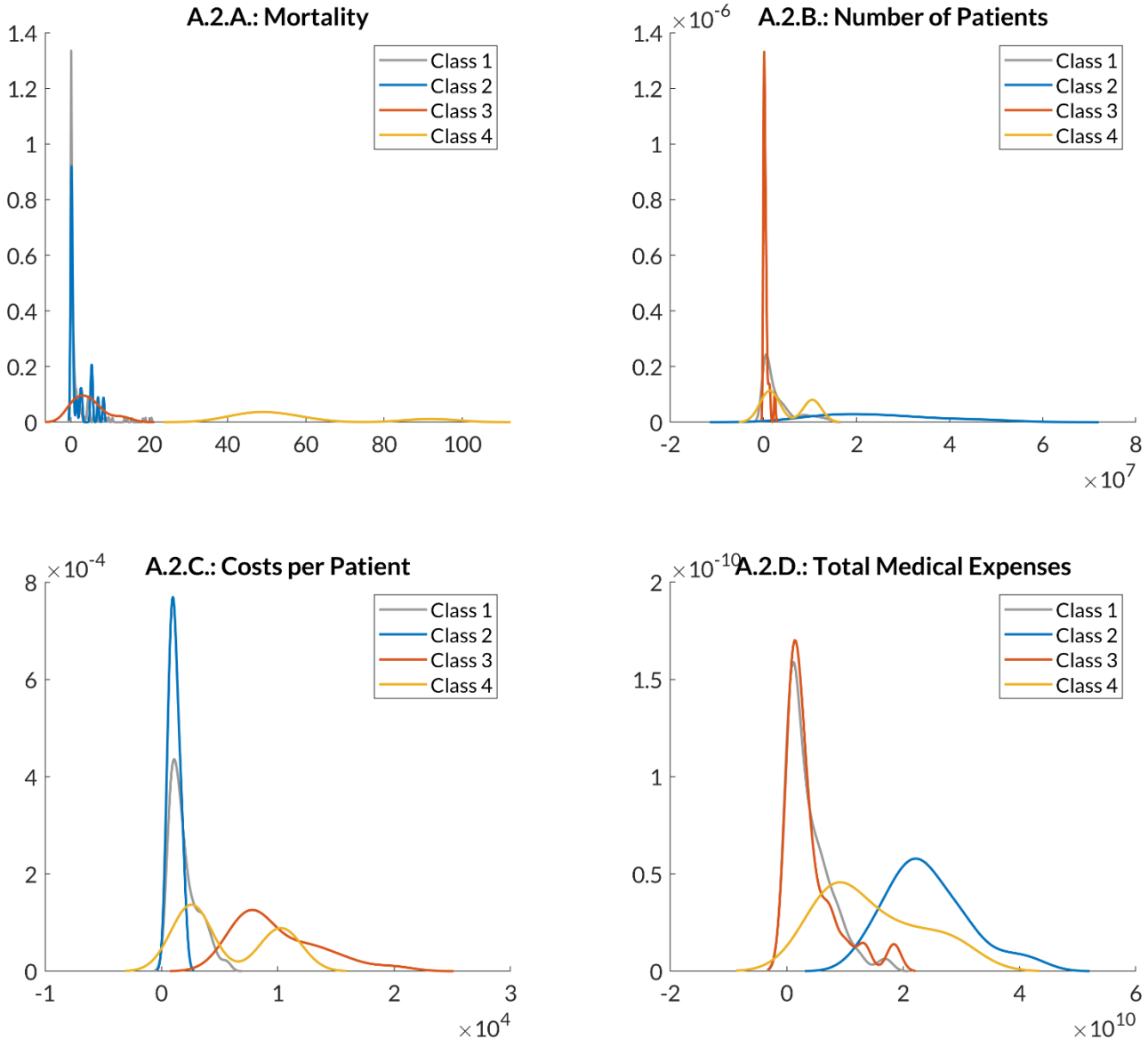
Appendix 1: Relationship between the Measures of Cost and Mortality



Note: Each panel plots comparisons between two variables among the four variables considered for this study (medical expenses per patient, the total number of patients, total medical expenses, and mortality rates). To get a better sense of what the classification based on the four variables entails, each data point in a class is denoted as a different color.

Source: Authors' calculation using CDC's WONDER and Blended Account database from BEA (2020)

Appendix 2: Density Functions



Note: Each figure in four panels indicates a kernel density of data in a class according to each of four variables considered for this study (mortality rates, the total number of patients, medical expenses per patient, and total medical expenses). The horizontal axis indicates a variable range, and the vertical axis denotes probability density. To get a better sense of what the classification based on the four variables entails, each line for a class is denoted as a different color.

Source: Authors' calculation using CDC's WONDER and Blended Account database from BEA (2020)

Appendix 3: List of CCS Disease Categories per Class

Class	CCS Disease Categories
2	Immunizations and screening for infectious disease; Other and unspecified benign neoplasm; Diabetes mellitus without complication; Disorders of lipid metabolism; Other nervous system disorders; Essential hypertension; Nonspecific chest pain; Cardiac dysrhythmias; Other upper respiratory infections; Other lower respiratory disease; Other upper respiratory disease; Other gastrointestinal disorders; Other skin disorders; Osteoarthritis; Other non-traumatic joint disorders; Spondylosis, intervertebral disc disorders, other back problems; Other connective tissue disease; Abdominal pain; Other screening for suspected conditions (not mental disorders or infectious disease); Residual codes, unclassified; Mood disorders
3	Septicemia (except in labor); HIV infection; Cancer of esophagus; Cancer of stomach; Cancer of colon; Cancer of rectum and anus; Cancer of liver and intrahepatic bile duct; Cancer of pancreas; Cancer of brain and nervous system; Non-Hodgkin's lymphoma; Leukemias; Multiple myeloma; Secondary malignancies; Cystic fibrosis; Sickle cell anemia; Multiple sclerosis; Aspiration pneumonitis, food/vomitus; Respiratory failure, insufficiency, arrest (adult); Appendicitis and other appendiceal conditions; Chronic kidney disease; Short gestation, low birth weight, and fetal growth retardation; Respiratory distress syndrome; Gangrene
4	Cancer of bronchus, lung; Acute myocardial infarction; Coronary atherosclerosis and other heart disease; Chronic obstructive pulmonary disease and bronchiectasis; Delirium dementia and amnestic and other cognitive disorders

Notes: Semicolons divide each CCS disease category, and commas separate related diseases within a category.

Source: Authors' calculation using CDC's WONDER and Blended Account database from BEA (2020)

Appendix 4: NIH Funding Percentage Distribution within Each CSS

Figure A.4.A.: Immunizations and Screening for Infectious Disease

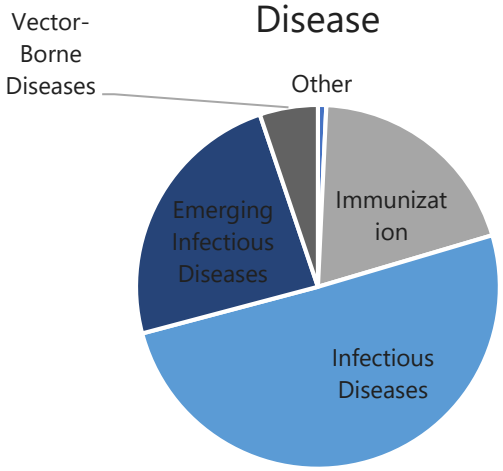


Figure A.4.B.: Other Nervous System Disorders

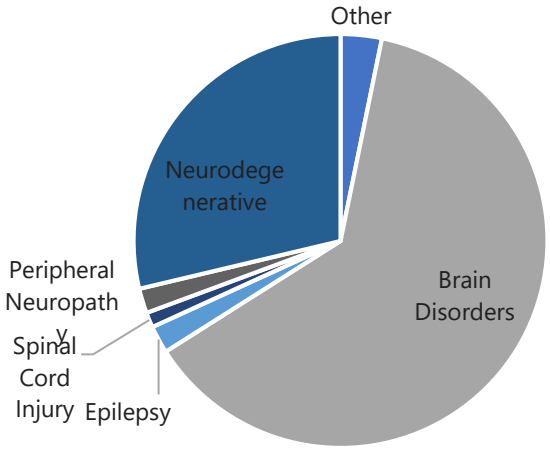


Figure A.4.C.: Coronary Atherosclerosis and Other Heart Disease

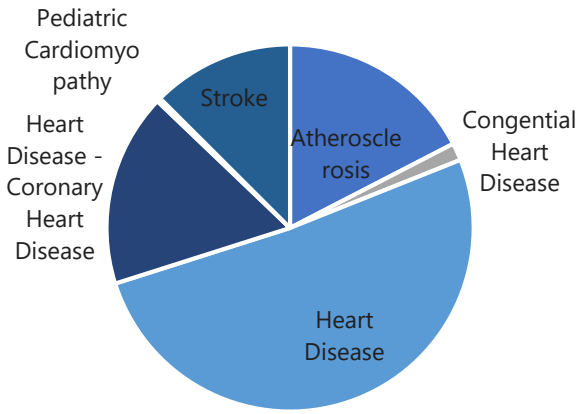


Figure A.4.D: Other Lower Respiratory Disease

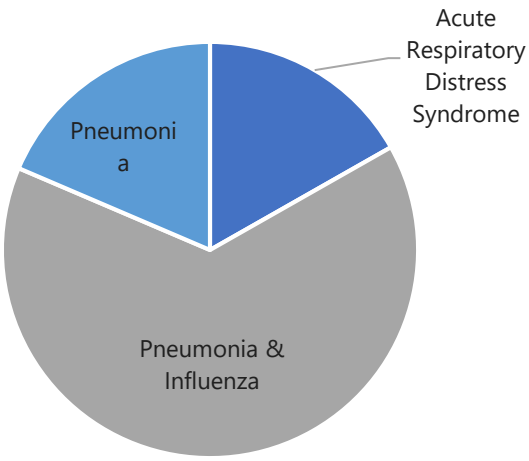


Figure A.4.E: Other Upper Respiratory Disease

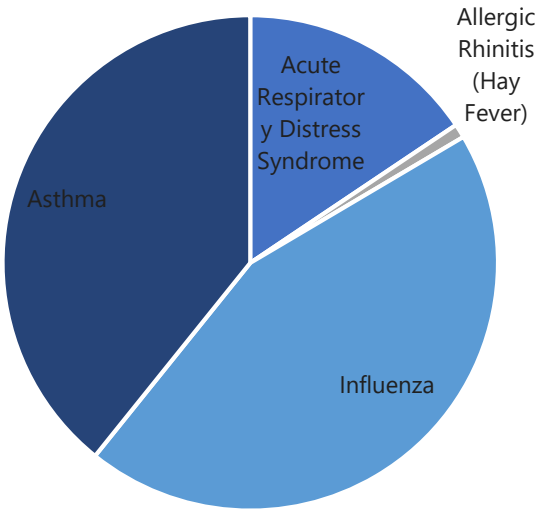


Figure A.4.F: Other Gastrointestinal Disorders

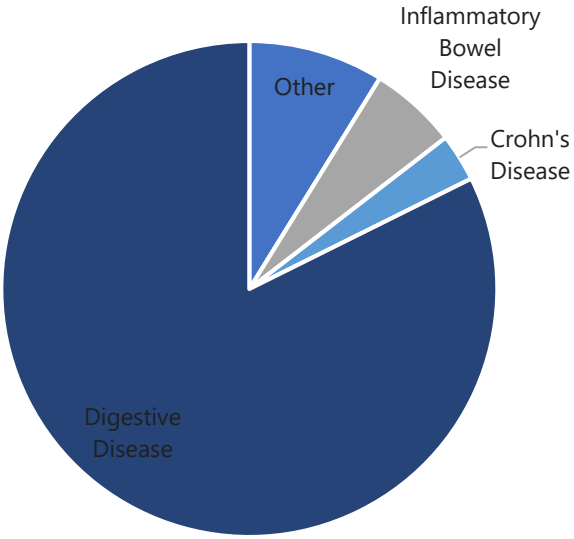


Figure A.4.G: Other Connective Tissue Disease

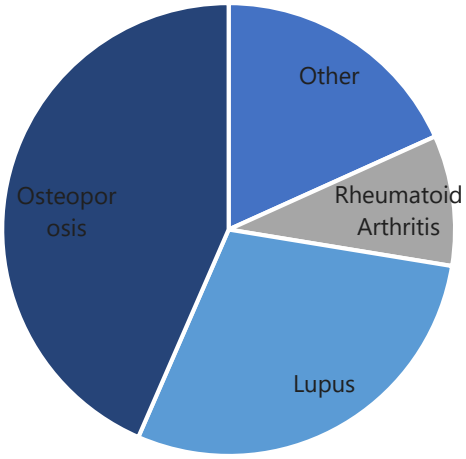


Figure A.4.H.: Delirium, Dementia, and Amnestic and Other Cognitive Disorders

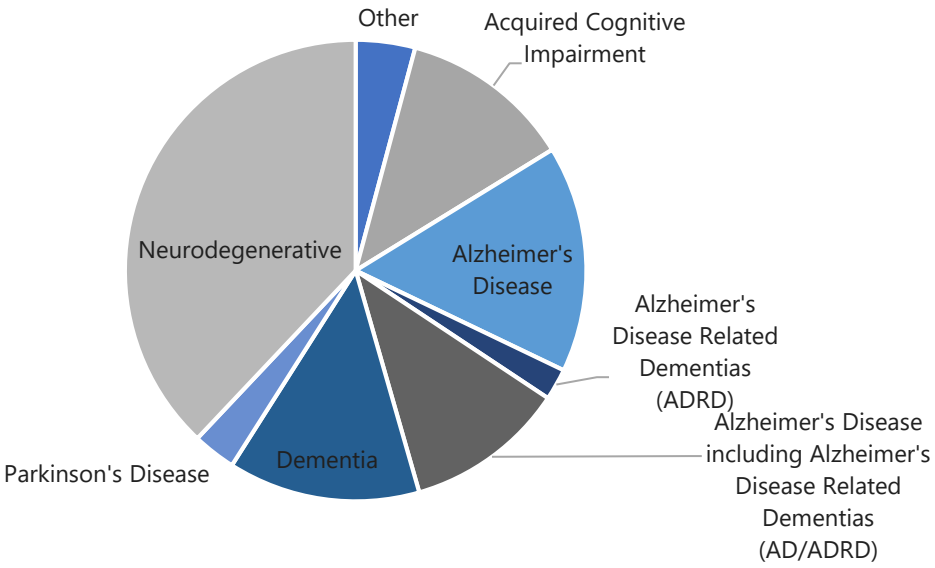


Figure A.4.I.: HIV Infection

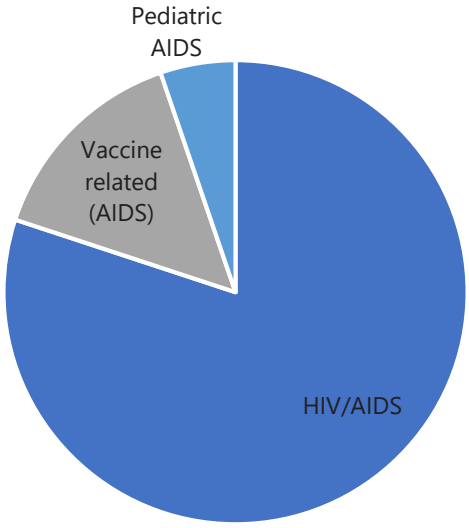


Figure A.4.J.: Mood Disorders

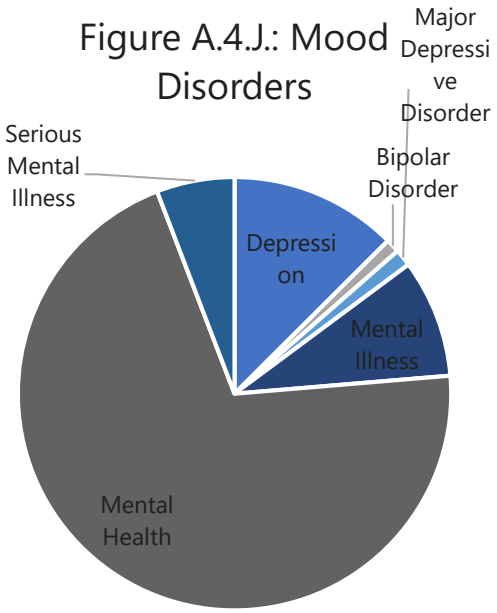


Figure A.4.K.: Respiratory Distress Syndrome

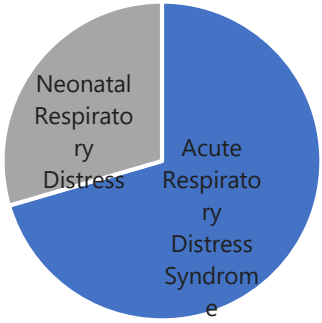


Figure A.4.L.: Spondylosis; Intervertebral Disc Disorders; Other Back Problems

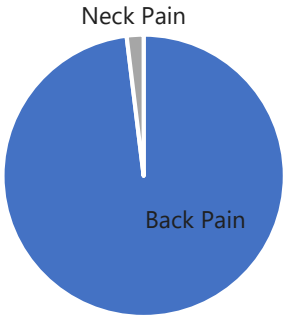


Figure A.4.M.: Chronic Kidney Disease

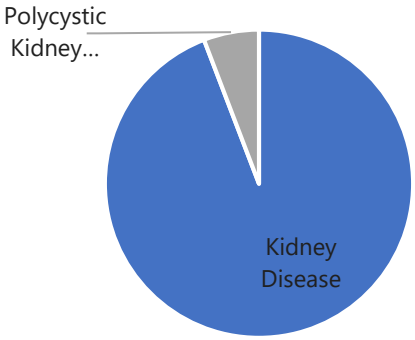


Figure A.4.O.: Chronic Kidney Disease, Other Non-Traumatic Joint Disorders

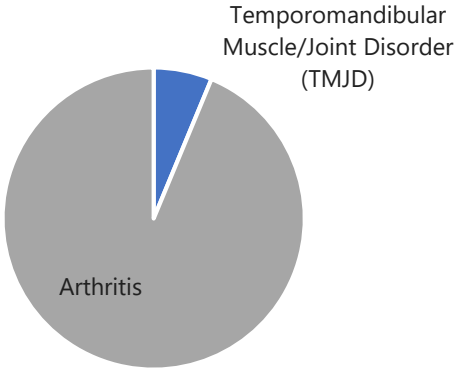
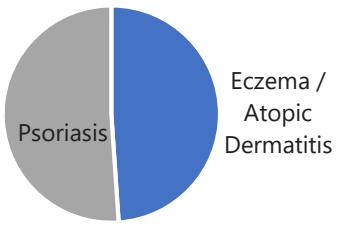


Figure A.4.N.: Chronic Kidney Disease Other Skin Disorders



Source for all figures in appendix: Bureau of Economic Analysis (BEA) and the National Institutes of Health (NIH)

About the Authors

Claude Lopez, PhD, is the head of the Research Department at the Milken Institute, where she leads data-driven efforts aimed at influencing global policy issues on International Finance, Health Economics, and Regional Economics. She is an active member of the T20 task force on international financial architecture for stability and development, and a contributor to W20 (Women 20), two advisory committees to the G20. Lopez has over 20 years of experience in academic and policy research in the US and abroad. Before joining the Institute, Lopez headed multiple research teams at the Banque de France, the nation's central bank, and was a professor of economics at the University of Cincinnati. She has an MS in econometrics from Toulouse School of Economics and a PhD in economics from the University of Houston.

Hyeongyul Roh, PhD, is a senior research analyst specialized in applied economics within the Research Department at the Milken Institute. The core of his research interests is to empirically assess the efficiency of markets and identify optimal economic decisions based on the empirical setting. His recent work focuses on energy industry organization and clean energy policy by employing economic theory and diverse machine learning methods. Before joining the Institute, Roh was a postdoctoral fellow at the Duke University Energy Initiative. Roh holds a PhD in economics from North Carolina State University and an MS in mathematical finance from Boston University.

Brittney Butler is a health economics research analyst within the Research Department at the Milken Institute. Her current work focuses on health equity and disparities within the United States, taking into account social, demographic, economic, and environmental factors. More specifically, she is identifying how the aforementioned factors interact with health outcomes through mapping and data analysis. Butler holds a bachelor's degree in integrative biology from the University of California, Berkeley and master of science in global health from the University of California, San Francisco.