Health Affairs Scholar, 2024, **2**(3), 1–7
https://doi.org/10.1093/haschl/qxae017
Advance access publication: February 14, 2024
Research Article

Health Affairs Scholar
EMERGING & GLOBAL HEALTH POLICY
OXFORD

American clusters: using machine learning to understand health and health care disparities in the United States

Diana M. Bowser[1,*], Kaili Maurico[1], Brielle A. Ruscitti[2], William H. Crown[2]

[1]Connell School of Nursing, Boston College, Chestnut Hill, MA 02467, United States
[2]Heller School for Social Policy and Management, Brandeis University, Waltham, MA 02454, United States
*Corresponding author: Connell School of Nursing, Boston College, Chestnut Hill, MA 02467. Email: bowserdi@bc.edu

Abstract

Health and health care access in the United States are plagued by high inequality. While machine learning (ML) is increasingly used in clinical settings to inform health care delivery decisions and predict health care utilization, using ML as a research tool to understand health care disparities in the United States and how these are connected to health outcomes, access to health care, and health system organization is less common. We utilized over 650 variables from 24 different databases aggregated by the Agency for Healthcare Research and Quality in their Social Determinants of Health (SDOH) database. We used *k*-means—a non-hierarchical ML clustering method—to cluster county-level data. Principal factor analysis created county-level index values for each SDOH domain and 2 health care domains: health care infrastructure and health care access. Logistic regression classification was used to identify the primary drivers of cluster classification. The most efficient cluster classification consists of 3 distinct clusters in the United States; the cluster having the highest life expectancy comprised only 10% of counties. The most efficient ML clusters do not identify the clusters with the widest health care disparities. ML clustering, using county-level data, shows that health care infrastructure and access are the primary drivers of cluster composition.

Key words: health disparities; machine learning; social determinants of health; American clusters.

Introduction

Health and health care access in the United States continue to be plagued by high levels of inequality. Despite the passage of one of the largest health care reforms in American history, the Affordable Care Act in 2010, inequality in key health metrics continues to increase in the United States. Access to formal health insurance in the United States increased to an all-time high of 91% in 2016. However, large disparities continue, with Texas still having 20% of their population uninsured while Massachusetts had an uninsurance rate of 3% in 2021.[1] Similarly, health status, as measured by life expectancy, has decreased for the third year in a row for the United States, and inequality, with regard to this same metric, has only become worse.[2-4] Most recently, these trends have been attributed mainly to COVID-19 and the aging population in the United States, yet the underlying causes and drivers of these inequalities remain complicated and are poorly understood.[5,6]

Murray et al[7] examined inequality in the United States in their 2006 paper titled, "Eight Americas: Investigating Mortality Disparities across Race, Counties, and Race-Counties in the U.S." Using data across 3144 US counties over the period 1982–2001, they showed significant disparity in the race-county combinations they created and termed these combinations the "Eight Americas." Since the publication of this paper, numerous other studies have been published describing similar disparities within and across the United States for chronic conditions,[8] life expectancy,[9] rural vs urban areas,[10,11] and most recently, structural racism and neighborhoods.[12] The COVID-19 pandemic exposed many additional inadequacies and a range of inequities in the US health care system, and deaths due to COVID-19 were disproportionately higher among Black individuals compared with the population overall.[13] While ML has been utilized to examine many aspects of clinical medicine and decision making in medicine in the United States, the use of ML as a research tool to understand health care disparities in the United States and how these are connected to health outcomes, access to health care, and health system organization is less common.[14] For the purposes of this paper, ML is defined as utilizing electronic data and computers to learn patterns and relationships to make prediction and decisions.[15]

Life expectancy and longevity vary widely by state as well as by race and ethnicity, and these inequalities continue to grow. For example, life expectancy in Connecticut and Oklahoma was 71.1 years in 1959. However, by 2017, Connecticut had gained 9.6 years in life expectancy whereas Oklahoma only gained 4.7 years.[16] Prior studies have also shown that estimated life expectancy was lower for Black populations than for White populations in 87% of counties, as of 2022.[17] The COVID-19 pandemic has had numerous consequences for population health, including a notable reduction in life expectancy. More specifically, the COVID-19 pandemic has resulted in an estimated 39% increase in the Black–White life expectancy gap and a reduction in the Latino life expectancy advantage.[5]

Inequalities in the United States are not only related to health status and the US health care system but to social determinants of health (SDOH) as well. Health disparities are found by education, race, ethnicity, sex, sexual orientation, and place of residence.[18] SDOH can be attributed to differences in access and use

Received: October 3, 2023; Revised: December 19, 2023; Accepted: February 12, 2024

of quality health care services; health risk behaviors, including tobacco use, obesity, and lack of access to healthy foods; social and built environment; and socioeconomic and living conditions.[19,20] Decades of research have found significant socioeconomic, racial/ethnic, and occupational disparities in mortality and life expectancy among Americans, identifying population groups[21,22] with social inequalities in health continuing to be related to key SDOH[23] as well as geographic variation in health care practice.[24,25]

Given the continued problem of inequality in the United States, the added complexity of the SDOH, and the magnitude of data that are now available across multiple domains, it would be impossible and most likely inaccurate to use a priori assumptions to create American clusters as developed by Murray et al. We initially intended to use ML to identify the 8 optimal ML clusters and compare these with those described by Murray et al. As described in this paper, we were surprised to find that the optimal number of clusters identified by ML was only 3. The results reveal useful new insights into the opportunities and challenges associated with using ML methods to examine health disparities questions. This paper presents an ML-based methodology that can be used as a research tool to explore inequalities in the United States.

Data and methods

Data and clustering

Machine learning clustering methodology was used to cluster over 650 variables from 24 different databases aggregated by the Agency for Healthcare Research and Quality (AHRQ) in their AHRQ SDOH database. The AHRQ SDOH database consisted of county-level data across 6 key SDOH domains (social, economic, geographic, education, physical infrastructure, and health care contexts) from 2009 to 2020. For the purposes of this analysis, the years 2019 and 2020 data were utilized. The number of data sources and variables fluctuated due to the diversity of collection periods for the supporting data, but a core set of variables supporting all 6 SDOH domains was consistently available across the years 2019 and 2020. The only variable extracted from another data source was life expectancy, as the AHRQ SDOH database did not have this variable for all time periods. Life expectancy was extracted from the US Small Area Life Expectancy Estimates Project from the Centers for Disease Control and Prevention for the years 2015–2020.[26]

A non-hierarchical clustering method, the *k*-means approach, was used to cluster the AHRQ SDOH into clusters, grouping the data by defined center points, using over 650 variables from 24 different databases.[27] *k*-Means is an unsupervised ML algorithm for clustering data into a set of *k* groups, minimizing the intracluster variation. The Canberra method, the sum of a series of fraction differences between a pair of objects, was used to calculate the distance between observations. While other methods were tested (Euclidean distance,[28] Manhattan, and Sorensen), the Canberra, a weighted version of the Manhattan distance,[29] was chosen as most appropriate for the AHRQ SDOH dataset due to the high dimensionality of the data and the presence of multiple outliers within the included variables.

Optimal clusters

The elbow method, a systematic evaluation of clustering solutions that analyzes the within-cluster sum of squares (WCSS) to identify a point of diminishing returns in WCSS reduction, was utilized to determine the optimal number of clusters in the context of the *k*-means clustering algorithm.[30] The optimal number of clusters was determined by running *k*-means and calculating the WCSS, representing the sum of square distances between datapoints and the calculated cluster centroids, for a range of *k*-values from 1 to 30. A standard elbow graph assists in determining the most efficient number of clusters. As shown in Supplemental Material 1, the most efficient number of clusters was determined at the "elbow" of 3 clusters and there was a flattening of the curve at 6 clusters. For the purpose of the analysis presented below, we report clustering results for 3 and 6 clusters. The 3 cluster results are presented in the main body of the paper and the 6 cluster results are presented in Supplemental Materials 2–5.

Principal factor analysis

In order to determine the main contributors to each of the clusters identified with the elbow method, an index value for each of the AHRQ SDOH domains was created for each county using principal factor analysis. The principal factor analysis identified linear combinations (factors) of the variables within each of the AHRQ SDOH domains, capturing the maximum amount of variance in domain variables. Within each factor, the principal factor analysis created an index value, based on the strength of the relationship between the factor and each variable using the loadings for that variable on the factor.

We used a modified health system framework[31] to guide the cluster analysis. In their framework, Roberts characterized a health system as composed of health system inputs referred to as control knobs, intermediate indicators of health system performance and health system outcomes. A comprehensive health system analysis would examine in detail these different aspects of the health system for the ML Americas. A comprehensive health system analysis was outside the scope of the present study. Consequently, the system analysis utilized for this paper examined 3 main domains with the AHRQ SDOH health care context: a health system input domain, an access to health care services domain, and 1 outcome indicator, life expectancy, capturing the level of health status for the population. Supplemental Material 6 summarizes the framework used in the analysis.

Each cluster was categorized in Table 1 using a variable for each of the SDOH of health domains and 2 variables to capture the health system: one for health system infrastructure and one for health care access. The variable with the highest loading from the principal factor analysis was chosen for each SDOH domain as follows: social (percentage of foreign-born residents), economic (percentage <137% of the poverty line), education (percentage with any postsecondary education), physical infrastructure (average temperature in March), and geographic (Rural-Urban Index, with 1 indicating the most urban and 9 indicating the most rural). Health care infrastructure was captured by the number of nurse practitioners with a National Provider Identifier per 10 000 population. Health care access was captured by the total number of marketplace enrollees with household incomes between >250% and 300% of the Federal Poverty Level per 10 000 population.

Logistic regression classification

Finally, a logistic regression classification approach, using a statistical model that predicted 2 categories of categorial outcomes, was used to determine the primary drivers of cluster

Table 1. Cluster descriptions and summary measures of social determinants of health variables for 3 clusters.

Cluster	General description	Average population (N)	Total counties (% total)	Percentage below 137% of the poverty line (%)	Percentage of population with any postsecondary education (%)	Number marketplace enrollees, household income >250% to 300% of FPL per 10 000	Nurse practitioners with NPI per 10 000	Average temperature in March (°F)	Percentage of foreign born residents (%)	Rural-urban	Life expectancy (years)
1	Rural, poorest, least educated, with lowest health access and health infrastructure, with the lowest life expectancy and least immigration	12 915	1673 (53%)	23.7%	50%	112.0	6.1	48.5	3.8	6.6	76.8
2	Urban, least share below poverty, most educated, most insured, best health infrastructure, and most immigration	696 752	310 (10%)	17.9%	65.40%	25.9	10.6	51.5	14.5	1.5	79.4
3	Suburban, average poverty, education, insurance and health infrastructure, relatively low immigration	76 721	1161 (37%)	20.9%	55.80%	40.9	8.9	49.0	5.2	3.7	77.4

Abbreviations: FPL, Federal Poverty Level; NPI, National Provider Identifier.
Cluster description with summary results for variables representing each of the social determinants of health domains and the 2 health system categories (health system infrastructure and health care access). Summary statistics for the variable with the highest loading as part of the principal factor analysis are included.

determination, utilizing the index value created for each county, for each of the 5 SDOH domains and 2 health system domains (inputs and access). For this analysis, the dichotomous categories were the membership in each individual predetermined cluster generated using *k*-means, tested against all nonmember counties.

Results

Cluster analysis results

Table 1 provides a summary of each of the 3 distinct clusters utilizing the most efficient number of clusters as determined with the elbow method. As shown in Table 1, cluster 2 comprises the largest counties in terms of population size, lowest share below the poverty line, most educated, most insured, best health infrastructure, most immigration, and highest life expectancy. Cluster 1 is the smaller, poorest, rural, lowest health care access and health infrastructure, with the least migration and the lowest life expectancy. The descriptive variables provided in Table 1 for each SDOH domain, provided through the principal factor analysis, show results for variables not normally chosen to describe geographic variation across the United States—for example, the most important contributor to physical infrastructure is average temperature of the county in March. In addition, the 3 ML clusters demonstrate a large inequality in distribution, with the largest cluster in terms of population comprising only 10% of all counties.

Figure 1 shows the result for the indicator for health status, life expectancy in years for each of the 3 America clusters for the single year, 2020 (panel A), as well as over time (panel B). As shown in Figure 1A, those in cluster 2 (only 10% of US counties), with the highest average life expectancy in 2020 (79.4 years), have an average life expectancy 2.6 years greater than those in cluster 1 (76.8 years). In addition, as shown in Figure 1B, this differential is consistent over time. The results for the 6-cluster analysis (Supplemental Material 2) show a wider level of inequality, with a 4.8-year differential in life expectancy between cluster 6 and cluster 3. As the clusters increase so does the inequality among clusters, with a life expectancy differential reaching 9.0 years for 50 clusters (see Supplemental Material 7). The most efficient results according to the elbow method do not identify the clusters with the largest disparities.

Logistic regression classification results

Table 2 shows the results of the logistic regression classification approach that is used to determine the primary drivers of each cluster, examining specifically the contribution of health care access and health system infrastructure indices. Table 2 shows that health care access and health infrastructure are the 2 most important factors in cluster composition, especially for cluster 2, with all other SDOH domains also contributing significantly. Neither health care domain, access, nor infrastructure contribute positively to clusters 1 or 3, clusters that make up 90% of the counties in the United States. Lower health care access is a contributing factor to cluster 3 and less

Figure 1. Life expectancy, in years. Panel A shows the average life expectancy, 2020. Panel B shows the average life expectancy over time from 2015 to 2020, by year and cluster. Life expectancy is the mean value for all of the counties included in the dataset for all years ($n = 1673$, cluster 1; $n = 310$, cluster 2; $n = 1161$, cluster 3).

Table 2. Logistic regression examining AHRQ SDOH domain contribution to each cluster.

Index value by AHRQ SDOH domain	Cluster 1	Cluster 2	Cluster 3
Health care access	0.80	4.97***	0.66***
Health care infrastructure	0.003***	19.97***	1.04
Economic	1.00	1.24*	1.04
Education	0.77**	0.64***	1.44***
Geography	2.14***	1.65***	0.45***
Physical infrastructure	0.92	1.16*	0.96
Social	0.85	1.78***	0.76***

Abbreviations: AHRQ, Agency for Healthcare Research and Quality; SDOH, social determinants of health.
Logistic regression model using AHRQ SDOH domain index values for each of the 3 clusters ($n = 3144$), reporting coefficients and significance level at $^{*}P = .05$; $^{**}P = .01$; $^{***}P = .001$.

health care infrastructure is a contributing factor to cluster 1. Other positively contributing factors to cluster definition include education for cluster 3 and geography for cluster 1. Negatively contributing factors include less education for cluster 1 and lower social and geographic index values.

Figure 2 shows the map of the US counties that correspond to the 3 ML American clusters. Cluster 2, composed of the highest health care access and infrastructure, comprises mainly urban areas. Cluster 1 represents the Midwest and parts of the deep South. Cluster 3 represents suburban areas and higher income rural areas.

Discussion

This is one of the few studies to utilize ML as a research tool to understand important patterns of health system utilization, infrastructure, and outcomes in the United States utilizing county-level data with over 650 variables. The results are important as we continue to understand the utility of ML as it relates to the health of a population. The results show that the most efficient ML cluster models do not fully capture the maximal health inequalities and that, for the clusters that are created, health care infrastructure and health care access are the primary drivers of cluster composition. This analysis has provided useful methods and information for researchers, policymakers, implementers, and government officials planning system changes in the United States.

While the analyses presented above are mainly driven by machines rather than by data-driven decisions by researchers, the results can be compared with several other important studies. Murray et al[7] in their landmark paper that outlined 8 distinct Americas found a much larger disparity in life expectancy between the group with the highest life expectancy ("Asian females") and lowest life expectancy ("high-risk urban Black"), a disparity of 20.7 years in 2001. The disparity in life expectancy shown in the 3 cluster ML results presented above is not as large as in Murray et al (2.6 years difference in life expectancy between cluster 2 and cluster 1). However, this is an artifact of choosing the most efficient clustering, and as we add more clusters (see 6 cluster results in Supplemental Materials 2–5), the disparity in life expectancy increases. Dyer et al[12] used a sophisticated methodology to create a structural racism index using census tract data and found a 9.6-year difference in average life expectancy (2010–2015). Like Murray et al, they used data-driven decisions to create their index measures. To better compare with the ML clusters created in this study it would be interesting to see geographic mapping of their indices and what contributes to the variation in the indices.

An innovative aspect of the analysis is a principal factor analysis that captures key drivers of cluster composition, focusing on 2 key drivers: health care inputs and health care access. The health care infrastructure index is the largest contributor to cluster composition and most important for the cluster with the highest life expectancy. The main drivers of the health care infrastructure index include health care system components that are clearly important to health outcomes, such as hospitals; health care facilities including federally qualified health centers; and workforce. Newer methods in ML are utilizing factor analysis as a data-reduction tool when building the clusters.[32]

Adding these health system inputs into the analysis is extremely important as the United States struggles with hospital closures

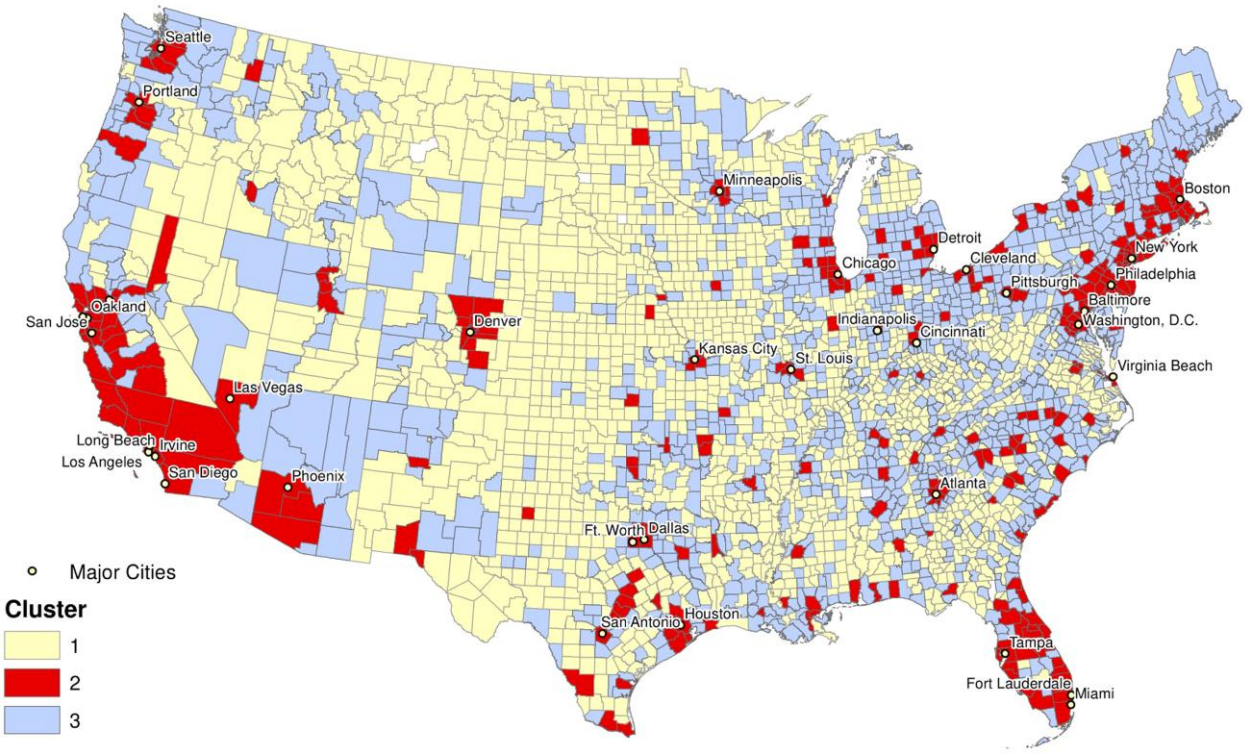

Figure 2. Map of US counties for the machine learning 3-clusters analysis. Note that the clusters, although showing some tendencies for geographic concentration, are not defined by geography. For example, some counties in the upper peninsula of Michigan and the panhandle of Texas are both assigned to cluster 1 based on the underlying patterns of the 650 county variables, not geography. Maps are created using data for all counties in the dataset ($n = 1673$, cluster 1; $n = 310$, cluster 2; $n = 1161$, cluster 3).

and workforce burnout.[33,34] It is even more important to note that one of the categories of health care workers that contributes to higher health care infrastructure are nurses and nurse practitioners, some of the main "heroes" during the COVID-19 pandemic and an essential part of patient care.[35,36] The authors do want to acknowledge, however, that efforts to improve health care inputs alone do not guarantee improved access or improved health outcomes for populations.[37]

The other key driver of cluster composition is the health care access index, mainly driven by access to either public, private, or marketplace insurance. This result is not surprising given the strong link between insurance coverage and health care access and outcomes.[38-41] It is important to note that the analyses related to health system drivers (infrastructure and access) are limited by data availability and, as such, there may be important variables that are missing in the analysis.

The methods and results described in this paper illustrate the usefulness of ML for understanding health disparities. But the results also illustrate the difficult tradeoff between efficiency and equity. The most efficient ML clusters do not identify the clusters with the widest health care disparities. The difference between the cluster 3 results (shown above) and the cluster 6 results (Supplemental Materials 2–5) shows that, as more clusters are created, minimal efficiency is gained but more inequality is seen. Future directions in ML equity analysis could use ML to identify dimensions underlying maximum inequality in health care access and outcomes.

There are a number of potential policy implications related to both the methods and the results presented above. First, the results above are presented to illustrate how ML methodology can be applied to understand disparities. In future research, the methodology can be utilized to identify and understand clusters with large inequalities. For example, a policymaker might want to create upwards of 100 clusters, which may be difficult to create manually, but which could be informative to understand disparities at the county level. Second, while this analysis has utilized county-level data, ML analysis conducted at a more granular geographic level such as zip codes or census tracts would be more useful to examine even more detailed results. Detailed SDOH data are also available at these levels, although the number of variables available declines with each level of increasing geographic granularity. Third, ML cluster analyses could also be linked with patient outcomes to enhance even further the link to health services. The methods presented above lay out the foundations for these next steps. We used unsupervised ML methods to identify the clusters in this paper. If understanding disparities in a particular measure, such as maternal mortality, were the goal, a combination of supervised and unsupervised methods might be particularly useful. For example, one could begin by using a lasso regression to model all the predictors of maternal mortality at a particular geographic level and then use the set of statistically significant variables as the starting point for the unsupervised cluster approach described in this paper. Such clusters would inherently be related to the variation in maternal mortality.

The analysis presented above combines ML with health services and health system research. As has been noted by many, we have a plethora of health care data in the United States, but they are often not utilized appropriately to understand our largest system issues. This is an attempt to use ML in the broader health system space to advance this field as others have done

within smaller systems[42,43] and to improve predictive, clinical analytics.[44,45] The importance of this work is that it begins to bridge the gap between systems research and clinical research, with a key next step focus on combining ML and cluster analysis at the system level with individual-level claims and/or electronic medical record data—analyses that the authors believe are essential to bridging the gap between clinical outcomes and key SDOH.

Conclusion

This paper builds upon the prior literature and applies principles of ML to move the research forward to better understand disparities in the United States and how these are connected to health outcomes, access to health care, and health system organization. The results show that using the most efficient ML cluster models does not fully capture the maximal health inequalities and that ML clustering, using county-level data, shows that health care infrastructure and health care access are the primary drivers of cluster composition.

Supplementary material

Supplementary material is available at *Health Affairs Scholar* online.

Conflicts of interest

Please see ICMJE form(s) for author conflicts of interest. These have been provided as Supplementary materials.

Notes

1. Tolbert J, Drake P, Damico A. Key facts about the uninsured population. KFF. Published 2022. Accessed July 24, 2023. https://www.kff.org/uninsured/issue-brief/key-facts-about-the-uninsured-population/
2. Dwyer-Lindgren L, Mokdad AH, Srebotnjak T, Flaxman AD, Hansen GM, Murray CJL. Cigarette smoking prevalence in US counties: 1996-2012. *Popul Health Metr*. 2014;12(1):5. https://doi.org/10.1186/1478-7954-12-5
3. Kulkarni SC, Levin-Rector A, Ezzati M, Murray CJL. Falling behind: life expectancy in US counties from 2000 to 2007 in an international context. *Popul Health Metr*. 2011;9(1):16. https://doi.org/10.1186/1478-7954-9-16
4. Wang H, Schumacher AE, Levitz CE, Mokdad AH, Murray CJL. Left behind: widening disparities for males and females in US county life expectancy, 1985–2010. *Popul Health Metr*. 2013;11(1):8. https://doi.org/10.1186/1478-7954-11-8
5. Andrasfay T, Goldman N. Reductions in 2020 US life expectancy due to COVID-19 and the disproportionate impact on the Black and Latino populations. *Proc Natl Acad Sci U S A*. 2021;118(5):e2014746118. https://doi.org/10.1073/pnas.2014746118
6. Woolf SH, Schoomaker H. Life expectancy and mortality rates in the United States, 1959-2017. *JAMA*. 2019;322(20):1996-2016. https://doi.org/10.1001/jama.2019.16932
7. Murray CJL, Kulkarni S, Ezzati M. Eight Americas: new perspectives on U.S. health disparities. *Am J Prev Med*. 2005;29(5):4-10.
8. Woolf SH, Aron LY, Dubay L, Simon SM, Zimmerman E, Luk K. How are income and wealth linked to health and longevity? Published online June 4, 2016. Accessed July 15, 2023. https://policycommons.net/artifacts/632420/how-are-income-and-wealth-linked-to-health-and-longevity/1613697/
9. Bor J, Cohen GH, Galea S. Population health in an era of rising income inequality: USA, 1980–2015. *Lancet*. 2017;389(10077):1475-1490. https://doi.org/10.1016/S0140-6736(17)30571-8
10. Abedi V, Olulana O, Avula V, et al. Racial, economic, and health inequality and COVID-19 infection in the United States. *J Racial Ethn Health Disparities*. 2021;8(3):732-742. https://doi.org/10.1007/s40615-020-00833-4
11. Anderson TJ, Saman DM, Lipsky MS, Lutfiyya MN. A cross-sectional study on health differences between rural and non-rural U.S. counties using the county health rankings. *BMC Health Serv Res*. 2015;15(1):441. https://doi.org/10.1186/s12913-015-1053-3
12. Dyer Z, Alcusky MJ, Galea S, Ash A. Measuring the enduring imprint of structural racism on American neighborhoods. *Health Aff (Millwood)*. 2023;42(10):1374-1382. https://doi.org/10.1377/hlthaff.2023.00659
13. van Dorn A, Cooney RE, Sabin ML. COVID-19 exacerbating inequalities in the US. *Lancet*. 2020;395(10232):1243-1244. https://doi.org/10.1016/S0140-6736(20)30893-X
14. Padula WV, Kreif N, Vanness DJ, et al. Machine learning methods in health economics and outcomes research—the PALISADE checklist: a good practices report of an ISPOR task force. *Value Health*. 2022;25(7):1063-1080. https://doi.org/10.1016/j.jval.2022.03.022
15. Mohri M, Rostamizadeh A, Talwalkar A. *Foundations of Machine Learning*. 2nd ed. MIT Press; 2018.
16. Montez JK, Beckfield J, Cooney JK, et al. US state policies, politics, and life expectancy. *Milbank Q*. 2020;98(3):668-699. https://doi.org/10.1111/1468-0009.12469
17. Dwyer-Lindgren L, Kendrick P, Kelly YO, et al. Life expectancy by county, race, and ethnicity in the USA, 2000–19: a systematic analysis of health disparities. *Lancet*. 2022;400(10345):25-38. https://doi.org/10.1016/S0140-6736(22)00876-5
18. Adler NE, Glymour MM, Fielding J. Addressing social determinants of health and health inequalities. *JAMA*. 2016;316(16):1641-1642. https://doi.org/10.1001/jama.2016.14058
19. Singh GK, Kogan MD, Slifkin RT. Widening disparities in infant mortality and life expectancy between Appalachia and the rest of the United States, 1990–2013. *Health Aff (Millwood)*. 2017;36(8):1423-1432. https://doi.org/10.1377/hlthaff.2016.1571
20. US Department of Health and Human Services. Healthy People 2020l. Published 2020. Accessed July 15, 2023. https://wayback.archive-it.org/5774/20210812065410/https://www.healthypeople.gov/2020
21. Grove RD, Hetzel AM. Vital statistics rates in the United States, 1940-1960. Published 1968. Accessed July 15, 2023. https://stacks.cdc.gov/view/cdc/6200
22. Kitagawa EM, Hauser PM. *Differential Mortality in the United States: A Study in Socioeconomic Epidemiology*. Harvard University Press; 2013. https://doi.org/10.4159/harvard.9780674188471
23. Singh GK, Daus GP, Allender M, et al. Social determinants of health in the United States: addressing major health inequality trends for the nation, 1935-2016. *Int J MCH AIDS*. 2017;6(2):139-164. https://doi.org/10.21106/ijma.236
24. Bronner K, Eliassen MS, King A, Leggett C, Punjasthitkul S, Skinner J. *The Dartmouth Atlas of Health Care: 2018 Data Update*. The Dartmouth Institute for Health Policy and Clinical Practice; 2021.
25. Cooper Z, Stiegman O, Ndumele CD, Staiger B, Skinner J. Geographical variation in health spending across the US among privately insured individuals and enrollees in Medicaid and Medicare. *JAMA Netw Open*. 2022;5(7):e2222138. https://doi.org/10.1001/jamanetworkopen.2022.22138
26. Centers for Disease Control and Prevention. NVSS—United States small-area life expectancy estimates project. Published 2020. Accessed December 14, 2023. https://www.cdc.gov/nchs/nvss/usaleep/usaleep.html
27. Al-Wakeel A, Wu J, Jenkins N. k-Means based load estimation of domestic smart meter measurements. *Appl Energy*. 2017;194:333-342. https://doi.org/10.1016/j.apenergy.2016.06.046
28. Kaur D. A comparative study of various distance measures for software fault prediction. *Int J Comput Trends Technol*. 2014;17(3):117-120. https://doi.org/10.14445/22312803/IJCTT-V17P122
29. Jurman G, Riccadonna S, Visintainer R, Furlanello C. Canberra distance on ranked lists. In: Agarwal S, Burges C, Crammer K, eds.

Proceedings of Advances in Ranking NIPS 09 Workshop. Citeseer; 2009:22-27.
30. Kodinariya TM, Makwana PR. Review on determining of cluster in K-means clustering. *Int J Adv Res Comput Sci Manag Stud*. 2013;1(6):90-95.
31. Roberts MJ, ed. *Getting Health Reform Right: A Guide to Improving Performance and Equity*. Oxford University Press; 2004.
32. Fop M, Murphy TB. Variable selection methods for model-based clustering. *Stat Surv*. 2018;12:18-65. https://doi.org/10.1214/18-SS119
33. Levinson Z, Godwin J, Hulver S. Rural hospitals face renewed financial challenges, especially in states that have not expanded Medicaid. KFF. Published February 23, 2023. Accessed July 17, 2023. https://www.kff.org/health-costs/issue-brief/rural-hospitals-face-renewed-financial-challenges-especially-in-states-that-have-not-expanded-medicaid/
34. Murthy VH. Confronting health worker burnout and well-being. *N Engl J Med*. 2022;387(7):577-579. https://doi.org/10.1056/NEJMp2207252
35. Fawaz M, Anshasi H, Samaha A. Nurses at the front line of COVID-19: roles, responsibilities, risks, and rights. *Am J Trop Med Hyg*. 2020;103(4):1341-1342. https://doi.org/10.4269/ajtmh.20-0650
36. National Academies of Sciences, Engineering, and Medicine; National Academy of Medicine; Committee on the Future of Nursing 2020–2030, et al. The role of nurses in improving health care access and quality. In: *The Future of Nursing 2020-2030: Charting a Path to Achieve Health Equity*. National Academies Press (US); 2021:99-126. Accessed December 14, 2023. https://www.ncbi.nlm.nih.gov/books/NBK573910/
37. Kruk ME, Gage AD, Arsenault C, et al. High-quality health systems in the sustainable development goals era: time for a revolution. *Lancet Glob Health*. 2018;6(11):e1196-e1252. https://doi.org/10.1016/S2214-109X(18)30386-3
38. Stevens GD, West-Wright CN, Tsai K-Y. Health insurance and access to care for families with young children in California, 2001–2005: differences by immigration status. *J Immigr Minor Health*. 2010;12(3):273-281. https://doi.org/10.1007/s10903-008-9185-8
39. Sommers BD, Gunja MZ, Finegold K, Musco T. Changes in self-reported insurance coverage, access to care, and health under the affordable care act. *JAMA*. 2015;314(4):366-374. https://doi.org/10.1001/jama.2015.8421
40. Antonisse L, Garfield R, Rudowitz R, Artiga S. The effects of Medicaid expansion under the ACA: updated findings from a literature review. 2017. Accessed December 14, 2023. https://nationaldisabilitynavigator.org/wp-content/uploads/news-items/KFF_Effects-of-Medicaid-Expansion-Lit-Review_Feb-2017-chart.pdf
41. Institute of Medicine (US) Committee on the Consequences of Uninsurance. Why health insurance matters. In: *Coverage Matters: Insurance and Health Care*. National Academies Press (US); 2001:19-34. Accessed December 14, 2023. https://www.ncbi.nlm.nih.gov/books/NBK223643/
42. Beam AL, Kohane IS. Big data and machine learning in health care. *JAMA*. 2018;319(13):1317-1318. https://doi.org/10.1001/jama.2017.18391
43. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. *N Engl J Med*. 2019;380(14):1347-1358. https://doi.org/10.1056/NEJMra1814259
44. Rabiei R, Ayyoubzadeh SM, Sohrabei S, Esmaeili M, Atashi A. Prediction of breast cancer using machine learning approaches. *J Biomed Phys Eng*. 2022;12(3):297-308. https://doi.org/10.31661/jbpe.v0i0.2109-1403
45. Ranapurwala SI, Alam IZ, Pence BW, et al. Development and validation of an electronic health records-based opioid use disorder algorithm by expert clinical adjudication among patients with prescribed opioids. *Pharmacoepidemiol Drug Saf*. 2023;32(5):577-585. https://doi.org/10.1002/pds.5591