nature mental health

Article

https://doi.org/10.1038/s44220-025-00424-4

Machine learning to investigate policy-relevant social determinants of health and suicide rates in the United States

Received: 7 May 2024

Accepted: 7 April 2025

Published online: 12 May 2025

Check for updates

Yunyu Xiao[1,8] ✉, Yuan Meng[1,8], Timothy T. Brown[2], Alexander C. Tsai[3,4], Lonnie R. Snowden[2], Julian Chun-Chung Chow[5], Jyotishman Pathak[1] & J. John Mann[6,7]

This study aimed to categorize county clusters of multidimensional social determinants of health (SDOH) using unsupervised machine learning and to analyze their association with county-level suicide rates, considering temporal, geographic and demographic variation. We analyzed aggregated SDOH data across 3,018 US counties for 2009, 2014 and 2019, which were linked to county-level suicide rates from the National Vital Statistics System. We identified three distinct SDOH clusters: 'REMOTE' (rural, elderly, marginalized environments, old housing, traditional systems, empty houses), 'COPE' (complex family dynamics, high consumption of health services, poverty, extreme heat) and 'DIVERSE' (dense, immigrant rich, environmentally challenged, economically unequal, racial/ethnic diversity, saturated health care, expensive housing). We used negative binomial regression after identifying clusters to estimate the associations between county-level SDOH clusters and suicide rates. Compared with other clusters, REMOTE was associated with higher overall suicide rates, particularly among men; COPE showed elevated suicide rates among whites; and DIVERSE exhibited increased rates among women and Black and Hispanic populations. The distribution of suicide rates across US states corresponded to the variations in SDOH cluster distribution within each state. These findings provide a foundation for designing more effective, data-driven suicide prevention strategies tailored to specific regional and demographic contexts.

For over two decades, the United States has experienced rising suicide rates, marked by notable disparities across race/ethnicity, sex, gender, age and geography[1–9]. From 2018 to 2021, suicide rates escalated notably among American Indian/Alaskan Native (26%) and non-Hispanic Black populations (19%), while rates among non-Hispanic whites declined by 4% (ref. 10). Among Black males aged 10–19 years, suicide rates surged by 116% from 2005 to 2021, a rise that more than doubles the increase observed among their white counterparts (43%)[6]. Suicide ranks as the second leading cause of death among younger adolescents aged 10–14 years and adults aged 20–34 years, yet the highest rates occur among individuals aged 85 years and above[7,8,11]. In addition, non-metropolitan areas have experienced a 46% rise in suicide rates compared with a 27.3% increase in metropolitan areas between 2000 and 2020, underscoring regional disparities[7–9]. The underlying reasons for these demographic and geographic differences in increasing suicide rates remain unclear.

Disparities in suicide rate may, in part, reflect differences in social determinants of health (SDOH), such as economic status, educational opportunities and access to health care, which may vary substantially across communities and over time[12–14]. For example, the higher poverty

A full list of affiliations appears at the end of the paper. ✉e-mail: yux4008@med.cornell.edu

among Black adults compared with white adults[15], or differences in health-care availability between rural and urban counties[16], may explain suicide-rate variations by race/ethnicity and geography[17–19]. Societal pressures and expectations also vary by sex/gender[20–22] and age[23–26], often exacerbating these disparities.

Early research focused on individual SDOH, a method insufficient for capturing the full spectrum of SDOH and across diverse demographics and regions[27,28]. More recent research has evaluated SDOH such as poverty[19], unemployment[29], social integration[23,26], rurality[30] and greenspace[31] as simple linear combinations. A step beyond this approach is the development of SDOH indices, including the 17-item Area Deprivation Index[32] and the 16-item Social Vulnerability Index (SVI)[33], which reduce multiple factors into a single score of socioeconomic deprivation. While these indices managed to measure the intensity of unfavorable SDOH with a single score, they cannot capture the multidimensional interrelationships between SDOH that drive distinct patterns and outcomes. This limitation restricts policymakers to identifying only which geographic areas are most deprived, without providing insights into the specific SDOH patterns most strongly associated with health outcomes—a critical gap for guiding effective policy decisions[34].

More recent advances[28] have overcome these limitations by leveraging machine-learning clustering techniques to categorize 284 SDOH characteristics into distinct multidimensional SDOH clusters, demonstrating strong associations with child mental health. However, no studies have applied county-level SDOH clusters to examine temporal trends and geographic differences in their associations with suicide rates. County-level SDOH clusters are impactful because they serve as place-based indicators, characterizing natural groupings of SDOH and enabling precise evaluation of their associations with county-level health outcomes to better policymaking.

In this study, we employed unsupervised machine learning to transform 284 SDOH characteristics into distinct multidimensional SDOH county clusters. We examined how these clusters correlated with suicide rates across US counties from 2009 to 2019 and analyzed how these relationships vary by sex, age and race/ethnicity across different temporal and geographic scales. The key innovation of this approach is that it moves beyond simple deprivation scores to rich county SDOH profiles for targeted suicide prevention strategies with known variation geographically and over time.

Results

Identification and characteristics of SDOH clusters

Analysis of 284 SDOH variables identified three county-level SDOH clusters (Fig. 1 and Supplementary Result 1): REMOTE, COPE and DIVERSE. REMOTE (rural, elderly, marginalized environments, old housing, traditional energy systems, empty houses) was the most prevalent throughout the study period, encompassing 1,867 (60.1%) counties in 2009, 1,696 (54.6%) in 2014 and 1,904 (61.3%) in 2019. COPE (complex family dynamics (for example, single-parent households), low health-care coverage, poverty, extreme heat) saw a decrease in geographic prevalence over the study period: 1,073 counties (34.5%) in 2009, 768 (24.7%) in 2014 and 688 (22.1%) in 2019. DIVERSE (dense urban populations, immigrant, environmental challenge, economic inequality, racial/ethnic diversity, saturated health-care systems, expensive housing) was the least widespread, including 169 counties (5.4%) in 2009, 644 (20.7%) in 2014 and 516 (16.6%) in 2019.

Economic challenges manifested differently across the clusters. Counties in the COPE cluster faced the highest poverty rates and lowest household incomes, whereas the DIVERSE cluster grappled with income disparities and high living costs, particularly in real estate. Workforce composition[10,35] also differed notably: manufacturing and construction dominated in REMOTE areas, retail and transportation were prevalent in COPE counties, and a broader range of industries—including the arts and armed forces—characterized DIVERSE areas.

Educational attainment differed substantially by cluster. COPE counties had the lowest levels of educational achievement, while DIVERSE counties had the highest proportion of individuals with bachelor's and postgraduate degrees.

Physical infrastructure characteristics also varied. Compared with the other two clusters, REMOTE counties were characterized by older housing (for example, built before 1979), owner-occupied units with inefficient heating systems (for example, coal/wood) and high vacancy rates. By contrast, COPE counties featured newer, larger housing units, often accommodating multi-generational families. DIVERSE areas exhibited high population densities, longer commute times and a mix of high-cost rental housing, with rental costs exceeding 50% of average resident income.

Health-care access and utilization also revealed stark disparities. COPE counties had the highest rates of uninsured individuals and were frequently designated as medically underserved areas. Conversely, DIVERSE counties experienced higher hospital admission rates and longer inpatient stays, reflecting strained health-care infrastructure.

Environmental challenges were pronounced across clusters. COPE counties faced more frequent extreme heat events (for example, days with a heat index >40.6 °C, which increased from 2009 to 2019), while DIVERSE areas struggled with poorer air quality (for example, elevated $PM_{2.5}$ (particulate matter less than 2.5 µm in diameter) levels).

There were moderate to high correlations between the three clusters and established indices, including the US Agency for Healthcare Research and Quality (AHRQ) SDOH (0.54 to 0.88; Supplementary Fig. 2 and Supplementary Table 3) and SVI indices (0.57 to 0.81; Supplementary Fig. 3 and Supplementary Table 4) based on geographically weighted regression[36]. This demonstrated that our approach not only captured much of what previous indices captured but also identified distinct, natural configurations of SDOH that drive health outcomes.

Sociodemographic patterns and disparities

Significant age, sex, racial, ethnic and geographical disparities were evident across the clusters ($P < 0.0001$; Table 1, Fig. 2, Supplementary Fig. 1 and SupplementaryTable 2). The REMOTE cluster was predominantly non-Hispanic white (87.5% in 2009, 85.8% in 2014, 84.8% in 2019) and rural, with rurality increasing over time (70.1% in 2009, 73.3% in 2014, 71.6% in 2019). DIVERSE areas had the highest racial/ethnic diversity, featuring larger proportions of Hispanic (23.2% in 2009, 19.9% in 2014, 20.9% in 2019) and non-Hispanic Black (14.5% in 2009, 14.5% in 2014, 15.8% in 2019) populations and more urban areas (100% in 2009, 90.2% in 2014, 89.5% in 2019) than the other clusters. The COPE cluster exhibited a mix of urban and rural characteristics, with noticeable shifts toward urbanization over time. Age demographics revealed minimal differences across clusters, yet the REMOTE cluster consistently had a higher proportion of older adults over 65 years of age (14.6% in 2009, 17.5% in 2014 and 19.1% in 2019) compared with the other clusters ($P < 0.0001$).

Temporal changes and geographic distributions

From 2009 to 2019, most counties (2,160, 69.5%) retained the same dominant SDOH cluster designations (Supplementary Fig. 4a). However, the percentage of counties designated as COPE counties decreased from 34.5% to 22.1% (Supplementary Fig. 4b,c). This shift was driven by increased health-care utilization (4.8% rise in DIVERSE counties) and worsening environmental conditions (37.3% increase in DIVERSE counties and 26.8% in REMOTE counties). Conversely, improvements in educational resources were observed in the DIVERSE cluster from 2014 to 2019 (Supplementary Fig. 5 and Supplementary Table 5).

Geographically, the REMOTE cluster was located predominantly in the northern and central United States (for example, Montana, Wyoming, South Dakota, Nevada, Utah and Maine). The COPE cluster consisted primarily of Southern and Central states (for example, Texas, Oklahoma, Mississippi, Alabama and Georgia). The DIVERSE cluster was

https://doi.org/10.1038/s44220-025-00424-4

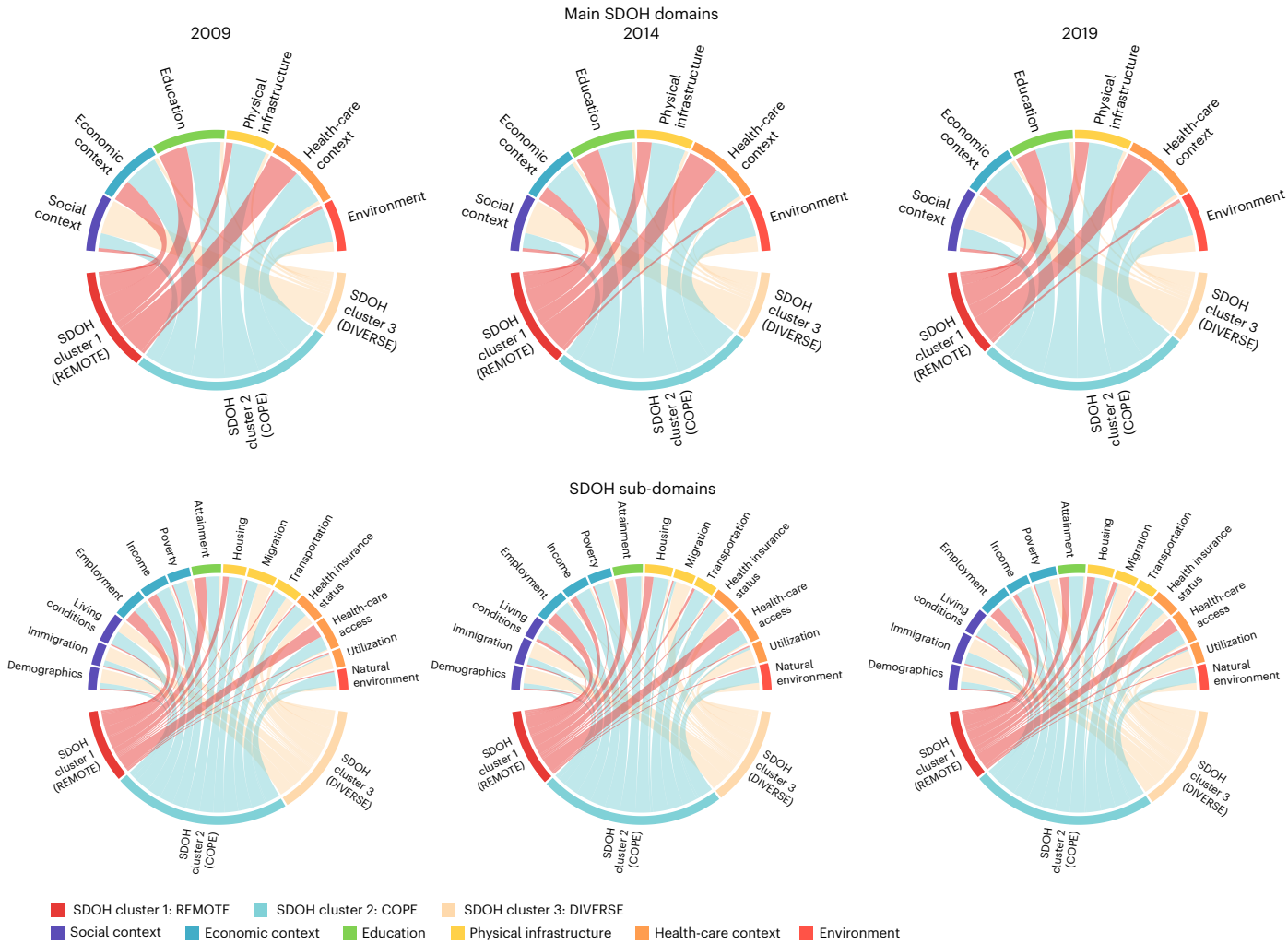


Fig. 1 | SDOH cluster distributions. Chord plot of main SDOH domains and SDOH sub-domains across SDOH clusters. Values in each SDOH domain and sub-domain were separately standardized within 1 to 10. Wider widths in the chord flow indicate more unfavorable SDOH. The exact widths of each SDOH cluster across years may change, while we allowed small variations and selected the SDOH clusters qualitatively for policymakers to facilitate interpretation.

found mainly along the east and west coasts (for example, California, Florida, New Jersey, Connecticut and Massachusetts).

Within-state heterogeneity analysis revealed that states such as California and North Carolina had a more diverse distribution of SDOH clusters compared with other states (Supplementary Fig. 6b and Supplementary Table 6). In 2009, most counties in California had the DIVERSE cluster as dominant (44.8%) compared with REMOTE (34.5%) and COPE (20.7%), but then shifted to predominantly REMOTE (39.7%) in 2019.

While less common, some states exhibited homogeneity in their dominant clusters. In 2009, Louisiana and Wyoming were entirely REMOTE, while Mississippi had only COPE counties. By 2014, Connecticut became entirely dominated by DIVERSE, a pattern that persisted throughout the decade in the District of Columbia. In 2019, Wyoming and Montana remained predominantly REMOTE, underscoring their stability in their SDOH cluster designations.

Associations between SDOH clusters and suicide rates

The geographical distribution of SDOH clusters correlated strongly with suicide rates (Fig. 3 and Supplementary Fig. 6c). Larger proportions of REMOTE cluster counties were associated with states reporting the highest suicide rates. For example, Rosebud County, Montana, exhibited suicide rates of 43.6 (2008–2010), 42.8 (2013–2015) and 41.0 (2018–2020) per 100,000 population. By contrast, DIVERSE areas, such as San Luis Obispo County, California, displayed intermediate suicide rates: 16.5 (2008–2010), 17.6 (2013–2015) and 17.2 (2018–2020) per 100,000 population.

REMOTE areas consistently reported higher suicide rates among males compared with other clusters. The mean suicide rates in REMOTE areas were 15.3 (2008–2010), 18.6 (2013–2015), and 19.9 (2018–2020) per 100,000 population, compared with 11.1, 14.0 and 15.1, respectively, for DIVERSE areas during the same periods (Fig. 3).

Between 2009 and 2019, suicide rates increased across all SDOH clusters, with REMOTE areas experiencing the steepest rise ($P = 0.000$, Supplementary Table 7). Counties in the DIVERSE cluster showed significant increases in suicide rates among youth aged 15–24 years and non-Hispanic Black populations. For youth, suicide rates rose by 2.84 (95% confidence interval (CI): 1.58–4.09) in 2013–2015 and by 5.30 (4.05–6.55) in 2018–2020, while suicide rates among non-Hispanic Blacks increased by 2.10 (1.13–3.07, $P = 0.000$) in 2018–2020 (Supplementary Result 2).

Negative binomial regression models, which better fit the suicide-rate data (Supplementary Table 8), underscored the strong associations between REMOTE cluster location and higher suicide rates among the whole population. Incidence rate ratios for the REMOTE cluster were 1.38 (95% CI: 1.32–1.45), 1.36 (1.31–1.40) and 1.34 (1.28–1.38) across different years compared with DIVERSE clusters. Slight differences with COPE clusters were observed, with incidence rate ratios of 1.04

https://doi.org/10.1038/s44220-025-00424-4

Table 1 | Demographic statistics of SDOH clusters

	Overall	SDOH cluster 1 (REMOTE)	SDOH cluster 2 (COPE)	SDOH cluster 3 (DIVERSE)	*P* value[b]	Adjusted *P*
	n[a] (%)	*n* (%)	*n* (%)	*n* (%)		
2009						
All	304,725,917 (100.0%)	85,213,326 (100.0%)	66,765,365 (100.0%)	15,2747,226 (100.0%)	0.000	0.000
Sex						
Male	149,771,919 (49.1%)	42,185,941 (49.5%)	32,832,916 (49.2%)	74,753,062 (48.9%)	0.000	0.000
Female	154,953,998 (50.9%)	43,027,385 (50.5%)	33,932,449 (50.8%)	77,994,164 (51.1%)	0.000	0.000
Age (yr)						
15–24	43,290,486 (14.2%)	11,815,639 (13.9%)	9,679,769 (14.5%)	21,795,078 (14.3%)	0.000	0.000
25–64	161,372,825 (53.0%)	44,711,741 (52.5%)	34,349,836 (51.4%)	82,311,248 (53.9%)	0.000	0.000
Over 65	39,379,917 (12.9%)	12,415,010 (14.6%)	9,128,456 (13.7%)	17,836,451 (11.7%)	0.000	0.000
Race/ethnicity						
Hispanic	49,170,152 (16.1%)	4,395,393 (5.2%)	9,384,941 (14.1%)	35,389,818 (23.2%)	0.000	0.000
Non-Hispanic Black	39,048,245 (12.8%)	3,872,608 (4.5%)	12,953,742 (19.4%)	22,221,895 (14.5%)	0.000	0.000
Non-Hispanic white	199,164,305 (65.4%)	74,552,950 (87.5%)	42,527,420 (63.7%)	82,083,935 (53.7%)	0.000	0.000
Others[c]	17,343,215 (5.7%)	2,392,375 (2.8%)	1,899,262 (2.8%)	13,051,578 (8.5%)	0.000	0.000
Metro/nonmetro						
Urban	1,161 (37.3%)	559 (29.9%)	433 (40.4%)	169 (100.0%)	0.000	0.000
Rural[d]	1,948 (62.7%)	1,308 (70.1%)	640 (59.6%)	0 (0.0%)	0.000	0.000
2014						
All	316,700,763 (100.0%)	55,804,993 (100.0%)	19,140,645 (100.0%)	241,755,125 (100.0%)	0.000	0.000
Sex						
Male	155,831,092 (49.2%)	27,776,424 (49.8%)	9,526,638 (49.8%)	118,528,030 (49.0%)	0.000	0.000
Female	160,869,671 (50.8%)	28,028,569 (50.2%)	9,614,007 (50.2%)	12,3227,095 (51.0%)	0.000	0.000
Age (yr)						
15–24	43,681,931 (13.8%)	7,224,423 (12.9%)	2,640,027 (13.8%)	33,817,481 (14.0%)	0.000	0.000
25–64	166,422,224 (52.5%)	28,615,902 (51.3%)	9,543,509 (49.9%)	128,262,813 (53.1%)	0.000	0.000
Over 65	45,945,644 (14.5%)	9,780,263 (17.5%)	3,136,818 (16.4%)	33,028,563 (13.7%)	0.000	0.000
Race/ethnicity						
Hispanic	55,194,122 (17.4%)	3,831,716 (6.9%)	3,369,425 (17.6%)	47,992,981 (19.9%)	0.000	0.000
Non-Hispanic Black	41,242,256 (13.0%)	2,533,362 (4.5%)	3,697,063 (19.3%)	35,011,831 (14.5%)	0.000	0.000
Non-Hispanic white	200,192,002 (63.2%)	47,872,810 (85.8%)	11,530,518 (60.2%)	140,788,674 (58.2%)	0.000	0.000
Others	20,072,383 (6.3%)	1,567,105 (2.8%)	543,639 (2.8%)	17,961,639 (7.4%)	0.000	0.000
Metro/nonmetro						
Urban	1,160 (37.3%)	453 (26.7%)	126 (16.4%)	581 (90.2%)	0.000	0.000
Rural	1,948 (62.7%)	1,243 (73.3%)	642 (83.6%)	63 (9.8%)	0.000	0.000
2019						
All	326,092,106 (100.0%)	73,302,477 (100.0%)	23,902,449 (100.0%)	228,887,180 (100.0%)	0.000	0.000
Sex						
Male	160,567,866 (49.2%)	36,465,996 (49.7%)	11,976,016 (50.1%)	11,2125,854 (49.0%)	0.000	0.000
Female	165,524,240 (50.8%)	36,836,481 (50.3%)	11,926,433 (49.9%)	116,761,326 (51.0%)	0.000	0.000
Age (yr)						
15–24	42,425,180 (13.0%)	9,307,822 (12.7%)	3,200,529 (13.4%)	29,916,829 (13.1%)	0.000	0.000
25–64	169,802,400 (52.1%)	37,024,617 (50.5%)	11,799,160 (49.4%)	120,978,623 (52.9%)	0.000	0.000
Over 65	53,698,227 (16.5%)	13,968,845 (19.1%)	3,971,932 (16.6%)	35,757,450 (15.6%)	0.000	0.000
Race/ethnicity						
Hispanic	60,368,161 (18.5%)	4,982,919 (6.8%)	7,463,443 (31.2%)	47,921,799 (20.9%)	0.000	0.000
Non-Hispanic Black	43,166,073 (13.2%)	3,715,852 (5.1%)	3,270,458 (13.7%)	36,179,763 (15.8%)	0.000	0.000
Non-Hispanic white	200,081,781 (61.4%)	62,139,727 (84.8%)	1,2019,158 (50.3%)	125,922,896 (55.0%)	0.000	0.000
Others	22,476,085 (6.9%)	2,463,979 (3.4%)	1,149,384 (4.8%)	18,862,722 (8.2%)	0.000	0.000
Metro/nonmetro						
Urban	1,160 (37.3%)	540 (28.4%)	158 (23.0%)	462 (89.5%)	0.000	0.000
Rural	1,948 (62.7%)	1,364 (71.6%)	530 (77.0%)	54 (10.5%)	0.000	0.000

[a]Total number of residents in the county. [b]*P* values were calculated by analysis of variance for continuous variables and $\chi 2$ test for categorical variables; adjusted *P* values were calculated using Bonferroni two-sided tests with no adjustment for multiple comparisons. [c]Others include Asian, Pacific Islander, American Indian and Alaska Native. [d]Rural areas, usually defined as nonmetro areas, are counties with populations less than 50,000 (ref. 76).

https://doi.org/10.1038/s44220-025-00424-4

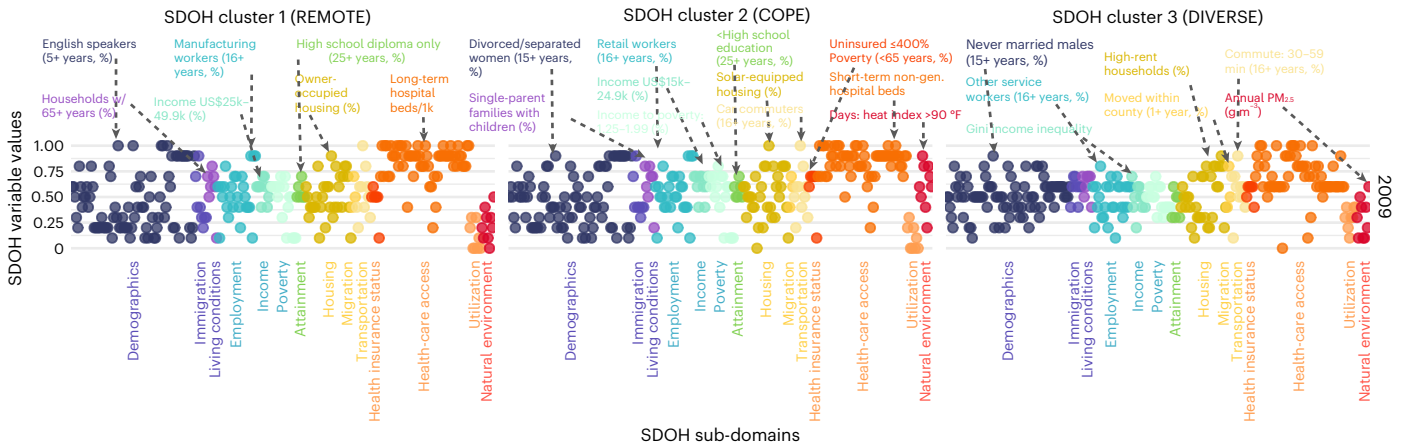


Fig. 2 | SDOH variable distributions by SDOH clusters in 2009. Distribution of SDOH variable values across SDOH clusters. Plots indicate the individual SDOH variable values grouped by SDOH sub-domains for each SDOH cluster. Individual SDOH variables were scaled within the range of 0–1. non-gen., non-general.

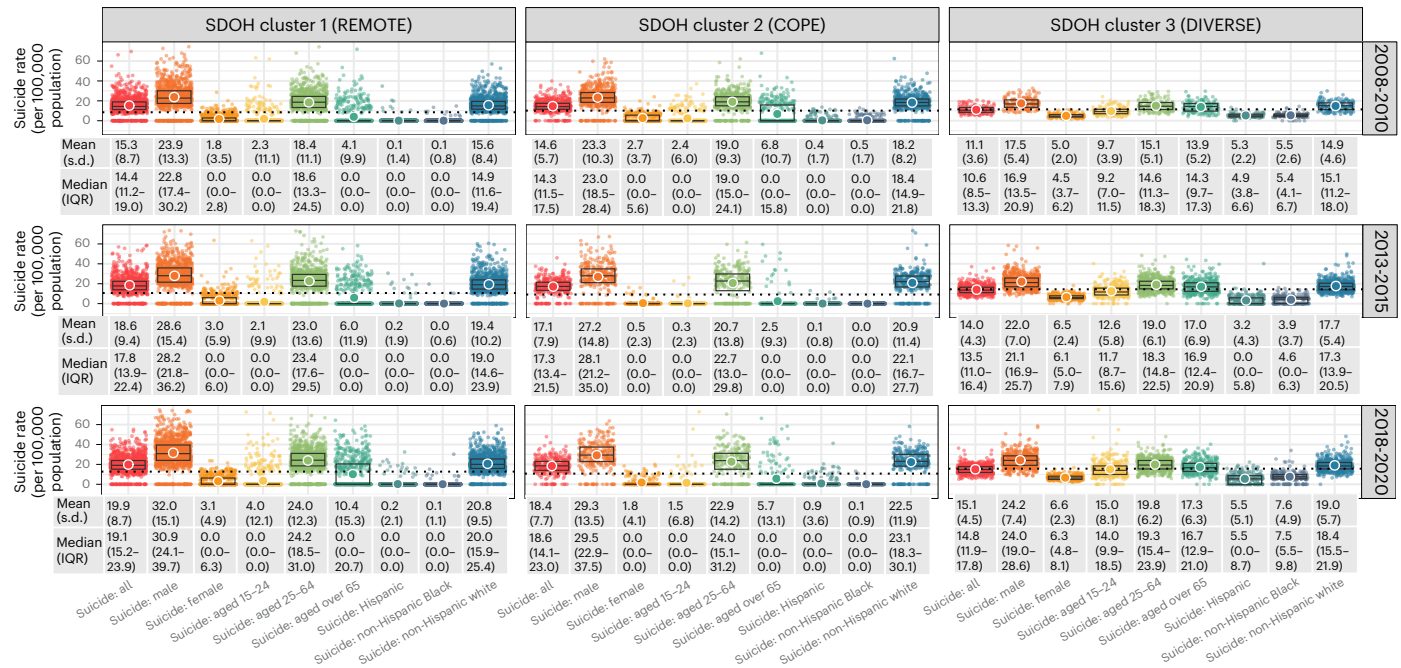


Fig. 3 | Suicide rates by SDOH clusters in 2009, 2014 and 2019. Plots indicate the aggregated 3-year suicide rates of each county by subgroups among SDOH clusters for 2008–2010, 2013–2015 and 2018–2020. The scattered transparent points represent the aggregated 3-year suicide rates of individual counties; the solid point represents the mean suicide rate of each subgroup; the box plot represents the interquartile range (IQR); and the mean is represented by the middle horizontal lines. Counties with suicide rates higher than 75 were not displayed in this figure.

(1.01–1.08), 1.06 (1.02–1.11) and 1.07 (1.02–1.12), respectively (Fig. 4 and Supplementary Table 9).

Racial/ethnic, sex and age disparities

There were racial/ethnic disparities within SDOH clusters. For example, in 2019, DIVERSE clusters had higher suicide rates among Hispanic individuals (mixed-effect linear regression coefficient: 2.59 (2.14 to 3.16)) and non-Hispanic Black individuals (3.22 (2.66 to 3.94)) compared with REMOTE counties. Age disparities were also evident: REMOTE counties reported higher suicide rates in individuals over 65 years of age compared with COPE areas (for example, mixed-effect linear regression coefficient: 1.36 (1.21–1.55) for REMOTE versus 1.15 (1.07–1.23) for DIVERSE in 2019). Detailed random effects of the SDOH cluster (Supplementary Table 10) illustrate the preceding associations of suicide rates.

Pairwise tests (Supplementary Table 11) revealed differences across the three SDOH clusters. Investigating the association between specific SDOH (sub)domains and suicide rates revealed demographic disparities (Supplementary Results 2). For example, non-Hispanic whites exhibited higher suicide rates (least absolute shrinkage and selection operator (LASSO) coefficients: 0.44 to 3.24) compared with Hispanics (LASSO coefficients: 0 to 1.39) and non-Hispanic Blacks (LASSO coefficients: 1.27 to 2.10). LASSO regression also identified SDOH variables with high-value coefficients (Supplementary Result 3), which aligned with the characteristics of each cluster. This alignment validates our clustering approach and underscores the robustness of the identified patterns.

Sex and age disparities were evident across the clusters. Males were more affected than females by poor infrastructure and low income, and those over 65 years faced higher rates partly because of

https://doi.org/10.1038/s44220-025-00424-4

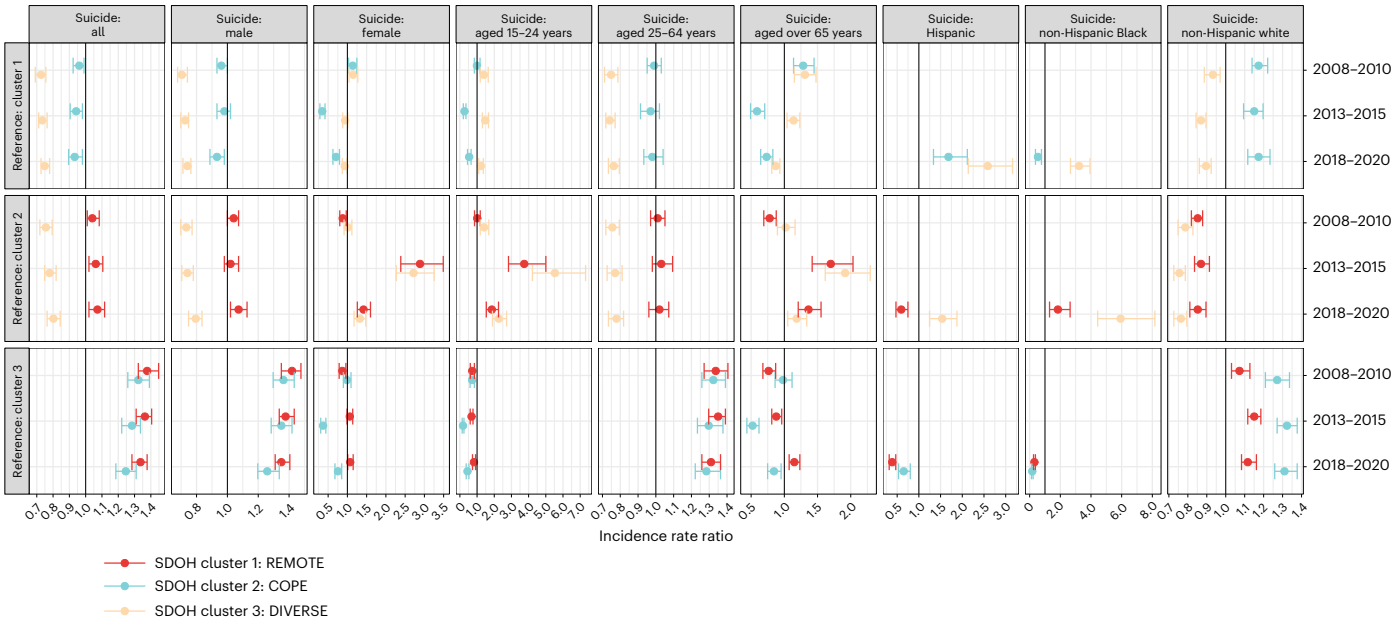


Fig. 4 | Coefficients (95% CI) of negative binomial model between SDOH clusters and suicide rates. Plots indicate the coefficients (95% CI) of SDOH clusters. Center of the error bars indicates incidence rate ratios, and lines indicate 95% CIs. The greater magnitude of coefficients represents stronger SDOH associations with suicide rates compared with the reference SDOH cluster. Subgroups with the proportion of counties with suicide rates lower than 10% or that had data from fewer than ten counties were removed from this study for robust estimation.

the lack of social support (Supplementary Fig. 7). For individuals aged 25–64 years, educational and infrastructure issues were prevalent SDOH concerns, explaining some of the variances in their suicide rates (Supplementary Fig. 8).

Discussion

The machine-learning approach identified three distinct SDOH clusters characterized by 284 SDOH variables—REMOTE, COPE and DIVERSE—revealing the dominant cluster across 3,018 US counties over a decade. These SDOH clusters exhibited unique temporal and geographical patterns, strongly correlating with county-level suicide rates overall and within subpopulations. This comprehensive analysis enhances our understanding of multidimensional SDOH and their association with both overall suicide rates and disparities.

Building on our previous work[28], which used individual children as the unit of analysis, this study expanded the scope to county-level units to create profiles and uncover temporal trends and geographic differences in SDOH clusters and their relationship to suicide rates. By moving beyond traditional, single metrics (for example, Area Deprivation Index[32], SVI[37], Social Vulnerability Metric[38]) and self-reported SDOH[39], our clustering approach captured the natural ecology of multidimensional SDOH clusters more comprehensively. Unlike LASSO, which selects individual predictors one at a time, cluster analysis identified data-driven and unbiased groupings of SDOH variables with minimal estimation bias that will serve as critical place-based measures for policymakers. This allows policymakers to design health policies targeted at geographically defined groups of SDOH, enabling tailored interventions that address regional needs more effectively. The breadth and richness of the 284 SDOH indicators across six domains measured over 10 years enabled robust, minimally biased SDOH cluster characterizations[40,41]. These findings provide nuanced insights into how these clusters explain variations in suicide rates geographically, temporally and across demographic subpopulations. Overall, the results permit tailoring SDOH-based suicide prevention policies to align with geographic and demographic variations in SDOH clusters.

Implications for suicide prevention

Each identified SDOH cluster revealed distinct challenges and opportunities for suicide prevention. The REMOTE cluster, with rural, elderly populations in marginalized environments, confirmed previous findings linking higher suicide rates to factors such as limited access to mental health services, social isolation and firearm ownership[42]. However, our study broadened these insights by incorporating housing quality and traditional energy systems, suggesting that interventions should also target housing and transportation infrastructure[43].

The COPE cluster, characterized by economic and environmental stressors, exhibited higher suicide rates among white individuals. This aligns with evidence linking economic strain[44], insufficient health-care resources[45] and 'deaths of despair'[46]. Community-based interventions addressing economic stress and substance use, including alcohol and opioid overdoses, are urgently needed in these areas.

The DIVERSE cluster, increasingly prevalent from 2009 to 2019, highlighted challenges in immigrant-dense, economically unequal areas[47]. Despite having more health-care resources, DIVERSE counties faced high population density, unmet needs, treatment gaps, low care quality and disproportionately higher suicide rates among Black, Hispanic, female and young populations[19,38,48]. Interventions in these areas should include improving culturally adapted mental health programs[49], urban health-care systems[50] and measures to enhance air quality[51].

Geographic and temporal tailoring

Our findings emphasize the need for geographically tailored suicide prevention strategies that reflect the unique challenges of SDOH clusters. REMOTE areas, predominantly in the north and central United States, could benefit from focused efforts to improve health-care access and reduce social isolation[52]. By contrast, states with diverse SDOH clusters, such as California, require varied approaches to maximize effectiveness and cost efficiency[39,53]. Further examination of whether these recommendations are effective is warranted before tailoring solutions to distinct community needs[54,55].

Shifts in some counties reflected broader socioeconomic trends, such as the 2014 expansion of health-care coverage or the 2008 economic downturn[39,56,57], which may subsequently increase suicide rates[58]. Testing

SDOH cluster-specific suicide prevention strategies rather than one-size-fits-all approaches is critical for addressing these shifts effectively.

Finally, recognizing that some demographic groups have higher suicide rates in specific clusters, integrating an intersectional approach to SDOH-based suicide prevention may further enhance the tailoring of interventions.

Strengths and limitations

The strengths of this study include the innovative generation of SDOH clusters, their temporal and regional variation analysis and intersectional insights into demographic factors, which advance our understanding of SDOH and suicide relationships. However, limitations include reliance on available SDOH data and excluding post-COVID-19 era trends, highlighting areas for future investigation.

While our SDOH clusters were robustly associated with suicide rates over a decade, replicating these methods for other outcomes, such as opioid overdose, would validate their broader applicability. To facilitate future research, we made our SDOH cluster label data available (Supplementary Table 18). Future studies should also explore how existing policies may work differently across the SDOH clusters, determine mediator and moderator effects, and prospectively evaluate targeted prevention policies and interventions. There is a need for follow-up observational or experimental studies on how policies based on our findings could improve suicide rates.

Conclusion

This study demonstrates a new, data-driven methodology for uncovering three distinct SDOH clusters that are significantly associated with suicide rates across the United States. By illuminating the complex interplay among demographic factors such as sex, age and race/ethnicity, our findings underscore the potential of employing the derived multifaceted SDOH clusters to guide targeted policies and interventions. This approach opens new avenues to inform testing of finely tuned, geographically specific suicide prevention policies and strategies, marking an important step forward in our collective effort to mitigate the factors contributing to suicide disparities nationwide.

Methods

Study design

This cohort study covered 2009 to 2019, focusing on US counties as the unit of analysis, each linked to their respective multidimensional SDOH data (see Supplementary Fig. 9a for study workflow). Counties that changed their boundaries were mapped to geographically consistent units (hereafter called 'counties' for consistency)[53,59], resulting in samples of 3,109 counties for 2009 and 3,018 for 2014 and 2019 (detailed in Supplementary Fig. 9b and Supplementary Method 1). This study was approved by the Weill Cornell Medicine Institutional Review Board. Informed consent was waived due to the use of publicly available, de-identified data.

SDOH data collection and normalization

County-level SDOH data for 2009, 2014 and 2019 were extracted from the AHRQ database[60], covering 48 states (excluding Alaska and Hawaii) and Washington, DC (Supplementary Method 2). Initially encompassing 173 SDOH variables, the dataset was expanded to 284 variables across 6 main domains and 14 sub-domains (Supplementary Table 20), with environmental factors explicitly categorized as a distinct domain to highlight their unique correlations with suicide[61,62].

Data normalization involved standardizing scales to ensure higher values reflected more unfavorable SDOH conditions. Two independent reviewers (Y.X. and J.J.M.) achieved a 95% consensus on the direction of associations between each SDOH variable and suicide rates, resolving discrepancies through discussion. Cube root and minimum–maximum transformations were applied to reduce skewness, scaling all data uniformly from 0 to 1.

Suicide-rate data

County-level suicide rates (per 100,000 persons per year), classified using the ICD-10 (International Classification of Diseases, Tenth Revision)[63] codes U03, X60–X84 and Y87.0, were obtained from the National Vital Statistics System Multiple Cause of Death database by the US Centers for Disease Control and Prevention (CDC)[64]. We aggregated suicide rates over 3 years to match SDOH time points (2008–2010 for 2009, 2013–2015 for 2014 and 2018–2020 for 2019), mitigating data suppression for counties with smaller populations, consistent with CDC methodologies (Supplementary Method 3)[2]. Counties reporting fewer than ten suicides across each 3-year period were excluded to maintain confidentiality and data integrity (Supplementary Fig. 1)[65]. We analyzed overall suicide rates and disaggregated by sex (male and female), age groups (15–24, 25–64, >65 years) and race/ethnicity (Hispanic, Black, white and other (including Asian, Pacific Islander, American Indian and Alaska Native)). Data for individuals under 15 years (2008–2010: 0.04%; 2013–2015: 0.1%; 2018–2020: 0.2%) and 'other' racial/ethnic subgroups (2008–2010: 2.2%; 2013–2015: 2.5%; 2018–2020: 3.1%) were not included due to small sample sizes and anonymity protection[65].

Statistical analysis

Identification and validation of SDOH clusters. We utilized hierarchical agglomerative clustering with Euclidean and Ward linkage to identify SDOH clusters (Supplementary Method 4)[66,67], evaluating models ranging from 1 to 15 clusters. The optimal cluster count was selected on the basis of 23 model fit indices (for example, CH index, DB index; Supplementary Method 5) and clinical significance. We used geographically weighted regression to compare with coefficients from two established SDOH indices (AHRQ SDOH original values, CDC/ATSDR (US Agency for Toxic Substances and Disease Registry) SVI)[37] and assess the accuracy of local SDOH condition representation (Supplementary Method 6).

Temporal changes and geographic distributions. The evolution of SDOH clusters was tracked across six primary domains from 2009 to 2019, assessing the consistency of county cluster memberships (Supplementary Method 7). Spatial distributions were visualized to highlight geographic variations.

Regression analysis. We employed negative binomial regression with incidence rate ratios to examine the association between SDOH cluster memberships and county-level suicide rates[68], adjusting for data overdispersion[69]. This approach was supplemented by a linear regression with coefficients assessing suicide-rate changes among all clusters, mixed-effect linear regression with coefficients accounting for random county-level effects[70] and pairwise comparisons with P values of suicide rates between clusters. Further, we used LASSO regression[71,72] to handle the high-dimensional nature of SDOH data while avoiding overfitting. We calculated the mean values of each main SDOH domain and sub-domain, as well as individual SDOH variables as independent variables in LASSO regression, implementing tenfold cross-validation to enhance precision in associational tests[73].

Sensitivity analyses

The robustness of our negative binomial regression findings was validated against alternative models (Poisson, zero-inflated Poisson and zero-inflated negative binomial models) using Akaike information criterion and Bayesian information criteria[74]. The accuracy of LASSO regression models was assessed by root mean square error evaluations, with lower values indicating improved performance[75]. Counties with suicide rates below 10% were excluded in sensitivity analyses to ensure robust estimations.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All raw data in this study are publicly available and can be retrieved from the US National Center for Health Statistics, US Centers for Disease Control and Prevention Mortality Data (https://www.cdc.gov/nchs/nvss/deaths.htm) and US Agency for Healthcare Research and Quality Social Determinants of Health Database (https://www.ahrq.gov/sdoh/data-analytics/sdoh-data.html).

Code availability

Data analysis was performed using R version 4.4.2, leveraging several specialized packages to ensure robust and reproducible results. The 'NbClust' package (version 3.0.1) was used to determine the optimal number of clusters, while 'cluster' (version 2.1.6) facilitated clustering techniques and visualization. Core statistical computations and modeling were conducted using the 'stats' package (version 3.6.2). For regression modeling, including elastic net and regularization, the 'glmnet' package (version 4.1.8) was employed. In addition, geographically weighted regression was implemented using the 'spgwr' package (version 0.6-37). Code is available at github.com/X-PLORE-Lab/sdoh_clustering_suicide.

References

1. Stone, D. M., Mack, K. A. & Qualters, J. Notes from the field: recent changes in suicide rates, by race and ethnicity and age group—United States, 2021. *MMWR Morb. Mortal. Wkly Rep.* **72**, 160–162 (2023).
2. *Provisional Suicide Deaths in the United States, 2022* (CDC, 2023); https://www.cdc.gov/media/releases/2023/s0810-US-Suicide-Deaths-2022.html
3. Bridge, J. A. et al. Suicide trends among elementary school-aged children in the United States from 1993 to 2012. *JAMA Pediatr.* **169**, 673–677 (2015).
4. Opara, I. et al. Suicide among Black children: an integrated model of the interpersonal–psychological theory of suicide and intersectionality theory for researchers and clinicians. *J. Black Stud.* **51**, 611–631 (2020).
5. Martínez-Alés, G. et al. Age, period, and cohort effects on suicide death in the United States from 1999 to 2018: moderation by sex, race, and firearm involvement. *Mol. Psychiatry* **26**, 3374–3382 (2021).
6. AACAP *AACAP Policy Statement on Increased Suicide Among Black Youth in the US* (AACAP, accessed 12 October 2023); https://www.aacap.org/aacap/Policy_Statements/2022/AACAP_Policy_Statement_Increased_Suicide_Among_Black_Youth_US.aspx
7. Xiao, Y. & Lu, W. Temporal trends and disparities in suicidal behaviors by sex and sexual identity among Asian American adolescents. *JAMA Netw. Open* **4**, e214498 (2021).
8. Xiao, Y., Cerel, J. & Mann, J. J. Temporal trends in suicidal ideation and attempts among US adolescents by sex and race/ethnicity, 1991–2019. *JAMA Netw. Open* **4**, e2113513 (2021).
9. Lindsey, M. A., Sheftall, A. H., Xiao, Y. & Joe, S. Trends of suicidal behaviors among high school students in the United States: 1991–2017. *Pediatrics* **144**, e20191187 (2019).
10. *Disparities in Suicide | Suicide Prevention* (CDC, accessed 19 January 2024); https://www.cdc.gov/suicide/facts/disparities-in-suicide.html
11. *About Multiple Cause of Death, 2018–2023 Single Race* (CDC, accessed 1 January 1 2024); http://wonder.cdc.gov/mcd-icd10-expanded.html
12. *Social Determinants of Health* (CDC, accessed 12 October 2023); https://www.cdc.gov/about/sdoh/index.html
13. *Social Determinants of Health* (NCHHSTP and CDC, accessed 19 January 2024); https://www.cdc.gov/nchhstp/socialdeterminants/index.html
14. VanderWeele, T. & Robinson, W. On the causal interpretation of race in regressions adjusting for confounding and mediating variables. *Epidemiology* **25**, 473–484 (2014).
15. Shrider, E. A., Kollar, M., Chen, F. & Semega, J. *Income and Poverty in the United States: 2020* (US Government Publishing Office, 2021).
16. Cyr, M. E., Etchin, A. G., Guthrie, B. J. & Benneyan, J. C. Access to specialty healthcare in urban versus rural US populations: a systematic literature review. *BMC Health Serv. Res.* **19**, 974 (2019).
17. Robertson, R. A., Standley, C. J., Gunn, J. F. III & Opara, I. Structural indicators of suicide: an exploration of state-level risk factors among Black and White people in the United States. *J. Public Ment. Health* **21**, 23–34 (2022).
18. Smith, N. D. L. & Kawachi, I. State-level social capital and suicide mortality in the 50 US states. *Soc. Sci. Med.* **120**, 269–277 (2014).
19. Hoffmann, J. A., Farrell, C. A., Monuteaux, M. C., Fleegler, E. W. & Lee, L. K. Association of pediatric suicide with county-level poverty in the United States, 2007–2016. *JAMA Pediatr.* **174**, 287–294 (2020).
20. Knipe, D., Padmanathan, P., Newton-Howes, G., Chan, L. F. & Kapur, N. Suicide and self-harm. *Lancet* **399**, 1903–1916 (2022).
21. Luby, J. & Kertz, S. Increasing suicide rates in early adolescent girls in the United States and the equalization of sex disparity in suicide: the need to investigate the role of social media. *JAMA Netw. Open* **2**, e193916 (2019).
22. Mustanski, B. & Espelage, D. L. Why are we not closing the gap in suicide disparities for sexual minority youth? *Pediatrics* **145**, e20194002 (2020).
23. Tsai, A. C., Lucas, M. & Kawachi, I. Association between social integration and suicide among women in the United States. *JAMA Psychiatry* **72**, 987–993 (2015).
24. de Leo, D. Ageism and suicide prevention. *Lancet Psychiatry* **5**, 192–193 (2018).
25. Van Orden, K. A. & Conwell, Y. Issues in research on aging and suicide. *Aging Ment. Health* **20**, 240–251 (2016).
26. Tsai, A. C., Lucas, M., Sania, A., Kim, D. & Kawachi, I. Social integration and suicide mortality among men: 24-year cohort study of US health professionals. *Ann. Intern. Med.* **161**, 85–95 (2014).
27. Perry, S. W. et al. Achieving health equity in US suicides: a narrative review and commentary. *BMC Public Health* **22**, 1360 (2022).
28. Xiao, Y. et al. Patterns of social determinants of health and child mental health, cognition, and physical health. *JAMA Pediatr*. https://doi.org/10.1001/jamapediatrics.2023.4218 (2023).
29. Skinner, A., Osgood, N. D., Occhipinti, J. A., Song, Y. J. C. & Hickie, I. B. Unemployment and underemployment are causes of suicide. *Sci. Adv.* **9**, eadg3758 (2023).
30. Graves, J. M., Abshire, D. A., Mackelprang, J. L., Amiri, S. & Beck, A. Association of rurality with availability of youth mental health facilities with suicide prevention services in the US. *JAMA Netw. Open* **3**, e2021471 (2020).
31. Bower, M. et al. The impact of the built environment on loneliness: a systematic review and narrative synthesis. *Health Place* **79**, 102962 (2023).
32. Kind, A. J. & Buckingham, W. R. Making neighborhood-disadvantage metrics accessible—The Neighborhood Atlas. *N. Engl. J. Med.* **378**, 2456–2458 (2018).
33. Flanagan, B. E., Gregory, E. W., Hallisey, E. J., Heitgerd, J. L. & Lewis, B. A social vulnerability index for disaster management. *J. Homel. Secur. Emerg. Manag.* **8**, 3 (2011).
34. Kirkbride, J. B. et al. The social determinants of mental health and disorder: evidence, prevention and recommendations. *World Psychiatry* **23**, 58–90 (2024).

35. Sussell, A. Suicide rates by industry and occupation—National Vital Statistics System, United States, 2021. *MMWR Morb. Mortal. Wkly Rep*. https://doi.org/10.15585/mmwr.mm7250a2 (2023).
36. Chen, Q. et al. Impacts of land use and population density on seasonal surface water quality using a modified geographically weighted regression. *Sci. Total Environ.* **572**, 450–466 (2016).
37. *CDC/ATSDR SVI Data and Documentation Download | Place and Health* (CDC and ATSDR, accessed 6 May 2023); https://www.atsdr.cdc.gov/placeandhealth/svi/data_documentation_download.html
38. Liu, S. et al. Social vulnerability and risk of suicide in US adults, 2016–2020. *JAMA Netw. Open* **6**, e239995 (2023).
39. Bundy, J. D. et al. Social determinants of health and premature death among adults in the USA from 1999 to 2018: a national cohort study. *Lancet Public Health* **8**, e422–e431 (2023).
40. Masulli, F. & Rovetta, S. Clustering high-dimensional data. In *Proc. Clustering High-Dimensional Data: First International Workshop, CHDD 2012, Naples, Italy, May 15, 2012, Revised Selected Papers* (eds Masulli, F. et al.) 1–13 (Springer, 2015).
41. Zhang, C.-H. & Huang, J. The sparsity and bias of the Lasso selection in high-dimensional linear regression. *Ann. Stat.* **36**, 1567–1594 (2008).
42. Mann, J. J., Michel, C. A. & Auerbach, R. P. Improving suicide prevention through evidence-based strategies: a systematic review. *Am. J. Psychiatry* https://doi.org/10.1176/appi.ajp.2020.20060864 (2021).
43. Mock, C., Quansah, R., Krishnan, R., Arreola-Risa, C. & Rivara, F. Strengthening the prevention and care of injuries worldwide. *Lancet* **363**, 2172–2179 (2004).
44. Thomson, R. M. et al. How do income changes impact on mental health and wellbeing for working-age adults? A systematic review and meta-analysis. *Lancet Public Health* **7**, e515–e528 (2022).
45. Blumenthal, D., Fowler, E. J., Abrams, M. & Collins, S. R. Covid-19—implications for the health care system. *N. Engl. J. Med.* **383**, 1483–1488 (2020).
46. Beseran, E. et al. Deaths of despair: a scoping review on the social determinants of drug overdose, alcohol-related liver disease and suicide. *Int. J. Environ. Res. Public Health* https://doi.org/10.3390/ijerph191912395 (2022).
47. Ramchand, R., Gordon, J. A. & Pearson, J. L. Trends in suicide rates by race and ethnicity in the United States. *JAMA Netw. Open* **4**, e2111563 (2021).
48. Kolak, M., Bhatt, J., Park, Y. H., Padrón, N. A. & Molefe, A. Quantification of neighborhood-level social determinants of health in the continental United States. *JAMA Netw. Open* **3**, e1919928 (2020).
49. Sit, H. F., Hall, B. J. & Li, S. X. Cultural adaptation of interventions for common mental disorders. *Lancet Psychiatry* **10**, 374–376 (2023).
50. Talukder, S. et al. Urban health in the post-2015 agenda. *Lancet* **385**, 769 (2015).
51. Romanello, M. et al. The 2021 report of the *Lancet* Countdown on health and climate change: code red for a healthy future. *Lancet* **398**, 1619–1662 (2021).
52. Bennett, K. J., Borders, T. F., Holmes, G. M., Kozhimannil, K. B. & Ziller, E. What is rural? Challenges and implications of definitions that inadequately encompass rural people and places. *Health Aff.* **38**, 1985–1992 (2019).
53. Dwyer-Lindgren, L. et al. Cause-specific mortality by county, race, and ethnicity in the USA, 2000–19: a systematic analysis of health disparities. *Lancet* https://doi.org/10.1016/S0140-6736(23)01088-7 (2023).
54. Imbens, G. Understanding and misunderstanding randomized controlled trials: a commentary on Deaton and Cartwright. *Soc. Sci. Med.* **210**, 50–52 (2018).
55. Deaton, A. & Cartwright, N. Understanding and misunderstanding randomized controlled trials. *Soc. Sci. Med.* **210**, 2–21 (2018).
56. Lund, C. et al. Social determinants of mental disorders and the Sustainable Development Goals: a systematic review of reviews. *Lancet Psychiatry* **5**, 357–369 (2018).
57. Frank, J. et al. Work as a social determinant of health in high-income countries: past, present, and future. *Lancet* **402**, 1357–1367 (2023).
58. Bhugra, D. & Ventriglio, A. Commercial determinants of mental health. *Int. J. Soc. Psychiatry* **69**, 1065–1067 (2023).
59. Dwyer-Lindgren, L. et al. Life expectancy by county, race, and ethnicity in the USA, 2000–19: a systematic analysis of health disparities. *Lancet* **400**, 25–38 (2022).
60. *Social Determinants of Health (SDOH)* (AHRQ, accessed 1 May 2023); https://www.ahrq.gov/sdoh/index.html
61. Hwong, A. R. et al. Climate change and mental health research methods, gaps, and priorities: a scoping review. *Lancet Planet. Health* **6**, e281–e291 (2022).
62. Helbich, M. et al. Natural environments and suicide mortality in the Netherlands: a cross-sectional, ecological study. *Lancet Planet. Health* **2**, e134–e139 (2018).
63. *International Statistical Classification of Diseases and Related Health Problems: 10th Revision (ICD-10)* (World Health Organization, accessed 10 January 2024); https://icd.who.int/browse10/2019/en
64. *About Multiple Cause of Death, 1999–2020* (CDC, accessed 4 September 2023); https://wonder.cdc.gov/mcd-icd10.html
65. *Multiple Cause of Death 1999–2020* (CDC, accessed 4 September 2023); https://wonder.cdc.gov/wonder/help/mcd.html#Assurance%20of%20Confidentiality
66. Müllner, D. Modern hierarchical, agglomerative clustering algorithms. Preprint at https://arxiv.org/abs/1109.2378 (2011).
67. Murtagh, F. & Legendre, P. Ward's hierarchical agglomerative clustering method: which algorithms implement Ward's criterion? *J. Classif.* **31**, 274–295 (2014).
68. Laszlo, A. M., Hulman, A., Csicsman, J., Bari, F. & Nyari, T. A. The use of regression methods for the investigation of trends in suicide rates in Hungary between 1963 and 2011. *Soc. Psychiatry Psychiatr. Epidemiol.* **50**, 249–256 (2015).
69. Lindén, A. & Mäntyniemi, S. Using the negative binomial distribution to model overdispersion in ecological count data. *Ecology* **92**, 1414–1421 (2011).
70. Bouwmeester, W. et al. Prediction models for clustered data: comparison of a random intercept and standard regression model. *BMC Med. Res. Methodol.* **13**, 19 (2013).
71. Santosa, F. & Symes, W. W. Linear inversion of band-limited reflection seismograms. *SIAM J. Sci. Stat. Comput.* **7**, 1307–1330 (1986).
72. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. B* **58**, 267–288 (1996).
73. Hastie, T., Friedman, J. & Tibshirani, R. *The Elements of Statistical Learning* (Springer, 2001); https://doi.org/10.1007/978-0-387-21606-5
74. Ismail, N. & Zamani, H. Estimation of claim count data using negative binomial, generalized Poisson, zero-inflated negative binomial and zero-inflated generalized Poisson regression models. *Casualty Actuarial Society E-Forum* **41**, 1–28 (2013).
75. Chai, T. & Draxler, R. R. Root mean square error (RMSE) or mean absolute error (MAE). *Geosci. Model Dev. Discuss.* **7**, 1525–1534 (2014).
76. *Rural Classifications* (USDA ERS, accessed 8 September 2023); https://www.ers.usda.gov/topics/rural-economy-population/rural-classifications/

Acknowledgments
This study was supported by funding from the US National Institute of Mental Health (RF1MH134649; Y.X.), American Foundation for Suicide Prevention (YIG-2-133-22; Y.X.), Mental Health Research Network (Y.X.), Google (Y.X.), the US National Institute on Drug Abuse Center for Health Economics of Treatment Interventions for Substance Use Disorder, HCV, and HIV (CHERISH, Y.X.), the US National Institute on Drug Abuse (K24DA061696-01; A.C.T.) and the Artificial Intelligence/Machine Learning Consortium to Advance Health Equity and Researcher Diversity (AIM-AHEAD), a program of the US National Institutes of Health (1OT2OD032581-02-259; Y.X., J.P.). The sponsor had no role in the design and conduct of the study, the collection, management, analysis and interpretation of the data; preparation, review or approval of the manuscript; or the decision to submit the manuscript for publication. We would like to acknowledge Dr L. Snowden for his invaluable contributions to this work. Dr Snowden was engaged in the conceptualization and early writing of this manuscript prior to his passing in January 2025. As a pioneering scholar and leader in mental health services research, his intellectual guidance shaped this paper and many others in the field. We are honored to carry forward his legacy through this work.

Author contributions
Y.X., Y.M. and J.J.M. carried out concept and design. Y.X., Y.M., T.T.B., A.C.T., L.R.S., J.C.-C.C., J.P. and J.J.M. carried out acquisition, analysis or interpretation of data. Y.X. and Y.M. drafted the paper. Y.X., Y.M., T.T.B., A.C.T., L.R.S., J.C.-C.C., J.P. and J.J.M. provided critical revision of the paper for important intellectual content. Y.X. and Y.M. carried out statistical analysis. Y.X. obtained funding. Y.X. provided administrative, technical or material support. Y.X. and Y.M. had full access to all of the data in the study and took responsibility for the integrity of the data and the accuracy of the data analysis. Y.X. and Y.M. have made equal contributions.

Competing interests
A.C.T. reports an honorarium from Elsevier for his work as co-editor in chief of the Elsevier-owned journal *SSM-Mental Health*. J.J.M. reports royalties for commercial use of the C-SSRS from the Research Foundation for Mental Hygiene and for the Columbia Psychiatry Pathways App from Columbia University. All other authors report no conflicts of interest.

Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s44220-025-00424-4.

Correspondence and requests for materials should be addressed to Yunyu Xiao.

Peer review information *Nature Mental Health* thanks the anonymous reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

© The Author(s), under exclusive licence to Springer Nature America, Inc. 2025

[1]Department of Population Health Sciences, Weill Cornell Medicine, NewYork-Presbyterian, New York, NY, USA. [2]School of Public Health, University of California, Berkeley, CA, USA. [3]Center for Global Health and Mongan Institute, Massachusetts General Hospital, Boston, MA, USA. [4]Harvard Medical School, Boston, MA, USA. [5]School of Social Welfare, University of California, Berkeley, CA, USA. [6]Departments of Psychiatry and Radiology, Columbia University Irving Medical Center, Columbia University, New York, NY, USA. [7]Division of Molecular Imaging and Neuropathology, New York State Psychiatric Institute, New York, NY, USA. [8]These authors contributed equally: Yunyu Xiao, Yuan Meng. ✉e-mail: yux4008@med.cornell.edu

nature portfolio

Corresponding author(s): Yunyu Xiao

Last updated by author(s): Mar 5, 2025

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used.
Data analysis	Data analysis was performed using R version 4.4.2, leveraging several specialized packages to ensure robust and reproducible results. The "NbClust" package (version 3.0.1) was used to determine the optimal number of clusters, while "cluster" (version 2.1.6) facilitated clustering techniques and visualization. Core statistical computations and modeling were conducted using the "stats" package (version 3.6.2). For regression modeling, including elastic net and regularization, the "glmnet" package (version 4.1.8) was employed. Additionally, geographically weighted regression was implemented using the "spgwr" package (version 0.6-37). These tools collectively supported comprehensive data processing and analysis. Code is available at github.com/X-PLORE-Lab/sdoh_clustering_suicide.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

All raw data in this study are publicly available and can be retrieved from the National Center for Health Statistics, Centers for Disease Control and Prevention Mortality Data (https://www.cdc.gov/nchs/nvss/deaths.htm) and U.S. Agency for Healthcare Research and Quality Social Determinants of Health Database (https://www.ahrq.gov/sdoh/data-analytics/sdoh-data.html).

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	Sex and age were considered as subgroups in the association analysis between SDOH clusters and suicide rates.
Reporting on race, ethnicity, or other socially relevant groupings	Race/ethnicity was considered as subgroups in the association analysis between SDOH clusters and suicide rates.
Population characteristics	The statistics of the population characteristics were reported in Table 1.
Recruitment	n.a.
Ethics oversight	This study was approved by the Weill Cornell Medicine Institutional Review Board. Informed consent was waived due to the use of de-identified data.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	This study used quantitative method, focusing on the multi-dimensional SDOH clustering and correlations between SDOH clusters and suicide rates.
Research sample	Research samples were 3109 counties in 2009 and 3018 in 2014 and 2019 as the unit of analysis. Samples included 149771919 males (49.1%) and 154953998 (50.9%) females in 2009, 155831092 males (49.2%) and 160869671 females (50.8%) in 2014, and 160567866 males (49.2%) and 165524240 females (50.8%) in 2019. A total of 49170152 Hispanic (16.1%), 39048245 non-Hispanic Black (12.8%), and 199164305 non-Hispanic White (65.4%) in 2009, 55194122 Hispanic (17.4%), 41242256 non-Hispanic Black (13.0%), and 200192002 non-Hispanic White (63.2%) in 2014, and 60368161 Hispanic (18.5%), 43166073 non-Hispanic Black (13.2%), and 200081781 non-Hispanic White (61.4%) in 2019. Data for individuals under 15 years (2008-2010: 0.04%; 2013-2015: 0.1%; 2018-2020: 0.2%) and "Other" racial/ethnic subgroups (2008-2010: 2.2%; 2013-2015: 2.5%; 2018-2020: 3.1%) were not included due to small sample sizes and anonymity protection.
Sampling strategy	The sample size were determined based on the availability of the AHRQ SDOH dataset and county-level suicide rates. All counties in U.S. mainland, including samples of 3109 counties for 2009 and 3018 for 2014 and 2019, were included.
Data collection	Data were downloaded directly from the Internet via websites (SDOH data: https://www.ahrq.gov/sdoh/data-analytics/sdoh-data.html; suicide rates: https://wonder.cdc.gov/mcd-icd10.html). We were unaware of the study hypothesis to prevent expectancy effects
Timing	Data were collected from 12/30/2022 to 01/06/2023.
Data exclusions	Our study involved a meticulous selection of Social Determinants of Health (SDOH) variables from the Agency for Healthcare Research and Quality (AHRQ) database, spanning three key years: 2009, 2014, and 2019. Initially, the database contained 651 SDOH

variables in 2009, 927 in 2014, and 884 in 2019. We focused on variables excluding geographical data provided by AHRQ. Our selection criteria were twofold: 1) Data completeness, ensuring that the SDOH variables were available for all counties in the US mainland, and 2) Data availability, requiring the presence of these variables across all three years. Variables that did not meet these criteria were excluded. This rigorous process led to the final selection of 284 SDOH variables that were consistently available across the specified years and counties. These variables were used to model SDOH clusters and assess their associations with suicide rates.

Non-participation

Individuals under 15 years (2008-2010: 0.04%; 2013-2015: 0.1%; 2018-2020: 0.2%) and "Other" racial/ethnic subgroups (2008-2010: 2.2%; 2013-2015: 2.5%; 2018-2020: 3.1%) were not included due to small sample sizes and anonymity protection.

Randomization

We used 3109 counties in 2009 and 3018 in 2014 and 2019 as research samples.
Suicide rates were calculated in each sample by sex (male and female), age groups (15-24, 25-64, >65 years), and race/ethnicity (Hispanic, Black, White, and Other [including Asian, Pacific Islander, American Indian, and Alaska Native.])

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Plants

Seed stocks

n.a.

Novel plant genotypes

n.a.

Authentication

n.a.