SSM - Population Health 24 (2023) 101511

Contents lists available at ScienceDirect

SSM - Population Health

journal homepage: www.elsevier.com/locate/ssmph

Validation of a geospatial aggregation method for congressional districts and other US administrative geographies

Ben R. Spoer [a,*], Alexander S. Chen [a], Taylor M. Lampe [a], Isabel S. Nelson [a], Anne Vierse [a], Noah V. Zazanis [a], Byoungjun Kim [a], Lorna E. Thorpe [a], Subu V. Subramanian [b], Marc N. Gourevitch [c]

[a] New York University Grossman School of Medicine, Department of Population Health, Division of Epidemiology, New York, NY, USA
[b] Harvard T.H. Chan School of Public Health, Department of Social and Behavioral Sciences, Boston, MA, USA
[c] New York University Grossman School of Medicine, Department of Population Health, New York, NY, USA

ARTICLE INFO

Keywords:
Geospatial analysis
Spatial data aggregation
Congressional districts
US administrative geographies

ABSTRACT

Stakeholders need data on health and drivers of health parsed to the boundaries of essential policy-relevant geographies. US Congressional Districts are an example of a policy-relevant geography which generally lack health data. One strategy to generate Congressional District heath data metric estimates is to aggregate estimates from other geographies, for example, from counties or census tracts to Congressional Districts. Doing so requires several methodological decisions. We refine a method to aggregate health metric estimates from one geography to another, using a population weighted approach. The method's accuracy is evaluated by comparing three aggregated metric estimates to metric estimates from the US Census American Community Survey for the same years: Broadband Access, High School Completion, and Unemployment. We then conducted four sensitivity analyses testing: the effect of aggregating counts vs. percentages; impacts of component geography size and data missingness; and extent of population overlap between component and target geographies. Aggregated estimates were very similar to estimates for identical metrics drawn directly from the data source. Sensitivity analyses suggest the following best practices for Congressional district-based metrics: utilizing smaller, more plentiful geographies like census tracts as opposed to larger, less plentiful geographies like counties, despite potential for less stable estimates in smaller geographies; favoring geographies with higher percentage population overlap.

1. Introduction

Most public health practice conceptual models begin with using data to describe a public health challenge, often with the premise that data should be parsed as close to the area of focus as possible (McNabb et al., 2002) to increase the timeliness and appropriateness of the proposed response. However, though health-related data are currently widely available for states, (America's Health Rankings, 2022) counties (County Health Rankings and Roadmaps, 2023) and—more recently—cities,(Gourevitch et al., 2019), US Congressional Districts (CDs) are a salient example of a policy-relevant geography that lacks data (Eberth et al., 2019; Mansfield et al., 2007; Siegel et al., 2015). Although sociodemographic and some social determinants of health data are available from the US Census for CD populations (U.S. Census Bureau, 2023), these data often lag the redistricting of new CD boundaries by a year or more, causing spatial and temporal misalignment in available CD data. Furthermore, recent ongoing shocks to the job market and educational systems related to the SARS COV-2 pandemic have caused more volatility in these metrics than is normal. Looking beyond the US Census, data on specific measures of health are not usually collected at the CD level. For example, vital statistics measures like cardiovascular disease death rate, cancer mortality, and firearm deaths, and measures of the prevalence of chronic diseases like diabetes or hypertension, are rarely available for CDs.

To fill this gap, researchers need generalizable methods to generate estimates at the CD level, as well as for other US administrative geographies. Generating these estimates requires geospatial methods that aggregate metric estimates originally calculated for other geographic areas into CD estimates using rigorous processes to maintain accuracy. These methods, which sum population-weighted estimates from

* Corresponding author. 180 Madison Ave, M-18, New York, NY, 10016, USA.
E-mail address: Benjamin.spoer2@nyulangone.org (B.R. Spoer).

https://doi.org/10.1016/j.ssmph.2023.101511
Received 19 May 2023; Received in revised form 1 September 2023; Accepted 3 September 2023
Available online 4 September 2023

component geographies, have their roots in demography research (Wright, 1936; Wu et al., 2005), and have been used to generate CD estimates for life expectancy and opioid overdose metrics, among others (Islami et al., 2023; Rolheiser et al., 2018; Takai et al., 2022; Zhang et al., 2014). Similar methods have also been used to aggregate index values from census tracts to ZIP codes (Noelke et al.). Still, they remain outside of mainstream methods for estimating and validating a wide range of policy-relevant health metrics for CDs and other administrative geographies. This may be in part because aggregating metric estimates from one geography to another requires numerous methodological decisions, including identifying the most appropriate component geographies, applying systematic approaches to accommodate missingness in component geographies, and addressing imperfect population overlap between component and target geographies.

To address this need for CD-level health-related data, and to support the release of an online Congressional District Health Dashboard (CDHD)(Dashboard Team, 2023b), we refined a method to aggregate census tract and county-level metric estimates into CD metric estimates using the most recent available years of county and census tract data. We have presented these metric estimates on the CDHD (www.congressionaldistricthealthdashboard.org). The goal of the present article is to summarize our methods and perform sensitivity analyses to evaluate the validity of aggregated estimates. We do so by comparing a subset of metric estimates, aggregated using ACS county and census tract data from the 116th congress, to 'gold standard' CD-level estimates drawn directly from ACS for the same timeframe.

2. Material and methods

2.1. Deriving CD estimates

Our method for aggregating CD-level estimates from imperfectly nested component geographies (census tracts and counties) accounts for heterogeneous population distributions better than methods based exclusively on geographic overlap (Holt et al., 2004; Liu & Martinez, 2019; Rolheiser et al., 2018; Schroeder, 2017; Wilson & Mansfield, 2012). Population counts from census blocks, the smallest available census geography, are used to population-weight estimates before then being re-aggregated to fit the boundaries of the target geography. Our sample includes all 435 congressional districts plus the District of Columbia (using boundaries from the 116th US Congress), as aggregated from county (n=3,143) and census tract (n=73,050) geographies defined by the 2010 Census.

To aggregate metric estimates, we first create two population crosswalks between component geographies and target geographies (CDs). One crosswalk was created to define the relationship between census tract and CD populations; the second was created to define the relationship between county and CD populations. We identified the smallest component geography that perfectly nests in tracts, counties and CDs, census blocks. We use census blocks to create a relationship between tracts and CDs, or counties and CDs. Then, we calculate the overlapping population count for a given tract-CD or county-CD relationship, which equals the sum of the population of the blocks assigned to both geographies.

Using these population crosswalks, we then create population weights. Different weights are calculated depending on whether the metric unit is count or percentage. In both instances, population counts from tracts or counties with missing metric estimates are dropped from the numerator and denominator in the weight calculation.

When aggregating count metrics, we create population weights by dividing the overlapping population count (from the population crosswalk) by the population count of full component geography (tract or county). Count values need only be weighted by the proportion of the component geography population that overlaps with the target geography population, to approximate what proportion of the individuals residing in that component geography contribute to the target geography (Equation One). When aggregating percent estimates, we create population weights by dividing the overlapping population count (from the population crosswalk) by the full target population count. For percent values it is necessary to weight by the proportion of the target geography contained in the component geography, to approximate what proportion of the target geography percentage estimate should come from the component geography (Equation Two).

We multiply the population weight by the tract or county metric estimate, then sum all weighted estimates to calculate the final CD estimate for the metric of interest (Equation Three).

Equation One: Calculating Population Weights for Aggregating Count Values

$$P_{component\ geography|CD\ i} = \frac{Pop_{comoponent\ geography|CD\ i}}{Pop_{component\ geography\ i}}$$

where:

- P represents the population weight for the component geography (i)
- $Pop_{component\ geography|CD\ i}$ represents the component geography population (i) that overlaps with the target geography (CD) population
- $Pop_{component\ geography}$ represents the full component geography population (i)

Equation Two: Calculating Population Weights for Aggregating Percent Values

$$P_{component\ geography|CD\ i} = \frac{Pop_{component\ geography|CD\ i}}{Pop_{CD}}$$

where:

- P represents the population weight for the component geography (i)
- $Pop_{component\ geography|CD\ i}$ represents the component geography population (i) that overlaps with the target geography (CD) population
- Pop_{CD} represents the full CD population (i)

Equation Three: Calculating Aggregated Estimates

$$Est_{CD} = \sum_{i=1}^{n} Est_{(component\ geography\ i)} * P_{component\ geography|CD\ i}$$

where:

- Est represents the metric estimate (count or percentage) for the component geography (i)
- n represents the number of component geographies overlapping with the CD
- P represents the population weight for the component geography (i)

2.2. Selection of comparison metrics

To evaluate the accuracy of the aggregated estimates, we focus on three metrics produced by ACS and compare our aggregation results to CD estimates from ACS. CD estimates for other metrics are not available for comparison. Though the ultimate goal of this work was to create estimates for 118th Congress geographies, the validation analyses presented here used metrics from 116th Congress geographies (and 2019 tracts or counties for aggregation calculations) to enable these comparisons; at the time of analysis, ACS had not released 118th Congress CD metric estimates. We also aggregate and validate metric estimates for racial/ethnic subgroup populations to validate methods for smaller population subgroups (see Supplemental Table 1 for definitions of race/ethnicity variables).

We utilize three metrics: percentage of population aged $\geq$ 25 with a high school diploma or higher degree (high school completion), percentage population aged $\geq$ 16 years that was unemployed but seeking

work (unemployment), and percentage population with a high-speed broadband internet connection (broadband access). See Supplemental Table 1 for specific metric definitions. These three metrics represent social determinants of health, and differences in their distributions help demonstrate the flexibility of the utilized aggregation methods. Specifically, among ACS metrics calculated for the CDHD, broadband access had the largest range between urban and rural areas, unemployment had the lowest prevalence, and high school completion had the largest range across racial/ethnic subgroups. Component data were provided by ACS for the 116th Congress, or the timeframe consistent with the 116th (2019). These data are not contemporaneous with our current congressional session (118th Congress as of this writing), but are the most recent data available from ACS for CDs, counties, and census tracts. The data available on the CDHD are aligned with the 118th CD boundaries, and we update the website's estimates as CD boundaries are redrawn.

3. Calculation

3.1. Sensitivity analyses

We perform sensitivity analyses to validate the aggregation method and optimize the rigor of our CD estimates. We calculate summary measures (mean, median, standard deviation (SD)) for derived estimates. Differences between ACS and derived estimates were assessed using mean absolute difference, median absolute difference, and root mean square error (RMSE). These measures of difference are frequently utilized when comparing geospatially aggregated estimates to ACS values(Liu & Martinez, 2019; McVeigh et al., 2016; Schroeder, 2007; Zoraghein & Leyk, 2018). We also calculate minimum and maximum error, and IQR of error, to characterize the full error distribution.

The first of four sensitivity analyses focuses on metric estimates expressed as percentages. For a subset of demographic subgroups in each of our selected measures, percentage variables are not directly available from ACS, requiring analysts to decide between aggregating numerator/denominator counts and calculating the percentage at the CD-level vs. aggregating percentages from the component geography. For each metric we first aggregate the numerator and denominator from tract to the CD-level, and then calculate the final percentage metric estimate at the CD-level. We then also calculate the percentage for each metric estimate at the tract-level and aggregate those tract percentages to the CD-level. Estimates from each calculation method were compared against ACS 116th CD values in an effort to validate metric estimates.

Our second sensitivity analysis assesses whether utilizing census tracts or counties as our component geography produced more accurate aggregated estimates. Our aggregation method assumes that the outcome we measure in a given metric is uniformly distributed across the start geography. Larger geographic areas (e.g., counties) with larger populations may demonstrate increased variation across space, potentially violating this assumption. Additionally, counties require more population weighting than do census tracts to fit into CDs (Fig. 1). Conversely, estimates for smaller geographies (e.g. tracts) may be less reliable because of statistical 'noise' introduced by modeled estimates or imprecision in estimates generated using smaller sample sizes. To determine which component geographies produced more accurate estimates, we compare estimates aggregated using either county or tract to ACS estimates (see Fig. 2).

Our third sensitivity analysis assesses whether the accuracy of aggregated estimates is affected by the extent to which component geography population (census tract or county) overlaps with CD population. We calculate population overlap by averaging the component geographies' population weights (i.e. proportion of component geography population contained in the target CD) across each CD(Rolheiser et al., 2018). A value of one indicates 100% of the component geography's population also resided in the CD. Lower values indicate the component geographies' population was less completely nested within the CD. We divide CDs into population overlap quartiles and compare CD estimates within each quartile to ACS estimates.

Our fourth, and final, sensitivity analysis evaluates whether missing component geography estimates affect the accuracy of aggregated CD estimates. Though missingness is rare in ACS data, it is more common in other publicly available health-related datasets. Missingness was evaluated for both tract- and county-to-CD aggregation. We induce random tract or county missingness, and then compare estimates aggregated with induced missingness to ACS estimates. For each metric, we sample 10% of CDs, representing an equal distribution of CDs across urbanicity categories (Dashboard Team, 2023a) that were not missing estimates for any component geographies (n=48). Component geography missingness is then randomly induced for each of these CDs so that *x%* of component geography estimates are artificially set to missing, *x%* being an increasing percentage of the component geography's population, ranging from 5% to 50%, in 5% increments, to simulate varying levels of population missingness. Then we aggregate a new estimate from the component geography estimates that includes induced missingness. This process was repeated 50 times for each *x%* threshold to create a sample of 50 aggregate estimates per CD in which *x%* of the component geography estimates are missing at random. The estimates calculated with induced missingness are compared to CD estimates calculated using ACS congressional district data.

4. Results

Table 1 presents two categories of CD estimates for the three ACS metrics; the first is drawn directly from ACS, the second displays

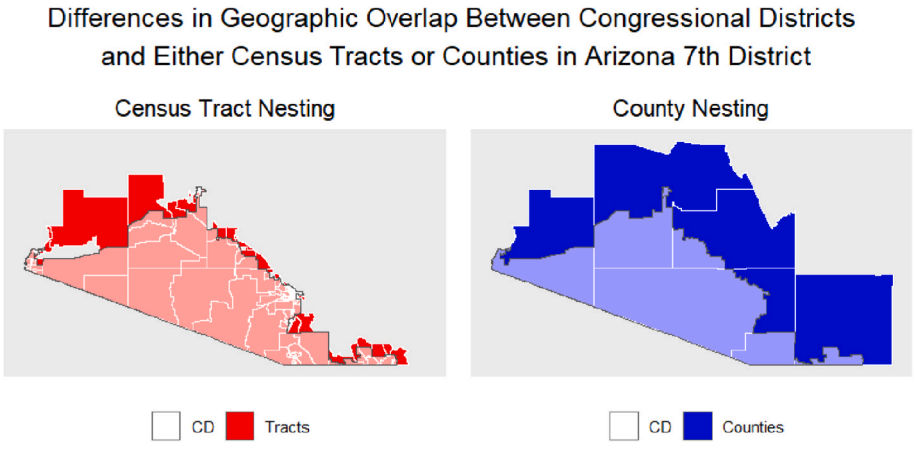

Fig. 1. County and Census Track nesting within Congressional Districts: an illustration

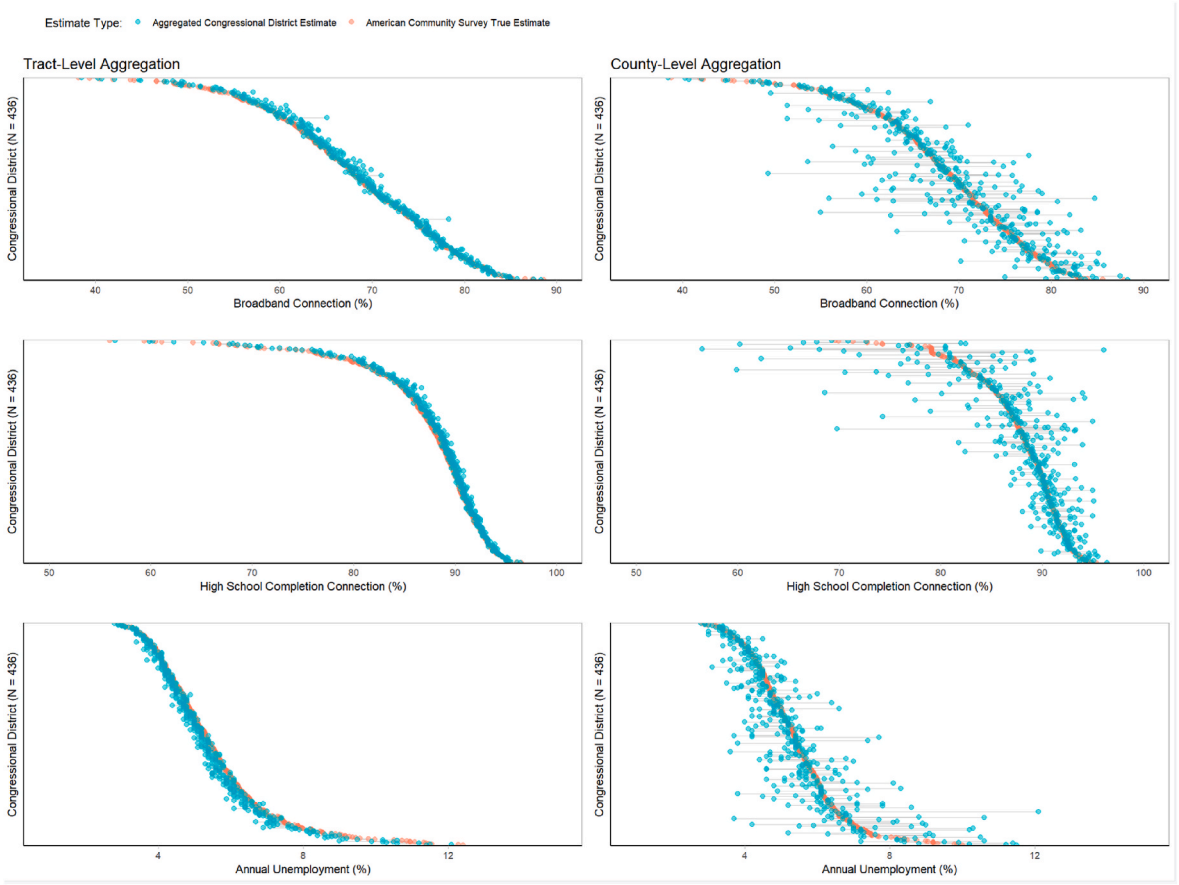

Fig. 2. Differences between American community survey true estimates and aggregated congressional district estimates by metric.

aggregated estimates produced using percent estimates, with tracts as the component geography. Aggregated estimates were very similar to ACS estimates, with a mean absolute error less than 1.4 percentage points, median absolute error less than 1.0, and RMSE less than 2 percentage points. Minimum and maximum error and IQR of error highlight that there are error outliers, especially among subgroups, but error values (those within the IQR) are generally small. Error values for each metric for the entire distribution of aggregated estimates are visible in Figure Two, using the percent aggregation method.

In the first sensitivity analysis we explored whether calculating percentage estimates at the CD level after aggregation of count values produced more or less accurate estimates than calculating percentage estimates at the tract level and then aggregating. The minimum and maximum error and IQR of error metrics follow a similar pattern to that described above-there are some large minimum and maximum values, especially among smaller population subgroups, and the IQR of error tends to be similar in size to measures of central tendency (Supplemental Table 2). Both methods produced estimates very similar to ACS estimates.

In the second sensitivity analysis we evaluated whether deriving CD estimates from tracts or counties produced values closer to ACS estimates. Summary statistics were similar to ACS estimates across metrics, subgroups, and component geographies (Table 2a). Across metrics, median absolute error was generally higher for county-derived estimates for broadband access and high school completion, but similar between tract and county for unemployment. Minimum and maximum error were sometimes high, especially among smaller subgroup populations, but IQR of error was generally smaller. IQR of error was larger when aggregating from counties than from tracts.

In the third sensitivity analysis, we explored whether population overlap between component geographies (census tracts and counties) and CDs impacted aggregated CD estimate accuracy, stratified by population overlap quartile. In general, tracts overlap very well with CDs, while there is more variation in population overlap between counties and CDs (Fig. 3). Results for tracts are displayed in Table 2b, and for counties in Table 2c. Differences across population percentage overlap quartiles were overall small for high school completion for tracts, but somewhat larger for counties. For unemployment, median absolute error and IQR of error decreased as population overlap quartile increased for county-derived estimates, and fluctuated across quartiles for tract estimates. As with high school completion, error for unemployment and broadband access was larger for county than for tract. In the lowest population overlap quartile for county-derived broadband access estimates, median absolute error was 4.2 and IQR of error was 0.73; these are among the largest error values produced by these analyses, yet are still below 5.

For all tract-derived metrics, measures of error tended to be low and similar across quartiles, with a slight decrease in error measurements as population percentage overlap quartiles increased. Mean absolute error and RMSE were not significantly larger than median standard error, indicating minimal outliers. Unlike tract-derived estimates, all measures of error for county-derived metrics exhibited a gradient, with CDs in lower quartiles of population percentage overlap (where component geography population comprised smaller proportions of the CD population) having larger measures of error. These patterns held true for minimum and maximum error and IQR of error. Mean absolute error and RMSE values were larger than median absolute error across all quartiles, indicating some outliers.

Results of the fourth sensitivity analysis identified that summary statistics of estimates with induced population missingness remained close to ACS estimates, even at high levels of missingness, for both tract and county component geographies (Tables 3a and 3b). CD summary

Table 1
Comparison of 2019 US Census American Community Survey (ACS) 116th Congressional District estimates to Congressional District estimates aggregated from tract percent estimates.

Metric	Demographic group	Method	Mean (SD)	Median	Mean Absolute Error	Median Absolute Error	Root Mean Square Error	Minimum and Maximum Error	IQR of error
Broadband access	Total	ACS	68.70 (9.24)	69.10	0.57	0.40	0.81	−1.54, 4.87	0.69
		Aggregated	68.25 (9.47)	68.58					
High school completion	Total	ACS	87.8 (5.9)	89.4	0.41	0.31	0.54	−0.54, 2.17	0.41
		Aggregated	87.5 (6.1)	89.0					
	Asian	ACS	86.21 (6.06)	87.27	1.30	0.86	1.99	−13.99, 5.30	1.75
		Aggregated	86.68 (6.15)	87.85					
	Black	ACS	86.77 (4.77)	87.03	0.80	0.61	1.26	−4.57, 13.98	0.76
		Aggregated	86.29 (4.99)	86.65					
	Hispanic	ACS	70.70 (8.90)	71.31	1.07	0.89	1.37	−4.47, 5.64	1.28
		Aggregated	70.24 (9.12)	70.59					
	Other	ACS	75.38 (8.18)	75.87	1.35	0.99	1.80	−5.42, 6.94	2.11
		Aggregated	75.12 (8.31)	75.57					
	White	ACS	92.93 (3.12)	93.39	0.16	0.10	0.28	−0.92, 2.76	0.19
		Aggregated	92.84 (3.16)	93.32					
	Female	ACS	88.49 (5.88)	90.00	0.44	0.33	0.58	−0.57, 2.36	0.45
		Aggregated	88.07 (6.09)	89.62					
	Male	ACS	87.14 (6.02)	88.65	0.36	0.27	0.50	−1.09, 2.22	0.48
		Aggregated	86.86 (6.18)	88.38					
Unemployment	Total	ACS	5.41 (1.55)	5.20	0.18	0.13	0.26	−1.87, 0.26	0.19
		Aggregated	5.58 (1.66)	5.30					
	Asian	ACS	4.27 (1.53)	4.17	0.54	0.32	0.88	−6.75, 3.60	0.63
		Aggregated	4.27 (1.64)	4.14					
	Black	ACS	9.07 (2.71)	8.90	0.87	0.55	1.41	−8.41, 3.59	0.84
		Aggregated	9.58 (3.07)	9.25					
	Hispanic	ACS	6.11 (1.83)	5.70	0.41	0.23	0.81	−9.89, 2.31	0.46
		Aggregated	6.22 (2.12)	5.74					
	Other	ACS	7.63 (2.36)	7.08	0.71	0.46	1.01	−3.01, 5.82	1.00
		Aggregated	7.35 (2.27)	6.89					
	White	ACS	4.44 (1.04)	4.30	0.15	0.11	0.22	−1.87, 0.26	0.19
		Aggregated	4.57 (1.09)	4.46					
	Female	ACS	4.90 (1.51)	4.60	0.15	0.11	0.22	−1.51, 0.37	0.18
		Aggregated	5.03 (1.58)	4.74					
	Male	ACS	5.10 (1.57)	4.80	0.22	0.13	0.35	−2.31, 0.26	0.23
		Aggregated	5.31 (1.74)	4.99					

Table 2a
Comparison of 2019 tract vs. county percent-derived Congressional District estimates to US Census American Community Survey (ACS) 116th Congressional District estimates (Sensitivity analysis #2).

Metric	Demographic group	Method	Mean (SD)	Median	Mean Absolute Error	Median Absolute Error	Root Mean Square Error	Minimum and Maximum Error	IQR of error
Broadband access	Total	Tract	68.25 (9.47)	68.58	0.57	0.4	0.81	−1.54, 4.87	0.69
		County	68.7 (8.48)	69.42	2.18	0.93	3.62	−19.66, 13.2	1.76
		ACS	68.7 (9.24)	69.10	–	–	–	–	–
High school completion	Total	Tract	87.45 (6.1)	89.01	0.41	0.31	0.54	−0.54, 2.17	0.41
		County	87.88 (4.38)	88.83	1.83	0.63	3.63	−22.71, 17.00	1.20
		ACS	87.8 (5.9)	89.4	–	–	–	–	–
	Asian	Tract	86.68 (6.15)	87.85	1.3	0.86	1.99	−13.99, 5.30	1.75
		County	86.31 (5.18)	87.39	2.02	1.12	3.25	−18.01, 11.49	2.17
		ACS	86.21 (6.06)	87.27	–	–	–	–	–
	Black	Tract	86.29 (4.99)	86.65	0.78	0.61	1.26	−4.57, 13.98	1.75
		County	86.46 (4.34)	87.02	1.24	0.67	1.85	−6.49, 6.49	1.41
		ACS	86.77 (4.77)	87.03	–	–	–	–	–
	Hispanic	Tract	70.24 (9.12)	70.59	1.07	0.89	1.37	−4.47, 5.64	1.28
		County	69.78 (8.01)	69.74	2.66	1.41	4.29	−13.10, 22.08	2.78
		ACS	70.70 (8.90)	71.31	–	–	–	–	–
	Other	Tract	75.12 (8.31)	75.57	1.35	0.99	1.8	−5.42, 6.94	2.11
		County	74.68 (7.28)	74.42	2.69	1.41	4.29	−19.30, 23.92	2.71
		ACS	75.38 (8.18)	75.87	–	–	–	–	–
	White	Tract	92.84 (3.16)	93.32	0.16	0.1	0.28	−0.92, 2.76	0.19
		County	93.68 (2.87)	93.68	0.69	0.3	1.25	−8.30, 4.50	0.61
		ACS	92.93 (3.12)	93.39	–	–	–	–	–
	Female	Tract	88.07 (6.09)	89.62	0.44	0.33	0.58	−0.60, 2.36	0.45
		County	88.54 (4.39)	89.51	1.77	0.6	3.55	−22.20, 16.80	1.20
		ACS	88.49 (5.88)	90.00	–	–	–	–	–
	Male	Tract	86.86 (6.18)	88.38	0.36	0.27	0.5	−1.09, 2.22	0.48
		County	87.21 (4.45)	88.18	1.91	0.67	3.73	−23.65, 17.10	1.38
		ACS	87.14 (6.02)	88.65	–	–	–	–	–
Unemployment	Total	Tract	5.58 (1.66)	5.3	0.18	0.13	0.26	−1.87, 0.26	0.19
		County	5.43 (1.24)	5.35	0.46	0.21	0.78	−3.05, 5.56	0.37
		ACS	5.41 (1.55)	5.20	–	–	–	–	–
	Asian	Tract	4.27 (1.64)	4.14	0.54	0.32	0.88	−6.75, 3.60	0.63
		County	4.23 (1.33)	4.09	0.53	0.35	0.82	−3.19, 7.62	0.69
		ACS	4.27 (1.53)	4.17	–	–	–	–	–
	Black	Tract	9.58 (3.07)	9.25	0.87	0.55	1.41	−8.41, 3.59	0.84
		County	9.41 (2.49)	9.4	0.99	0.50	1.61	−10.34, 4.24	1.00
		ACS	9.07 (2.71)	8.90	–	–	–	–	–

(*continued on next page*)

Table 2a (*continued*)

Metric	Demographic group	Method	Mean (SD)	Median	Mean Absolute Error	Median Absolute Error	Root Mean Square Error	Minimum and Maximum Error	IQR of error
	Hispanic	Tract	6.22 (2.12)	5.74	0.41	0.23	0.81	−9.89, 2.31	0.46
		County	6.21 (1.77)	5.87	0.51	0.3	0.77	−3.10,4.09	0.58
		ACS	6.11 (1.83)	5.70	–	–	–	–	–
	Other	Tract	7.35 (2.27)	6.89	0.71	0.46	1.01	−3.01,5.82	1.00
		County	7.65 (2.16)	7.16	0.63	0.43	0.94	−4.45, 5.92	0.84
		ACS	7.63 (2.36)	7.08	–	–	–	–	–
	White	Tract	4.57 (1.09)	4.46	0.15	0.11	0.22	−1.87, 0.26	0.19
		County	4.41 (0.9)	4.3	0.27	0.13	0.46	−1.60, 3.40	0.25
		ACS	4.44 (1.04)	4.30	–	–	–	–	–
	Female	Tract	5.03 (1.58)	4.74	0.15	0.11	0.22	−1.51, 0.37	0.18
		County	4.92 (1.23)	4.8	0.43	0.2	0.72	−3.05, 4.65	0.36
		ACS	4.90 (1.51)	4.60	–	–	–	–	–
	Male	Tract	5.31 (1.74)	4.99	0.22	0.13	0.35	−2.31, 0.26	0.23
		County	5.13 (1.25)	5.0	0.47	0.221	0.8	−2.67, 6.08	0.41
		ACS	5.10 (1.57)	4.80	–	–	–	–	–

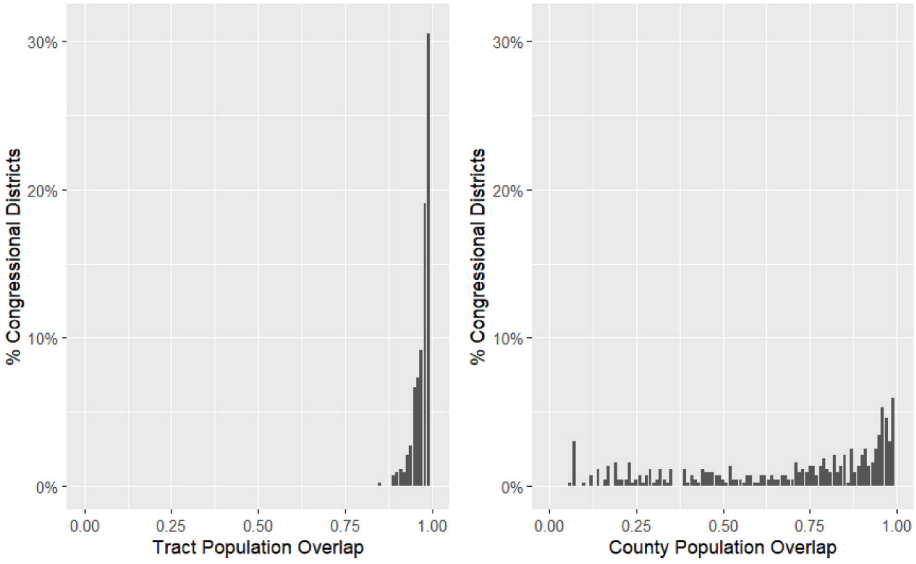

Fig. 3. Percentage of Congressional Districts by Proportion Population Overlap with Tracts and Counties, 2019 (Sensitivity analysis #3).

statistics across increasing levels of induced population missingness and for both component geographies shift, but differences are not dramatic-within 1% for high school completion and unemployment, and 5% for broadband access. Across metrics, error measures generally increase as the percentage missing population increases. Similarly, minimum and maximum error, and IQR of error, also increase as population missingness increases; for the broadband access metric at the tract level, for example, IQR of error increased from 0.65 in the lowest missingness strata to 1.55 in the highest missingness strata.

More pronounced differences were observed when utilizing county as the start geography, as opposed to tract. For broadband access at the county level, IQR of error increased from 1.69 in the lowest missingness strata to 7.74 in the highest missing strata. This metric provides the most extreme example among the three metrics analyzed. Median absolute error remains relatively low, demonstrating that even with high levels (>50%) of induced population missingness, CD estimates are overall similar to ACS estimates. The corresponding low levels of mean absolute error and RMSE across component geographies suggest there are few extreme outliers.

5. Discussion

We describe a systematic approach to validating a population weighted method for aggregating US Census metric estimates from census tracts and counties to CDs. Overall, aggregated estimates were very similar to estimates drawn from ACS, suggesting the approach is

Table 2b
Comparison of 2019 tract percent-derived total population Congressional District estimates to US Census 116th Congressional District estimates, by quartile of population overlap (Sensitivity analysis #3).

Metric	Tract Population Overlap Quartile	Mean (SD)	Median	Mean Absolute Error	Median Absolute Error	Root Mean Square Error	Minimum and Maximum Error	IQR of Error
Broadband access	1 (0.85–0.97)	72.36 (8.51)	73.01	0.6	0.45	0.92	−0.86, 4.87	0.73
	2 (0.97–0.98)	67.9 (9.23)	67.97	0.63	0.48	0.85	−1.54, 2.24	0.82
	3 (0.98–0.99)	66.77 (10)	67.48	0.54	0.36	0.75	−0.66, 3.6	0.67
	4 (0.99–1.00)	65.98 (8.88)	65.58	0.51	0.38	0.70	−1.15, 2.82	0.57
High school completion	1	86.37 (6.89)	87.98	0.48	0.42	0.6	−0.54, 1.84	0.43
	2	86.51 (6.88)	88.32	0.45	0.38	0.58	−0.33, 1.91	0.43
	3	87.86 (5.88)	89.44	0.37	0.27	0.5	−0.19, 1.89	0.41
	4	89.07 (3.94)	89.94	0.23	0.23	0.46	−0.28, 2.17	0.34
Unemployment	1	5.66 (1.67)	5.32	0.17	0.1	0.3	−1.87, 0.1	0.19
	2	5.51 (3)	5.51	0.19	0.15	0.27	−1.02, 0.26	0.22
	3	5.33 (3)	5.33	0.18	0.13	0.26	−1.05, 0.05	0.17
	4	5.14 (1.16)	5.03	0.16	0.12	0.22	−1.06, 0.06	0.14

Table 2c
Comparison of 2019 county percent-derived total population Congressional District estimates to US Census 116th Congressional District estimates, across quartile of population overlap (Sensitivity Analysis #3).

Metric	County Population Overlap Quartile	Mean (SD)	Median	Mean Absolute Error	Median Absolute Error	Root Mean Square Error	Minimum and Maximum Error	IQR of Error
Broadband access	1 (0.06–0.47)	73.1 (4.94)	71.99	4.8	4.2	6.19	−19.66, 13.2	8.54
	2 (0.47–0.8)	72.21 (7.04)	73.67	2.36	1.86	3.23	−12.06, 8.57	3.18
	3 (0.8–0.95)	67.54 (8.38)	67.24	1.07	0.66	1.67	−7.9, 6.67	1.34
	4 (0.95–1.0)	61.95 (8.15)	62.84	0.49	0.27	0.84	−1.56, 4.89	0.51
High school completion	1	85.53 (4.38)	86.1	4.88	3.6	6.75	−22.71, 17.0	6.49
	2	88.65 (4.42)	89.6	1.75	1.3	2.57	−12.41, 8.32	2.33
	3	88.98 (4.27)	90.05	0.54	0.4	0.74	−1.8, 2.76	0.76
	4	88.38 (3.59)	88.86	0.16	0.1	0.26	−0.98, 0.94	0.15
Unemployment	1	5.81 (1.16)	5.8	0.99	0.77	1.33	−3.05, 5.56	1.47
	2	5.46 (1.28)	5.3	0.54	0.39	0.75	−1.92, 2.46	0.75
	3	5.18 (1.11)	5.22	0.21	0.14	0.31	−1.43, 1.07	0.27
	4	5.27 (1.31)	5.22	0.11	0.06	0.16	−0.66, 0.44	0.14

generally robust. Given that most observed differences between ACS estimates and aggregated estimates are small and do not exhibit consistent patterns (with the exception of component geography size/population overlap, and among racial/ethnic subgroups), it is likely that observed differences reflect statistical noise, potentially introduced as a result of small differences in block population proportions compared to ACS block population count/percentages, or random error in ACS estimates, rather than properties of the underlying metric estimate distributions.

Aggregated total population percentage estimates were broadly similar when aggregating either counts or percentages, using either tract or county as the component geography, and by variation in percentage population overlap between component geography and CD. Generally, county-derived estimates had larger errors and more and larger outliers, and counties with low percentage population overlap with the target CD produced estimates more different than ACS estimates, indicating tract-derived estimation is likely to yield more precise results. This may be because lower population overlap tends to occur in higher density, more urban, more heterogeneous areas where CDs are small and numerous, and higher population overlap tends to occur in lower density, more suburban or rural, more homogenous areas where CDs are large. As such, wherever possible we utilize census tracts as our component geography. Total population results were robust to induced component geography missingness, in some cases up to 50% induced missingness.

Racial/ethnic subgroup estimates generally had larger error measures than did total population estimates. This is likely driven by many

Table 3a
Descriptive statistics & measures of error of aggregated Congressional District estimates with percentage thresholds of total population missingness artificially induced at the tract level in a sample of 48 Congressional Districts (Sensitivity analysis #4).

Metric Name	% Population Missing Threshold	Mean (SD)	Median	Mean Absolute Error	Median Absolute Error	Root Mean Square Error	Minimum and Maximum Error	IQR of Error
Broadband Access (%)	<=10% pop missing	68.41 (8.61)	69.66	0.54	0.43	0.37	−2.44, 2.12	0.65
	>10–20% pop missing	68.65 (8.63)	69.83	0.61	0.49	0.38	−2.76, 3.32	0.85
	>20–30% pop missing	68.60 (8.54)	69.87	0.69	0.56	0.38	−3.14, 3.38	0.99
	>30–40% pop missing	68.51 (8.72)	69.64	0.83	0.65	0.40	−4.24, 3.70	1.23
	>40–50% pop missing	68.56 (8.69)	69.98	0.91	0.74	0.35	−3.80, 4.64	1.40
	>50% pop missing	68.41 (8.60)	69.57	1.02	0.85	0.46	−3.23, 4.36	1.55
	US Census ACS	68.93 (8.51)	70.30	NA	NA	NA	NA	NA
High School Completion (%)	<=10% pop missing	87.62 (5.93)	88.61	0.42	0.34	0.39	−0.59, 2.12	0.45
	>10–20% pop missing	87.61 (5.97)	88.59	0.44	0.34	0.39	−0.93, 2.67	0.49
	>20–30% pop missing	87.62 (5.98)	88.61	0.48	0.36	0.39	−1.27, 2.94	0.60
	>30–40% pop missing	87.63 (5.97)	88.62	0.54	0.41	0.39	−1.73, 3.74	0.71
	>40–50% pop missing	87.70 (5.90)	88.64	0.58	0.43	0.38	−1.78, 3.59	0.82
	>50% pop missing	87.37 (6.18)	88.56	0.66	0.49	0.40	−1.88, 4.67	0.92
	US Census ACS	88.0 (5.81)	89.05	NA	NA	NA	NA	NA
Unemployment (%)	<=10% pop missing	5.41 (1.52)	5.35	0.17	0.12	0.15	−0.88, 0.29	0.20
	>10–20% pop missing	5.40 (1.52)	5.32	0.18	0.13	0.15	−0.97 0.48	0.23
	>20–30% pop missing	5.40 (1.51)	5.31	0.20	0.15	0.16	−1.23, 0.54	0.26
	>30–40% pop missing	5.38 (1.53)	5.24	0.22	0.17	0.16	−1.27, 0.68	0.30
	>40–50% pop missing	5.42 (1.56)	5.31	0.25	0.19	0.15	−1.65, 0.77	0.36
	>50% pop missing	5.42 (1.50)	5.21	0.28	0.21	0.17	−1.51, 0.90	0.39
	US Census ACS	5.25 (1.46)	5.20	NA	NA	NA	NA	NA

factors, including random error due small sample sizes, and that racial/ethnic groups may be less likely to be uniformly distributed across geographies - due to racial residential segregation, ethnic enclave effects, and other social forces - violating an assumption of this aggregation method. We are exploring ways to highlight potentially unstable estimates, including racial/ethnic subgroup estimates, on the CDHD.

On the CDHD we elected to aggregate percentage estimates, use tract as the start geographies whenever possible, and apply censorship criteria for estimates with high missingness (when 10% of county population, or 25% of tract population, is missing). We are also using the results of the present analysis to identify CDs that may consistently exhibit high error. Preliminary results indicate that CDs which are smaller than counties, and CDs that include component geographies with both high and low population density, are more likely to exhibit high error. We plan to continue this investigation, identifying patterns among CDs that have larger errors, and flag or censor those CDs on our website.

In conducting these sensitivity analyses we addressed numerous important questions about the validity of the described aggregation method. Robust population-weighted aggregation methods can unlock the ability of researchers and analysts to provide public health metrics for salient geographies for which geospatially specific data were previously unavailable. Importantly, these methods can be used to create metric estimates for other geographies, such as school districts, city council districts, and other spatially delimited administrative and policy-making entities. To that end, the geographic aggregation functions we developed are available, free of charge, upon request to the authors.

Though this research focused on social determinant of health metrics derived from ACS, we have used these methods to aggregate metric estimates from multiple other health-related data sources, including the Centers for Disease Control and Preventions' PLACES and USA Life Expectancy Estimation Program projects, among others (PLACES recommends the method we have described (PLACES: Local Data for Better Health, 2021)). These other data sources were not featured in this research because there are no comparator data available against which to validate aggregated estimates. This underscores a potential limitation of the present analysis, that it was not possible to conduct validation analyses on non-ACS or modeled metrics, which may be noisier than ACS metrics. Nonetheless, the variety of metrics and component geographies from which our metric estimates were aggregated demonstrates the versatility of this method.

We believe these aggregated metric estimates can be immediately useful in identifying and addressing federal public health priorities. CDHD allows congresspersons, lobbyists, and voters to compare health, health equity and their drivers in their districts to outcomes and drivers in districts throughout the country. These data can help policy makers and others identify areas in which their districts are doing well, and

Table 3b
Descriptive statistics & measures of error of aggregated Congressional District estimates with total percentage thresholds of population missingness artificially induced at the county level in a sample of 48 Congressional Districts (Sensitivity analysis #4).

Metric Name	% Population Missing Threshold	Mean (SD)	Median	Mean Absolute Error	Median Absolute Error	Root Mean Square Error	Range of Error (min, max)	IQR of Error
Broadband (%)	<=10% pop missing	67.84 (6.79)	67.72	1.76	0.73	0.32	−8.88, 5.13	1.69
	>10–20% pop missing	65.49 (8.27)	67.05	1.70	1.05	0.19	−10.49, 6.08	2.00
	>20–30% pop missing	63.32 (8.1)	62.61	1.63	1.29	0.32	−3.61, 8.15	2.65
	>30–40% pop missing	63.23 (7.92)	62.56	2.37	1.86	1.17	−5.60, 10.06	3.37
	>40–50% pop missing	63.44 (9.02)	63.39	2.26	1.49	1.48	−5.11, 9.97	2.65
	>50% pop missing	62.45 (9.71)	62.19	5.73	4.10	4.36	−11.40, 45.40	7.74
	US Census ACS	68.93 (8.51)	70.30	NA	NA	NA	NA	NA
High School Completion (%)	<=10% pop missing	89.17 (2.68)	89.35	1.44	0.52	0.00	−5.40, 5.46	1.01
	>10–20% pop missing	89.32 (2.78)	90.17	0.88	0.34	0.59	−1.17, 6.05	0.88
	>20–30% pop missing	88.84 (3.28)	89.43	1.18	0.61	0.07	−5.23, 6.20	1.31
	>30–40% pop missing	88.5 (3.46)	88.78	1.15	0.78	0.26	−3.81, 5.97	1.38
	>40–50% pop missing	88.7 (3.28)	89.47	1.21	0.73	0.64	−4.05, 7.12	1.42
	>50% pop missing	88.78 (3.27)	89.07	1.90	1.29	0.67	−6.00, 14.60	2.44
	US Census ACS	88.0 (5.81)	89.05	NA	NA	NA	NA	NA
Unemployment (%)	<=10% pop missing	5.19 (0.99)	5.36	0.32	0.14	0.04	−0.93, 1.65	0.27
	>10–20% pop missing	5.20 (1.07)	5.09	0.26	0.17	0.19	−1.03, 0.84	0.34
	>20–30% pop missing	5.17 (1.13)	5.17	0.38	0.23	0.13	−1.24, 1.85	0.66
	>30–40% pop missing	5.16 (1.28)	4.81	0.50	0.34	0.14	−1.66, 2.10	0.63
	>40–50% pop missing	4.91 (1.27)	4.63	0.37	0.27	0.16	−2.17, 1.37	0.47
	>50% pop missing	5.48 (1.42)	5.46	0.64	0.47	0.34	−3.90, 2.48	0.95
	US Census ACS	5.25 (1.46)	5.20	NA	NA	NA	NA	NA

areas that could benefit from intervention. Metrics of particular interest that are available on the website include firearm homicide and suicide rates, opioid overdose mortality rate, frequent mental distress, and many others.

The described method has limitations. One major limitation is that aggregated metric estimates are subject to, and in some cases may amplify, limitations of the source data. If the source data oversampled or undersampled certain populations, or suffered from response or selection bias or other challenges common to survey research, the aggregated estimates will also suffer these limitations. Rare events are difficult to measure accurately at any level, and when aggregating estimates, a few missed rare events can have a substantial impact on the final metric estimate. Similarly, the data we use to validate these estimates - US Census ACS measures - are themselves estimates with margins of error, which tend to be larger in small geographies.

Also important is an assumption inherent in area-based metric estimates drawn from survey data. Such estimates implicitly assume a uniform distribution of the outcome of interest across the geographic unit in question. Given that populations are not uniformly distributed across geographic units, and that outcomes/exposures are not uniformly distributed across populations, it is unlikely any metric meets this assumption. Other researchers are currently innovating statistical techniques that can address this assumption in regression models (Nethery et al., 2023). We attempt to reduce the impact of this limitation by utilizing the smallest analytic geography available, the census block, to calculate population weights. We have also recently added census tracts, a smaller geographic unit, to the Congressional District Health Dashboard, partially to display within-CD variation in metric estimates. Finally, this analysis was conducted prior to the release of 2020 decennial census block population counts, instead relying on 2010 data. These results cannot account for population changes between 2010 and 2019 (the metric data year).

The present research also has numerous strengths. The sensitivity analyses described here have, to our knowledge, not been published previously. The analyses were conducted using publicly available data, and the authors will share relevant code upon request, making the aggregation method easy to reproduce (analytic code for a similar method has recently become available via the tidycensus R package). In demonstrating its validity and reliability, this research describes a geographic aggregation method that researchers, policy makers, and members of the public health workforce can use to produce needed data for geographies that otherwise lack data parsed to their boundaries. This is an important step towards empowering data-driven decision making and holding accountable policy makers, health officials, and others responsible for the public's health. Put another way, these methods can extend "what gets measured", in the hopes that doing so will empower public health professionals and policy makers to extend "what gets done". Finally, the metric estimates generated using these methods are

available to view and download via the CDHD (https://www.congressionaldistricthealthdashboard.org/), and the website's technical documentation has further detail on the methods described here (Dashboard Team, 2023c).

6. Conclusion

The present research describes and validates a geographic aggregation method that can be used to generate metric estimates for health- and policy-relevant geographies that have previously lacked location-specific data. In general, aggregated estimates were similar to ACS derived estimates. Large geographies and those with lower population overlap produced more error, and more outliers, when aggregated. These methods can be used to provide data for public health decision making, and in doing so, can contribute to improving population health and health equity across the nation. Future research should continue to explore what drives high error measures in the geographies that produce the largest errors.

Works Cited.

Declarations of interest

No financial or other conflicts of interest.

Funding

This work was supported by the Robert Wood Johnson Foundation, grant number 79461.

Ethical statement

Declaration of interests: none.

AI was not used at any point, in any capacity, in the production of this research.

Author statement

Declarations of interest: no financial or other conflicts of interest.

Funding: This work was supported by the Robert Wood Johnson Foundation, grant number 79461.

Data availability

Data are freely available via the source websites

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ssmph.2023.101511.

References

America's Health Rankings. (2022). America's health Rankings. Retrieved April 6th from https://www.americashealthrankings.org/.

County Health Rankings and Roadmaps. (2023). *County health Rankings and Roadmaps.* Retrieved April 6th from countyhealthrankings.org.

Dashboard Team. (2023a). 2022 CDHD district density index. Retrieved April 6th from https://www.congressionaldistricthealthdashboard.org/2022-cdhd-district-density-index.

Dashboard Team. (2023b). *Congressional district health dashboard.* Department of Population Helath, NYU Langone Health. Retrieved March 6th from https://www.congressionaldistricthealthdashboard.org/.

Dashboard Team. (2023c). *Congressional district health dashboard technical document.* http://www.congressionaldistricthealthdashboard.org/technical-documentation.

Eberth, J. M., Zahnd, W. E., Adams, S. A., Friedman, D. B., Wheeler, S. B., & Hebert, J. R. (2019). Mortality-to-incidence ratios by US Congressional District: Implications for epidemiologic, dissemination and implementation research, and public health policy. *Preventive Medicine, 129*, Article 105849.

Gourevitch, M. N., Athens, J. K., Levine, S. E., Kleiman, N., & Thorpe, L. E. (2019). City-level measures of health, health determinants, and equity to foster population health improvement: The city health dashboard. *American Journal of Public Health, 109*(4), 585–592.

Holt, J. B., Lo, C., & Hodler, T. W. (2004). Dasymetric estimation of population density and areal interpolation of census data. *Cartography and Geographic Information Science, 31*(2), 103–121.

Islami, F., Wiese, D., Marlow, E. C., Kratzer, T. B., Massey, J., Sung, H., & Jemal, A. (2023). *Progress in reducing cancer mortality in the United States by congressional district, 1996–2003 to 2012–2020.* Cancer.

Liu, X., & Martinez, A. (2019). Areal interpolation using parcel and census data in highly developed urban environments. *ISPRS International Journal of Geo-Information, 8*(7), 302.

Mansfield, C. J., Wilson, J. L., & Kirk, D. (2007). *Health statistics by congressional district: A foundation for political epidemiology to inform health policy*.

McNabb, S. J., Chungong, S., Ryan, M., Wuhib, T., Nsubuga, P., Alemu, W., Carande-Kulis, V., & Rodier, G. (2002). Conceptual framework of public health surveillance and action and its application in health sector reform. *BMC Public Health, 2*(1), 1–9.

McVeigh, K. H., Newton-Dame, R., Chan, P. Y., Thorpe, L. E., Schreibstein, L., Tatem, K. S., Chernov, C., Lurie-Moroni, E., & Perlman, S. E. (2016). Can electronic health records Be used for population health surveillance? Validating population health metrics against established survey data. *EGEMS (Wash DC), 4*(1), 1267. https://doi.org/10.13063/2327-9214.1267

Nethery, R. C., Testa, C., Tabb, L. P., Hanage, W. P., Chen, J. T., & Krieger, N. (2023). Addressing spatial misalignment in population health research: A case study of US congressional district political metrics and county health data. *medRxiv, 2023*, 2001.2010.23284410.

Noelke, C., Ressler, R. W., McArdle, N., Hardy, E., & Acevedo-Garcia, D. COI 2.0 ZIP CODE DATA..

PLACES: Local Data for Better Health. (2021). Using the data. Retrieved April 6th from https://www.cdc.gov/places/faqs/using-data/index.html.

Rolheiser, L. A., Cordes, J., & Subramanian, S. (2018). Opioid prescribing rates by congressional districts, United States, 2016. *American Journal of Public Health, 108*(9), 1214–1219.

Schroeder, J. P. (2007). Target-density weighting interpolation and uncertainty evaluation for temporal analysis of census data. *Geographical Analysis, 39*(3), 311–335.

Schroeder, J. P. (2017). Hybrid areal interpolation of census counts from 2000 blocks to 2010 geographies. *Computers, Environment and Urban Systems, 62*, 53–63.

Siegel, R. L., Sahar, L., Portier, K. M., Ward, E. M., & Jemal, A. (2015). Cancer death rates in US congressional districts. *CA: A Cancer Journal for Clinicians, 65*(5), 339–344.

Takai, A., Kumar, A., Kim, R., & Subramanian, S. (2022). Life expectancies across congressional districts in the United States. *Social Science & Medicine, 298*, Article 114855.

U.S. Census Bureau. (2023). *Congressional and state legislative districts.* Retrieved April 6th from https://www.census.gov/acs/www/data/congressional-and-state-legislative-districts/.

Wilson, J. L., & Mansfield, C. J. (2012). Disease, death, and the body politic: An areal interpolation example for political epidemiology. In *Geospatial technologies and advancing geographic decision making: Issues and trends* (pp. 253–272). IGI Global.

Wright, J. K. (1936). A method of mapping densities of population: With Cape Cod as an example. *Geographical Review, 26*(1), 103–110.

Wu, S.-s., Qiu, X., & Wang, L. (2005). Population estimation methods in GIS and remote sensing: A review. *GIScience and Remote Sensing, 42*(1), 80–96.

Zhang, X., Holt, J. B., Lu, H., Wheaton, A. G., Ford, E. S., Greenlund, K. J., & Croft, J. B. (2014). Multilevel regression and poststratification for small-area estimation of population health outcomes: A case study of chronic obstructive pulmonary disease prevalence using the behavioral risk factor surveillance system. *American Journal of Epidemiology, 179*(8), 1025–1033.

Zoraghein, H., & Leyk, S. (2018). Enhancing areal interpolation frameworks through dasymetric refinement to create consistent population estimates across censuses. *International Journal of Geographical Information Science, 32*(10), 1948–1976. https://doi.org/10.1080/13658816.2018.1472267