

Strategies to Deal with Data Uncertainty and Limitations

For the National Neighborhood Indicators Partnership



Community Information Now (CINow) is a nonprofit local data intermediary and the San Antonio partner in the National Neighborhood Indicators Partnership (NNIP). This document is a compilation of strategies CINow uses to communicate uncertainty in and limitations of data in its online data platforms and narrative reports, illustrated with examples where readily available.

Contents

Offer explanatory text in call-out boxes	2
Embed explanatory text about error and bias in the narrative itself	4
Aggregate up to reduce uncertainty because people often ignore MOEs	5
Visualize uncertainty with error bands (line chart) or bars (bar chart)	7
Show only the MOE/CI when it's "too large" relative to estimate (coded threshold)	9
Mark values where error is "too large" relative to estimate (coded threshold)	10
Suppress or visually disconnect problematic values from trend lines.....	11
Include other measures for context.....	12
Warn when the numerator or denominator is small but not suppressed.....	13
Decline to use datasets or variables that are worse than nothing	14
Use the data limitations to make the case for investing in better data	15

Offer explanatory text in call-out boxes

Section 1: Population



Growth and distribution

Population growth and geographic distribution – where they live in each county – are both extremely important drivers of needs for physical infrastructure, human services assets like schools and health clinics, and businesses and amenities. Strong population growth in Bexar is old news, but Atascosa is growing as well. While the population size and growth in absolute numbers are vastly different, the rate of growth from 2012 to 2017 differs much less (Fig. 1.1). Bexar added about 173,000 people over the past five years alone – a 10% increase – while Atascosa added an estimated 1,200, a roughly 2.5% increase. While Atascosa’s rate of growth is not as steep as that of Comal, Guadalupe, and other historically rural and semi-rural counties in metro areas along and east of the I-35 corridor, it does buck the overall trend of continued population decline in rural and semi-rural Texas counties.¹

Box 1

Is it just a wild guess?

No data is ever perfect, but some things can be counted one by one – housing units, deaths, hospitalizations. For others the effort and expense of a count is often very high, so instead we look only at a sample, or subset of the total. Wherever there’s a sample, there’s always an open question about the estimates that came from it. The smaller the sample relative to the total, the less confident we can be that the estimate holds true for the total. So no, we’ll never give you an estimate that’s just a wild guess, but know that some estimates can get a bit wild. Box 3 shows how to spot those right away.

The population is not evenly distributed in either county (Fig. 1.1.1). Bexar’s most populous zip codes are those radiating from the near Westside to the northwest Loop 1604 corridor and beyond, as well as 78223 on the I-37 corridor to the southeast. Unsurprisingly, Atascosa’s most populous zip code is 78064 on the I-37 corridor, home to Jourdanton and adjacent to Pleasanton. For reasons discussed below, many Bexar zip codes – not just those on the county outskirts, but also the military bases and San Antonio’s city center – also appear sparsely settled.

Zip code size and shape vary tremendously in both counties, though. Looking at population density – number of people per square mile – controls for that variation in zip code size (Fig. 1.1.2). As a result, the near Eastside and areas south of downtown and west of King William join the near Westside and northwest zip codes as having among the highest population densities in the county. Population density in Atascosa and Bexar’s outlying zip codes still reflects a semi-rural character, and despite a meaningful uptick over the past decade, downtown San Antonio still has a low number of residents. The other less-dense zip codes are JBSA Lackland and Kelly Field Annex southwest of downtown, JBSA Fort Sam Houston northeast of downtown, and the more industrial areas bounded by Highway 87 to the north and I-10 East to the south.

maps in this report since the data by ZCTA, but for readability we just say “zip code.”


Box 3

Spotting uncertainty

Box 1 explains that there’s always some degree of uncertainty about numbers that come from samples rather than actual counts. In this report, we usually call that uncertainty the margin of error or MOE. Wherever possible we’ve used methods that reduce the MOE, such as using Census American Community Survey’s combined five-year dataset instead of the one-year dataset. High or low, though, you’ll always see that MOE in bar charts and line (time trend) charts. In bar charts, MOE is represented with an error bar, a gray line that overlaps the end of the bar. In line charts, MOE is shown as a band of color on either side of the line. In general, the wider the error bar or the color band for an estimate, the more we need to take that estimate with a grain of salt.

ABOUT ERROR BARS

Any time an estimate is created from a survey sample rather than a 100% count of the population, as is the case for many of the indicators in this report, that estimate has an associated margin of error. That margin of error (MOE) is a measure of how much uncertainty there is. For example, if the estimate is 27% ± 3%, we can feel relatively confident that the true value is between 24% and 30%. Generally, the smaller the sample size, the wider the margin of error. Estimates with wide margins of error should be interpreted with caution, or "taken with a grain of salt." The charts in this report show margins of error with error bars, small gray lines in a bar chart, or with error bands, which are shaded bands along a trend line in a line chart. Note, if the error bars or error bands of two estimates overlap, we cannot be sure there is any real difference between the two values, even if the estimates themselves are far apart.




STATISTICAL SMALL AREAS (SSAS) IN MAPS

SSAs, or Statistical Small Areas, are clusters of adjacent census tracts grouped together using statistical methods. CI:Now recently developed SSAs to retain the size advantages of ZIP codes – less data suppression and smaller margins of error – but with more meaningful boundaries than those created for efficient mail delivery. Much more information about SSAs can be found in [Appendix C: Technical Notes and Reference Maps](#).

SOUTH BEXAR COUNTY - 2023

Example from [South Bexar County 2023 Community Health Needs Assessment Report](#)



[ENGLISH](#) | [ESPAÑOL](#)

[FREQUENTLY ASKED QUESTIONS](#)

[ABOUT THE TOOLS](#)

[GUIDES AND TRAININGS](#)

[DATA CORRECTIONS](#)

Why do some indicators show numbers for "Range" and others don't?

The "range" is the margin of error (or standard error, depending on the dataset) for an estimate that came from a survey rather than an actual count. The smaller the survey sample relative to the total universe – of people, households, whatever – the less confident we can be that the estimate holds true for the total universe. That uncertainty is the margin of error (MOE) or confidence interval. It's common to see a margin of error talked about as, for example, "35% plus or minus 5%". Instead of the "plus or minus" format, we're showing it as the range so the user doesn't have to do the mental math to figure out that the true value for the total universe is somewhere between 30% and 40%.

Example from [Bexar Data Dive - FAQ](#)

Embed explanatory text about error and bias in the narrative itself

**PREVENTIVE CARE
MEDICAL AND DENTAL VISITS**

Many of the charts that follow in this section represent data from the Behavioral Risk Factor Surveillance System (BRFSS), a household telephone survey of adults with a very small sample size for Bexar County. Because the sample size is so small relative to the size of the adult population, even with multiple years of data combined, each BRFSS estimate has a good bit of uncertainty. The true value may lie anywhere in the range of the estimate's confidence interval, which is represented as a horizontal gray line in each bar of the chart. When the confidence intervals (gray lines) for two estimates overlap, one cannot be sure that there is truly any difference between the two estimates. That issue will arise over and over again in the narrative describing these charts.

Fig. 2.7 Percent of adults who visited a doctor last year, by race, 2015-2020

Example from [2022 Bexar County Community Health Needs Assessment](#)

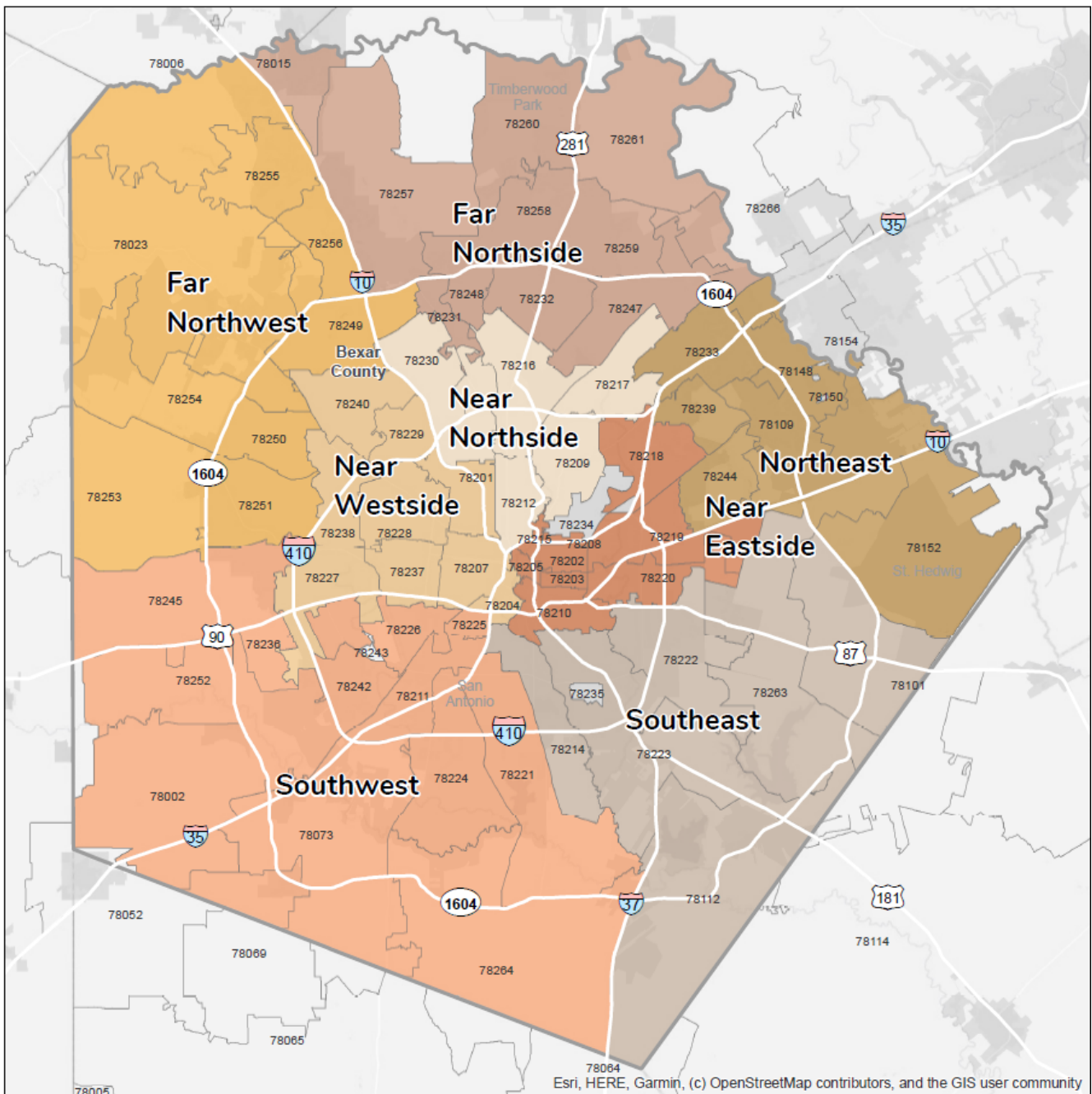
ASTHMA

Another issue with the BRFSS dataset is that the survey is by self-report, and people may or may not report accurately. Many questions are phrased as "Have you ever been told by a doctor, nurse, or other health professional that you have...?" a disease. Answering yes to that question requires that the person had access to care, utilized care, was formally diagnosed with the disease (regardless of the reason for the visit), understood the diagnosis, and remembered the diagnosis months or years later. For that reason, the estimates in these next several BRFSS charts should likely be considered underestimates.

Example from [2022 Bexar County Community Health Needs Assessment](#)

Aggregate up to reduce uncertainty because people often ignore MOEs

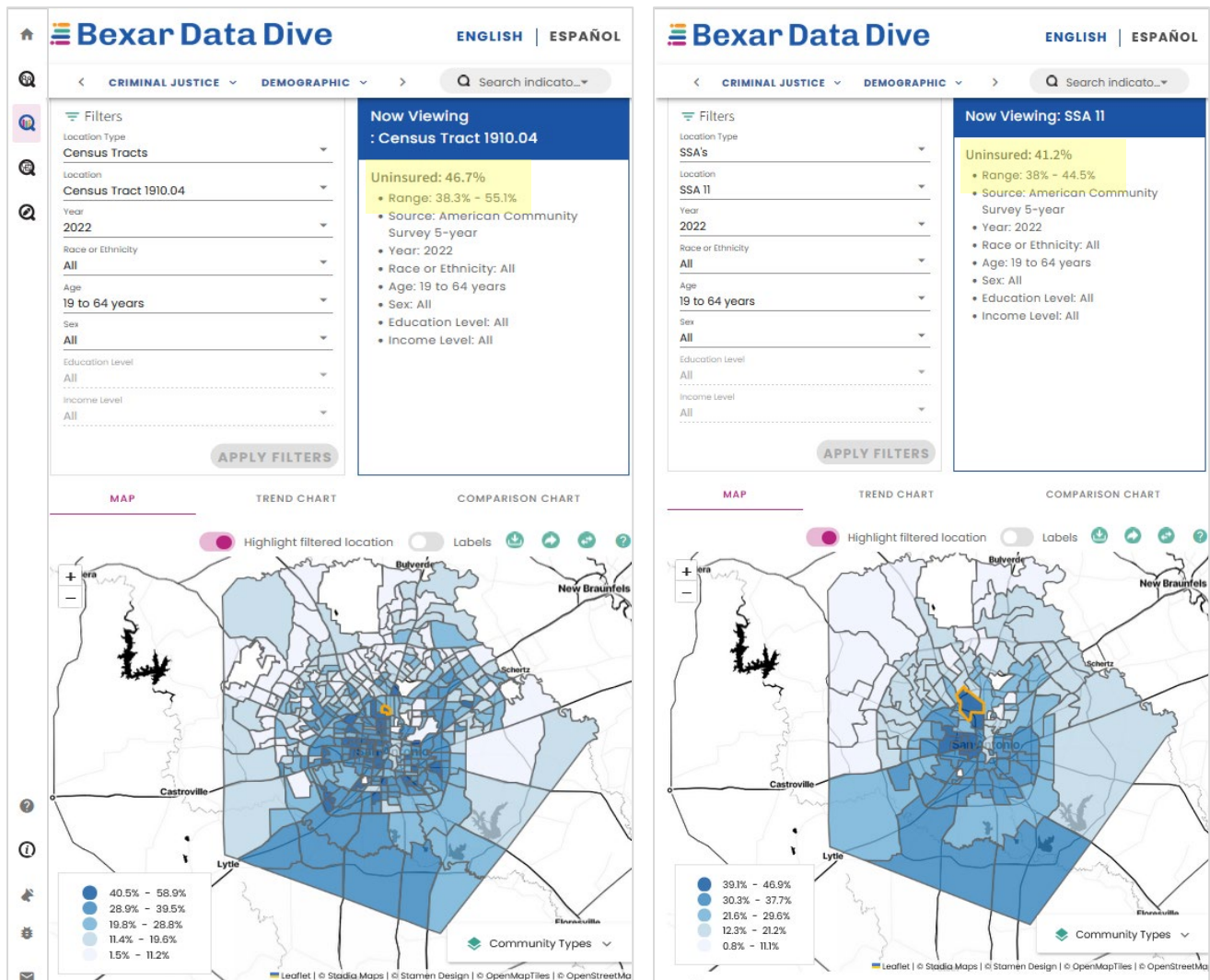
This strategy is so basic that we sometimes forget about it, but the desire to maximize granularity while minimizing uncertainty was the catalyst for creation of both our Sector and SSA geographies, which have both been invaluable to us. The Sector geography (map below) groups ZIP codes and was created in 2016 specifically to cope with uncertainty deriving from a small BRFSS sample. State BRFSS staff have our ZIP-sector crosswalk and are able to aggregate for us for key indicators for which we couldn't otherwise access data.



Example from [2022 Bexar County Community Health Needs Assessment](#)

Statistical Small Areas (SSAs) are statistically-grouped census tracts.¹ Created in 2022, SSAs are intended to achieve the size advantage of ZIP codes but with boundaries that are more logical and meaningful than mail delivery routes. The side-by-side images below show an example of the tradeoff between small-area granularity and MOE size.

As with sectors and BRFS, San Antonio Metro Health staff have our tract-SSA crosswalk and are able to share aggregate data for key indicators (e.g., teen births, prenatal care) for which virtually all tract-level values would be suppressed for privacy.



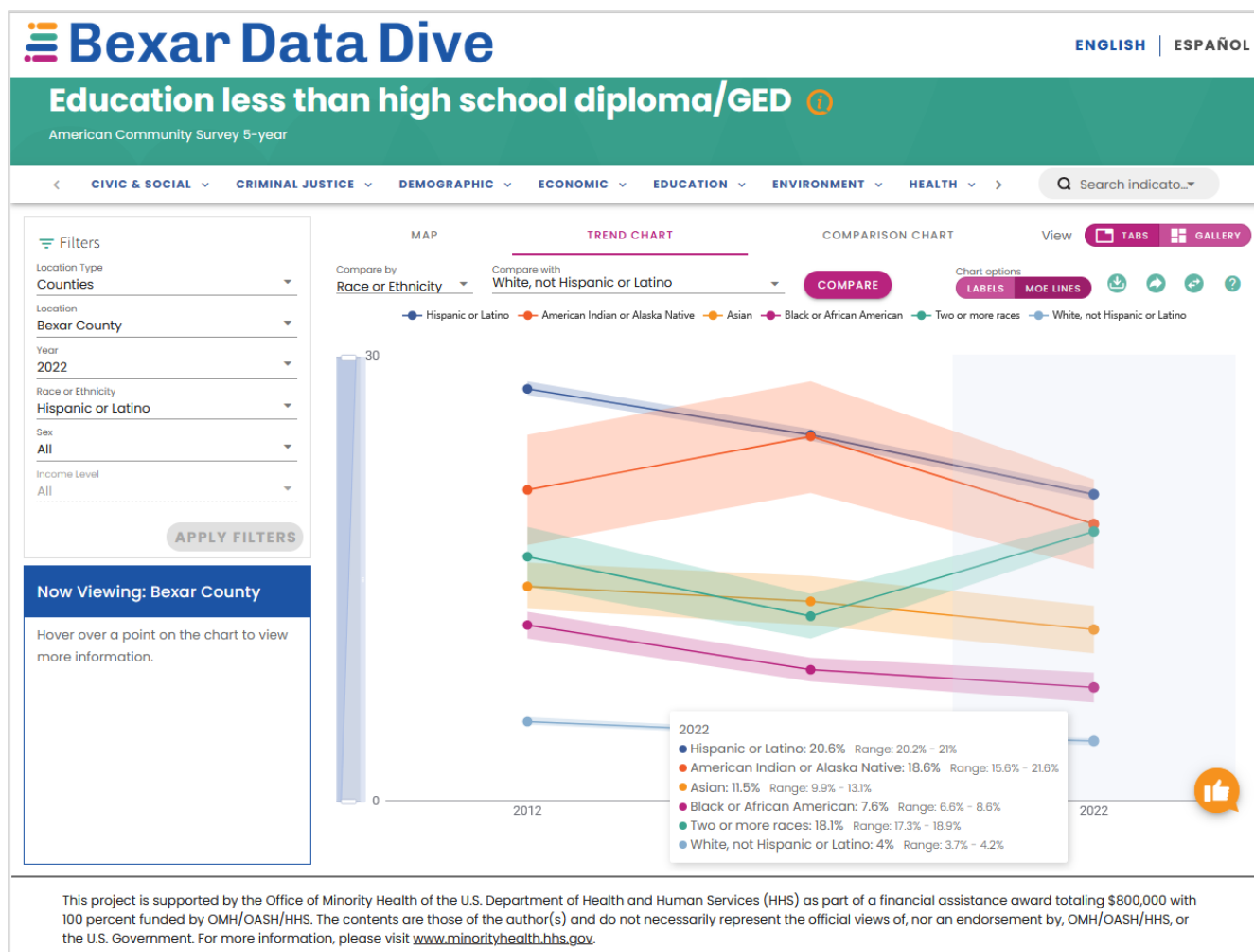
Examples from [Bexar Data Dive](#)

¹ We're happy to share our clustering method and code.

Visualize uncertainty with error bands (line chart) or bars (bar chart)

In Bexar Data Dive we allow people to toggle back and forth between showing the MOE range values and showing the error bands or bars. In some reports we've done basically the same thing by showing both the line/band and the actual internal min/max (see pneumonia vaccination chart on p. 9 of this document), but in charts we usually only show the line or band to keep things readable.

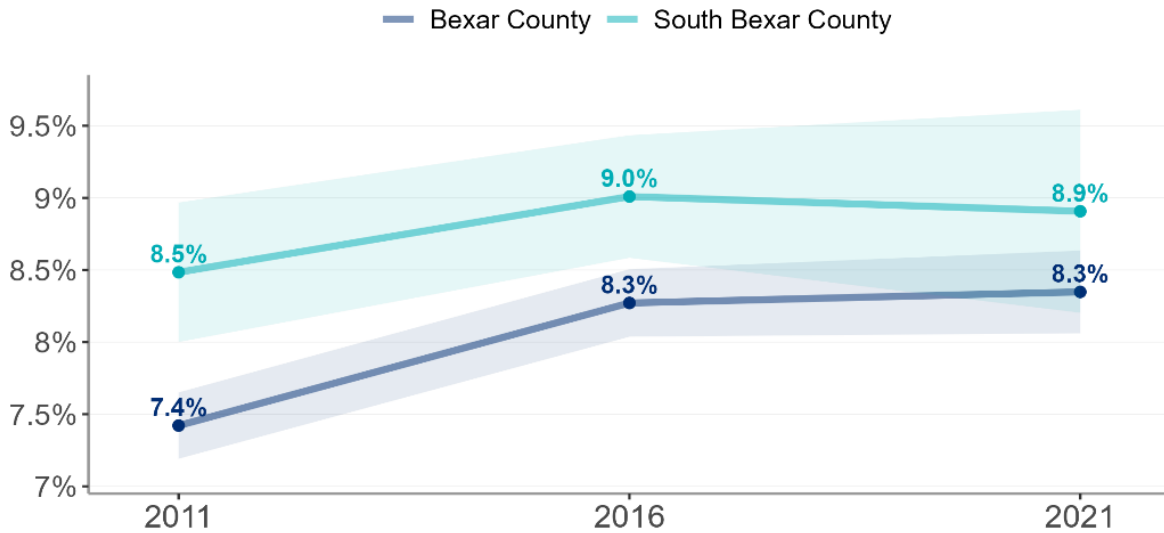
Visualizing the uncertainty as bands or bars also helps make it obvious where MOEs overlap, meaning we can't be sure there's a difference among groups even though the estimates are different, and also makes it clear where there *are* definitely differences.²



Example from [Bexar Data Dive](#)

² We're happy to share our method and R code for generation of chart images with error bars/bands.

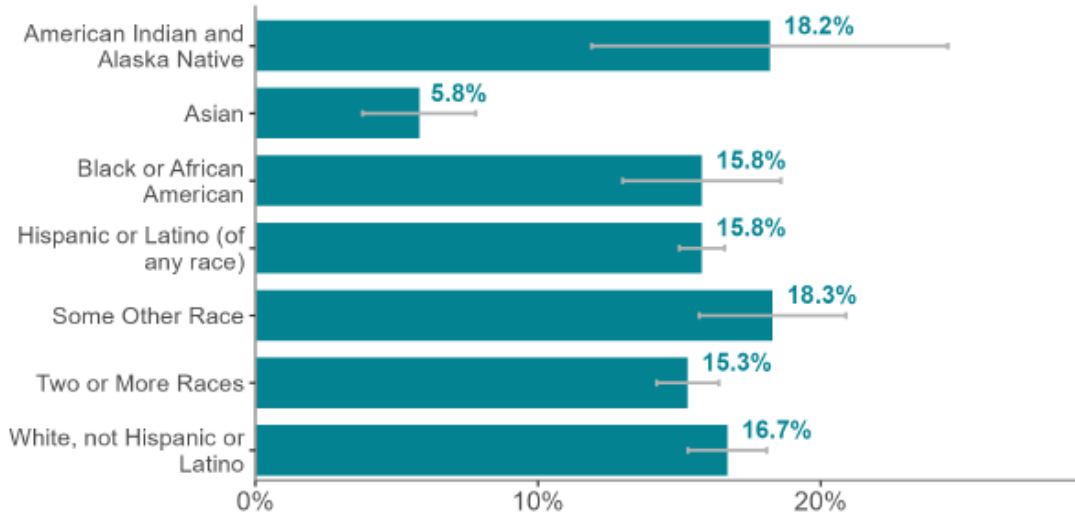
Fig. 1.18 Percent of households where householder is 65 or older and living alone



Source: ACS 5-Year Estimates. Table: B11007
Prepared by CI:Now for University Health

Example from [South Bexar County 2023 Community Health Needs Assessment Report](#)

Fig. 1.3 Percent of population with a disability by race/ethnicity, 2022
San Antonio, Texas

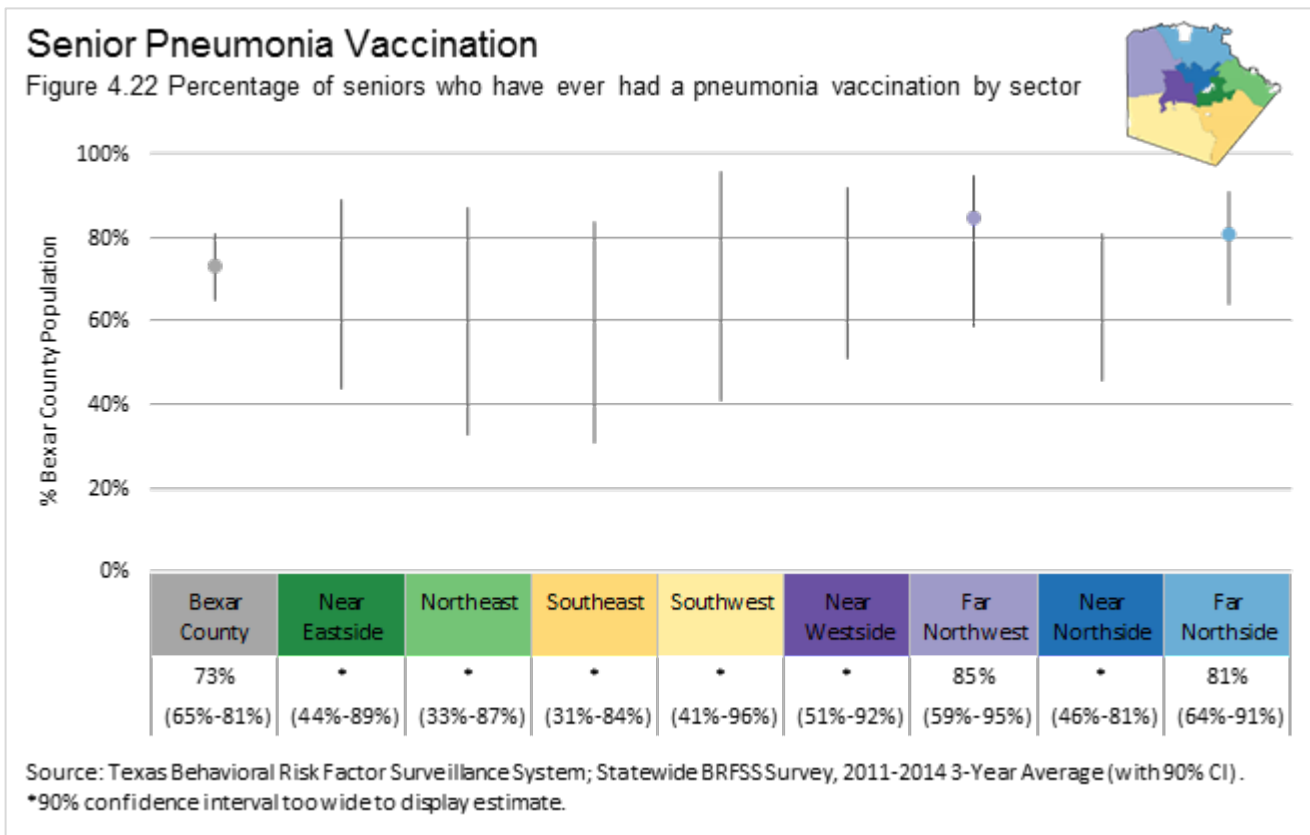


Source: ACS 1-year Estimates. Table: S1810
Prepared by CINow for the City of San Antonio

Example from *2024 Disability Access in San Antonio Indicator Report* (as yet unpublished)

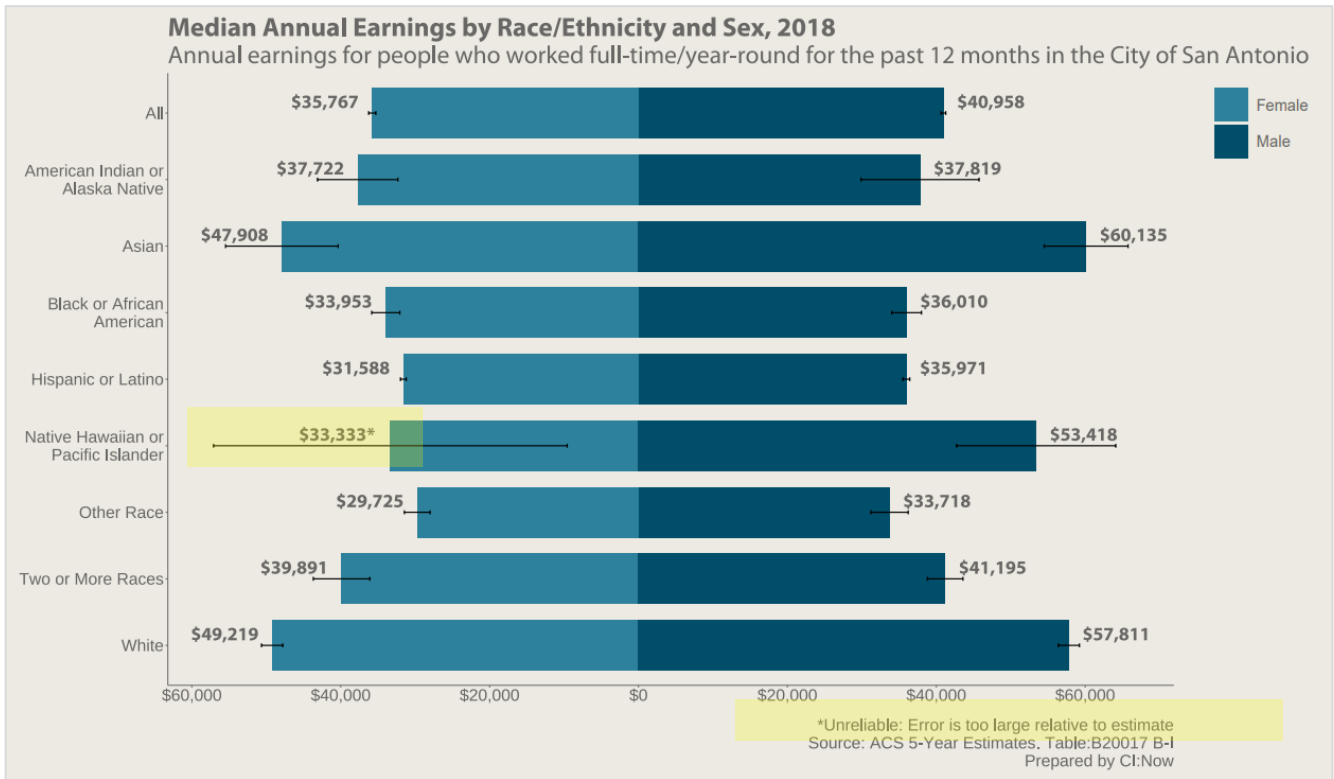
Show only the MOE/CI when it's "too large" relative to estimate (coded threshold)

In the 2016 Bexar County needs assessment we suppressed estimates where the uncertainty was too great, while still letting the reader see how the CIs compare across sectors. This strategy works especially well when the estimate isn't at the exact midpoint of the interval.

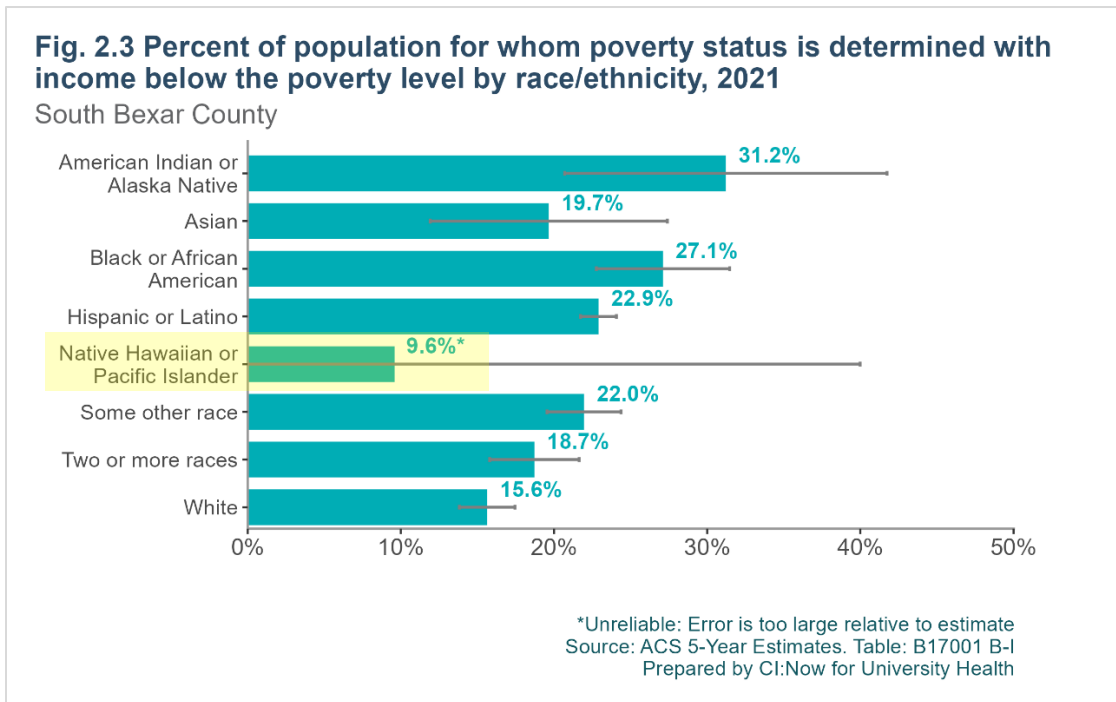


Example from [2016 Bexar County Community Health Needs Assessment](#)

Mark values where error is “too large” relative to estimate (coded threshold)

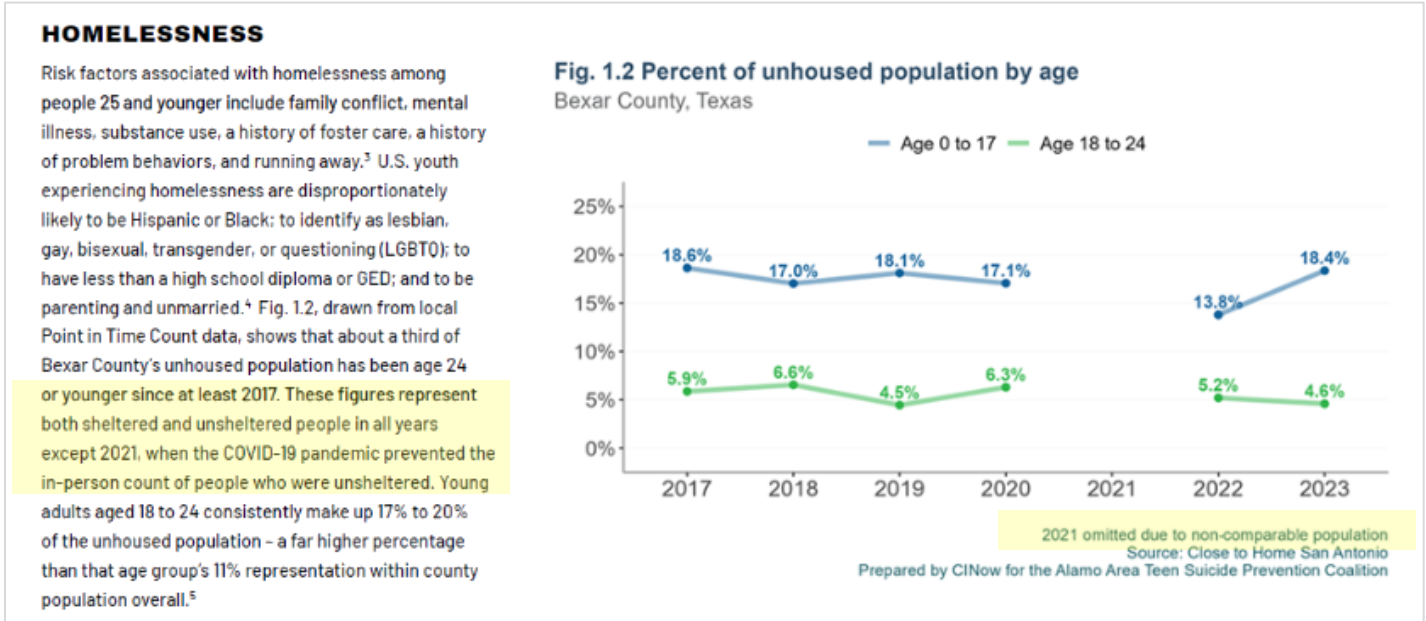


Example from [2019 Racial Equity Indicator Report](#)



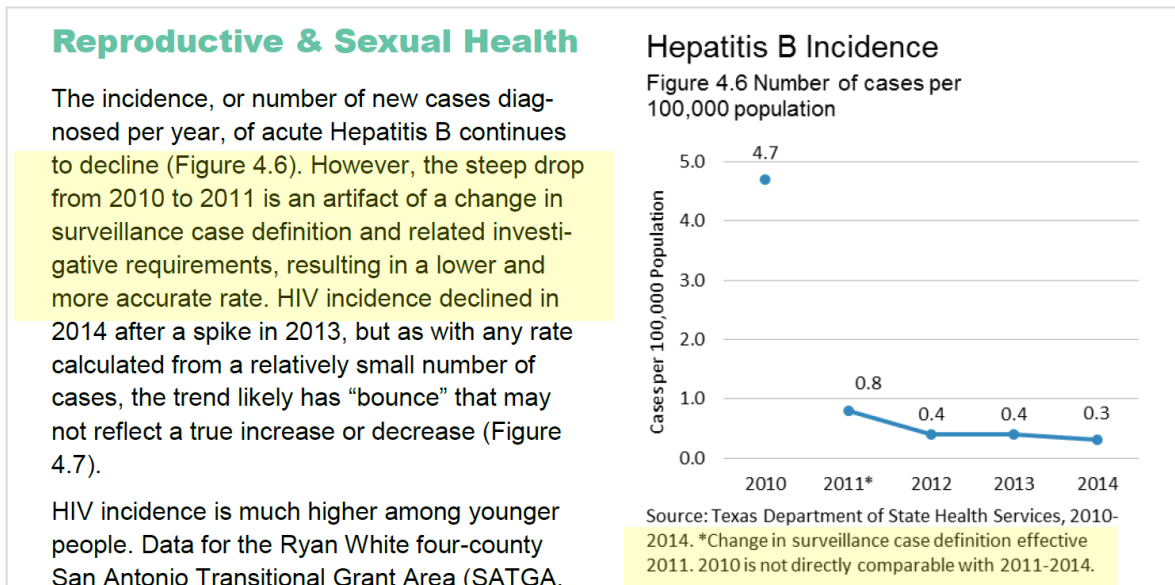
Suppress or visually disconnect problematic values from trend lines

The Point-in-Time Count couldn't be conducted in Jan 2021 because of COVID, so the 2021 Homeless CoC data only included people in the HMIS, and they were different from the non-HMIS unsheltered.



Example from 2024 Bexar County Teen Suicide Prevention Community Needs Assessment Report (as yet unpublished)

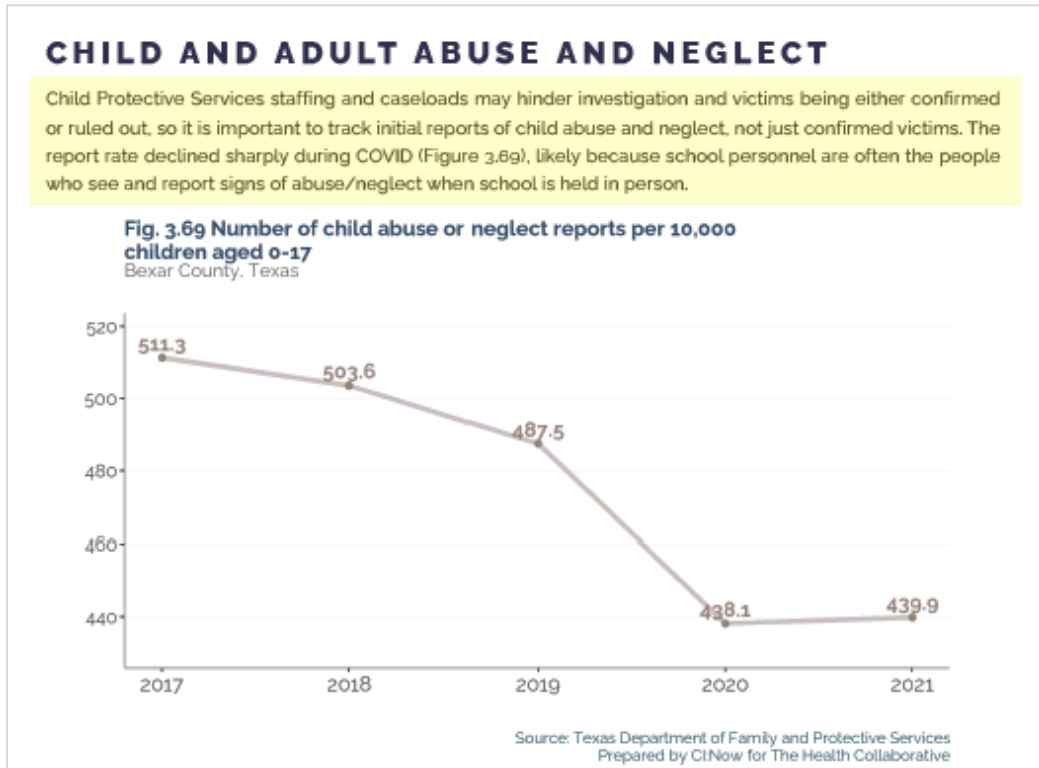
We broke the Hep B trend line to make it clear we're trending four apples and one orange. After researching we found out the case definition changed from 2010 to 2011, but if we hadn't gotten that explanation, we probably would have excluded 2010 from the chart entirely. That strategy would have raised different red flags about why we were showing only four time points for that indicator rather than the five time points shown for all other indicators.



Example from 2016 Bexar County Community Health Needs Assessment

Include other measures for context

We used to only report rate of confirmed victims of child abuse and neglect, but then there was this huge drop that there was no reason to believe had occurred. We dug into the data and discovered that there had indeed been a 10% drop in report rate (“initial intakes alleging”), but there were bigger drops in the percent of intakes assigned for investigation and in the percent of investigations completed, and you can’t confirm or rule out abuse without the investigation completed. Now we include both report rate and confirmed victim rate, but once COVID hit we had to talk about the impact of remote schooling on report rate.



Other applications

- We could think about a combo chart showing a trended indicator of interest on one axis and a data quality indicator on the other – data completeness, survey sample size, survey response rate, etc.
- In a line chart, bar chart, or scatterplot, hue or saturation could signal data quality or trustworthiness.
 - In a single-color scatterplot, the hue or saturation of the dots could be a color ramp representing categories of uncertainty, e.g.,
 - MOE is <10% of estimate
 - MOE is 10%-19% of estimate
 - MOE is 20%-29% of estimate
 - MOE is 30%+ of estimate
- If data quality decreased markedly in a frequently-used dataset like Census ACS, the data labels for 2025 and later years could be shown in gray while 2024 and prior years’ data labels are in black.

Warn when the numerator or denominator is small but not suppressed

In our now-retired Viz-a-lyzer data platform we suppressed rates when the population denominator was <100, because those rates were very misleading and confusing. In Bexar Data Dive we don't suppress them, but we provide a warning in our About the Data and in trainings. (We do still suppress for privacy based on the numerator, of course.)



Bexar Data Dive

Drug poisoning hospitalization rate - 3 year average (per 10K)

Definition: Rate of people hospitalized with a primary diagnosis of drug poisoning, for every 10,000 people. A primary diagnosis of drug poisoning is defined as an inpatient record having a principal diagnosis code of ICD-10-CM T36-T50 with a 7th character of A, or missing. This includes poison by, adverse effects of, and underdose effects of drugs, medicines, and biological agents.

Data Notes: To keep abreast of medical knowledge, the International Classification of Diseases (ICD) is revised periodically. Large increases or decreases in hospitalizations are typically indicative of such changes.


Population counts to calculate rates come from the U.S. Census Bureau's American Community Survey 5-year estimates.

Rates based on total counts less than 20 and population counts less than 100 may be unstable; use with caution.

Results have been suppressed when the estimated hospitalization count is less than 5.

Data Source: Texas Department of State Health Services - THCIC Hospital Discharge PUDF; U.S. Census Bureau's American Community Survey 5-year estimates

CITE SOURCE

GO TO SOURCE 

Example from [Bexar Data Dive](#)

Decline to use datasets or variables that are worse than nothing

As a general rule we believe in going ahead and showing problematic data while doing our best to communicate its limitations and minimize the likelihood of misunderstanding or misuse. There are some cases though where we feel the available data is more harmful than no data.

- Business owner/applicant race/ethnicity was missing for 87% of Bexar County records in the PPP loan dataset, so we declined to use it in an analysis of access to capital and related issues for Black and Hispanic business owners

Business Access to Credit and Capital

Years of national data show a clear pattern of inequity in access to credit and capital for Black or African, Hispanic or Latino, and female business owners.¹ Unfortunately, no good local data broken down by race/ethnicity for business loan number, amount, source, or denial rate was found for this report. Paycheck Protection Program data is publicly available but unusable, as race/ethnicity data was available for only about 10% of loan records.

Related data on mortgage loan applications and denials can be found in the Population/Housing section. Although not specific to business owners, individual consumer credit score data is presented in Figures 4.22 and 4.23 above.

Example from [State of the African American Community in San Antonio and Bexar County, 2022](#)

- We discovered the sex field is empty for a sizable number of records in certain areas of the state hospital discharge dataset, which yields odd and misleading results. For that reason we don't enable disaggregation by sex for some hospital discharge indicators (e.g. [drug poisoning](#)) in Bexar Data Dive (or other uses) where disaggregation by sex is typically available (e.g. [asthma](#))

Use the data limitations to make the case for investing in better data

We often include a call for investments in local data and data capacity in Executive Summaries, introduction sections, and findings/conclusions.

90

Criminal and juvenile justice. Justice system involvement for even low-level offenses dramatically reduces the universe of possible employment, exposes people to violence in justice facilities, and greatly increases the risk of PTSD and other mental illness. Parental incarceration can harm families through increased stress, reduced income, and other factors.

Adverse childhood experiences (ACE). Child sexual abuse and physical and emotional abuse or neglect; family violence; loss of a parent to divorce, abandonment, or incarceration; and exposure to mental illness or substance are all common in varying degrees in Bexar County. All have been shown to have an effect on health and well-being and on health-related behaviors, mostly especially depression, alcoholism, substance abuse, suicide, smoking, and risky sexual behaviors. Although not assessed in the ACE Study itself, childhood poverty is also a critical adverse childhood experience common in Bexar County.

System-Level Barriers to Effective Action

A common thread across both the quantitative data and the qualitative data – interviews and discussion groups – is how little improvement the Bexar County community has been able to achieve on key important health issues ranging from early and adequate prenatal care to vaccination across the lifecycle to healthy eating, physical activity, and overweight/obesity issues. Poverty has not budged, and it's unlikely that child abuse has decreased by even a fraction of the degree reflected in the most commonly-used indicator.

There is no single solution that will make the needle move in the right direction for any health-related problem. But addressing two key

system-level issues would likely help: improving the data available to inform decisions and actions, and developing more effective ways of working together across organizations and sectors.

Systemic, persistent underfunding of prevention and interventions targeting root causes. Despite knowing that the relative contribution of medical care to health and well-being is small – an estimated 10% to 20%¹¹, very little funding is available for prevention and other interventions to address the “upstream” factors that contribute the remaining 80% to 90%. For example, personal health care spending per capita in 2014 was \$9,523. In contrast, combined federal and state public health funding per capita was \$248 – just 2.5% of health-related spending.¹² And the U.S. spends proportionally less on social root causes than other nations with better population health outcomes.¹³

Gaps and disparities in data quality. One pattern that emerges very clearly throughout this assessment is the disparity not just in health determinants and outcomes, but also in the quality of the data about those determinants and outcomes. The limitations of surveys – at least, as they are currently administered – are so great that the data are frequently unusable below the county level. Race/ethnicity, sex, age group, income level, and neighborhood are all critical factors in both understanding the nature of the issue and in deploying appropriate interventions to address it. Very little information about the health of Bexar County's African American and other non-Hispanic non-white populations can be gleaned from BRFSS dataset; the same is true for geographic areas with lower population densities. Even at the county level, confidence intervals are wide enough that short-term trends cannot be accurately identified. Unless the margin of error or

91

confidence interval is very narrow, small and moderate change in a point estimate cannot be trusted to mean a true change in the scope or severity of the issue.

Survey administrators and researchers also know that specific groups of people are systematically missing in the data collected. The list of reasons why those most likely to have health-related disparities are the people least likely to respond to 40- to 60-minute long surveys via any method of administration – self or interview, in person or online or by landline or mobile phone – is long. Just a few reasons relevant to this assessment are mental illness, including depression and PTSD; alcohol and substance use; lack of time because of multiple jobs and three-hour bus rides; fear of discovery of undocumented status or other justice system involvement; a disability related to sight, hearing, mobility, or cognition; general distrust and privacy concerns; fear of answering the door as a person living alone; and not speaking the language of the survey.

Changes in survey approach can help, although there is no magic bullet. Given that change occurs slowly in most issues, perhaps resources would be better spent if administrations were less frequent, enabling samples to be larger or incentives offered. The Bexar County community could certainly make much better use than it does of administrative data – data generated in the everyday course of doing business – like health care visit information and school absence information.

The legal and regulatory protections around data are important and the barriers they erect are significant – and also surmountable. Healthcare Access San Antonio (HASA), the region's health information exchange, has been working since 2006 to establish data-sharing agreements among local health care providers

and to solve the technical challenges of integrating health care data. The Eastside Promise and Wheatley Choice Neighborhoods, as well as other education-related initiatives, have begun doing the same kind of work with education, housing, and other human service data. Absent broad community buy-in and political will, progress on both fronts will be slow.

Working effectively across organizations and sectors. Bexar County has a reputation for being much more collaborative than most major cities, with turf and competition often taking a backseat – if temporarily – to cooperation to solve specific problems. Here and across the country, though, many practitioners and policy-makers are coming to the conclusion that collaboration as it usually looks is not sufficient. Again, there is no magic bullet. And unfortunately, without a robust evidence base like that for many clinical interventions, “best practices” is too often code for “things other communities are doing that are getting good press.”

Having said that, certain principles and practices do appear to make a real difference. Several of these principles have been bundled and adopted in communities across the country as the *collective impact* approach to solving complex, adaptive problems that do not have a clear and straightforward technical solution. In 2011 Kania and Kramer proposed that initiatives that achieve meaningful results have five conditions in common: a common agenda, shared measurement systems, mutually reinforcing activities, continuous communication, and backbone support organizations.¹⁴ In many ways the collective impact approach resembles time-tested quality and performance improvement approaches that, when deployed well, can make a tremendous impact on outcomes within an organization.

Example from [2016 Bexar County Community Health Needs Assessment](#)