# Doing Applied Research

Mixtape Track

MIXTAPE SESSIONS

## Bonus Section. Be careful when collecting your data

# Collecting unique data sets

- The data collection process is where the empirical component of your project begins.

- It is also one of the most important.

- It does not matter if you are using the newest and fanciest estimator, when you are using it on a data set filled with systematic errors.

- You don't want to publish a paper, have it cited 1,000 times, and find out 15 years later that your estimates were a product of data transcription errors…

# Re-examining the Contribution of Public Health Efforts to the Decline in Urban Mortality

D. Mark Anderson

Kerwin Charles

Daniel Rees

*AEJ: Applied* (2022) 14: 126-157

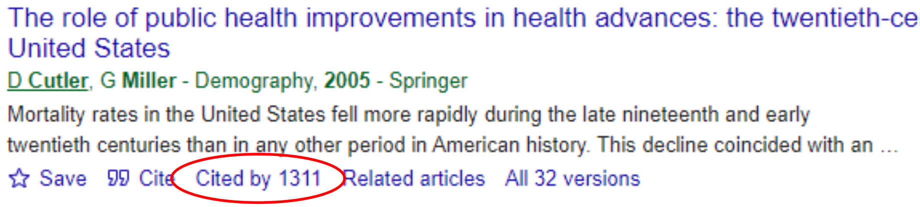
- Using data on 25 major American cities for the years 1900-1940, we revisited the causes of the urban mortality decline at the turn of the 20th century

- Following previous researchers, we explored the extent to which filtering and chlorinating water supplies contributed to the decline in urban mortality

- In addition, we explored the effects of other municipal-level efforts that were, at the time, viewed as critical in the fight against food- and water-borne diseases but have not received as much attention
    - Clean water projects
    - Sewage treatment plants
    - Milk-related interventions

# What did we find?

- Most interventions did not matter
- Water filtration was the exception
  - Water filtration is associated with a 36-41 percent decease in typhoid mortality and an 11-13 percent decrease in infant mortality. This latter estimate, however, is much smaller than those found by previous researchers, including the seminal work by Cutler and Miller (2005).
- Using data from 13 major American cities for the period 1900-1936, Cutler and Miller (2005) found that filtration led to a 15% reduction in total mortality and a 35% reduction in infant mortality



The role of public health improvements in health advances: the twentieth-century United States

D Cutler, G Miller - Demography, 2005 - Springer

Mortality rates in the United States fell more rapidly during the late nineteenth and early twentieth centuries than in any other period in American history. This decline coincided with an …

- Using the original data provided by Cutler and Miller (2005) and their specification, we find that the estimated effect of filtration on infant mortality shrinks by two-thirds when a series of data transcription errors are corrected.

- For infant mortality, 79 city-year observations in the C&M data set are recorded incorrectly (out of N = 415)
- In 9 out of the 13 cities in their sample, C&M make systematic transcription errors for the years 1910-1917
  - Differences are not large for some cities (e.g., Chicago)
  - Differences for other cities, however, are substantial (e.g., New Orleans)
  - In all cases, the values recorded by C&M are less than the correct counts from *Mortality Statistics*
- In addition to the systematic errors for the years 1910-1917, C&M make several other transcription mistakes, some of which can be easily explained.

**Differences in Recorded Infant Mortality Counts between Cutler and Miller (2005) and the U.S. Census Bureau's *Mortality Statistics***

| City | Year | C&M's recorded infant mortality count[a] | Correct infant mortality count from *Mortality Statistics*[b] | Reason for difference (when known) |
|---|---|---|---|---|
| Chicago, IL | 1910 | 6595.52 | 6844 | |
| | 1911 | 6017.86 | 6252 | |
| | 1912 | 6394.31 | 6678 | |
| | 1913 | 6649.87 | 6939 | |
| | 1914 | 6571.52 | 6878 | |
| | 1915 | 5942.99 | 6219 | |
| | 1916 | 6566.35 | 6910 | |
| | 1917 | 6246.72 | 6664 | |
| | 1931 | 766 | 2992 | To calculate, one needs to add white infant mortality (=2,617) and nonwhite infant mortality (=375). It appears as if C&M incorrectly added mortality for one-year-olds, rather than infants, for whites (=391) and nonwhite infant mortality, which gives their recorded total of 766. |
| Cleveland, OH | 1924 | 2366 | 1386 | To calculate, one needs to add white infant mortality (=1,219) and nonwhite infant mortality (=167). It appears as if C&M incorrectly added overall nonwhite mortality (=1,147) and white infant mortality (=1,219), which gives their recorded total of 2,366. |
| New Orleans, LA | 1910 | 571.931 | 1061 | |
| | 1911 | 595.471 | 1071 | |
| | 1912 | 416.903 | 774 | |
| | 1913 | 500.74 | 934 | |
| | 1914 | 477.419 | 883 | |
| | 1915 | 492.79 | 927 | |
| | 1916 | 404.008 | 757 | |
| | 1917 | 446.364 | 866 | |
| Pittsburgh, PA | 1901 | 6578 | 1580 | C&M incorrectly recorded the overall mortality count instead of the infant mortality count. |
| | 1904 | 771 | 1771 | C&M incorrectly entered "1771" as "771" |

## Comparing our Infant Mortality Estimates to those of Cutler and Miller (2005)

| | (1) | (2) | (3) | (4) | (5) |
|---|---|---|---|---|---|
| | Replicating C&M | Column (1) + cluster SEs | Column (2) + correct mortality counts | Column (3) + correct dates | Our specification limited to C&M’s city-years |
| *Filtration* | -.429$^{***}$ | -.429$^{***}$ | -.125$^{*}$ | -.067 | -.100$^{**}$ |
| | (.090) | (.138) | (.068) | (.057) | (.045) |
| N | 415 | 415 | 410 | 410 | 415 |
| Years | 1905-1936 | 1905-1936 | 1905-1936 | 1905-1936 | 1905-1936 |

Notes: The dependent variable is equal to the natural log of the infant mortality rate in city *c* and year *t*. Controls for the C&M regressions include those listed in Appendix Table 5, municipality fixed effects, year fixed effects and municipality-specific linear trends. Controls for our regression include those listed in Table 5, municipality fixed effects, year fixed effects and municipality-specific linear trends.