

Measuring the quality of mortality data in high-income settings

Amy Mann, Mathew Kiang, Monica Alexander

Toronto Data Science Workshop, January 14, 2026



UNIVERSITY OF
TORONTO

Motivation

Motivation

Motivation



[NASEM; 2021]

Motivation



[NASEM; 2021]

- More than 5,000 papers have cited US CDC's vital statistics system in the last ten years

Motivation



[NASEM; 2021]

- More than 5,000 papers have cited US CDC's vital statistics system in the last ten years
- Often taken for granted in high-income countries that mortality data is reliable

Motivation



[NASEM; 2021]

- More than 5,000 papers have cited US CDC's vital statistics system in the last ten years
- Often taken for granted in high-income countries that mortality data is reliable
- Most data quality papers focus on low-income settings and national registries. We are interested in sub-national differences in data quality

Death certificate coding

Death certificate coding

U.S. STANDARD CERTIFICATE OF DEATH	
1. DECEASED'S NAME (Print or Type)	
2. SEX	
3. DATE OF DEATH (Month, Day, Year)	
4. SOCIAL SECURITY NUMBER	
5. UNDER 1 YEAR	
6. DATE OF BIRTH (Month, Day, Year)	
7. BIRTHPLACE (City and State or Foreign Country)	
8. WAS DECEASED EVER IN U.S. ARMED FORCES?	
9. PLACE OF DEATH (Print or Type)	
10. FACILITY NAME (If not permanent, give street and number)	
11. CITY, TOWN, OR LOCATION OF DEATH	
12. COUNTY OF DEATH	
13. MARITAL STATUS (Married, Single, Widowed, Divorced, Separated, etc.)	
14. SURVIVING SPOUSE (If wife, give maiden name)	
15. DECEASED'S USUAL OCCUPATION (If not, give usual activity)	
16. END OF BUSINESS/INDUSTRY	
17. RESIDENCE - STREET	
18. CITY, TOWN, OR LOCATION	
19. STREET AND NUMBER	
20. INSURE CITY - ZIP CODE	
21. WAS DECEASED OF HISPANIC ORIGIN? (Specify to or from: P, M, S, C, G, A, B, H, I, J, K, L, N, O, R, T, U, V, W, X, Y, Z, No)	
22. DECEASED'S EDUCATION (Specify with highest grade completed)	
23. FATHER'S NAME (Print or Type)	
24. MOTHER'S NAME (Print or Type)	
25. INFORMANT'S NAME (Print or Type)	
26. MAILING ADDRESS (Street and Number or Rural Route Number, City or Town, State, Zip Code)	
27. METHOD OF DISPOSITION	
28. PLACE OF DISPOSITION (Name of cemetery, crematorium, or other place)	
29. LOCATION - City or Town, State	
30. SIGNATURE OF FUNERAL SERVICE LICENSEE OR PERSON ACTING AS SUCH	
31. LICENSE NUMBER (If licensed)	
32. NAME AND ADDRESS OF FACILITY	
33. Complete items 34a-c only. Do not complete if death is a result of homicide or suicide.	
34a. DATE OF DEATH	
34b. DATE REPORTED TO MEDICAL EXAMINER/DOCTOR (If not)	
34c. DATE REPORTED TO MEDICAL EXAMINER/DOCTOR (If not)	
35. PART I: State the disease, disorder, or complication that caused the death. Do not enter the mode of dying, such as cardiac or respiratory arrest, shock, or heart failure. List only one cause on each line.	
36. IMMEDIATE CAUSE (If not disease or condition, specify resulting in death)	
37. DUE TO OR AS A CONSEQUENCE OF	
38. DUE TO OR AS A CONSEQUENCE OF	
39. DUE TO OR AS A CONSEQUENCE OF	
40. PART II: State conditions contributing to death but not resulting in the underlying cause given in Part I.	
41. WAS AN AUTOPEY PERFORMED? (Yes or no)	
42. WERE AUTOPEY FINDINGS AVAILABLE PRIOR TO COMPLETION OF CAUSE OF DEATH? (Yes or no)	
43. MANNER OF DEATH	
44. DATE OF INJURY	
45. TIME OF INJURY	
46. INJURY AT SCENE? (Yes or no)	
47. DESCRIBE HOW INJURY OCCURRED	
48. PLACE OF INJURY (Home, Tavern, School, Office, etc.)	
49. LOCATION (Street and Number or Rural Route Number, City or Town, State)	
50. CERTIFIER (Print or Type)	
51. SIGNATURE OF CERTIFIER	
52. DATE SIGNED (Month, Day, Year)	
53. NAME AND ADDRESS OF PERSON WHO COMPLETED CAUSE OF DEATH ITEM 35 (Print or Type)	
54. DATE SIGNED (Month, Day, Year)	

[NAS; 2003]

[illegible]

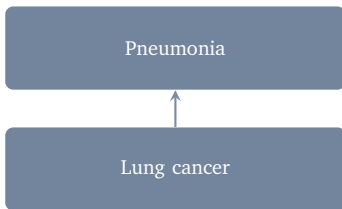
Underlying cause of death

Death certificate coding example

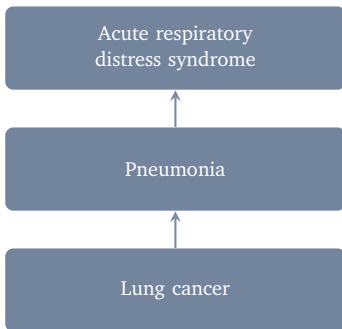
Death certificate coding example

Lung cancer

Death certificate coding example



Death certificate coding example



Mortality data collection process

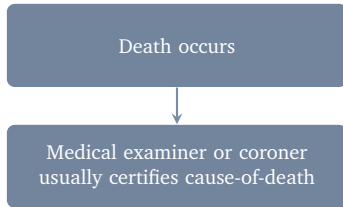
Mortality data collection process

Mortality data collection process

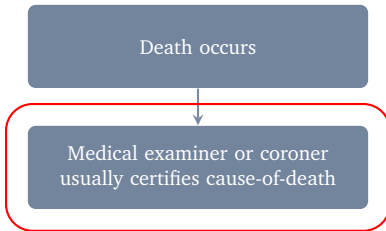
```
graph LR; A[Death occurs] --> B[Death is reported to the health department]; B --> C[The health department reports the death to the vital records office];
```

Death occurs

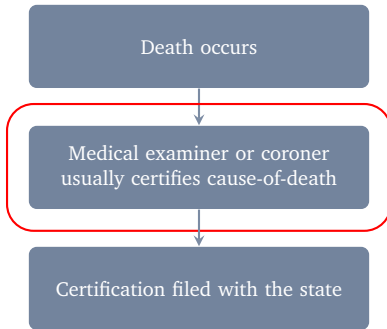
Mortality data collection process



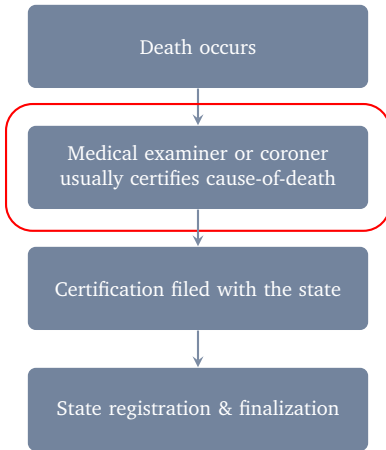
Mortality data collection process



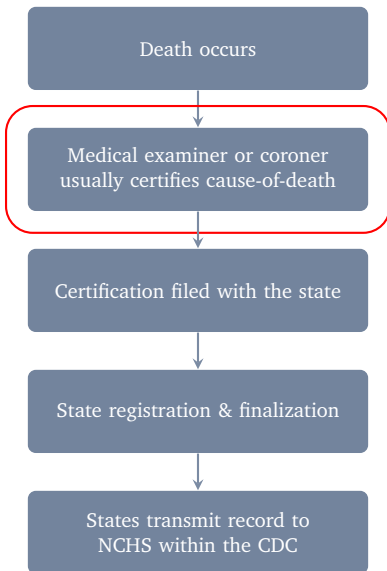
Mortality data collection process



Mortality data collection process



Mortality data collection process



Data quality problems

Data quality problems

- Missing information

Data quality problems

- Missing information
- Lack of detail

Data quality problems

- Missing information
- Lack of detail
- Garbage codes: useless or wrong codes

Data quality problems

- Missing information
- Lack of detail
- Garbage codes: useless or wrong codes

In the paper, we look at 3 aspects of data quality and then apply these metrics to the US at a county-level.

Methods

Proportion of garbage codes

Proportion of garbage codes

- First, we look at the proportion of garbage codes in each county:

$$\text{Prop. garbage} = \frac{\text{Number of garbage-coded deaths}}{\text{Number of deaths}}$$

Proportion of garbage codes

- First, we look at the proportion of garbage codes in each county:

$$\text{Prop. garbage} = \frac{\text{Number of garbage-coded deaths}}{\text{Number of deaths}}$$

- IHME has led work on garbage codes and efforts to re-classify them.

Proportion of garbage codes

- First, we look at the proportion of garbage codes in each county:

$$\text{Prop. garbage} = \frac{\text{Number of garbage-coded deaths}}{\text{Number of deaths}}$$

- IHME has led work on garbage codes and efforts to re-classify them.
- Re-assignment methods: (1) Proportional redistribution by age and sex; (2) Statistical modelling; (3) Expert judgment (for example, reconstructing the chain of events)

Proportion of garbage codes

- First, we look at the proportion of garbage codes in each county:

$$\text{Prop. garbage} = \frac{\text{Number of garbage-coded deaths}}{\text{Number of deaths}}$$

- IHME has led work on garbage codes and efforts to re-classify them.
- Re-assignment methods: (1) Proportional redistribution by age and sex; (2) Statistical modelling; (3) Expert judgment (for example, reconstructing the chain of events)
- Lozano et al, 2012 show that re-distributing garbage codes changes the top 10 leading causes of death worldwide.

Proportion of garbage codes

- First, we look at the proportion of garbage codes in each county:

$$\text{Prop. garbage} = \frac{\text{Number of garbage-coded deaths}}{\text{Number of deaths}}$$

- IHME has led work on garbage codes and efforts to re-classify them.
- Re-assignment methods: (1) Proportional redistribution by age and sex; (2) Statistical modelling; (3) Expert judgment (for example, reconstructing the chain of events)
- Lozano et al, 2012 show that re-distributing garbage codes changes the top 10 leading causes of death worldwide.
- In 2023, about 10% of deaths in the United States were garbage coded.

Level of detail

Level of detail

Goal is to measure how specific ICD-10 coding is for non-garbage codes.

Level of detail

Goal is to measure how specific ICD-10 coding is for non-garbage codes.

- We need to control for the underlying epidemiological differences in cause of death to capture the variation in detail driven by coding differences rather than the differences in the (true) distribution of causes of death. To do so, we weight deaths so that each county set matches the national average for the broad distribution of causes of death

Level of detail

Goal is to measure how specific ICD-10 coding is for non-garbage codes.

- We need to control for the underlying epidemiological differences in cause of death to capture the variation in detail driven by coding differences rather than the differences in the (true) distribution of causes of death. To do so, we weight deaths so that each county set matches the national average for the broad distribution of causes of death
- Another way to think about this is that we are interested in the diversity within each cause of death category.

Level of detail

- For each time period t , we compute the national share of deaths in each broad cause group. These national shares are used as weights.

Level of detail

- For each time period t , we compute the national share of deaths in each broad cause group. These national shares are used as weights.
- Using these weights, we construct a standardized ICD-10 distribution for each county set k :

$$p_{k,t}^*(d) = \sum_c s_t(c) w_{k,t}(d | c)$$

Level of detail

- For each time period t , we compute the national share of deaths in each broad cause group. These national shares are used as weights.
- Using these weights, we construct a standardized ICD-10 distribution for each county set k :

$$p_{k,t}^*(d) = \sum_c s_t(c) w_{k,t}(d | c)$$

- Here, $w_{k,t}(d | c)$ is the within-cause distribution of ICD-10 codes in county set k , and $s_t(c)$ is the national share of cause group c .

Level of detail

- For each time period t , we compute the national share of deaths in each broad cause group. These national shares are used as weights.
- Using these weights, we construct a standardized ICD-10 distribution for each county set k :

$$p_{k,t}^*(d) = \sum_c s_t(c) w_{k,t}(d | c)$$

- Here, $w_{k,t}(d | c)$ is the within-cause distribution of ICD-10 codes in county set k , and $s_t(c)$ is the national share of cause group c .
- $p_{k,t}^*(d)$ represents the probability that a death in county set k would be coded as ICD-10 code d , *if the county had the national cause-of-death mix*.

Level of detail

We then measure how spread out this standardized ICD-10 distribution is using Shannon entropy.

$$H_{k,t} = - \sum_d p_{k,t}^*(d) \log p_{k,t}^*(d)$$

- Entropy is a measure of diversity:
 - high entropy implies deaths spread across many ICD-10 codes
 - low entropy implies deaths clumped into a few codes

Level of detail

We then measure how spread out this standardized ICD-10 distribution is using Shannon entropy.

$$H_{k,t} = - \sum_d p_{k,t}^*(d) \log p_{k,t}^*(d)$$

- Entropy is a measure of diversity:
 - high entropy implies deaths spread across many ICD-10 codes
 - low entropy implies deaths clumped into a few codes
- We rescale entropy to a 0-100 score to obtain the level of detail.

Re-assignability Index

Re-assignability Index

The goal is to quantify how re-assignable garbage codes are in each county.

Re-assignability Index

The goal is to quantify how re-assignable garbage codes are in each county.

$$h_i = \frac{-\sum_{k \in K_g} p_i(k) \log p_i(k)}{\log |K_g|} \in [0, 1]$$

Re-assignability Index

The goal is to quantify how re-assignable garbage codes are in each county.

$$h_i = \frac{-\sum_{k \in K_g} p_i(k) \log p_i(k)}{\log |K_g|} \in [0, 1]$$

h_i := normalized Shannon entropy for record i

K_g = candidate underlying causes of death

$p_i(k)$ = probability candidate k is the true cause of death

Re-assignability Index

The goal is to quantify how re-assignable garbage codes are in each county.

$$h_i = \frac{-\sum_{k \in K_g} p_i(k) \log p_i(k)}{\log |K_g|} \in [0, 1]$$

Re-assignability Index

The goal is to quantify how re-assignable garbage codes are in each county.

$$h_i = \frac{-\sum_{k \in K_g} p_i(k) \log p_i(k)}{\log |K_g|} \in [0, 1]$$

- We calculate p_i using a model trained on the deaths with candidate i listed as the underlying cause of death and the garbage code g listed in the multiple cause of death, building on the work of Foreman et al., 2016

Re-assignability Index

The goal is to quantify how re-assignable garbage codes are in each county.

$$h_i = \frac{-\sum_{k \in K_g} p_i(k) \log p_i(k)}{\log |K_g|} \in [0, 1]$$

- We calculate p_i using a model trained on the deaths with candidate i listed as the underlying cause of death and the garbage code g listed in the multiple cause of death, building on the work of Foreman et al., 2016

$$\text{RI} := 1 - \frac{\sum_{i \in G_{c,t}} h_i}{N_{c,t}} \in [0, 1]$$

Aggregating data quality indices

Aggregating data quality indices

- Constructed an aggregate data quality index by averaging the z-scores of the three metrics for each county set

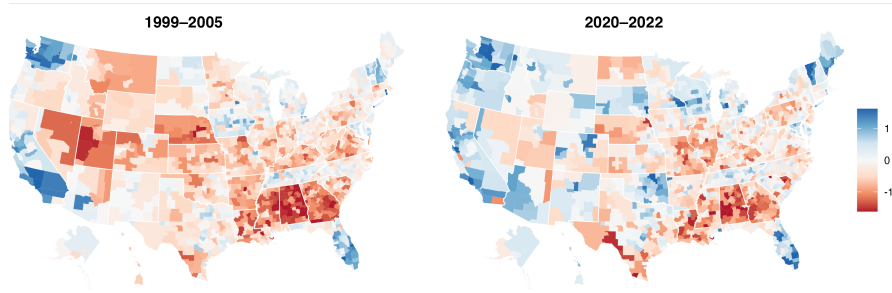
Aggregating data quality indices

- Constructed an aggregate data quality index by averaging the z-scores of the three metrics for each county set
- For example, $z_{\text{avg}} = 1$ means the county set has, on average, one standard deviation better data quality than the mean

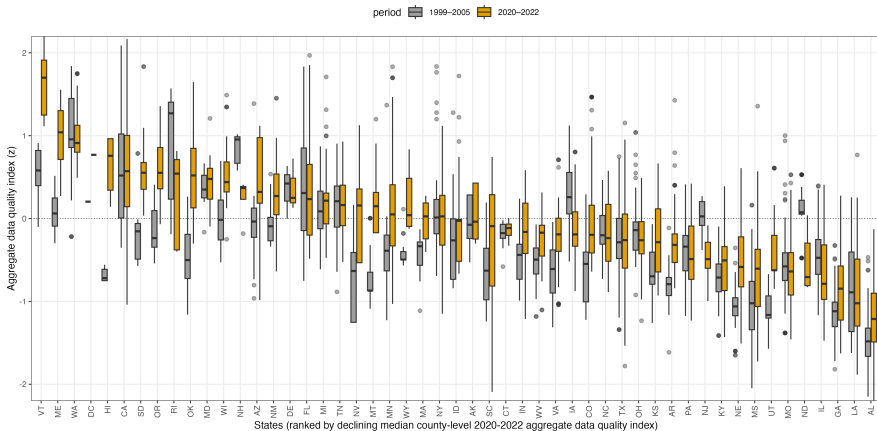
Results

Aggregate data quality index

Aggregate data quality index



Aggregate data quality index



Aggregate data quality indices

Aggregate data quality indices

- In addition to socio-economic effects, there are clear state-specific effects

Aggregate data quality indices

- In addition to socio-economic effects, there are clear state-specific effects
- Correlation with median county income is 0.31 in 1999-2005 and 0.37 in 2020-2022

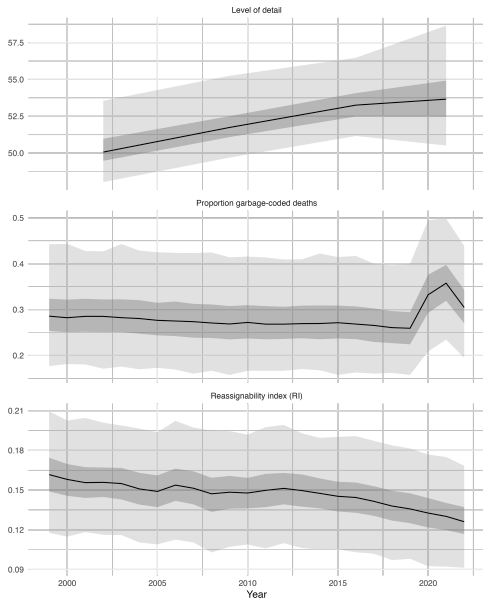
Aggregate data quality indices

- In addition to socio-economic effects, there are clear state-specific effects
- Correlation with median county income is 0.31 in 1999-2005 and 0.37 in 2020-2022
- Per-capita public health spending correlation is 0.17 in 1999-2005 and 0.14 in 2020-2022

Aggregate data quality indices

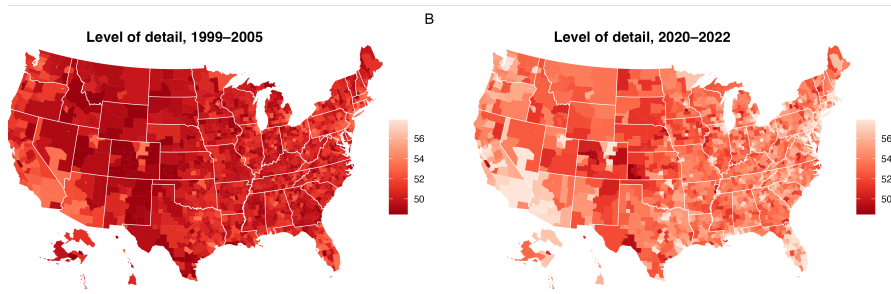
- In addition to socio-economic effects, there are clear state-specific effects
- Correlation with median county income is 0.31 in 1999-2005 and 0.37 in 2020-2022
- Per-capita public health spending correlation is 0.17 in 1999-2005 and 0.14 in 2020-2022
- Association with reporting type is 0.33 in 1999-2005 and 0.34 in 2020-2022

Data quality over time

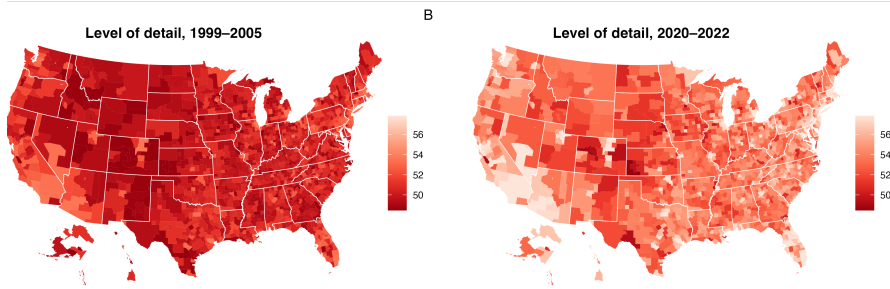


Level of detail

Level of detail



Level of detail



- Increased from 1999-2022; unclear to what extent it is driven by increasing diversity in cause of death in general (the underlying cause mixture is controlled for by period, not globally) or by improved specificity in coding

Level of detail

- Higher level of detail associated with higher median county income ($\rho = 0.23$ in 1999-2005; 0.26 in 2020-2022),

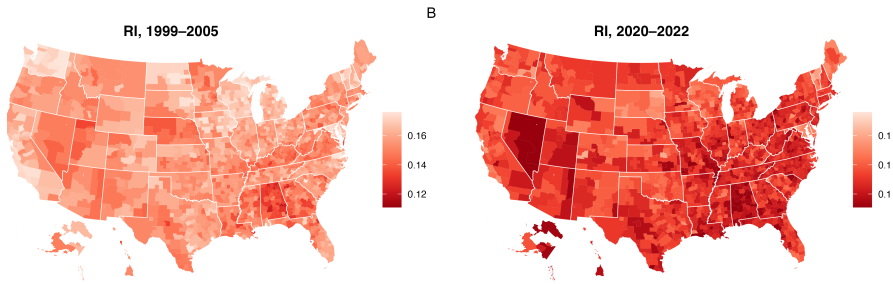
Level of detail

- Higher level of detail associated with higher median county income ($\rho = 0.23$ in 1999-2005; 0.26 in 2020-2022),
- Higher level of detail with death investigation system type ($\rho = 0.24$ in 1999-2005; 0.27 in 2020-2022).

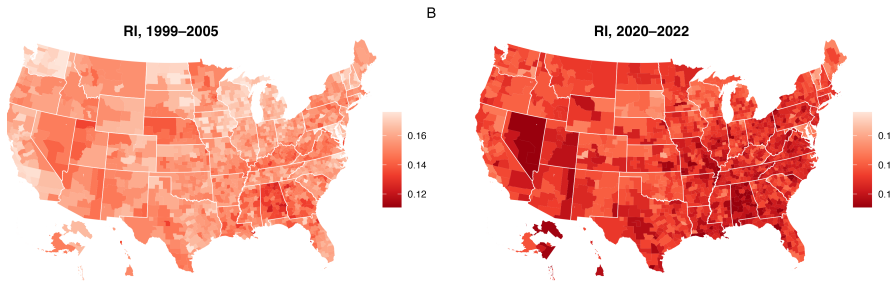
Level of detail

- Higher level of detail associated with higher median county income ($\rho = 0.23$ in 1999-2005; 0.26 in 2020-2022),
- Higher level of detail with death investigation system type ($\rho = 0.24$ in 1999-2005; 0.27 in 2020-2022).
- Small association between detail and per-capita public health spending ($\rho = 0.08$ in 1999-2005; 0.05 in 2020-2022).

Re-assignability Index

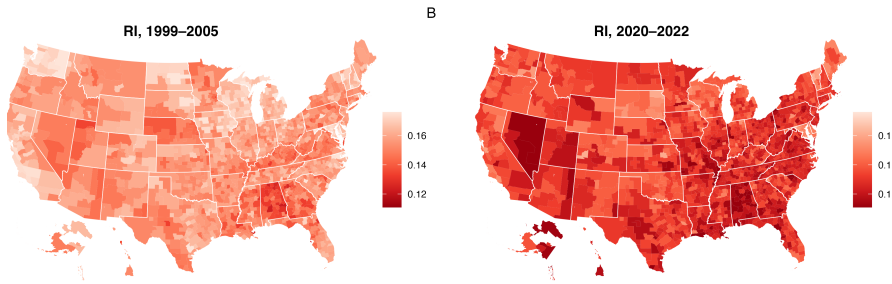


Re-assignability Index



- Positively correlated with mean county income ($\rho = 0.14$ in 1999-2005; 0.13 in 2020-2022) and with per-capita public health spending ($\rho = 0.09$ in 1999-2005; 0.08 in 2020-2022)

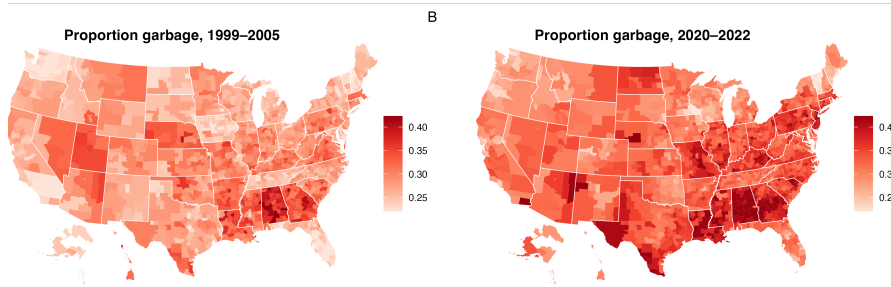
Re-assignability Index



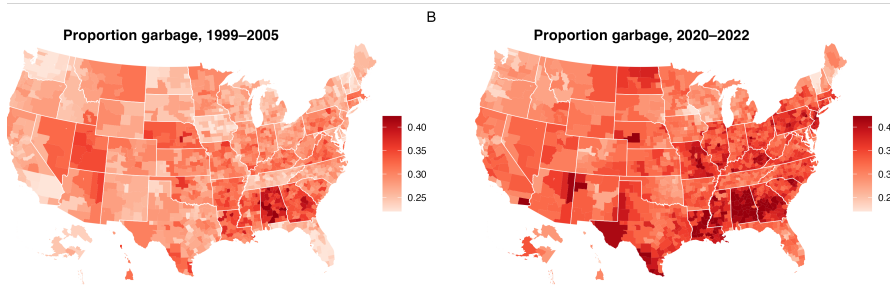
- Positively correlated with mean county income ($\rho = 0.14$ in 1999-2005; 0.13 in 2020-2022) and with per-capita public health spending ($\rho = 0.09$ in 1999-2005; 0.08 in 2020-2022)
- Small differences by reporting type ($\rho = 0.11$ in 1999-2005; 0.06 in 2020-2022).

Proportion of garbage codes

Proportion of garbage codes

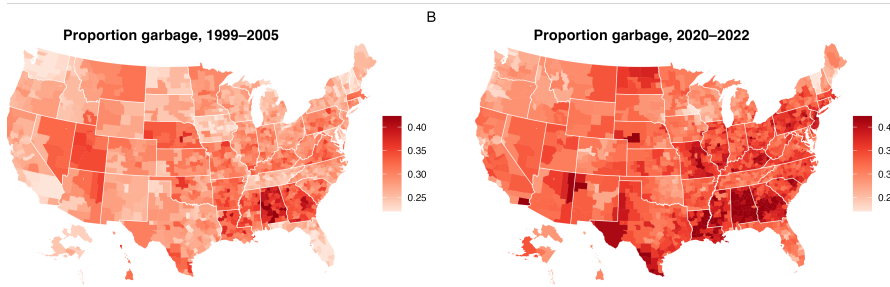


Proportion of garbage codes



- Decreased from 1999-2019 and increased during the COVID-19 pandemic
- The proportion of garbage-coded deaths was lower in higher-income counties ($\rho = -0.25$ in 1999-2005; -0.29 in 2020-2022)

Proportion of garbage codes



- Decreased from 1999-2019 and increased during the COVID-19 pandemic
- The proportion of garbage-coded deaths was lower in higher-income counties ($\rho = -0.25$ in 1999-2005; -0.29 in 2020-2022)
- Lower in counties using medical examiners ($\rho = -0.31$ in 1999-2005; -0.34 in 2020-2022) and associated with per-capita public health spending ($\rho = -0.21$ in 1999-2005; -0.15 in 2020-2022)

Relationship between data quality metrics and COVID-19 undercounting

Relationship between data quality metrics and COVID-19 undercounting

- Correlation between COVID-19 underreporting and proportion of higher garbage-coded deaths and lower level of detail is 0.27 and 0.28.

Relationship between data quality metrics and COVID-19 undercounting

- Correlation between COVID-19 underreporting and proportion of higher garbage-coded deaths and lower level of detail is 0.27 and 0.28.
- Correlation between lower RI and underreporting is 0.13

Relationship between data quality metrics and COVID-19 undercounting

- Correlation between COVID-19 underreporting and proportion of higher garbage-coded deaths and lower level of detail is 0.27 and 0.28.
- Correlation between lower RI and underreporting is 0.13
- Correlation between lower aggregate data quality index and COVID-19 underreporting is 0.33

Discussion

Discussion

- Different data quality metrics point to different potential problems

Discussion

- Different data quality metrics point to different potential problems
- More centralized reporting practises as well as use of medical examiners rather than coroners might help reduce garbage codes

Discussion

- Different data quality metrics point to different potential problems
- More centralized reporting practises as well as use of medical examiners rather than coroners might help reduce garbage codes
- Low level of detail might suggest that diagnostic specificity is constrained and could perhaps benefit from more medicolegal capacity, toxicology and autopsy access

Discussion

- Different data quality metrics point to different potential problems
- More centralized reporting practises as well as use of medical examiners rather than coroners might help reduce garbage codes
- Low level of detail might suggest that diagnostic specificity is constrained and could perhaps benefit from more medicolegal capacity, toxicology and autopsy access
- Low RI suggests incomplete or generic MCODE reporting and is likely most affected by jurisdictional differences in death certificate coding practises

Discussion

- Different data quality metrics point to different potential problems
- More centralized reporting practises as well as use of medical examiners rather than coroners might help reduce garbage codes
- Low level of detail might suggest that diagnostic specificity is constrained and could perhaps benefit from more medicolegal capacity, toxicology and autopsy access
- Low RI suggests incomplete or generic MCODE reporting and is likely most affected by jurisdictional differences in death certificate coding practises
- The strong state-specific effects on data quality is something to be optimistic about

Thank you!

Contact information:

Amy Mann

amy.mann@chch.ox.ac.uk

Monica Alexander

monica.alexander@utoronto.ca

Mathew Kiang

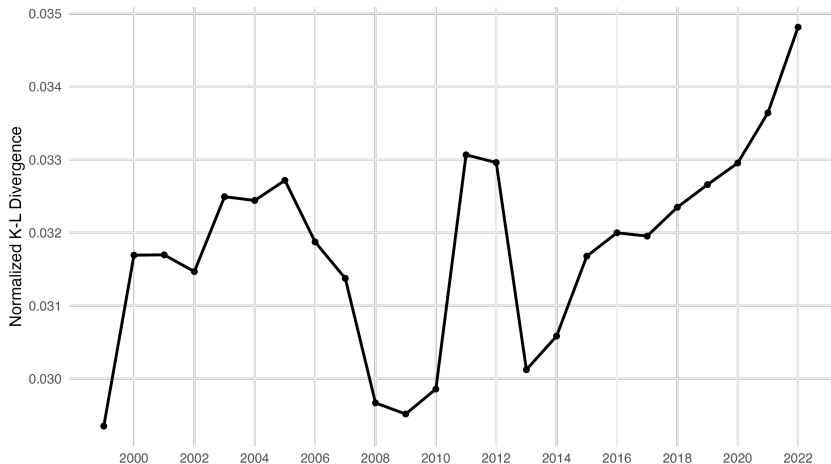
mkiang@stanford.edu



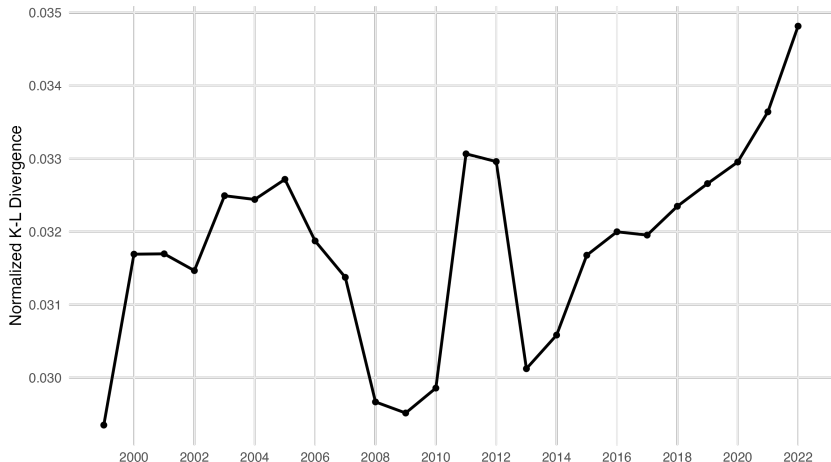
Supported by Data Sciences Institute

Supplementary Material

Why is it getting harder to re-assign garbage codes?



Why is it getting harder to re-assign garbage codes?



It's probably not because death certificates are becoming less descriptive.