Improving race and ethnicity completeness in reportable disease surveillance data by matching to external data sources

Maryam Iqbal

Bureau of Communicable Diseases

NYC Department of Health and Mental Hygiene

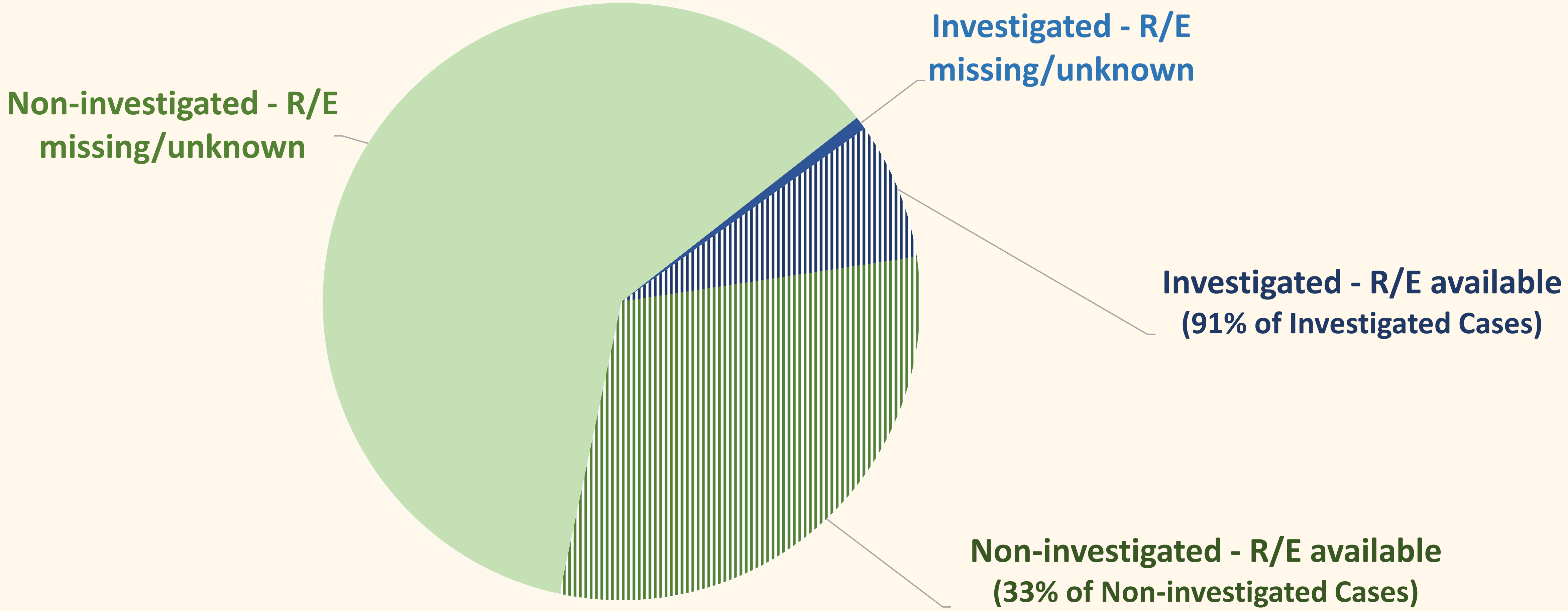
Background

The availability of race and ethnicity (R/E) data for cases of reportable disease is critical for

- understanding disparities in disease incidence and outcomes
- developing and evaluating the success of strategies to address these disparities

Availability of race & ethnicity data in the NYC Communicable Disease Surveillance System (Maven)

Race Ethnicity is Complete for 38% of Total Cases*



Confirmed and probable cases N=63,777

Purpose

- Determine the quality of R/E data found in the external sources
- Inform the decision on which external source can be used to improve the completeness of R/E data in Maven

Data sources

NYC Communicable disease (Maven - CD)

Confirmed, probable, suspect

Diagnosis years 2000 - 2016

External Sources

Combined surveillance database

- HIV, TB, STD, >30 communicable diseases, death records, A1C
- Diagnosis years 2000 – 2013

Hospitalization/Healthcare database (SPARCS)

- Inpatient, outpatient, emergency room visits
- Discharge years 2000 – 2015

Health Information Exchange (HIE)

- Any patient with a Hepatitis C ICD 9/10, lab result, or medication reported to HIE between July and December 2016

Methods - Record matching and QA

- SPARCS dataset
  - Matching done on keys: parts of first name, last name, SSN, DOB, sex
  - QA – random sample manually reviewed

- Combined surveillance database
  - Common person ID exists in matched dataset and Maven-CD

- HIE
  - Matching done on key: parts of first name, last name, DOB

Methods - Race & ethnicity value cleaning

Collapsed race & ethnicity values
Hispanic
American Indian/Alaska Native, non-Hispanic
Asian, non-Hispanic
Black/African American, non-Hispanic
Native Hawaiian/Pacific Islander, non-Hispanic
White, non-Hispanic
Multiracial (>1 race), non-Hispanic
Other, non-Hispanic
Missing/Unknown

Methods - Race & ethnicity value matching

- The “gold standard” was used to evaluate concordance
  - Investigated cases from Maven used as a proxy “gold standard”
  - R/E values that were Missing/Unknown were not used

Type of R/E value or match	Maven	External data source
Exact informative	Asian	Asian
Exact non-informative	Other	Other
Mismatch	Asian	Black/African American

Methods - Measures

- Concordance: Exact informative only

	Maven - CD			
External source		Match	Mismatch	Concordance
	Match	A	B	(A/A+B)*100

- Concordance of R/E: Number of persons categorized as a specific R/E category in the external source that matched to the same R/E category in Maven over the total number of persons categorized as that same R/E in Maven

- Improvement: Replacing Maven R/E values that were Missing/Unknown with external source values that were not missing/unknown, multiracial or other
  - If post-match completeness is greater than pre-match completeness, we consider this an improvement

$$\textbf{R/E Completeness} = \left(\frac{\text{No. known R/E}}{\text{Total N Maven}}\right) * 100$$

Results – Person record match with Maven

Data source	Percent
Combined surveillance (2000 – 2013)	16%
SPARCS (2000 – 2015)	81%
HIE (July – December 2016)	73%

Results - Race & ethnicity value match

Data source	N	Concordance	R/E completeness	
			Pre-match (Maven only)	Post-match (Maven + Data source)
Combined surveillance	401,360	91%	44%	51%
SPARCS	579,124	91%	46%	76%
HIE*	7,005	NA	71%	89%

* Concordance not calculated because Hepatitis C cases are not investigated

Results - Concordance in the combined surveillance dataset

Combined surveillance database	Race & ethnicity categories	Concordance
	Hispanic	95%
	Black/African American, non-Hispanic	94%
	White, non-Hispanic	91%
	Asian, non-Hispanic	87%
	Other, non-Hispanic	19%
	American Indian/Alaska Native , non-Hispanic	17%
	Multiracial, non-Hispanic	7%
	Native Hawaiian/Pacific Islander, non-Hispanic	6%

Results - Concordance in the SPARCS match

	Race & ethnicity categories	Concordance
SPARCS	Black/African American, non-Hispanic	89%
	White, non-Hispanic	85%
	Asian, non-Hispanic	65%
	Hispanic	64%
	Other, non-Hispanic	33%
	American Indian/Alaska Native , non-Hispanic	13%
	Multiracial, non-Hispanic	0%
	Native Hawaiian/Pacific Islander, non-Hispanic	0%

Conclusions

- These preliminary results suggest we can improve our R/E data through importation of data from external sources
  - High person record match rate
    - 81% Maven to SPARCS
    - 73% HIE to Maven
  - High concordance
    - 91% SPARCS and combined surveillance database
  - > 10% improvement in R/E completeness
    - 13% combined surveillance database
    - 65% SPARCS
    - 25% HIE

Limitations

- R/E data from SPARCS and HIE sourced from electronic medical records (EMRs) which may not be accurate [1,2]
    - Perceived R/E by provider → Misclassification of R/E
- "Gold standard"
    - Approximately 11% of the cases that are investigated are by medical chart review, sourced from EMRs
    - Approximately 80% of investigated cases are by phone interview – quality of R/E data may vary across staff that collect these data

[1] West CN, Geiger AM, Greene SM, et al. Race and ethnicity: comparing medical records to self-reports. J Natl Cancer Inst Monographs. 2005;(35):72-4.

[2] Klinger EV, Carlini SV, Gonzalez I, et al. Accuracy of race, ethnicity, and language preference in an electronic health record. J Gen Intern Med. 2015;30(6):719-23.

Next steps

- Additional matches
    - Enhanced QA on matches
- HIE will conduct a person match to BCD’s reportable diseases for 2016
- Evaluate Multiracial category – Can it be more informative?
- Multivariable analysis to evaluate factors associated with missingness or mismatches

Acknowledgements

- Jennifer Baumgartner
- Annie Fine
- Sharon Greene
- General Surveillance Unit, Bureau of Communicable Disease

Questions?

miqbal1@health.nyc.gov