The association of demographic factors with negative desriptors in medical text

Andrew Zimolzak, Traber Giardina, Darius Dawson, Terri Fletcher, Taylor Scott, Debra Choi

April, 2022

Introduction

Recent work used an expert panel to develop a list of negative patient descriptors and found that Black patients had disproportionately higher odds of negative patient descriptors appearing in the history and physical notes of their medical records compared with White patients. Sun *et al.* used the following 15 descriptors:

> (non-)adherent, aggressive, agitated, angry, challenging, combative, (non-)compliant, confront, (non-)cooperative, defensive, exaggerate, hysterical, (un-)pleasant, refuse, and resist.
>
> **Source:** Sun M, Oliwa T, Peek ME, Tung EL. Negative Patient Descriptors: Documenting Racial Bias In The Electronic Health Record. *Health Aff (Millwood).* 2022 Feb;41(2):203-211. PMID: 35044842.

We measured the association of race, ethnicity, and sex with these negative descriptor words (and word variants), in the clinical notes of 100 BCM patients who had an ER visit and inpatient stay.

Data cleaning, analysis, and reporting was performed using R Markdown software. All code (including code used to generate this report) can be found at https://github.com/zimolzak/datathon-2022

Tables and figures

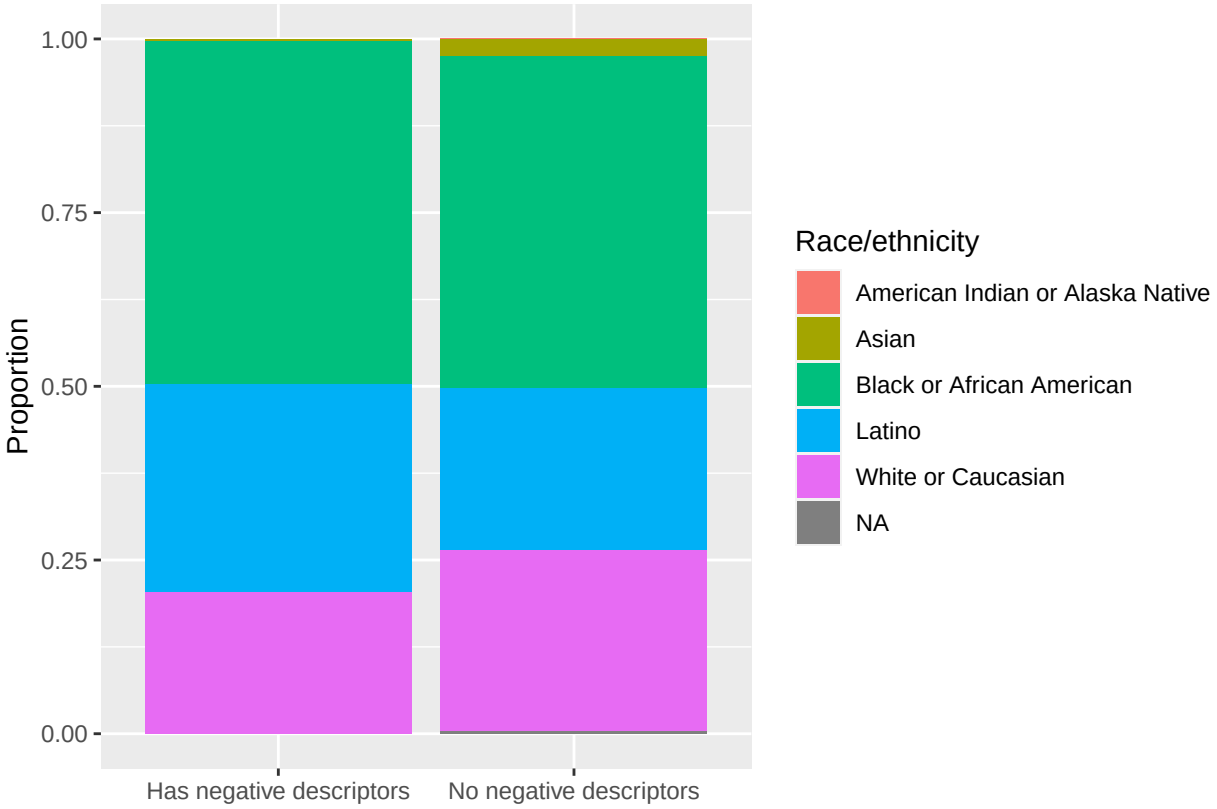

Figure. Distribution of race/ethnicity, in notes with negative descriptors, compared to notes without negative descriptors. Latino patients are overrepresented among notes that contain one or more negative descriptor words. Here, the Latino grouping comprises *any* race, whereas the other groupings comprise *non-Latino* patients of the specified race.

	Has negative descriptors	No negative descriptors
American Indian or Alaska Native	0	1
Asian	2	314
Black or African American	239	6259
Latino	145	3032
White or Caucasian	99	3415

Table 1. Distribution of race/ethnicity, in notes with negative descriptors, compared to notes without negative descriptors. The population with negative descriptors is significantly different from the population without negative descriptors (in terms of race/ethnicity distribution). $P = 2.424595 \times 10^{-5}$, Fisher's exact test.

	Has negative descriptors	No negative descriptors
Female	268.000	8506.000
Male	217.000	4515.000
Proportion.women	0.553	0.653

Table 2. Distribution of sex, in notes with/without negative descriptors. There are proportionally

fewer women / more men represented among notes containing negative descriptors. $P = 7.7124405 \times 10^{-6}$

	Has negative descriptors	No negative descriptors
Hispanic or Latino	145.000	3032.000
Not Hispanic or Latino	340.000	9989.000
Proportion.Latino	0.299	0.233

Table 3. Distribution of ethnicity, in notes with/without negative descriptors. There are proportionally more Latino patients represented among notes containing negative descriptors. $P = 0.0010466$

	Has negative descriptors	No negative descriptors
American Indian or Alaska Native	0.000	1.000
Asian	2.000	314.000
Black or African American	239.000	6259.000
Native Hawaiian or Other Pacific Islander	0.000	15.000
Unable to Determine	0.000	17.000
White or Caucasian	244.000	6415.000
Total	485.000	13021.000
Proportion.black	0.493	0.481

Table 4. Distribution of race, in notes with/without negative descriptors. There are more black / African-American patients represented among notes containing negative descriptors. Here, Latino patients can be in any grouping, although the majority are white Latino. $P = 0.0374947$

Discussion

There are significantly **more Latino patients** (23.3 vs. 29.9 percent) than non-Latinos represented among notes containing any negative descriptor ($P = 0.00105$). There are significantly **more black patients** (48.1 vs. 49.3 percent) represented among notes with any negative descriptor ($P = 0.0375$). There are significantly **fewer women** (65.3 vs. 55.3 percent) represented among notes containing any negative descriptor, compared to notes without a negative descriptor ($P = 7.71 \times 10^{-6}$). Considering the *combination* of race and ethnicity fields, (that is, with the major groups being black, Latino, and white non-Latino,) there is a significant difference in the overall distribution of groups, between notes with vs. without a negative descriptor ($P = 2.42 \times 10^{-5}$).

Data lessons learned

- Data retrieval and cleaning takes time!
- Handing off a dataset is not trivial (receiving end does not know how it was made, or column meanings, and may need explanations for unexpected associations).
- Project management: often linear not parallel, critical path, tricky to involve whole team.
- It is possible to do things fast.
- Record keeping is important, even when going fast.