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 patients who had an ER visit and inpatient stay.

Methods

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 . There were 115 edits to the code from April 5 to April 20 (average 7.7 edits per day). As of April 20, the project comprises about 1000 lines of code (average 68 lines per day). All contingency tables were analyzed with the two-sided Fisher's exact test.

Tables and figures

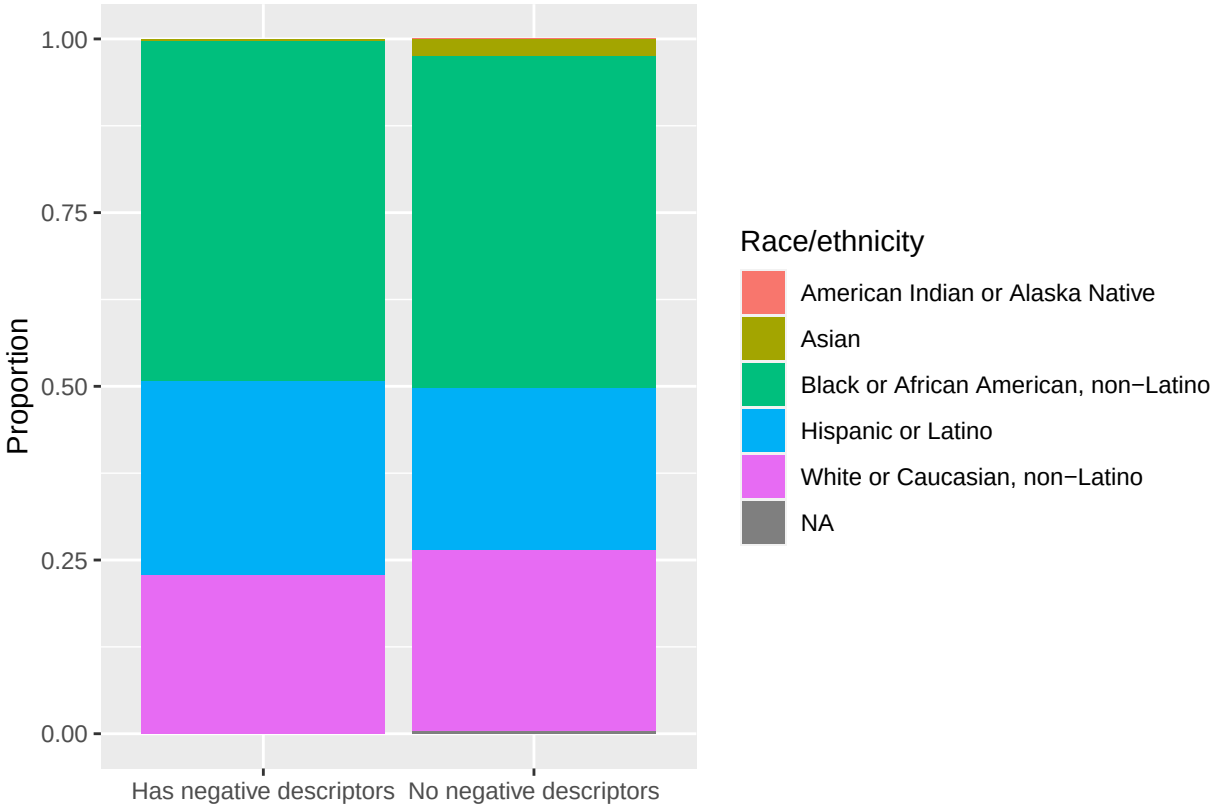

Figure. Distribution of race/ethnicity, in notes with negative descriptors, compared to notes without negative descriptors. Latino patients are overrepresented among notes that contain one or more negative descriptor words. Here, the Latino grouping comprises *any* race, whereas the other groupings comprise *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, non-Latino	310	6188
Hispanic or Latino	177	3000
White or Caucasian, non-Latino	144	3370

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), $P \ll 0.05$. (Actual value $P = 3.3105144 \times 10^{-5}$, Fisher's exact test.)

	Has negative descriptors	No negative descriptors
Female	393.000	8381.000
Male	240.000	4492.000
Proportion.women	0.621	0.651

Table 2. Distribution of sex, in notes with/without negative descriptors. For the difference

association of sex with notes containing negative descriptors, $P = 0.1245313$.

	Has negative descriptors	No negative descriptors
Hispanic or Latino	177.00	3000.000
Not Hispanic or Latino	456.00	9873.000
Proportion.Latino	0.28	0.233

Table 3. Distribution of ethnicity (alone), in notes with/without negative descriptors. There are proportionally more Latino patients represented among notes containing negative descriptors, $P \ll 0.05$. (Actual value $P = 0.0082391$.)

	Has negative descriptors	No negative descriptors
American Indian or Alaska Native	0.00	1.000
Asian	2.00	314.000
Black or African American	310.00	6188.000
Native Hawaiian or Other Pacific Islander	0.00	15.000
Unable to Determine	0.00	17.000
White or Caucasian	321.00	6338.000
Total	633.00	12873.000
Proportion.Black	0.49	0.481

Table 4. Distribution of race (alone), 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, $P < 0.05$. (Actual value $P = 0.0027114$.)

Discussion

There are significantly **more Latino patients** (23.3 vs. 28 percent) than non-Latinos represented among notes containing any negative descriptor ($P = 0.00824$). There are significantly **more Black patients** (48.1 vs. 49 percent) represented among notes with any negative descriptor ($P = 0.00271$). For the proportion of women (65.1 vs. 62.1 percent) among notes with a negative descriptor, vs. without, $P = 0.125$. 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 ($P = 3.31 \times 10^{-5}$).

Limitation: Negative descriptors are really only *potentially* negative descriptors. Sun *et al.* were able to classify usages as negative, positive, or neutral. We have not implemented this technique.

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: team members often work in tandem, not in parallel.
- Sometimes challenging to involve all team members.
- It is possible to do things fast.
- Record keeping is important, even when going fast.

- With “secondary use,” researchers make do with the available data. The warehouse may not have every data element needed to answer certain questions.