Drew Walker
Texts as Data Final Paper

Background

Provider biases are widely believed to contribute to discrimination and health disparities, and leaders across provider specialties and disciplines have long called for investigation of the processes by which provider bias forms and is cultivated among medical teams.[1,2] Implicit biases are unconscious negative evaluations about others on the basis of irrelevant characteristics such as race, gender, or SES, which are widely held by medical professionals and carry the potential to negatively impact care and patient relationships.[3,4] Patients who identify with frequently marginalized groups may experience multiple complex levels of bias, which can be investigated through utilizing intersectional approaches offered by feminist and critical race theories.[5]

Clinical notes play a key role in the transmission of information and attitudes towards patients that impact decision-making from one provider to another.[6] Through this transmission of information, patients can quickly develop "reputations" among the provider team, and different perceptions of patients can take on a self-fulfilling momentum of their own.[7] Within the context of clinical charts, specific features of language commonly used by providers have been identified that may perpetuate biases across providers and impact clinical decision-making through the induction of doubt, mistrust, or stigmatizing language about patients and the validity of their reported pain or symptoms.[8–11] In a recent study by Goddu and colleagues, medical students were exposed to patient charts comprised of real clinical note language comprised of both neutral and prejudiced language describing a hypothetical 28-year-old sickle cell patient. Students exposed to the prejudiced language were more likely to report negative attitudes toward the patient and endorse less aggressive pain management strategies. [8] Charts of black patients were found by Beach and colleagues to have 48% higher odds of containing inappropriate "scare quotes", and 32% higher odds of including evidentials, both of which have been studied as linguistic features used to specifically connotate distrust or invalidation of point-of-view.[10] Research on the transmission of provider bias through clinical notes has identified many common manifestations of both negative and positive bias within patient charts[9], and this text and language-based transmission of bias towards patients is thought to drive disparities in care across race, gender, drug use, and other marginalized conditions.[3,9,10,12] While some trainings utilizing Implicit Association Tests (IAT) and Diversity, Equity, and Inclusion (DEI) workshops have been implemented in provider education to combat this, these workshops typically lack long-term engagement, accountability, evaluation, or follow-up, and fail to catch and intervene on bias transmission in real-time and in real-life situations.[3]

Detecting bias in patient charts in real-time could offer healthcare teams opportunities to evaluate and intervene in biased chart language. This task requires development of scalable, computationally intensive systems to keep up with the high-volume data from live clinical provider teams.[13] State-of-the-art advances in natural language processing (NLP) with transformer-based contextual word embeddings have demonstrated high model performance in categorization texts for large amounts of medical data, [14] and others have been developed specifically for the detection of biased text.[15,16] This project aims to apply advanced methods in natural language processing to detect and assess the presence of scare quotes language among quoted text in provider clinical notes for a subset of patients who report significant stigmatization. We will seek to evaluate the ability of a mix of regex (matching for any full pair of quotation marks), and supervised machine learning algorithm approaches to classify the

presence of "scare quotes" that arise in clinical notes of patient records within the MIMIC-III database.

Methods

Data Collection Steps

The corpus for this project was MIT's MIMIC-III Database, a collection of medical charts for over 40,000 patients receiving in-hospital care at Beth Israel Deaconess Medical Center in Boston, MA from 2001-2012. We received authorization to access the data for this project on August 31, 2021. Initially, we planned to examine free-text clinical text notes from the MIMIC-III database for patients with Sickle Cell Disease (determined by ICD-9 codes) during hospital admissions. However, this strategy yielded prohibitively low sample sizes that were not conducive to NLP research. With this in mind, we expanded the sample to select patient charts for individuals who had the following diagnoses associated with opioid use that have been noted by previous literature to endure stigmatization by providers:

- Sickle Cell Anemia Types (Codes: 282.60-282.69, and 282.40-282.49)
- Chronic Pain (Code 338.2)[17] (Carr 2016)
- Opioid Use Disorders (Codes: 304.0, 304.7)[18], (Shah Methadone 2010)
- HIV/AIDs (Code 042) [19,20]

We removed notes from Radiology, ECG, Respiratory, or Echo from the sample, due to these being primarily laboratory results which typically contain less patient narrative text. We then tokenized the sentences, and filtered for all that contained quotation marks using the following regex: "\"(.+?)\"" . After filtering, the initial corpus collection task resulted in a dataframe of a total of 3842 quoted sentences in 2,909 unique charts of 928 patients written by 545 caregivers over 1479 admissions. To develop the test set, we randomly assigned numbers to each sentence, sorted by increasing value, and coded until successfully classifying 1200 quoted sentences.

After sections of text using quotations were identified using regex, we conducted sentence-level analyses for sentences including full quotation sets to examine presence of scare quotes at the sentence first, and then aggregate at patient and provider levels. We will aim to follow a coding ontology previously identified by Beach, Saha and colleagues in order to detect types of quotes. (Beach & Saha, 2021). A description of the ontology for categorizing patient charts into categories of "Helpful" (not a Scare Quote), "Harmful" (Likely Scare Quote), “Possibly Harmful” (Possible Scare Quote) is presented by the table previously published in their original paper and our adaptation for this study is included in Appendix 1. We also added “Not Applicable” for text fragments that did not involve direct quotations of patients.

Analyses

We sought to detect the presence of scare quotes in clinical notes by following a coding ontology previously identified by Beach, Saha and colleagues. (Beach & Saha, 2021). A description of the ontology for categorizing patient charts into categories of "Probably Useful" (not a Scare Quote), "Probably Harmful" (Likely Scare Quote), Potentially Misinterpreted (Potential Scare

Quote) is presented by the table previously published in their original paper. Following qualitative coding, we conducted multiple logistic regression analyses modeling the outcomes of patient quotation types (Helpful, Harmful, Possibly Harmful, Not Applicable) with gender as predictors. We also examined the presence of the quotation types within and between patients, providers, genders, and note types (Discharge Summary, Nursing, Physician, Social Work, etc (to determine the extent to which counts vary within and between, or cluster among these groups. We utilized the *R* package *iccCount* to conduct group-level Concordance Correlation Coefficients (A count Poisson-based intraclass correlation coefficient measure). [21] we employed supervised machine learning methods using a multinomial naive Bayes classifier in python's scikit learn library[22] to predict categories for sentences involving provider types of quotation usage within clinical free text. Prior to supervised learning, we converted text to lowercase, and tokenized by word, bi-gram, and tri-gram count-level vectorization. Punctuation was kept to allow for detection of use of specific quoted words and inclusion of exclamation points and question marks to potentially denote scare quotes.

Results

Qualitative Results

This sample was comprised of 1200 randomized sentences, where quotes were labeled as "Probably Useful" (n=535), "Harmful" (n = 191), "Possibly Harmful" (n=78), and "Not Applicable" (n=396) in which indicated the quotation was not a direct patient quote. Coding was completed by one researcher (lead author), and thus no interrater reliability scores are available. 34 of the sentences were identified to be duplicates, wherein the coder continued to categorize each following randomly generated sentence until successfully labeling 1200.

Table 1: Characteristics of Test set of quoted notes

	N (%)
Number of quoted sentences	1200
Number of charts	1092
Number of patients	567
Number of providers	334
Number of admissions	777
Gender	
Female	222 (39.2%)
Male	343 (60.5%)
Type of Quote	
Helpful	535 (44.6%)
Harmful	191 (15.9%)

Possibly Harmful	78 (6.5%)
Not Applicable	396 (33.0%)
Type of Note	
Nursing/other	539 (44.9%)
Discharge Summary	454 (37.8%)
Physician	113 (9.4%)
Nursing	5 (.4%)
Rehab Services	2 (.2%)
Social Work	16 (1.3%)
Nutrition	3 (.3%)
Case Management	1 (.08%)
General	1 (.08%)

Regression Analyses
Following qualitative coding, we ran regression analyses to determine any impact of patient gender on counts of types of quotations. We found that male patients had a significant increase of .07 (.03,.12) harmful note instances in comparison to female patients.

Table 3. Regression test using Gender as Predictor for Quote Type

Quote Label Counts	**Regression coefficient for Gender (Male) (95% CI)**
Helpful	-.037 (-.095, .021)
Harmful	**.072 (.030, .12)****
Possibly Harmful	.007 (-.021, .036)
Not Applicable	-.043 (-.098, .012)

****p<.001**

Supervised Learning
To account for class imbalance issues, we utilized stratified k-folds sampling, finding highest model performances on 3 k-folds. This was found after testing the number of folds one by one, where performance peaked at 3 and then began to decline due to 0-size class countin some of the samples.After sampling, we took two approaches to vectorization. The first was a count-based

vectorization which created unigrams, bigrams, and trigram tokenizations of each quoted, labeled sentence in the detecting bias dataset. Due to the differential usages of quotation marks and other punctuation, we included all instances of !, ?, and “ ‘ quotation marks in the tokens. Additionally, we tested this approach against the vectorization method of tokenizing by word, bigram, and tri-gram, as well as by calculating term frequency-inverse document frequency at the word level. Hyperparameters included: number of k-folds, maximum features, and vectorization techniques. We developed code that saved file plots and model summaries according to which hyperparameters were changed. Future analyses here could benefit from grid search strategies for hyperparameter optimization. To reduce data sparsity, we limited features to the top 3000 frequent tokens. Ultimately, count-based vectorization approaches proved to show higher performance and were utilized in final models. Model class-level performance metrics are displayed in Table 4, along with a Receiver Operating Characteristics graph modeling precision and recall tradeoffs in Figure 1, and our confusion matrix in Figure 2. Our model achieved an AUC score of .55.

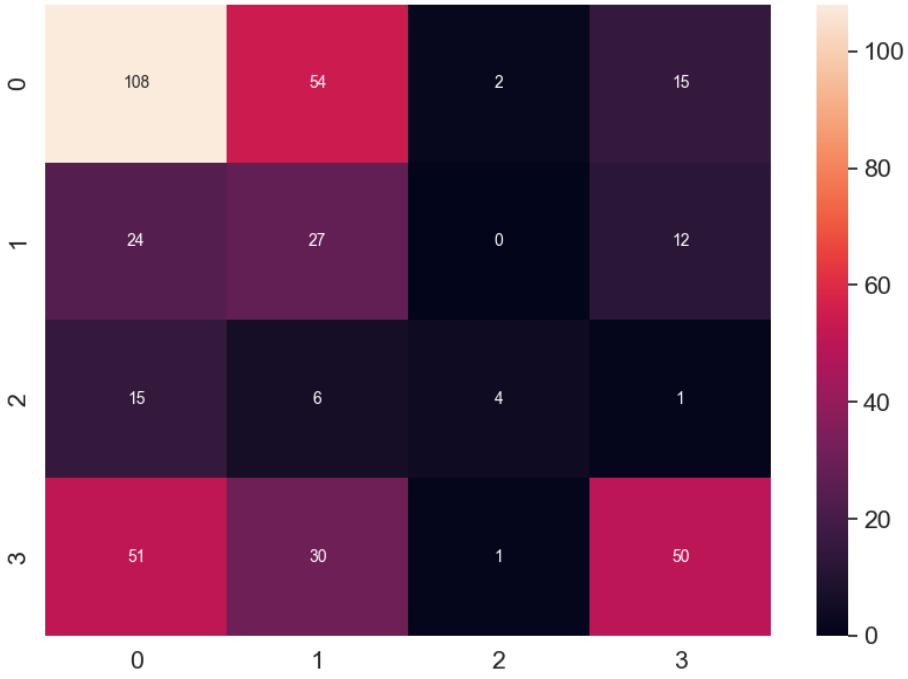
Table 4: Naive Bayes classifier accuracy

Class	**Precision**	**Recall**	**F-1 Score**	**Support**
0 “Helpful”	0.54545	0.60335	0.57294	179
1 “ Harmful”	0.23077	0.42857	0.30000	63
2 “Possibly Harmful”	0.57143	0.15385	0.24242	26
3 “Not Applicable”	0.64103	0.37879	0.47619	132
Macro Avg	**0.49717**	**0.39114**	**0.39789**	**400**
Weighted Avg	**0.52912**	**0.47250**	**0.47654**	**400**

Figure 1: ROC Curve for Naive Bayes classifier

Drew Walker
Texts as Data Final Paper

Figure 2: Confusion matrix for Naive Bayes classifier



Concordance Correlation Coefficients

We conducted CCCs on counts of different quote types grouped among patients, providers, types of notes, and gender, but found little to no evidence of clustering variances within these groups among both the test set (Table 5) and the held-out set (Table 6).

Table 5. ICCs from Test Sample for quote-type count data within Patients, Providers and Note Types

	Patients	**Providers**	**Types of Notes**	**Gender**
Helpful	<.00001	<.00001	.0056	<.00001
Harmful	<.00001	<.00001	.0002 (-.004, .005)	.007 (-.012, .026)
Possibly Harmful	.072 (-.074, .215)	<.00001	<.00001	<.00001
Not Applicable	.022 (-.009, .053)	.012 (-.037, .060)	.015 (-.0119, .041)	<.00001

Table 6: ***ICCs from Held-out Sample for quote-type count data within Patients, Providers and Note Types***

	Patients (n=809)	**Providers (n=451)**	**Types of Notes (n=9)**	**Gender**
Helpful	.02 (.002, .037)	.038 (.008, .067)	.016 (-.009, .041)	<.00001
Harmful	<.00001	.005 (-.014, .025)	.0002 (-.002,.002)	<.00001
Possibly Harmful	.050 (.015, .083)	.155 (.052, .256)	.073 (-.05, .194)	<.00001
Not Applicable	.034 (-.005, .074)	.012 (-.004, .030)	.008 (-.008, .024)	<.00001

Discussion

This project was an exploration of detecting biased use of quotation in provider clinical notes in which novel natural language processing methods were employed. Although

performance metrics were overall low for our classifier, we believe this project has laid out fundamental next-steps driving bias detection in clinical charts. Nevertheless, the project has significant limitations which we will discuss along with potential implications from the study below.

Qualitative Coding

Ultimately this study is limited in that all labels were assigned by a single coder, and thus data on inter-rater reliability scores are not available. Future investigation in this area will need multiple coders, and may benefit from incorporation of coders with previous medical training to detect subtleties in use of quotation marks. Implementation of more sophisticated duplication detection methods utilizing similarity assessments may be needed to also reduce coder burden in identifying duplicates manually. Additionally, longer-text sentence segments may need to be filtered out to aid in coding burden and assist in differentiation with bag of words models needing to incorporate more words for longer uninterrupted patient histories. We noted a specific difficulty in determining scare quotes from quotes detailing psychiatric symptoms and notes related to drug use in this sample. Psych-related notes were often aimed at demonstrating patient states of psychosis, which can be a borderline implementation of scare quotes that aim to reduce trust or credibility in patients. However, quotes describing patient state of mind may be critically relevant to the clinical team. Nevertheless, quotes in these settings run the risk of contributing to patient negative reputation development that prior studies have indicated is common on psychiatric units.[7] Additionally, quotations describing drug use often used quotation marks for slang, but also for patient responses detailing the frequency of use. In this situation, providers may quote patients as " patient has not drank in 'one year'"", which may be interpreted as scare quotes or be an honest attempt to represent patient-reported use. Regression analyses indicated differences in gender among the outcome of a number of harmful quotations.The difference observed among genders here may also be due to a higher proportion of males and thus quoted sentences within this sample. Of note, there were no nonbinary individuals indicated in this dataset, which may be a result of the database's binary gender classification system rather than any actual lack of nonbinary-identifying patients in this sample.

Supervised Learning

Overall, we achieved low performance metrics for detecting harmful types of scare quotes, with an F-1 performance of .30, indicating poor ability to correctly identify positive labels and successfully identify all positive labels. For this case, we would want the former, or a high recall level, due to a priority in identifying and preventing bias outweighing false-positive identification. It is clear at this point that increasing recall may assist in further labeling efforts as this area of research is unexplored and most data is currently unlabeled. Our model performances are higher in helpful quotes (F1=.57) and not applicable quotes (F1=.47), likely due to much higher sample size. While the priority of the study is to enhance bias detection, greater efforts in identification of not applicable quotes can be used to filter redundant text prior to future iterations of qualitative coding, which would increase coder efficiency for identifying more harmful or possibly harmful quotes. We recognize there is a lot of room for improvement in

model performance likely stemming from the need for a larger training dataset and grid-search hyperparameter optimization strategies.

ICCs in this sample were extremely low, which may be a result of measurement error here or potentially lack of count variance clustering at patient, chart, or gender levels. Ultimately, we'd want to estimate the variance components from a model allowing for multiple hierarchies of patient and provider levels to be estimated simultaneously in a cross-structured framework to be able to compare relative clustering. Additionally, results from the held-out test set are likely to be inaccurate due to the increased measurement error introduced by the classifier.

Overall, our findings represent room for model performance improvement and an important first step in bridging existing qualitative research in detecting provider biases into the advancements in NLP that have the potential to enable provider bias detection on a broader scale.

References

1. Institute of Medicine. *Unequal Treatment: Confronting Racial and Ethnic Disparities in Health Care*. (Smedley BD, Stith AY, Nelson AR, eds.). The National Academies Press; 2003. doi:10.17226/10260
2. Narayan MC. CE: Addressing Implicit Bias in Nursing: A Review. *AJN The American Journal of Nursing*. 2019;119(7):36-43. doi:10.1097/01.NAJ.0000569340.27659.5a
3. FitzGerald C, Hurst S. Implicit bias in healthcare professionals: a systematic review. *BMC Medical Ethics*. 2017;18(1):19. doi:10.1186/s12910-017-0179-8
4. Maina IW, Belton TD, Ginzberg S, Singh A, Johnson TJ. A decade of studying implicit racial/ethnic bias in healthcare providers using the implicit association test. *Social Science & Medicine*. 2018;199:219-229. doi:10.1016/j.socscimed.2017.05.009
5. Cho S, Crenshaw KW, McCall L. Toward a Field of Intersectionality Studies: Theory, Applications, and Praxis. *Signs: Journal of Women in Culture and Society*. 2013;38(4):785-810. doi:10.1086/669608
6. Hafferty FW, Franks R. The hidden curriculum, ethics teaching, and the structure of medical education. *Academic Medicine*. 1994;69(11):861-871.
7. Fontana AF. Patient Reputations: Manipulator, Helper, and Model. *Archives of General Psychiatry*. 1971;25(1):88-93. doi:10.1001/archpsyc.1971.01750130090011
8. P. Goddu A, O'Conor KJ, Lanzkron S, et al. Do Words Matter? Stigmatizing Language and the Transmission of Bias in the Medical Record. *J GEN INTERN MED*. 2018;33(5):685-691. doi:10.1007/s11606-017-4289-2
9. Park J, Saha S, Chee B, Taylor J, Beach MC. Physician Use of Stigmatizing Language in Patient Medical Records. *JAMA Network Open*. 2021;4(7):e2117052-e2117052. doi:10.1001/jamanetworkopen.2021.17052
10. Beach MC, Saha S, Park J, et al. Testimonial Injustice: Linguistic Bias in the Medical Records of Black Patients and Women. *J GEN INTERN MED*. Published online March 22, 2021. doi:10.1007/s11606-021-06682-z
11. Martin K, Ricciardelli R, Dror I. How forensic mental health nurses' perspectives of their patients can bias healthcare: A qualitative review of nursing documentation. *Journal of Clinical Nursing*. 2020;29(13-14):2482-2494. doi:10.1111/jocn.15264
12. Ashford RD, Brown AM, Curtis B. The Language of Substance Use and Recovery: Novel Use of the Go/No–Go Association Task to Measure Implicit Bias. *Health Communication*.

2019;34(11):1296-1302. doi:10.1080/10410236.2018.1481709
13. Dash S, Shakyawar SK, Sharma M, Kaushik S. Big data in healthcare: management, analysis and future prospects. *Journal of Big Data*. 2019;6(1):54. doi:10.1186/s40537-019-0217-0
14. [1904.03323] Publicly Available Clinical BERT Embeddings. Accessed September 22, 2021. https://arxiv.org/abs/1904.03323
15. Recasens M, Danescu-Niculescu-Mizil C, Jurafsky D. Linguistic models for analyzing and detecting biased language. In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. ; 2013:1650-1659.
16. Zhang H, Lu AX, Abdalla M, McDermott M, Ghassemi M. Hurtful words: quantifying biases in clinical contextual word embeddings. In: *Proceedings of the ACM Conference on Health, Inference, and Learning*. CHIL '20. Association for Computing Machinery; 2020:110-120. doi:10.1145/3368555.3384448
17. Carr DB. Patients with Pain Need Less Stigma, Not More. *Pain Medicine*. 2016;17(8):1391-1393. doi:10.1093/pm/pnw158
18. Shah S, Diwan S. Methadone: does stigma play a role as a barrier to treatment of chronic pain. *Pain Physician*. 2010;13(3):289-293.
19. Jackson-Best F, Edwards N. Stigma and intersectionality: a systematic review of systematic reviews across HIV/AIDS, mental illness, and physical disability. *BMC Public Health*. 2018;18(1):919. doi:10.1186/s12889-018-5861-3
20. Cunningham CO. Opioids and HIV Infection: From Pain Management to Addiction Treatment. *Top Antivir Med*. 2018;25(4):143-146.
21. Carrasco JL. A Generalized Concordance Correlation Coefficient Based on the Variance Components Generalized Linear Mixed Models for Overdispersed Count Data. *Biometrics*. 2010;66(3):897-904. doi:10.1111/j.1541-0420.2009.01335.x
22. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*. 2011;12:2825-2830.

Drew Walker
Texts as Data Final Paper

Appendix 1: Scare quotes in medical charts coding ontlology

Following the identification of sample size, inappropriate examples will be removed after assessment of results, and then a random sample of the quote-containing chart segments (sentence-level) will be subsequently qualitatively coded as Probably Useful, Probably Harmful, or Potentially Misinterpreted labeling derived from Park et al.'s 2021 study.[9]

Quote_use CODES:

Probably Useful (CODE AS 0)

- **Clinical Info**, **Effect on Life**, **Values or Preferences**
- Provides important contextual clues for **clinical info**
    - Examples:
        - Chest pain that "feels like an elephant is on my chest"
        - Reported that "this is the worst headache I've had in my life"
- Conveys **effect of illness on patient's life**
    - Examples:
        - Has a persistent low mood, endorsing "I don't want to live like this."
- Conveys Patient **Values or Preferences (PVP)**
    - Examples:
        - When discussing treatment goals, she said "if I cannot breathe without a tube, I don't want to live. I do not want to suffer. I want to make sure that my family are with me at the end."

Drew Walker
Texts as Data Final Paper

Probably Harmful (CODE AS 1)

- Cast **doubt on** integrity of patient to provide **reliable testimony**
    - Examples:
        - Reports she had a “reaction” to the medication.
        - Stated “migraine” was due to “stress”
- **Convey ridicule, contempt, or frustration** by highlighting **unsophisticated language** or **limited knowledge**
    - Examples:
        - Does not believe he has prostate cancer because “his bowels are working fine”

Potentially Misinterpreted (CODE AS 2)

- Neutral phrases where quotes serve no clear purpose but could be read as scare quotes conveying doubt or judgment.
- On the fence use of quotes
    - Examples:
        - She reports she has been off of cigarettes for “a year”
        - She states her living situation is “less than ideal”

Isn’t quoting the Patient (Code as 3)

- I.e. “Respiratory stated patient was ‘cooperative’ “
- These are times where quotes are used to describe words from other providers or family members, not the patient’s own words.
- Patient answers “Yes or No” questions
- Per PT: “patient is ambulatory”

not_applicable CODES:

Unsure (code as 2)

- If not sure which of the 3 categories the quote falls under. Please provide comments as to why in the Third column.

Is a repeat of another sentence (code as 3)

- When the quote of a patient is the exact same or wording seems copied from another chart verbatim or near-verbatim. This may indicate multiple

Nonsense: 4

THIRD COLUMN: Comments

- **Free-text**
    - If there are any issues where the text is difficult to understand or uninterpretable (like just a bunch of random characters) or if you cannot categorize based on the

ontology. This is very helpful for me to know, but try to code this section only if it can absolutely not be coded one way or the other.

Notes for me:

- Make sure that annotation columns fit new ontology
- Make sure columns are formatted correctly
- Format as table
- Chop at number of occurrences

First round Gold Standard coding notes

- There were several charts (~12) out of the x number found that were just pasted into the csv file where it seemed like each line of the note was its own row in the csv file. May be an error in the preprocessing or a weird character causing this
- I had to go manually and delete the ~12 notes i found that were chopping up the dataframe in this way.
    - Dataset: gold_standard_bias_annotation_doc.xlsx
    - I should create new excel column Rand() and sort by the random number

Second round Gold Standard coding notes

- Managing situations where patient is violent or threatening
- Quoting to demonstrate psychosis
- Where do family histories come into play? May be just indicating hearsay of patient on someone else's condition, but may also be ridiculing or used to highlight inaccuracies in testimony
- Longer text entries may be making classification difficult-- may want to restrict range to certain number of words around quotations to account for the long sentences we get sometimes using spaCy
- Patient shakes head/indicates "Yes" or "No" coded as 0