Towards Fairness in Classifying Medical Conversations into SOAP Sections

Elisa Ferracane,[1] **Sandeep Konam** [1]

[1]Abridge AI Inc.
elisa@abridge.com, san@abridge.com

Abstract

As machine learning algorithms are more widely deployed in healthcare, the question of algorithmic fairness becomes more critical to examine. Our work seeks to identify and understand disparities in a deployed model that classifies doctor-patient conversations into sections of a medical SOAP note. We employ several metrics to measure disparities in the classifier performance, and find small differences in a portion of the disadvantaged groups. A deeper analysis of the language in these conversations and further stratifying the groups suggests these differences are related to and often attributable to the type of medical appointment (e.g., psychiatric vs. internist). Our findings stress the importance of understanding the disparities that may exist in the data itself and how that affects a model's ability to equally distribute benefits.

Introduction

Disparities in healthcare in the U.S. have existed long before machine learning algorithms were introduced. An early seminal report from the Institute of Medicine (IOM) concludes racial and ethnic minorities receive poorer healthcare (Smedley et al. 2003), a finding still observed in recent work and extending to other disadvantaged groups (Jusot and Tubeuf 2019; Williams, Lawrence, and Davis 2019).

Applying machine learning solutions to healthcare problems holds great promise, but also poses challenges to understand disparities that it may introduce or amplify (Chen et al. 2020). For example, Obermeyer et al. (2019) examine an algorithm which identifies high-need patients in order to allocate more medical resources for them. The algorithm uses cost of care as a proxy for high-need, which the authors show is a poor and unfair construct since Black patients with the same cost of care (as Whites) have greater health needs but are being denied the benefit of the program.

Most machine learning solutions in healthcare are based on metrics for medical interventions and outcomes often sourced from electronic health records (Chen et al. 2020). Considerably less work makes use of doctor-patient conversations, even though these are known to play an important role in patient health (Ong et al. 1995). Our work is the first to focus on the algorithmic disparities of a classifier that operates over doctor-patient conversations.

We examine a classifier that categorizes utterances of a medical conversation into different sections of a SOAP note (see example Figure 1 in the Appendix). SOAP is an acronym for Subjective, Objective, Assessment, and Plan, referring to the four major sections of the problem-oriented medical note. A SOAP note is produced by a medical provider in order to summarize an encounter with a patient. The Subjective section includes information reported by the patient (such as past illnesses, current symptoms), the Objective consists of what the provider measures and observes (patient's temperature, an EKG), the Assessment is the provider's diagnosis of the patient (e.g., recovering well), and the Plan is the provider's plan for treating the problem (which can include drugs and therapeutics or further tests and appointments). In addition to the above 4 classes, we consider 'None' as the fifth class to denote utterances not relevant to the SOAP note, such as chit-chat. Output of this classifier can improve recall and understanding of care plans for patients (Schloss and Konam 2020). The Plan section is the most important of all sections from a patient adherence standpoint, as it can enable patients to better follow through on their care (e.g., to fill out a new prescription, or schedule a new appointment).

Motivated by concerns of fairness for the distribution of this patient benefit, we analyze the performance characteristics of the SOAP note classifier in order to detect possible disparities towards disadvantaged groups. Guided by the IOM report (Smedley et al. 2003) and recent healthcare studies on unequal treatment, we examine 7 protected attributes and define 18 disadvantaged groups (including intersectional ones that cross multiple protected attributes).

Using three carefully selected metrics most appropriate for our scenario, our findings show small but statistically significant disparate outcomes for fewer than half of the disadvantaged groups examined, as measured by one of the metrics. Probing deeper into these groups, we find differences in the language used in their conversations and what kind of medical provider these groups are visiting (e.g., psychiatrist vs. internist). These results suggest the disparate outcomes often stem from the type of medical visit, and underline the importance of understanding how disadvantaged groups differ in their use of medical care from the general population (Collins, Hall, and Neuhaus 1999; Hegarty et al. 2000).

Metric	Formula	Thresholds
AOD: average odds difference	$\frac{(FPR_{\text{disadv}} - FPR_{\text{adv}}) + (TPR_{\text{disadv}} - TPR_{\text{adv}})}{2}$	equal: 0, <-0.1: higher benefit for advantaged group, >0.1: higher benefit for disadvantaged group
EOR: equal opportunity ratio	$\frac{TPR_{\text{disadv}}}{TPR_{\text{adv}}}$	equal: 1, <0.8: higher benefit for advantaged group, > 1.25: higher benefit for disadvantaged group
FORR: false omission rate ratio	$\frac{FOR_{\text{adv}}}{FOR_{\text{disadv}}}, FOR = \frac{FN}{FN+TN}$	equal: 1, <0.8: higher benefit for advantaged group, > 1.25: higher benefit for disadvantaged group

Table 1: Group disparity metrics. FPR=false positive rate; TPR=true positive rate; FOR=false omission rate.

SOAP section	# utterances (%)	in conv	F1
Subjective	190,914 (16.6%)	4,975	51.3
Objective	23,741 (2.1%)	2,962	38.7
Assessment	175,326 (15.3%)	4,830	31.6
Plan	38,130 (3.3%)	3,999	24.5
none	720,478 (62.7%)	4,992	27.7
Total	1,148,589	5,000	37.6

Table 2: Statistics and classifier performance on the validation set.

Data and Classifier

The dataset consists of 63,000 doctor-patient conversations that have been collected with full consent, de-identified, and transcribed by humans. The data is split into 52,000/5,000/6,000 for train/validation/test. Our work focuses exclusively on the validation set, summarized in Table 2. The dataset includes sociodemographic data about the patient, the physician, and the location of the medical facility, which we detail in the following section.[1]

We use a classifier based on Schloss and Konam (2020) which classifies each utterance into one of five classes in accordance with the SOAP note: Subjective, Objective, Assessment, Plan, or none of these. The classifier encodes each utterance using ELMo word embeddings (Peters et al. 2018), attention and a stacked bi-LSTM. A decoder LSTM then predicts the class for the utterance.

Evaluation of Group Disparities

Group disparities exist when outcomes differ systematically between population groups (Barocas, Hardt, and Narayanan 2019). A group disparity could lead to allocative harm, by withholding benefit, or representational harm, by reinforcing the subordination of a group (Crawford 2017). The SOAP classifier provides potential benefits to patients by highlighting important parts of their conversation (usually utterances marked as Plan) in order to better follow through on their care. With an eye towards fairness, our work focuses on possible allocative harm: whether these benefits are withheld from disadvantaged groups.

We measure disparities between groups under the (naïve) assumption that the dataset distribution reflects the true pop-

---

ulation distribution. We recognize this is likely not the case and, further, that biases have likely been introduced earlier in the dataset creation process, starting from our choice of variable (examining medical conversations disfavors groups with fewer medical visits stemming from cultural or economic reasons, such as distrust of doctors or risk of job loss). We choose to focus on the disparities of the classifier as a starting point because we have control over the model. In future work, we intend to examine other sources of bias.

We view our classification task as assistive and we thus choose three metrics focused on true positives and false negatives (Table 1): average odds difference (AOD), equal opportunity ratio (EOR), false omission rate ratio (FORR). For the last two metrics, we compute the ratio between the two groups (as opposed to the difference) in order to better capture smaller differences. We focus on false omission rate instead of false negative rate because the former captures the conditional probability of the ground truth based on the predicted outcome, which is known to the decision maker, while latter is conditioned on the usually unknown ground truth (Rodolfa et al. 2020; Mitchell et al. 2018).

To judge whether the magnitude of a difference is disparate, we adopt the thresholds in Zhang, Bellamy, and Varshney (2020), which are based on legal guidelines from the U.S. Equal Employment Opportunity Commission. Statistical significance is determined using a permutation test as detailed in DiCiccio et al. (2020) which calculates the studentized metric on all possible permutations of the data split into two groups (approximated by shuffling the data N=1000 times). The null hypothesis assumes the two populations are equal, which we reject at $\alpha < 0.01$.

The groups we choose to focus on are based on recommendations from the IOM report and prior work showing bias against these groups specifically in the medical field. See Table 4 in the Appendix for coverage of the metadata (not every conversation has all the attributes studied) and detailed group statistics. We next explicitly define the groups within each protected attribute, and give a brief overview of medical bias against the groups.

Race/Ethnicity The dataset includes a single field for both race and ethnicity. We recognize this protected attribute conflates two notions which are both ill-defined and subject to reporting bias (Kressin et al. 2003). We define one advantaged group ('White') and three disadvantaged groups: 'Black', 'Hispanic', 'Asian'.

Historically, Black patients are less likely to be recom-

mended for operations, less likely to receive treatments and more likely to die (Ford et al. 1989; Lafata et al. 2001; Howard, Penchansky, and Brown 1998). At the same time, algorithmic bias against Black patients has surfaced by both causing harm (language models recommending "prison" instead of "hospitals" for violent Black patients; Zhang et al. 2020) and by denying benefit (excluding Black patients from a beneficial health program by using health costs as a discriminatory proxy for health needs; Obermeyer et al. 2019).

Hispanic and Asian patients face many of the same medical disadvantages as Blacks (Ramsey et al. 1997; Hannan et al. 1999; Carlisle, Leake, and Shapiro 1995; Thamer et al. 2001), in addition to linguistic barriers for patients with limited or low English proficiency and the lowest insurance coverage among minorities often as a result of immigration status (Smedley et al. 2003). Algorithms that predict outcomes such as disease susceptibility using electronic health record data fail to account for these confounding factors (Gianfrancesco et al. 2018). Furthermore, word embeddings used pervasively in NLP-based healthcare algorithms are shown to capture ethnic stereotypes and biases (Garg et al. 2018).

Gender The reported gender in the dataset is limited to 'male' or 'female', and does not include transgender or noncisgender choices. We define the advantaged group as 'male' and the disadvantaged as 'female'. Prior work shows women experience longer delays in diagnosis (Lyratzopoulos et al. 2013); doctors take the pain of women less seriously, often attributing it to psychogenic causes (Samulowitz et al. 2019). Medical-domain word embeddings, which can be helpful for medical NLP tasks, nevertheless also reflect these gender biases (Rios, Joshi, and Shin 2020).

Socio-economic status The dataset includes socioeconomic information about the patient: employment status (the advantaged group is 'full-time job', disadvantaged are 'unemployed' and 'retired'), living circumstance (advantaged is 'living at home', disadvantaged are 'nursing home' and 'incarcerated'), and medical insurance (advantaged is 'private insurance', disadvantaged are 'Medicaid', 'uninsured'). Lack of employment and insurance are associated with poorer healthcare, and are more prevalent among ethnic and racial minorities (Smedley et al. 2003). Living in a nursing home has shown mixed results for patient health. Studies show improved medical quality of life and social support for those in nursing homes, but also increased mortality risk when compared to living at home (Stevens, Mahoney, and Ehrenreich 2014). Patients who are incarcerated are more likely to receive fewer treatment options and overall lower quality of healthcare (Fuller and Eves 2017).

Age The dataset includes patient age, ranging from 0-98 years old. We bin the ages into three groups: youth (0-17), adult (18-64), older adult (65-98). The advantaged group is 'adult', and disadvantaged are 'youth' and 'older adult'. Less money is invested in children's health despite often poorer outcomes (Leatherman and McCarthy 2004), while older adults are less likely to receive kidney transplants and treatment for pain (Kjellstrand 1988; Bernabei et al. 1998).

Obesity The dataset reports the weight of the patient, which we use as a crude proxy for obesity ($>=250$lbs. is the disadvantaged obese group). Studies show doctors often have explicit and implicit biases against overweight patients, and the patients experience poorer care and worse outcomes (Sabin, Marini, and Nosek 2012; Tomiyama et al. 2018).

Mental health The dataset reports the physician's specialty, which we use as a proxy for patients with mental health issues. The disadvantaged group is patients seeing psychiatrists, and the advantaged group is those seeing all other physicians. Mental illness patients are often subject to 'diagnostic overshadowing' where physical illness signs are misattributed to their psychological disorder, resulting in poorer quality of care (Thornicroft, Rose, and Kassam 2007).

Location The dataset includes the U.S. state where the medical visit occurred. A recent study uncovered many machine learning algorithms are trained on a geographically limited cohort of patients (Florida, California and New York; Kaushal, Altman, and Langlotz 2020). Because many health conditions correlate with geographic location, this imbalance could result in unfair treatment for patients from other areas. The dataset we examine is not as geographically restricted (45 of the 50 states are included), but is slightly skewed (the three states comprise a third of the data).

Intersectional We further explore groups that intersect multiple protected attributes. Specifically, we examine Black females and Hispanic females, based on prior work showing these groups experience poorer care and outcomes compared to White males (Smedley et al. 2003).

Results

We evaluate 18 groups across 5 SOAP note sections (total 90 combinations) using the metrics and thresholds in Table 1, and report results only on groups that exceed the thresholds in favor of the advantaged group (see Tables 5 and 6 in the Appendix for all results). As shown in Table 3, we find 7 cases (out of 90) with a statistically significant higher benefit for the advantaged group. Importantly, we note that in all these cases, only one of the three metrics reaches the disparate threshold. In datasets where unfairness is well-established, the algorithms exceed the disparate thresholds on multiple metrics (Bellamy et al. 2018). We next analyze each group in more detail.

Analysis

To understand whether the disparities could be related to differences in the language of the doctor-patient conversations, and whether the differences are attributable to other confounding factors, we conduct two analyses. First, we perform a conditional word frequency analysis to identify lexical cues that are strongly indicative of the class, but that are absent from the disadvantaged group conversations. Specifically, we calculate the local mutual information (LMI) between the n-grams of the conversations and the class, and hypothesize the classifier would perform more poorly if the top-scoring n-grams in the overall conversations are absent (or much less prominent) in the disadvantaged group conversations. Second, because differences in conversation are

SOAP section	Protected attribute	Disadvantaged group	Advantaged group	AOD [-0.1, 0.1]	EOR [0.8, 1.25]	FORR [0.8, 1.25]	new FORR (omitted medical providers)
Plan	Race	Asian	White	0.00	0.97	0.69	0.83 (clinical cardiologist, ophthalmologist)
	Race+Gender	Hispanic female	White male	-0.01	0.95	0.71	0.82 (allergist)
	Socioeconomic	Incarcerated	Living at home	-0.05	0.95	0.71	0.82 (inf. disease specialist)
Objective	Socioeconomic	Uninsured	Private insur.	-0.02	0.94	0.67	1.71 (inf. disease specialist)
	Age	Older adult	Adult	0.00	0.99	0.69	0.75 (ophthalmologist)
None	Race	Black	White	-0.04	0.78	1.01	-
	Race+Gender	Black female	White male	-0.04	0.76	1.01	-

Table 3: Groups with higher benefit for the advantaged group (as measured by at least one metric). Numbers in green are within the thresholds, red are disparate favoring the advantaged group, blue favoring the disadvantaged group. The new FORR is recalculated after omitting the listed medical provider type. AOD=average odds difference, EOR=equal opportunity ration, FORR=false omission rate ratio.

expected and warranted in different types of medical visits, we experiment with omitting visit types from the group and observe the effect on the classifier error rates.

We first discuss disparities for the 'Plan' section, which provides the greatest benefit to the patient. In the Asian group with more false omissions, the LMI analysis shows 'blood' and 'blood work' are strong lexical cues for the class, but are less prominent in the Asian conversations, suggesting a different distribution of medical visit types. Indeed, visits to the clinical cardiologist and ophthalmologist are more frequent in the Asian group, though they comprise a small proportion (5.3% of Asian 'Plan' utterances vs. 3.6% of White, and 1.0% vs. 0.2%, respectively). If we omit these cases from the Asian group, the FORR no longer exceeds the thresholds, as seen in the last column of Table 3. Furthermore, when comparing these two specialties to all other specialties across the entire population, we find these two are inherently and significantly harder for the model to classify, regardless of race or ethnicity.

For the Hispanic female group with more false omissions, the LMI analysis shows results similar to the Asian group: lexical cues in the overall conversations are less prominent in the Hispanic ones (n-grams 'blood' and 'blood work'). We again explore the type of medical appointment and find omitting the more common visits to the allergist (5.1% vs. 2.3% in White) eliminates the observed disparities. However, we find the allergist visits are harder to classify only within the Hispanic group, and not for the entire population.

Incarcerated patients experience more false omissions, and the LMI analysis suggests a wider gap in the nature of the conversations: in addition to 'blood' and 'blood work' attested in other groups, other n-grams suggestive of future appointments are also less prominent (e.g., 'next', 'months', ''re going',). Omitting the slightly more common visits to the infectious disease specialist (5.7% vs. 1.9% for patients living at home) eliminates the disparities. However, we do not observe this specialty is harder for the model to classify across the two living condition groups.

We next explore errors in the 'Objective' section, although we note there are fewer benefits and thus lower harm from potential disparities. In the uninsured patient group with more false omissions, the LMI analysis reveals results similar to the Asian group: 'blood' is a strong lexical clue but is less prominent in the uninsured group. We again find visits to the infectious disease specialist are more common for uninsured vs. privately insured patients (6.3% vs. 1.7%), and omitting these cases erases the disparities. We also observe these types of visits are inherently harder for the model to classify, regardless of insurance status.

The older patient group with more false omissions has a more ambiguous LMI analysis, where 'liver' is less important. Omitting the internist and ophthalmologist visits mitigates but does not eliminate the disparities, and we do not find these visits to be inherently harder across all ages.

For the 'none' section, we posit little to no benefit for the patient when utterances are correctly categorized as chit-chat. We find fewer true positives for Black and Black female groups, with no differences in the LMI analysis, showing lexical cues are similar across groups. Because there is no clear harm, we do not pursue these differences.

In summary, our analysis finds considerable variation in the conversations of the different disadvantaged groups. These differences are often related to the types of medical visits and sometimes the type of visit is the true factor underlying the group disparities.

Conclusion

Motivated by concerns for fairness, we analyze the performance characteristics of a classifier that categorizes doctor-patient conversations in order to identify and understand disparate outcomes for disadvantaged groups. Our results show small but statistically significant differences for a portion of the groups, though only as measured by one of the three metrics. We further analyze the conversation within these disadvantaged groups to find variations that can be traced back to different types of medical visits. The type of visit can sometimes entirely account for the attested disparities. This finding highlights the importance of understanding the differences already present in datasets, and how these can affect a model's ability to allocate equal benefit. In future work, we aim to understand the longer term effects of disparities, as in the simulated studies of D'Amour et al. (2020).

References

Barocas, S.; Hardt, M.; and Narayanan, A. 2019. *Fairness and Machine Learning*. fairmlbook.org. http://www.fairmlbook.org.

Bellamy, R. K.; Dey, K.; Hind, M.; Hoffman, S. C.; Houde, S.; Kannan, K.; Lohia, P.; Martino, J.; Mehta, S.; Mojsilovic, A.; et al. 2018. AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *arXiv preprint arXiv:1810.01943* .

Bernabei, R.; Gambassi, G.; Lapane, K.; Landi, F.; Gatsonis, C.; Dunlop, R.; Lipsitz, L.; Steel, K.; Mor, V.; Group, S. S.; et al. 1998. Management of pain in elderly patients with cancer. *Jama* 279(23): 1877–1882.

Carlisle, D. M.; Leake, B. D.; and Shapiro, M. F. 1995. Racial and ethnic differences in the use of invasive cardiac procedures among cardiac patients in Los Angeles County, 1986 through 1988. *American Journal of Public Health* 85(3): 352–356.

Chen, I. Y.; Pierson, E.; Rose, S.; Joshi, S.; Ferryman, K.; and Ghassemi, M. 2020. Ethical machine learning in health care. *arXiv e-prints* arXiv–2009.

Collins, K. S.; Hall, A. G.; and Neuhaus, C. 1999. US minority health: A chartbook. *New York: The Commonwealth Fund* .

Crawford, K. 2017. The trouble with bias. In *Conference on Neural Information Processing Systems, invited speaker*.

D'Amour, A.; Srinivasan, H.; Atwood, J.; Baljekar, P.; Sculley, D.; and Halpern, Y. 2020. Fairness is not static: deeper understanding of long term fairness via simulation studies. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 525–534.

DiCiccio, C.; Vasudevan, S.; Basu, K.; Kenthapadi, K.; and Agarwal, D. 2020. Evaluating fairness using permutation tests. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 1467–1477.

Ford, E.; Cooper, R.; Castaner, A.; Simmons, B.; and Mar, M. 1989. Coronary arteriography and coronary bypass survey among whites and other racial groups relative to hospital-based incidence rates for coronary artery disease: findings from NHDS. *American Journal of Public Health* 79(4): 437–440.

Fuller, L.; and Eves, M. M. 2017. Incarcerated Patients and Equitability: The Ethical Obligation to Treat Them Differently. *The Journal of Clinical Ethics* 28(4): 308.

Garg, N.; Schiebinger, L.; Jurafsky, D.; and Zou, J. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences* 115(16): E3635–E3644.

Gianfrancesco, M. A.; Tamang, S.; Yazdany, J.; and Schmajuk, G. 2018. Potential biases in machine learning algorithms using electronic health record data. *JAMA internal medicine* 178(11): 1544–1547.

Hannan, E. L.; Van Ryn, M.; Burke, J.; Stone, D.; Kumar, D.; Arani, D.; Pierce, W.; Rafii, S.; Sanborn, T. A.; Sharma, S.; et al. 1999. Access to coronary artery bypass surgery by race/ethnicity and gender among patients who are appropriate for surgery. *Medical care* 68–77.

Hegarty, V.; Burchett, B. M.; Gold, D. T.; and Cohen, H. J. 2000. Racial differences in use of cancer prevention services among older Americans. *Journal of the American Geriatrics Society* 48(7): 735–740.

Howard, D. L.; Penchansky, R.; and Brown, M. B. 1998. Disaggregating the effects of race on breast cancer survival. *FAMILY MEDICINE-KANSAS CITY-* 30: 228–235.

Jusot, F.; and Tubeuf, S. 2019. Equality of opportunity in health and healthcare. In *Oxford Research Encyclopedia of Economics and Finance*. Oxford University Press. doi:10.1093/acrefore/9780190625979.013.3.

Kaushal, A.; Altman, R.; and Langlotz, C. 2020. Geographic Distribution of US Cohorts Used to Train Deep Learning Algorithms. *JAMA* 324(12): 1212–1213.

Kjellstrand, C. M. 1988. Age, Sex, and Race Inequality in Renal Transplantation. *Archives of Internal Medicine* 148(6): 1305–1309. ISSN 0003-9926. doi:10.1001/archinte.1988.00380060069016. URL https://doi.org/10.1001/archinte.1988.00380060069016.

Kressin, N. R.; Chang, B.-H.; Hendricks, A.; and Kazis, L. E. 2003. Agreement between administrative data and patients' self-reports of race/ethnicity. *American journal of public health* 93(10): 1734–1739.

Lafata, J. E.; Johnson, C. C.; Ben-Menachem, T.; and Morlock, R. J. 2001. Sociodemographic differences in the receipt of colorectal cancer surveillance care following treatment with curative intent. *Medical care* 361–372.

Leatherman, S. T.; and McCarthy, D. 2004. *Quality of health care for children and adolescents: a chartbook*. Commonwealth Fund New York, NY.

Lyratzopoulos, G.; Abel, G. A.; McPhail, S.; Neal, R. D.; and Rubin, G. P. 2013. Gender inequalities in the promptness of diagnosis of bladder and renal cancer after symptomatic presentation: evidence from secondary analysis of an English primary care audit survey. *BMJ open* 3(6).

Mitchell, S.; Potash, E.; Barocas, S.; D'Amour, A.; and Lum, K. 2018. Prediction-based decisions and fairness: A catalogue of choices, assumptions, and definitions. *arXiv preprint arXiv:1811.07867* .

Obermeyer, Z.; Powers, B.; Vogeli, C.; and Mullainathan, S. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366(6464): 447–453.

Ong, L. M.; De Haes, J. C.; Hoos, A. M.; and Lammes, F. B. 1995. Doctor-patient communication: a review of the literature. *Social science & medicine* 40(7): 903–918.

Peters, M.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; and Zettlemoyer, L. 2018. Deep Contextualized Word Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2227–2237. New Orleans, Louisiana: Association for Computational Linguistics. doi:10.18653/v1/N18-1202. URL https://www.aclweb.org/anthology/N18-1202.

Ramsey, D. J.; Goff, D. C.; Wear, M. L.; Labarthe, D. R.; and Nichaman, M. Z. 1997. Sex and ethnic differences in use of myocardial revascularization procedures in Mexican Americans and non-Hispanic whites: the Corpus Christi Heart Project. *Journal of clinical epidemiology* 50(5): 603–609.

Rios, A.; Joshi, R.; and Shin, H. 2020. Quantifying 60 Years of Gender Bias in Biomedical Research with Word Embeddings. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*, 1–13. Online: Association for Computational Linguistics. doi:10.18653/v1/2020.bionlp-1.1. URL https://www.aclweb.org/anthology/2020.bionlp-1.1.

Rodolfa, K. T.; Salomon, E.; Haynes, L.; Mendieta, I. H.; Larson, J.; and Ghani, R. 2020. Case study: predictive fairness to reduce misdemeanor recidivism through social service interventions. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 142–153.

Sabin, J. A.; Marini, M.; and Nosek, B. A. 2012. Implicit and explicit anti-fat bias among a large sample of medical doctors by BMI, race/ethnicity and gender. *PloS one* 7(11): e48448.

Samulowitz, A.; Gremyr, I.; Eriksson, E.; and Hensing, G. 2019. "Brave men" and "emotional women": A literature review on gendered norms towards patients with pain. *European Journal of Public Health* 29(Supplement_4): ckz185–075.

Schloss, B.; and Konam, S. 2020. Towards an Automated SOAP Note: Classifying Utterances from Medical Conversations. In *Machine Learning for Healthcare Conference*, 610–631. PMLR.

Smedley, B. D.; Stith, A. Y.; Nelson, A. R.; et al. 2003. Institute of Medicine. Unequal treatment: confronting racial and ethnic disparities in health care.

Stevens, J. A.; Mahoney, J. E.; and Ehrenreich, H. 2014. Circumstances and outcomes of falls among high risk community-dwelling older adults. *Injury epidemiology* 1(1): 5.

Thamer, M.; Hwang, W.; Fink, N. E.; Sadler, J. H.; Bass, E. B.; Levey, A. S.; Brookmeyer, R.; Powe, N. R.; et al. 2001. US NEPHROLOGISTS' ATTITUDES TOWARDS RENAL TRANSPLANTATION: RESULTS FROM A NATIONAL SURVEY. *Transplantation* 71(2): 281–288.

Thornicroft, G.; Rose, D.; and Kassam, A. 2007. Discrimination in health care against people with mental illness. *International review of psychiatry* 19(2): 113–122.

Tomiyama, A. J.; Carr, D.; Granberg, E. M.; Major, B.; Robinson, E.; Sutin, A. R.; and Brewis, A. 2018. How and why weight stigma drives the obesity 'epidemic'and harms health. *BMC medicine* 16(1): 123.

Williams, D. R.; Lawrence, J. A.; and Davis, B. A. 2019. Racism and health: evidence and needed research. *Annual review of public health* 40: 105–125.

Zhang, H.; Lu, A. X.; Abdalla, M.; McDermott, M.; and Ghassemi, M. 2020. Hurtful words: quantifying biases in clinical contextual word embeddings. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, 110–120.

Zhang, Y.; Bellamy, R.; and Varshney, K. 2020. Joint optimization of AI fairness and utility: A human-centered approach. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 400–406.

Appendix A: Example

Figure 1 illustrates an example of a fictitious doctor-patient conversation and the corresponding classification of each utterance into one of the four SOAP sections, or None.

Appendix B: Dataset

Table 4 summarizes all the protected attributes, disadvantaged and advantaged groups that were analyzed, with additional statistics describing their size relative proportions.

Appendix C: Results

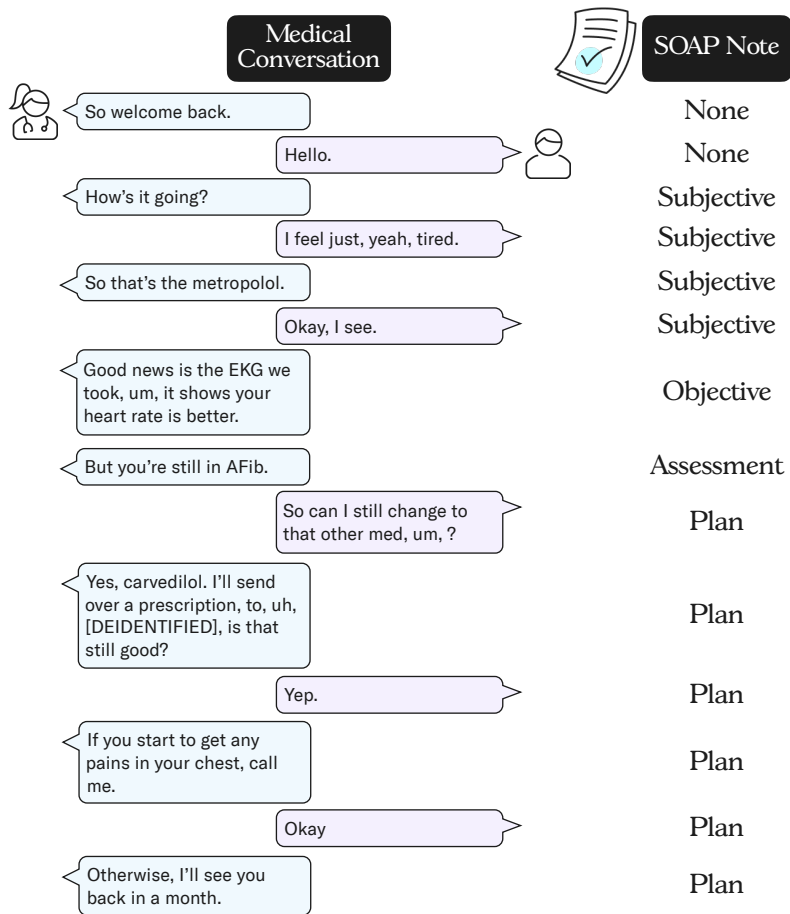Table 5 and Table 6 list results across all groups, all SOAP note sections and all metrics.

Figure 1: A fictitious doctor-patient conversation. Each utterance is classified into one of the four SOAP note sections, or None.

Protected attribute	Disadvantaged group	Advantaged group	% of Total	Size disadv/adv (%/%)	
Race/Ethnicity	Black	White	13/77	152,044/885,263	(15%/85%)
	Hispanic	White	6/77	66,729/885,263	(7%/93%)
	Asian	White	3/77	36,497/885,263	(4%/96%)
Gender	Female (Patient)	Male (Patient)	55/45	635,376/512,368	(55%/45%)
	Female (Physician)	Male (Physician)	22/78	248,366/900,223	(22%/78%)
Race/Ethn.+Gender	Black female	White male	8/35	87,779/398,528	(18%/82%)
	Hispanic female	White male	3/35	39,942/398,528	(9%/91%)
Socio-economic status	Unemployed	Full-time job	12/29	134,666/328,500	(29%/71%)
	Retired	Full-time job	35/29	398,910/328,500	(55%/45%)
	Nursing home	Living at home	2/94	21,120/1,083,642	(2%/98%)
	Incarcerated	Living at home	1/94	7,970/1,083,642	(1%/99%)
	Medicaid	Private insurance	9/15	103,260/176,920	(37%/63%)
	Uninsured	Private insurance	1/15	12,518/176,920	(7%/93%)
Age	Youth (0-17)	Adult (18-64)	4/58	40,478/660,750	(6%/94%)
	Older adult (65-98)	Adult (18-64)	38/58	431,843/660,750	(40%/60%)
Obesity	>= 250lbs.	< 250lbs.	8/84	97,576/962,080	(9%/91%)
Mental health	Psychiatrist (Physician)	Other specialty (Physician)	8/92	86,643/1,061,946	(8%/92%)
Location	Other U.S. state	FL, CA, and NY	68/32	785,731/362,858	(68%/32%)

Table 4: Protected attributes and their respective disadvantaged and advantaged groups. % of Total is the percentage each group forms of the entire validation set; these may not sum to 100% as not every conversation is tagged with every protected attribute. Size is the size of each group within this subset.

Protected attribute	Disadvantaged group	Advantaged group	SOAP section	AOD [-0.1,0.1]	EOR [0.8,1.25]	FORR [0.8,1.25]
Race/Ethn.	Black	White	Subjective	0.02	1.03	0.96
			Objective	0.01	1.03	1.10
			Assessment	-0.02	0.94	0.92
			Plan	0.02	1.02	0.91
			None	-0.04	0.78	1.01
	Hispanic	White	Subjective	0.01	1.01	0.98
			Objective	0.02	1.03	1.16
			Assessment	-0.01	0.98	0.99
			Plan	0.00	0.99	0.83
			None	-0.02	0.88	1.00
	Asian	White	Subjective	-0.03	0.93	0.85
			Objective	-0.01	0.99	0.83
			Assessment	0.00	1.03	1.06
			Plan	0.00	0.97	0.69
			None	-0.01	0.94	1.03
Gender	Female (Patient)	Male (Patient)	Subjective	0.00	1.00	0.99
			Objective	-0.01	0.99	1.16
			Assessment	0.01	1.02	0.94
			Plan	-0.01	0.98	0.98
			None	0.00	0.98	1.00
	Female (Physician)	Male (Physician)	Subjective	-0.02	0.95	1.02
			Objective	-0.01	0.97	0.97
			Assessment	0.01	1.05	1.14
			Plan	-0.01	0.98	1.02
			None	0.02	1.15	0.97
Race+Gender	Black female	White male	Subjective	0.02	1.04	0.98
			Objective	0.01	1.02	1.17
			Assessment	-0.01	0.94	0.88
			Plan	0.02	1.01	0.93
			None	-0.04	0.76	1.01
	Hispanic female	White male	Subjective	0.01	1.00	0.98
			Objective	0.02	1.06	1.66
			Assessment	0.02	1.03	0.89
			Plan	-0.01	0.95	0.71
			None	-0.03	0.82	1.02
Socio-economic	Unemployed	Full-time	Subjective	0.03	1.03	0.88
			Objective	-0.02	0.97	1.16
			Assessment	-0.01	0.97	0.96
			Plan	-0.01	1.00	1.19
			None	0.00	0.97	1.02
	Retired	Full-time	Subjective	-0.01	0.98	1.12
			Objective	0.00	1.01	1.22
			Assessment	0.06	1.13	0.95
			Plan	0.00	1.01	1.05
			None	-0.01	0.92	0.97
	Nursing home	Living at home	Subjective	0.05	1.05	0.9
			Objective	-0.01	1.00	2.35
			Assessment	-0.03	0.88	0.96
			Plan	-0.02	1.01	1.32
			None	0.02	1.03	0.98
	Incarcerated	Living at home	Subjective	-0.04	0.93	1.02
			Objective	-0.17*	0.63*	0.75*
			Assessment	0.18	1.38	0.94
			Plan	-0.05	0.95	0.71
			None	-0.04	0.81	1.00

Table 5: Group disparities metrics, organized by protected attribute and SOAP note section. Numbers in green are within the thresholds, red are disparate favoring the advantaged group, blue favoring the disadvantaged group. Numbers with an asterisk indicate the size of the group was too small to analyze (75 utterances). AOD=average odds difference, EOR=equal opportunity ratio, FORR=false omission rate ratio.

Protected attribute	Disadvantaged group	Advantaged group	SOAP section	AOD [-0.1,0.1]	EOR [0.8,1.25]	FORR [0.8,1.25]
Socio-economic	Medicaid	Private insurance	Subjective	0.04	1.08	1.02
			Objective	-0.01	0.98	1.23
			Assessment	-0.01	0.97	0.96
			Plan	0.01	1.05	1.69
			None	-0.01	0.97	1.00
	Uninsured	Private insurance	Subjective	0.04	1.07	0.86
			Objective	-0.03	0.94	0.67
			Assessment	-0.01	0.99	1.11
			Plan	-0.01	0.99	0.82
			None	-0.02	0.96	1.07
Age	Youth	Adult	Subjective	-0.02	0.91	1.01
			Objective	-0.06	0.87	1.43
			Assessment	0.00	0.99	0.97
			Plan	-0.04	0.91	1.01
			None	0.04	1.15	0.93
	Older adult	Adult	Subjective	0.00	0.98	0.88
			Objective	0.00	0.99	0.69
			Assessment	-0.06	0.89	1.09
			Plan	0.00	0.98	0.81
			None	0.02	1.11	1.03
Obesity	>=250lbs.	<250lbs.	Subjective	0.01	1.02	1.01
			Objective	0.01	1.02	1.09
			Assessment	-0.01	0.96	0.88
			Plan	0.02	1.04	1.23
			None	-0.01	0.96	1.03
Mental health	Psychiatrist	Other specialty	Subjective	-0.02	0.93	0.97
			Objective	-0.05	0.93	4.15
			Assessment	0.08	1.16	1.02
			Plan	-0.04	1.00	1.54
			None	0.06	1.30	0.97
Location	Other U.S. state	FL, CA, and NY	Subjective	-0.01	0.98	0.96
			Objective	0.00	1.00	1.17
			Assessment	0.01	1.03	1.08
			Plan	-0.01	0.99	0.98
			None	0.01	1.10	0.99

Table 6: (Continued) Group disparities metrics, organized by protected attribute and SOAP note section. Numbers in green are within the thresholds, red are disparate favoring the advantaged group, blue favoring the disadvantaged group. AOD=average odds difference, EOR=equal opportunity ratio, FORR=false omission rate ratio.