Modeling Mistrust in End-of-Life Care

Willie Boag[1] Harini Suresh[1] Leo Anthony Celi[1] Peter Szolovits[1] Marzyeh Ghassemi[1 2 3 4]

Abstract

In this work, we characterize the doctor-patient relationship using a machine learning-derived trust score. We show that this score has statistically significant racial associations, and that by modeling trust directly we find stronger disparities in care than by stratifying on race. We further demonstrate that mistrust is indicative of worse outcomes, but is only weakly associated with physiologically-created severity scores. Finally, we describe sentiment analysis experiments indicating patients with higher levels of mistrust have worse experiences and interactions with their caregivers. This work is a step towards measuring fairer machine learning in the healthcare domain.

1. Introduction

There are well-established gaps in the American healthcare system for minority populations. Groups that have been historically marginalized have also had worse treatment options and longitudinal health outcomes. Biases are especially troubling in the context of machine learning applied to clinical data. Bias can be replicated and exacerbated in the model's future recommendations (Ensign et al., 2017). For example, black and Hispanic patients are often given less pain medication for equivalent injuries and reported pain levels (Goyal et al., 2015; Singhal et al., 2016). If this pattern is present in the training data for a model built to recommend treatment, it would learn to associate race with pain medication dosage.

Differences in care have also been established during end-of-life (EOL), when critically ill patients are confronting death (Muni et al., 2011; Lee et al., 2016). Previous work has suggested that medical disparities can reflect higher levels of mistrust for the healthcare system among black patients. It is said that blacks are more suspicious of the clinical motives in advance directives and do-not-resuscitate (DNR) orders (Wunsch et al., 2010), and believe that the healthcare system was controlling which treatments they can receive (Perkins et al., 2002). When the doctor-patient relationship lacks trust, patients may believe that limiting any intensive treatment is unjustly motivated, and demand higher levels of aggressive care. While there are clinical examples of exemplary end-of-life care, studies have highlighted that aggressive care can lead to painful final moments, and may not improve patient outcomes (Cipolletta & Oprandi, 2014).

Prior works in the FATML community have attempted to mask out features that may lead to disparate treatment (Zemel et al., 2013), but including information about race may be important for some clinical tasks (e.g., if there are differences in recommended care by genetic makeup). In such a setting, quantifying bias and establishing proxy measures for medical trust is particularly important.

In this work, we present three contributions.

- We present a trust metric derived from coded doctor-patient interactions.
- We demonstrate that our trust score captures treatment differences by showing disparities in end-of-life care are pronounced when patients are stratified on trust.
- We validate our mistrust metric using sentiment analysis of patients' notes.

For further analysis – including the analysis of three different proxy scores for trust – see (Boag, 2018).

2. Background and Related Work

The quantity of health-related data is increasing rapidly, from genetic data to medical images like x-rays (Kruse et al., 2016; Raghupathi & Raghupathi, 2014). These rapid advancements have facilitated large-scale machine learning methods to guide care. Ferryman & Pitcan give an overview of fairness issues that may arise with such advances in personalized medicine. However, further research into these risks and the feasibility of applying existing FATML work to healthcare domains is limited.

Socialized mistrust of the medical community in minority groups has been established as a factor in care differences (Washington, 2007). Family members of African American patients are more likely to cite absent or problematic communication with physicians about EOL care (Hauser et al., 1997). Similarly, in surveys, African Americans report lower rates of satisfaction with the quality of care that they received by physicians (Hanchate et al., 2009). In end-of-life care, a mistrustful patient could be more resistant to a doctor's recommendation of comfort-based care, and instead insist receiving all possible treatments even if they are overly aggressive (Garrett et al., 1993; Hopp & Duffy, 2000).

Trust is difficult to quantify, and shaped by subtle interactions such as perceived discrimination, racial discordance, poor communication, language barriers, unsatisfied expectations, cultural stigmas and reputations, and more (L. Whaley, 2001). Trust is very important to success of a hospital stay; previous work has found that increased levels of doctor-patient trust were associated with stronger adherence to a physician's advice, increased patient satisfaction and improved health status (Gelb-Safran et al., 1998).

Previous efforts to create trust-based measures that correlate with outcomes have relied on surveys, which can be difficult to conduct for both theoretical (selection bias) and practical (cannot be done for retrospective, de-identified data) concerns (Lee et al., 2016).

3. Data

We use the publicly-available Medical Information Mart for Intensive Care (MIMIC-III) v1.4 (Johnson et al., 2016). This database contains de-identified EHR data from over 58,000 hospital admissions for nearly 38,600 adult patients. The data was collected from Beth Israel Deaconess Medical Center from 2001–2012. We examine a cohort of black and white patients in end-of-life care. We examine patients who have a hospital stay which lasted at least 6 hours, and have either died in the hospital, were discharged to hospice, or were discharged to a skilled nursing facility[1] (SNF). These experiments are repeated on a stricter definition of an EOL cohort (which excludes SNF patients) in (Boag, 2018), shows the same trends as this work, but with less statistical power because of smaller sample sizes. Both our data extraction and modelling code are made available[2] to enable reproducibility and further study (Johnson et al., 2017).

In order to measure disparities in aggressive end-of-life procedures, we extracted treatment durations (in minutes) from MIMIC's derived mechanical ventilation (*ventdurations*) and vasopressor (*vasopressordurations*) tables [3]. Due to the noisiness of clinical measurements,[4] we merge any treatment spans that occurred within 10 hours of each other.[5] If a patient had multiple spans, such as an intubation-extubation-reintubation, then we consider the patient's treatment duration to be the sum of the individual spans.

In this work, we wish to better understand and quantify the nuances of a patient's interactions with their nurses and doctors. We accomplish this using two sources: clinical notes and coded chart events. We obtain the notes of every patient who had a stay of at least 12 hours in the ICU. This resulted in 48,273 admissions and over 800,000 notes. Most notes are nursing notes, discharge summaries, physician notes, and social worker notes. To supplement this narrative prose, we also extract coded information from the MIMIC *chartevents* table, which records many interpersonal aspects of the patient's stay, including: code status, health literacy (e.g. whether there is a healthcare proxy), behavioral and mental status assessments, family communications, pain management, whether the patient was restrained, whether the patient wanted help bathing, support services, and more.

4. Methods and Experiments

We aim to replicate previously demonstrated racial disparities in end-of-life care using MIMIC-III (Johnson et al., 2016). We take as reference a set of three recent papers which examined the racial disparities in end-of-life care for nonwhite or minority populations (Yarnell et al., 2017; Muni et al., 2011; Lee et al., 2016). We compared the differences of patient outcomes between white and black populations using Mann-Whitney analysis for non-normally distributed variables (treatment durations, mistrust metric scores). In accordance with prior work, we consider p-values $< .05$ to be statistically significant.

4.1. Establishing a Medical Mistrust Metric

Ideally, the gold standard for measuring trust would be a survey where the patient – in their own words – describes their feelings and relationship with their caregiver (Gelb-Safran et al., 1998). However, such surveys were not recorded for MIMIC patients. But we believe that there is still a useful signal of the trust which can be inferred from the EHR. We quantify mistrust in a novel way by seeding a supervised machine learning task with labels which serve as a proxy for mistrust. Our goal is to model the underlying relationship and create a trust score which aims to explain treatment disparities better than race does.

We extract coded interpersonal features from the *chartevents* table for all MIMIC patients. Some of the information extracted includes: indication of family meetings, patient education, whether the patient needed to be restrained, how thoroughly pain is being monitored and treated, healthcare literacy (e.g. whether the patient has a healthcare proxy), whether the patient has a support system (such as family, social workers, and religion), and agitation scales (Riker-SAS and Richmond-RAS). In total, we extract 620 unique binary indicators.

We use a simple rule-based search through the notes to determine whether the patient has non-compliance documented somewhere in their notes (e.g. medical advice, regimens, follow-ups, etc.). Noncompliance indicates a very overt mistrust; rather than just holding an unspoken resentment, the patient actually defies their doctor's orders. Because crossing this line explicitly demonstrates that the patient is willing to disregard physician decisions, it is a suitable prediction target for training the model to quantify mistrust.

Of the 48,273 hospital admissions, we find 464 with notes that document noncompliance, and fit an L1-regularized Logistic Regression[6] model with chartvents features to predict whether the patient was noncompliant. Once the model is trained, we use the classifier's predicted probability for a new patient as a proxy for their degree of mistrust.

4.2. Validating the Mistrust Metric

Throughout a patient's stay, caregivers write narrative prose notes to document administered care, family meetings, patient preferences, reminders, warnings, and the patient's quality of care. In documenting their impressions of how to best understand and interact with their patients, caregivers can give clues into their relationship with the patient and family.

Clinical notes have been used for prediction tasks in previous work (Ghassemi et al., 2014) but not for investigating

[1] This was done because the notes indicate some SNF patients are discharged on hospice without coding that into the EHR.

[2] https://github.com/wboag/eol-mistrust

[3] Available freely at https://github.com/MIT-LCP/mimic-code/tree/master/concepts/durations.

[4] for instance, when one treatment span is erroneously coded as two back-to-back smaller spans

[5] This heuristic was suggested by MIMIC staff because 10 hours is approximately the shift of a nurse, and treatment duration events might get recorded once at the beginning of each shift.

[6] http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

Table 1. Top-3 most positively and negatively informative chartevent features for tuning the mistrust metric.

feature	weight
state: alert	-1.0156
riker-sas scale: agitated	0.7013
pain: none	-0.5427
richmond-ras scale: 0 alert and calm	-0.3598
education readiness: no	0.2540
pain level: 7-mod to severe	0.2168

mistrust. Sentiment analysis of clinical notes has also been used to measure whether one group of patients has a better experience, on average, than another group (McCoy et al., 2015). We use the Pattern software package for sentiment analysis (De Smedt & Daelemans, 2012).

For a given hospital admission, we compute the sentiment score of the concatenation of that patient stay's notes. Once we compute the score for each stay in our population, we scale the distribution of scores to be zero-mean and unit-variance in order to give better sense to the differences in sentiment, relative to average. Of particular interest is the differences in sentiment between different groups: white-and-black; trustful-and-mistrustful. As a sanity check, we hypothesize that sick patients have more negative stays than healthy patients, so we also stratify into low- and high-risk subpopulations using the Oxford Acute Severity of Illness Score (OASIS) score (E W Johnson et al., 2013). We create subpopulations to be same size as black/white split, i.e. the white:black dataset size ratio is the same as the trustful:mistrustful dataset size ratio.

5. Results

5.1. Creation of a Mistrust Metric

Table 1 shows the three most positively and most negatively informative weights used to predict a mistrust metric (Section 4.1). The features align well with our intuitive notion of mistrust: patients who are agitated and not receptive to education are more likely to be mistrustful, whereas calm, pain-free patients are more willing to trust their doctor.

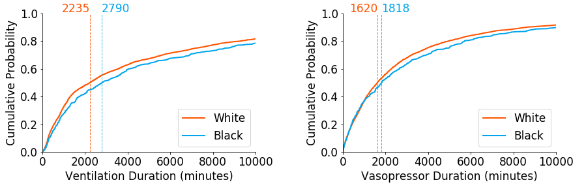
We observe a statistically significant racial disparity in the mistrust metric, where the median black patient has a higher level of mistrust than the median white patient using the Mann-Whitney test (p=0.003). This is not surprising, given the extensive literature investigating differences in iatrophobia by race (Washington, 2007).

5.2. Significant Disparities in EOL Care

5.2.1. Race-Based Disparities

We demonstrate racial treatment disparities in the MIMIC dataset. Figure 1 highlights the differences in white and black populations for aggressive treatment durations. Figures 1a and 1b show that for both mechanical ventilation and vasopressors, the median black patient receives a longer duration of treatment, perhaps suggesting a reluctance to transition to palliative care. While these results only show statistical significance for ventilation, the same trends are also observable for vasopressor administration.

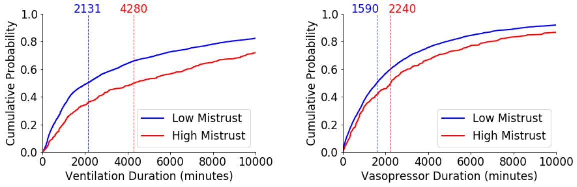
Figure 1. We observe racial disparities in for black patients (when compared to white patients) for the duration of aggressive interventions (vasopressors and ventilation). Medians are indicated by dotted lines; differences are significant ($p < 0.05$) for ventilation but not for vasopressors.



(a) CDF of ventilation duration by race ($p = .005$).

(b) CDF of vasopressor duration by race ($p = 0.12$).

Figure 2. Stratifying patients by mistrust directly shows starker disparities in care than race. Medians for ventilation and vasopressor durations are indicated with dotted lines and all differences between the groups are significant.



(a) CDF of ventilation duration ($p < .0001$).

(b) CDF of vasopressor duration ($p < .0001$).

5.2.2. Trust-Based Disparities

Using the mistrust metric, we can rank the patients by trust score and stratify them into two groups: low- and high-mistrust. Figure 2 revisits the experiments from Figure 1 except stratified into low and high mistrust instead of white and black populations.[7] We can see from Figure 2b that trust-based disparities in vasopressor durations are significant. The difference between medians of each group is 650 minutes for vasopressors (whereas the difference stratified by race was 200 minutes). This gap is even larger for ventilation durations, as shown in Figure 2a: the trust-based stratification shows a 2150-minute difference between medians, in contrast to the 550-minute gap for the race split in Figure 1a.

5.3. Low Trust Patients Have The Most Negative Notes

Table 2 shows the differences in sentiment analysis scores between race, severity of illness, and trust.[8] As a reminder, the scores were normalized to be zero-mean and unit-variance. It is interesting that every subpopulation

[7] For each treatment, we preserve the same size difference of stratified groups in order to maintain consistency in sample sizes for significance testing, e.g,. because the black group contains 510 patients for ventilation, we compare the 510 lowest trust patients against the 4811 highest trust patients.

[8] Note that a naive application of tokenization is misleadings, as even positive notes containing "Date:[**5-1-18**]" were tagged as negative because the tool's string-matching algorithm was identifying ":[" as negative emoticon use.

Table 2. Median sentiment analysis of cohorts stratified by race, severity, and trust.

population	N	median
White	9629	-0.064
Black	1164	-0.110
Low Severity	9629	-0.058
High Severity	1164	-0.167
High Trust	9629	-0.049
Low Trust	**1164**	**-0.242**

Table 3. Pairwise Pearson correlations between severity scores and mistrust score.

	OASIS	SAPS II	Mistrust
OASIS	1.0	0.680	0.095
SAPS II	0.680	1.0	0.045
Mistrust	0.095	0.045	1.0

(and indeed the full population) median score is at least slightly negative, indicating that the distribution has a positive skew.

We observed statistically significant differences in the population means ($p < .05$) for all three stratifications using the Mann-Whitney test. In particular, we see that black patients, high risk patients, and low trust patients all have stronger levels of negative sentiment in their notes. However, the low trust cohort had the most extreme negative sentiments. The median low trust sentiment (-0.242) was more than twice as far from the center as the median black sentiment (-0.110), further suggesting that the mistrust metric is able to tease out the cases with poor caregiver interactions and impressions.

5.4. Not Just Some Severity Score Proxy

One initial concern we had was that this mistrust metric might have actually been more similar to a severity score like OASIS than intended. Certainly, high-risk patients are treated differently than the general population. To dispel this concern, we compared the pairwise correlations between the mistrust score, OASIS, and SAPS II – another severity measure (Le Gall et al., 1993). Table 3 shows that the two well-established acuity scores, OASIS and SAPS II, have a strong correlation of 0.68. On the other hand, the mistrust score does not seem to simply recapitulating severity of illness, as indicated by its weak (0.095 and 0.045) correlations with the other two scores.

6. Limitations

The primary limitation of this study is that the labels for tuning the weights for the mistrust metric were generated with a simple rule-based search for the word ”noncompliant” in a patient’s clinical notes. Not only does this narrow definition of mistrust fail to capture some of the more subtle interactions in unhealthy doctor-patient relationships, but it also could falsely attribute malice to logistic issues such as being noncompliant with home medications because of a lack of access to prescriptions. In practice, however, we did not observe many false positive examples, and the mistrust metric – both in feature weights and in analysis of treatment/sentiment disparities – indicates that it is a sufficient first-attempt proxy to capture the more difficult-to-measure quantity of ”trust.”

7. Conclusion

In this work, we demonstrate that black patients receive – sometimes significantly – longer durations of invasive treatments in the MIMIC database. Though these trends have been studied in private datasets, we present our replicable analysis on a public dataset.

We create a mistrust score by using coded interpersonal features to predict patient noncompliance. This mistrust metric is a better identifier than race to show disparities in both end-of-life care and sentiment. However, this score also indicates a higher level of mistrust held by black patients than white patients.

Medical machine learning is moving forward at an exciting pace; we hope that this work will be a step towards creating models of human physiology that serve everyone, and do not propagate existing disparities in care. In order to achieve that goal, we must make better efforts to measure and understand these disparities.

References

Boag, William. Quantifying racial disparities in end-of-life care. Master’s thesis, MIT, 2018.

Cipolletta, Sabrina and Oprandi, Nadia. What is a good death? health care professionals narrations on end-of-life care. *Death studies*, 38(1):20–27, 2014.

De Smedt, T. and Daelemans, W. Pattern for python. 13: 20312035, 2012.

E W Johnson, Alistair, Kramer, Andrew, and D Clifford, Gari. A new severity of illness scale using a subset of acute physiology, age, and chronic health evaluation data elements shows comparable predictive accuracy. 41, 05 2013.

Ensign, Danielle, Friedler, Sorelle A, Neville, Scott, Scheidegger, Carlos, and Venkatasubramanian, Suresh. Runaway feedback loops in predictive policing. *arXiv preprint arXiv:1706.09847*, 2017.

Ferryman, Kadija and Pitcan, Mikaela. Fairness in precision medicine, 2018.

Garrett, Joanne Mills, Harris, Russell P., Norburn, Jean K., Patrick, Donald L., and Danis, Marion. Life-sustaining treatments during terminal illness - who wants what? *Journal of General Internal Medicine*, 8(7):361–368, 7 1993. ISSN 0884-8734. doi: 10.1007/BF02600073.

Gelb-Safran, Dana, Taira, DA, Rogers, William, Kosinski, Mark, Ware, John, and Tarlov, AR. Linking primary care performance to outcome of care. 47:213–20, 10 1998.

Ghassemi, Marzyeh, Naumann, Tristan, Doshi-Velez, Finale, Brimmer, Nicole, Joshi, Rohit, Rumshisky, Anna, and Szolovits, Peter. Unfolding physiological state: Mortality modelling in intensive care units. In *Proceedings of the 20th ACM SIGKDD international conference* *on Knowledge discovery and data mining*, pp. 75–84. ACM, 2014.

Goyal, Monika K., Kuppermann, Nathan, Cleary, Sean D., Teach, Stephen J., and Chamberlain, James M. Racial disparities in pain management of children with appendicitis in emergency departments. *JAMA Pediatrics*, 169 (11):996–1002, 2015.

Hanchate, Amresh, Kronman, Andrea C., Young-Xu, Yinong, Ash, Arlene S., and Emanuel, Ezekiel. Racial and ethnic differences in end-of-life costs: Why do minorities cost more than whites? *Archives of Internal Medicine*, 169(5):493–501, 2009.

Hauser, Joshua M., Kleefield, Sharon F., Brennan, Troyen A., and Fischbach, Ruth L. Minority populations and advance directives: Insights from a focus group methodology. *Cambridge Quarterly of Healthcare Ethics*, 6(1):58–71, 1997. ISSN 0963-1801. doi: 10.1017/S0963180100007611.

Hopp, Faith P. and Duffy, Sonia A. Racial variations in end-of-life care. *Journal of the American Geriatrics Society*, 2000.

Johnson, Alistair E. W., Pollard, Tom J., and Mark, Roger G. Reproducibility in critical care: a mortality prediction case study. In *MLHC 2018*, volume 68, pp. 361–376, Boston, Massachusetts, 18–19 Aug 2017.

Johnson, Alistair EW, Pollard, Tom J, Shen, Lu, Lehman, Li-wei H, Feng, Mengling, Ghassemi, Mohammad, Moody, Benjamin, Szolovits, Peter, Celi, Leo Anthony, and Mark, Roger G. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3, 2016.

Kruse, Clemens Scott, Goswamy, Rishi, Raval, Yesha, and Marawi, Sarah. Challenges and opportunities of big data in health care: a systematic review. *JMIR medical informatics*, 4(4), 2016.

L. Whaley, Arthur. Cultural mistrust: An important psychological construct for diagnosis and treatment of african americans. 32:555–562, 12 2001.

Le Gall, J.R., Lemeshow, S., and Saulnier, F. A new simplified acute physiology score (saps ii) based on a european/north american multicenter study. *JAMA*, 270(24): 2957–2963, 1993.

Lee, Janet J., Long, Ann C., Curtis, J. Randall, and Engelberg, Ruth A. The influence of race/ethnicity and education on family ratings of the quality of dying in the icu. *Journal of Pain and Symptom Management*, 51(1):9 – 16, 2016. ISSN 0885-3924. doi: https://doi.org/10.1016/j.jpainsymman.2015.08.008. URL http://www.sciencedirect.com/science/article/pii/S0885392415004558.

McCoy, Thomas H., Castro, Victor M., Cagan, Andrew, Roberson, Ashlee M., Kohane, Isaac S., and Perlis, Roy H. Sentiment measured in hospital discharge notes is associated with readmission and mortality risk: An electronic health record study. *PLOS ONE*, 10(8):1–10, 08 2015. doi: 10.1371/journal.pone.0136341. URL https://doi.org/10.1371/journal.pone.0136341.

Muni, Sarah, Engelberg, Ruth A., Treece, Patsy D., Dotolo, Danae, and Curtis, J. Randall. The influence of race/ethnicity and socioeconomic status on end-of-life care in the icu. *Chest*, 139(5):1025–1033, 2011. ISSN 0012-3692. doi: 10.1378/chest.10-3011.

Perkins, Henry S, Geppert, Cynthia, Gonzales, Adelita, Cortez, Josie D, and Hazuda, Helen P. Cross-cultural similarities and differences in attitudes about advance care planning. *Journal of General Internal Medicine*, 17(1):48–57, 2002.

Raghupathi, Wullianallur and Raghupathi, Viju. Big data analytics in healthcare: promise and potential. *Health information science and systems*, 2(1):3, 2014.

Singhal, Astha, Tien, Yu-Yu, and Hsia, Renee Y. Racial-ethnic disparities in opioid prescriptions at emergency department visits for conditions commonly associated with prescription drug abuse. *PLOS ONE*, 11(8):1–14, 08 2016. doi: 10.1371/journal.pone.0159224. URL https://doi.org/10.1371/journal.pone.0159224.

Washington, Harriet. *Medical Apartheid: The Dark History of Medical Experimentation on Black Americans from Colonial Times to the Present*. 2007. ISBN 978-0385509930.

Wunsch, Hannah, Guerra, Carmen, Barnato, Amber E, Angus, Derek C, Li, Guohua, and Linde-Zwirble, Walter T. Three-year outcomes for medicare beneficiaries who survive intensive care. *Jama*, 303(9):849–856, 2010.

Yarnell, Christopher J., Fu, Longdi, Manuel, Doug, Tanuseputro, Peter, Stukel, Theres, Pinto, Ruxandra, Scales, Damon C., Laupacis, Andreas, and Fowler, Robert A. Association between immigrant status and end-of-life care in ontario, canada. In *JAMA*, 2017.

Zemel, Rich, Wu, Yu, Swersky, Kevin, Pitassi, Toni, and Dwork, Cynthia. Learning fair representations. In *International Conference on Machine Learning*, pp. 325–333, 2013.