

Racial Disparities and Mistrust in End-of-Life Care

Willie Boag

MIT CSAIL

Cambridge, MA. USA

WBOAG@MIT.EDU

Harini Suresh

MIT CSAIL

Cambridge, MA. USA

HSURESH@MIT.EDU

Leo Anthony Celi, MD, MPH

MIT IMES

Cambridge, MA. USA

LCELI@BIDMC.HARVARD.EDU

Peter Szolovits, PhD

MIT CSAIL

Cambridge, MA. USA

PSZ@MIT.EDU

Marzyeh Ghassemi, PhD

University of Toronto, Vector Institute

Toronto, ON. CA

MARZYEH@CS.TORONTO.EDU

Abstract

There are established racial disparities in healthcare, including during end-of-life care, when poor communication and trust can lead to suboptimal outcomes for patients and their families. In this work, we find that racial disparities which have been reported in existing literature are also present in the MIMIC-III database. We hypothesize that one underlying cause of this disparity is due to mistrust between patient and caregivers, and we develop multiple possible trust metric proxies (using coded interpersonal variables and clinical notes) to measure this phenomenon more directly. These metrics show even stronger disparities in end-of-life care than race does, and they also tend to demonstrate statistically significant higher levels of mistrust for black patients than white ones. Finally, we demonstrate that these metrics improve performance on three clinical tasks: in-hospital mortality, discharge against medical advice (AMA) and modified care status (e.g., DNR, DNI, etc.).

1. Introduction

There are well-established gaps in the American healthcare system for minority populations: white Americans live 4.5 years longer than African Americans (Arias, 2014), infant mortality rates are twice as high for African Americans even after adjusting for socioeconomic status (Ely et al., 2017), African American men are 50% more likely to develop prostate cancer than white men and are twice as likely to die from it (CDC, 2014). Differences in care also persist during end-of-life (Muni et al., 2011; Lee et al., 2016).

Previous work has suggested that some treatment and outcome disparities are related to higher levels of mistrust: some African Americans have shown suspicion of the clinical motives in advance directives and do-not-resuscitate (DNR) orders (Wunsch et al., 2010), and some report that they suspect that the healthcare system was limiting which treatments they could receive (Perkins et al., 2002). Mistrust between patient and clinician can be detrimental to patient care, especially in end-of-life (EOL), when patients might defy physician recommendations and insist on higher levels of aggressive care. This would be especially problematic because aggressive EOL care can lead to painful final moments and

may not improve patient outcomes (Cipolletta and Oprandi, 2014). In this work, we turn to a novel source for estimating a patient and clinician’s trust relationship: clinical notes and documented interpersonal features. Prior work has established the importance of notes in prediction tasks (Ghassemi et al., 2014, 2015; Suresh et al., 2017), but not in quantifying mistrust. Further, while others have worked to establish variations in care in private datasets (Pritchard et al., 1998; Levinsky et al., 2001; Gessert et al., 2001), we target the publicly-available MIMIC dataset (Johnson et al., 2016) to emphasize reproducibility and encourage future work in this area ¹.

We provide three contributions:

- We quantify the racial disparities in EOL care in the publicly-available MIMIC dataset, and make data extraction and modeling code available for reproduction.²
- We propose multiple proxy trust scores to study the inherent mistrust that patients have in clinical staff providing their care.
- We demonstrate that our trust scores improve performance in three care-based classification tasks: in-hospital mortality, leaving the hospital against medical advice (AMA), and care status (e.g. Do Not Resuscitate).

2. Background and Related Work

2.1 Medical Treatment Gap

There are well-established biases in clinical care that follow social biases. For instance, women (Hoffmann and Tarzian, 2001) and obese patients (Phelan et al., 2015) often have worse treatment options and worse outcomes. Racial disparities have been demonstrated in many care settings; African Americans are less likely to receive pain medication than their white counterparts, even when controlling for covariates such as age, sex, and time of treatment Goyal et al. (2015); Singhal et al. (2016). Doctors are more likely to diagnose African Americans with more severe disorders (e.g., schizophrenia and other psychotic disorders), yet less likely to diagnose them with depression Simon et al. (1973); Adebimpe (1981); Ruiz (1983). Biases are especially troubling when applying machine learning because the model might identify and exacerbate biases in a feedback loop (Ensign et al., 2017). However, some select clinical tasks do benefit from knowing the patient’s race (e.g. when there are differences in recommended care by genetic makeup). In such settings, race should not be ignored, but algorithms must take steps to reduce unnecessary bias.

2.2 End-of-Life Care Differences

During EOL care, minorities are more likely to receive high-intensity, life-sustaining treatments (Pritchard et al., 1998; Levinsky et al., 2001; Gessert et al., 2001) and have fewer advance directives³ (Smith et al., 2008). White patients are more likely to utilize hospice care and are less likely to disenroll in it than non-white patients (Garrett et al., 1993; Hopp and Duffy, 2000). While some of these differences may be attributed to cultural preferences, many issues can also be the result of poor communication or unclear expectations. These

¹We note that further analysis is available in Boag (2018).

²<https://github.com/wboag/eol-mistrust>

³An advance directive is a legal document that can help ensure patient preferences for various medical treatments are followed if they become incapable of making healthcare decisions.

imbalances are potentially harmful because aggressive care does not always lead to improved patient outcomes (Cipolletta and Oprandi, 2014).

2.3 Medical Mistrust Among Minority Communities

Recent work has explored the multi-faceted history of mistrust between the African Americans community and medical institutions (Washington, 2007). Poor trust can specifically impact end-of-life care; 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, surveyed African Americans report lower rates of satisfaction with received quality of care (Hanchate et al., 2009). Mistrust can potentially help understand racial disparities in aggressive treatments such as mechanical ventilation and vasopressors. When further invasive procedures are unlikely to succeed or return the patient to a normal lifestyle, doctors may recommend withdrawing treatment and transitioning to comfort-based measures to ensure the patient does not suffer. However, mistrust may lead a patient or healthcare proxy to question the intention of the assessment (e.g. the hospital doesn't want to use resources), and instead demand more aggressive interventions (Garrett et al., 1993; Hopp and Duffy, 2000).

2.4 Quantifying Trust

Trust is shaped by subtle interactions in perceived discrimination, racial discordance, poor communication, language barriers, unsatisfied expectations, cultural stigmas and reputations, and is therefore difficult to quantify (L. Whaley, 2001). However, trust is a crucial part of medical care; increased levels of doctor-patient trust have been associated with stronger adherence to a physician's advice 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 methodological (e.g., selection bias) and practical (e.g., de-identification) reasons (Lee et al., 2016).

3. Data

3.1 Data Source

We use the MIMIC-III v1.4 (Johnson et al., 2016) database, consisting of 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 define two cohorts from MIMIC: **EOL** (11,000 admissions) and **ALL** (50,000 admissions). We use the EOL cohort to replicate racial disparities in MIMIC that other studies have observed (Perry et al., 2013). We use the ALL cohort to develop a metric to score the signs of mistrust in each patient's hospital stay and to predict non-end-of-life clinical tasks. Both our data extraction and modeling code are made available to promote reproducibility and further study (Johnson et al., 2017).

A patient is added to the EOL cohort if they have a hospital stay which lasted at least 6 hours and they either died in the hospital, were discharged to hospice, or were discharged to a skilled nursing facility⁴. Table 1 displays summary statistics of the cohort by race.

⁴See the extended masters thesis (Boag, 2018) for additional analysis. Disparities are also replicated on the eICU database, and all results are also performed on a stricter-but-smaller cohort that excludes skilled nursing facilities.

Table 1: Population characteristics by race. Parenthetical numbers for categorical variables denote % membership. Bracketed numbers for continuous variables denote 95% confidence intervals.

Variable	Value	Black	White	p-value
Population Size		1214	9987	-----
Insurance	Private	141 (11.61%)	1594 (15.96%)	< 0.001
	Public	1062 (87.48%)	8356 (83.67%)	
	Self-Pay	11 (0.91%)	37 (0.37%)	
Discharge Location	Deceased	401 (33.03%)	3869 (38.74%)	< 0.001
	Hospice	40 (3.29%)	421 (4.22%)	
	Skilled Nursing Facility	773 (63.67%)	5697 (57.04%)	
Gender	F	733 (60.38%)	5012 (50.19%)	< 0.001
	M	481 (39.62%)	4975 (49.81%)	
Length of stay		13.90 [5.55,19.56]	14.08 [6.45,19.45]	0.222
Age		71.31 [60.21,80.36]	77.87 [66.61,84.93]	< 0.001

A χ^2 test shows significant differences for insurance type, discharge location, and gender ($p < 0.001$ for all three). In particular, we see that the black population has both higher rates of uninsurance and publicly-funded insurance than their white counterparts. In lieu of other coded data, this often serves as a proxy for socio-economic status. In addition, white patients have higher in-hospital mortality and hospice rates, whereas a larger percent of black patients are discharged to skilled nursing facilities. Finally, there is a large difference between the black gender ratio (60-40 women) and white gender ratio (50-50). Using the Mann-Whitney test, the two populations have comparable lengths-of-stay ($p=0.222$), but significantly different population ages ($p < 0.001$).

3.2 Treatment Extraction

The main focus for this work is measuring disparities in aggressive end-of-life procedures, so we extract treatment durations (in minutes) from MIMIC’s derived mechanical ventilation (`ventdurations`) and vasopressor (`vasopressordurations`) tables. Due to the noisiness of clinical measurements – for instance, when one treatment span is erroneously coded as two back-to-back smaller spans – 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.

3.3 Patient-Provider Interaction Extraction

In this work, we quantify the patient’s interactions with their nurses and doctors using two sources: clinical notes and coded chart events. We obtain the notes of any patient who had a stay at least 12 hours in the ICU. This resulted in 48,273 admissions and over 800,000 notes. Throughout a patient’s stay, caregivers write narrative prose notes to document

⁵This 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.

Table 2: Coded interpersonal feature types from chartevents.

1:1 sitter present?	baseline pain level (0 to 10)	received bath?	bedside observer
behavioral intervent	currently experiencing pain	disease state	consults
education barrier	education learner	education method	family meeting?
education readiness	harm by partner?	education topic	judgement
follows commands?	family communication method	gcs - verbal response	informed?
hair washed?	goal richmond-ras scale	headache?	health care proxy?
pain management	non-violent restraints?	orientation	pain (0 to 10)
pain assess method	understand & agree with plan?	pain level acceptable?	reason for restraint
restraint device	richmond-ras scale (-5 to +4)	rsbi deferred	riker-sas scale
safety measures	violent restraints ordered?	security	security guard
side rails	status and comfort	sitter	skin care?
spiritual support	behavior during application	support systems	stress
verbal response	teaching directed toward	wrist restraints?	social work consult?

administered care, record patient preferences, issue reminders and warnings, and comment on the patient’s quality of care. In documenting their impressions, caregivers can give clues into the level of trust in their relationship with their patient. To supplement this narrative prose, we also extract coded information from the MIMIC `chartevents` table. Table 2 shows the `chartevents` information types, with categories including: 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 Riker-SAS (RR et al., 1999) and Richmond-RAS (Sessler et al.) agitation scales.

4. Methods

4.1 Quantifying Racial Disparities in End-of-Life Care

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

4.2 Establishing a Medical Mistrust Metric

We aim to quantify trust in the doctor-patient relationship. Because it is novel to study the effects of algorithmically-derived trust scores, we employ three metrics to avoid the impression that any single one tells the whole story. Much like fairness, trust may eventually prove to be impossible to fully characterize with a single score (Gajane and Pechenizkiy, 2018). Either way, we recommend that these scores be taken as a collective proxy for further-refined notions of trust. Two of the scores come from training a model to predict

Figure 1: An example of a nursing note documenting mistrust (in red). Situation-specific identifying information has been blacked out.

Social: Pt refused to sign ICU consent and expressed wishes to be DNR/DNI, seemingly very frustrated and mistrusting of healthcare system in relation to [redacted]. Also, w/ hx of poor medication compliance and follow-up

trust-associated labels using interpersonal doctor-patient features, and the third score is out-of-the-box sentiment analysis of the patient’s clinical notes.

For the first two metrics, we use an L1-regularized logistic regression to predict labels derived from clinical notes using structured event data. The features in this predictive framework were extracted for the ALL cohort from the `chartevents` table, as shown in Table 2. In total, we extract 620 binary features.

Figure 1 shows an example of the various signals of mistrust that are documented in nursing notes. We define two sets of trust-associated labels to train the above classifier using simple rule-based searches to detect:

1. **Noncompliance:** Noncompliance (e.g. refusing to adhere to follow-ups, regimens, take medicine, etc) indicates a very overt mistrust; rather than just holding an unspoken resentment, the patient actually defies their doctor’s orders. It explicitly demonstrates that the patient is willing to disregard physician decisions
2. **Autopsy:** One obvious benefit of an autopsy is quality assurance: did patient receive the proper treatment for the proper disease? Often times, families decline autopsies because they feel that dissecting a loved one would not be worth it when they trust that the doctor did everything they could. Conversely, higher autopsy rates could indicate patients suspect the doctor made a mistake. In this dataset, black patients (38.5% autopsy rate) are autopsied much more often than white patients (24.3% autopsy rate).

Once these model are trained, we use each classifier’s predicted probability as a measure of their mistrust for a new patient. This score defines a spectrum of trust based on how many indicators the patient has that are associated with typically poor interactions vs typically good interactions.

To encourage diversity of possible metrics, the third notion of mistrust is defined as the negative of the sentiment score from the patient’s clinical notes in the caregiver’s own words. We use the Pattern software package (De Smedt and Daelemans, 2012) and concatenate all of the notes from the stay into one document and tokenized using whitespace as a delimiter.⁶

Finally, we normalize each of the mistrust score distributions to be zero-mean and unit-variance, which helps for comparison.

⁶This step is actually important because a naive application of tokenization results in even positive notes which contain identified tags like “Date:[**5-1-18**]” to be tagged as negative because the tool’s string-matching algorithm was identifying “:|” as negative emoticon.

4.3 Prediction of Downstream Clinical Outcomes

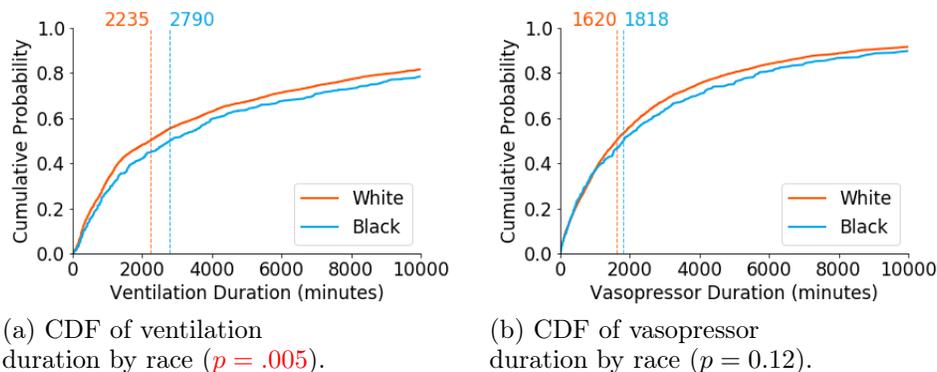
Trust is vital to a healthy doctor-patient relationship. A mistrustful patient might be reluctant to share sensitive, but potentially important information with their doctor. To further explore the impact of modeled trust, we examine two trust-associated outcomes (Code Status⁷ and Whether the patient leave Against Medical Advice (AMA)) and one more standard outcome (in-hospital mortality). We are interested to see how much value race and trust add as features to a baseline model which uses the patient’s age, gender, length-of-stay, and insurance type. We take the average AUCs of 100 runs from randomly chosen 60/40 train/test splits with an L1-regularized logistic regression model.

5. Results

5.1 Racial Treatment Disparities in EOL Care are Significant

We demonstrate racial treatment disparities in the MIMIC dataset do exist. Figure 2 highlights the differences in white and black populations for aggressive treatment durations. Figures 2a and 2b show that for both mechanical ventilation and vasopressors, the median black patient receives a longer duration of treatment, suggesting a reluctance to transition to palliative care. While these results only show statistical significance for ventilation ($p=0.005$), the same trends are also observable for vasopressor administration ($p=0.12$).

Figure 2: We observe racial disparities in aggressive interventions for black patients compared to white patients. Medians are indicated by dotted lines; differences are significant ($p < 0.05$) for ventilation but not for vasopressors.



5.2 Quantifying Mistrust Better Captures EOL Treatment Disparities

5.2.1 CREATION OF A MISTRUST METRIC

Table 3 shows the three most positively and most negatively informative weights in each model learned while fitting a mistrust metric. The features align well with an intuitive notion of mistrust: patients who are agitated, restrained, and unreceptive to education are

⁷either ‘Full Code’ or ‘DNR / DNI / Comfort Measures Only’

more likely to be mistrustful. Conversely, we see that calm, pain-free patients with good communication are more willing to trust their doctor.

5.2.2 BLACK PATIENTS HAVE HIGHER LEVELS OF MISTRUST

We observe statistically significant racial disparities in the two-out-of-three mistrust metrics. For both noncompliance-derived mistrust and negative sentiment, the median black patient has a higher level of mistrust than the median white patient using the Mann-Whitney test ($p=0.003$ and $p=0.007$, respectively). This is not surprising, given the extensive literature investigating differences in iatrophobia by race (Washington, 2007). Interestingly, there were virtually no racial disparities in the autopsy-derived metric ($p=0.13$), which is especially unexpected given the higher rate of autopsies among African Americans.

5.2.3 TRUST-BASED DISPARITIES IN END-OF-LIFE CARE

We hypothesize that if trust were a contributing factor to EOL treatment disparities, then stratifying the data into low mistrust and high mistrust⁸ would yield an even larger disparity than white and black.

Figure 3 shows significant disparities for both ventilation ($p < 0.001$) and vasopressor ($p=0.001$) durations using the noncompliance-derived metric to stratify the cohort. The difference between the medians of the mistrustful and trustful groups is 650 minutes for vasopressors (as opposed to 200 minutes between black and white). This gap is even larger for ventilation durations, as shown in Figure 3a: the noncompliance stratification shows a 2580-minute difference between medians, in contrast to the 832-minute gap for the race split in Figure 2a. This threefold-increase in the treatment gap suggests that trust might be one of the contributing factors for the original racial disparity.

For autopsy-derived mistrust, we can see that Figure 4 exhibits the same conclusions as race-based stratifications in Figure 2: mechanical ventilation has significant disparities ($p < 0.001$) whereas vasopressors do not ($p=0.059$). However, just as noncompliance-based

Table 3: Top-3 most positively and negatively informative chartevent features for each 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

(a) Noncompliance-derived Mistrust

Feature	Weight
pain present: no	-0.2689
spokesperson is healthcare proxy	-0.2271
family communication: talked to m.d.	-0.1184
reapplied restraints	0.1153
restraint type: soft limb	0.0980
orientation: oriented 3x	0.0363

(b) Autopsy-derived Mistrust

⁸For each treatment, we preserve the same size difference of stratified groups in order to maintain consistency in sample sizes for significance testing (e.g. since the black group contains 510 patients for ventilation, we compare the 510 lowest trust patients against the 4811 highest trust patients).

Figure 3: **Noncompliance Cohort Disparities:** A cohort of noncompliance-derived mistrust admissions yields significant differences in both ventilation and vasopressor duration.

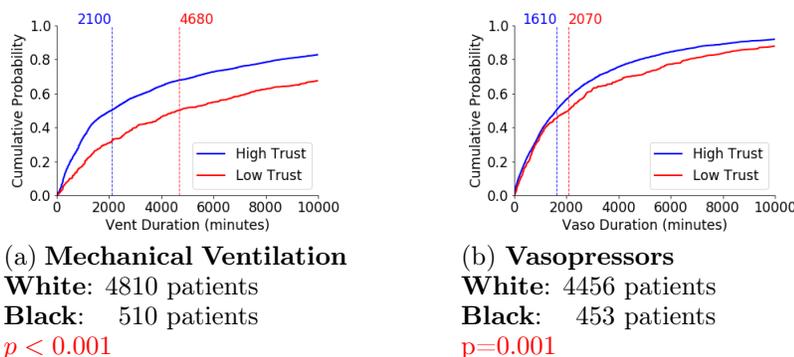
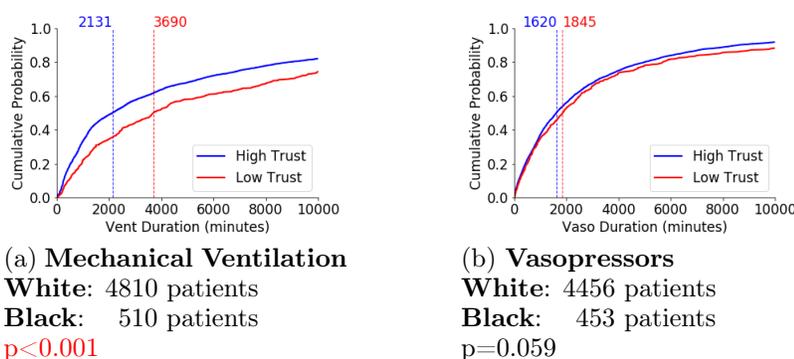


Figure 4: **Autopsy Cohort Disparities:** A cohort of autopsy-derived mistrust admissions yields significant differences in ventilation, but a non-significant difference in vasopressor duration.



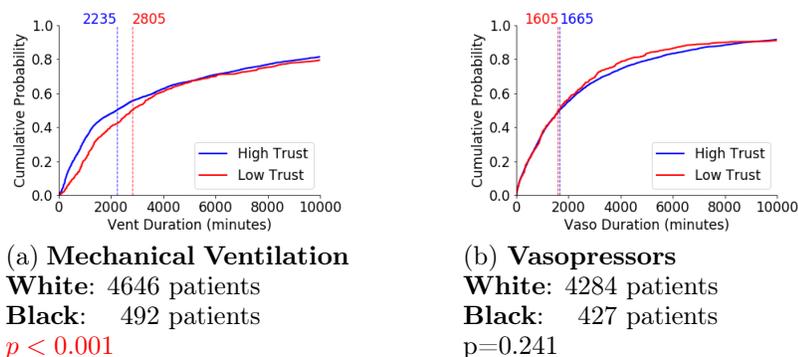
mistrust had a threefold increase in the treatment gap, this autopsy-derived metric has a twofold increase from the racial disparities found in ventilation (1,559 vs. 832 minutes) and vasopressors (245 vs 106 minutes).

Negative sentiment analysis exhibits the same trend for ventilation ($p < 0.001$) but a surprising result for vasopressor usage, as shown in Figure 5. There seems to be virtually no sentiment-based difference at all in vasopressor duration ($p = 0.241$). In fact, even the ventilation gap is smaller than with the other mistrust-based cohorts: 570 minutes (gaps for noncompliance and autopsy were 2,580 and 1,559, respectively). These results show that sentiment analysis is a bit of an outlier from the other two mistrust metrics. Nonetheless, we believe even this metric’s results are a useful contribution for exploring the space of algorithmically-defined trust.

5.2.4 NOT JUST SOME ACUITY SCORE PROXIES

These mistrust metrics are typically more effective than race at stratifying the data to show treatment disparities. However, one possible concern is the possibility that the scores are capturing severity-of-illness rather than mistrust: certainly, high-risk patients would be

Figure 5: **Sentiment Cohort Disparities:** A cohort of negative sentiment analysis admissions yields significant differences in ventilation, but virtually no differences in vasopressor duration.



treated differently than the general population. To address this concern, we compare the pairwise Pearson correlation coefficients between the three mistrust scores and two well-established acuity scores (OASIS and SAPS II). Table 4 shows that the severity scores have a strong correlation with one another (0.68). On the other hand, every mistrust scores has a very weak of a correlation with these risk scores; the largest severity-mistrust correlation being 0.086 between Sentiment and SAPS II. Interestingly, the autopsy-based mistrust metric is actually negatively correlated with the two severity scores (i.e. sicker patients are less likely to get autopsies) while still remaining positively correlated with the other two mistrust metrics. The noncompliant and autopsy metrics have the strongest intra-mistrust correlation (0.262). This is not surprising because these two metrics are both derived from Machine Learning on the `chartevents` features.

5.3 Mistrust Metric is Predictive in Downstream Tasks

We demonstrate that the mistrust score captures meaningful information by evaluating its contribution as a feature to predict three tasks: whether the patient is DNR, whether the patient left AMA, and in-hospital mortality.

The results can be found in Table 5, which show that race and trust both improve outcome prediction. Performance is variable across the tasks: no single feature is most useful for all three tasks. As is often the case, combining all of the features achieves the best results on each task — sometimes even with statistical significance, as for in-hospital

Table 4: Pairwise Pearson correlation coefficients between scores.

	OASIS	SAPS II	Noncompliance	Autopsy	Sentiment
OASIS	1.0	0.679	0.050	-0.012	0.075
SAPS II	0.679	1.0	0.013	-0.013	0.086
Noncompliance	0.050	0.013	1.0	0.262	0.058
Autopsy	-0.012	-0.013	0.262	1.0	0.044
Sentiment	0.075	0.086	0.058	0.044	1.0

Table 5: Effect of race and mistrust features on various binary classification tasks. Performance is measured by AUC and averaged over 100 runs.

Features	Left AMA (n=48,071)	Code Status (n=39,815)	In-Hospital Mortality (n=48,071)
Baseline	0.859 ± .014	0.763 ± .013	0.600 ± .011
Baseline + Race	0.861 ± .014	0.766 ± .014	0.614 ± .011
Baseline + Noncompliant	0.869 ± .012	0.767 ± .013	0.614 ± .010
Baseline + Autopsy	0.861 ± .012	0.773 ± .011	0.603 ± .012
Baseline + Negative-Sentiment	0.859 ± .013	0.765 ± .014	0.615 ± .010
Baseline + ALL	0.873 ± .012	0.782 ± .012	0.635 ± .010

mortality. Each mistrust metric achieves the top individual performance on one of the tasks: noncompliance-score for **Left AMA**, autopsy-score for **Code Status**, and negative-sentiment-score for **In-Hospital Mortality**. Race proves itself to be a very useful feature for all three tasks, outperforming at least one of the mistrust metrics in each category. Noncompliance-derived mistrust proves to be the most performant mistrust metric, achieving top-2 results for each task (excluding the ALL run).

Average classifier weights for the Baseline + ALL model are shown in Table 6. The two features most strongly associated with in-hospital mortality were the patient’s mistrust scores followed by the patient’s age. This is not surprising, because the highest-noncompliance-mistrust quartile has a 13.7% mortality rate, which is over three times as high as the lowest-noncompliance-mistrust quartile’s 4.4% mortality rate.

We also observe that noncompliance-derived mistrust, autopsy-derived mistrust, and race:black are the only three features positively associated with leaving the hospital AMA. Noncompliance (average coefficient of .52) is significantly more informative than autopsy and race:black (.01 and .03, respectively). As we saw in earlier experiments, mistrust is an even stronger indicator than race. In general, however, race tends to be a poor predictor for some of these outcomes because it is too coarse-grained to capture all of the different ways healthcare delivery can fail. For most tasks, the weights of racial features add little predictive value and are zero’d out during training. On the other hand, age is a very powerful predictor of these various outcomes, though not consistently indicating breakdowns in the doctor-patient relationship. While older patients are more likely to expire in-hospital, they are less likely to leave the hospital against medical advice. The mistrust score is the only feature positively associated with each outcome and consistently demonstrates predictive value.

6. Conclusion

One aim of this work is to establish existing findings of racial disparities in end-of-life care on a publicly available dataset, namely MIMIC III. We demonstrate that black patients receive – sometimes significantly – longer durations of invasive treatments and are significantly more likely to leave the hospital against medical advice. Though these trends have been studied in private datasets, we present our analysis on a public dataset and make our code available.

Table 6: Average regularized weights for BASELINE+ALL model on various tasks.

feature	Left AMA	Code Status	Mortality
noncompliant	0.52 ± 0.09	0.27 ± 0.04	0.16 ± 0.03
autopsy	0.01 ± 0.03	-0.44 ± 0.05	0.02 ± 0.02
negative sentiment	0.00 ± 0.02	0.09 ± 0.03	0.16 ± 0.03
race: asian	0.00 ± 0.00	0.00 ± 0.00	-0.05 ± 0.03
race: black	0.03 ± 0.12	-0.22 ± 0.19	-0.53 ± 0.31
race: hispanic	0.00 ± 0.00	-0.17 ± 0.21	-0.58 ± 0.34
race: other	-0.15 ± 0.19	-0.12 ± 0.17	0.15 ± 0.30
race: white	-0.02 ± 0.06	0.06 ± 0.15	-0.26 ± 0.30
race: native american	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
gender: male	0.00 ± 0.00	-0.85 ± 1.40	-0.67 ± 0.99
gender: female	-0.40 ± 0.20	-0.49 ± 1.39	-0.59 ± 0.99
insurance: private	-1.01 ± 0.21	-0.94 ± 0.29	-0.96 ± 0.95
insurance: public	0.00 ± 0.00	-0.02 ± 0.28	-0.50 ± 0.95
insurance: self-pay	0.00 ± 0.00	-0.02 ± 0.24	-0.21 ± 0.68
length-of-stay	-1.44 ± 0.37	-0.70 ± 0.10	0.08 ± 0.03
age	-2.10 ± 0.21	0.42 ± 0.02	0.20 ± 0.02

To further investigate this phenomenon, we propose multiple proxy mistrust scores using coded interpersonal data and clinical notes. We find that stratifying patients by trust score, instead of by race, more fully separates those patients who persist in aggressive interventions from those who do not. We show that the mistrust scores add value to multiple predictive tasks. While the scores partially capture racial differences (the median black patient had higher level of mistrusts than the median white patient), it is multifaceted, and feature analysis suggests that it captures information that agrees with our intuition of relationships and trust.

Medical machine learning is moving forward at an exciting pace; this work is a first step towards creating models that serve everyone, and do not propagate existing disparities in care. This work does have shortcomings, most notably that the small dataset size likely reduced statistical power and that the mistrust metrics were approximations of an intangible concept. Each metric had its flaws: there are non-trust-based reasons why a patient might be noncompliant (e.g. couldn't afford to refill their prescription), not all autopsies suggest skepticism of the doctor's performance, and although sentiment is a related concept it is ultimately not the same thing as trust. We encourage forming inter-disciplinary collaborations among machine learning, healthcare, and social science communities to refine algorithmic notions of mistrust, quantify potential biases and disparities in models, and identify potential factors that lead to patient dissatisfaction.

Acknowledgments

The authors would like to thank Matthew McDermott for his input and suggestions. This research was funded in part by the National Science Foundation Graduate Research Fellowship Program under Grant No. 1122374 as well as the National Institute of Health under

grants NIBIB Grant R01 EB017205, U54-HG007963, R01-EB017205, and 1R01MH106577 (National Institute of Mental Health).

References

V.R. Adebimpe. Overview: White norms and psychiatric diagnosis of black patients. 138: 279–85, 04 1981.

Elizabeth Arias. United states life tables, 2009. 6 no 7, 2014.

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

CDC. Prostate cancer rates by race and ethnicity. <https://www.cdc.gov/cancer/prostate/statistics/race.htm>, 2014.

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

T. De Smedt and W. Daelemans. Pattern for python. In *Journal of Machine Learning Research*, volume 13, page 2031–2035, 2012.

Danielle M. Ely, Anne K. Driscoll, and T.J. Matthews. Infant mortality rates in rural and urban areas in the united states, 2014. 2017. URL <https://www.cdc.gov/nchs/data/databriefs/db285.pdf>.

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

Pratik Gajane and Mykola Pechenizkiy. On formalizing fairness in prediction with machine learning. 2018.

Joanne Mills Garrett, Russell P. Harris, Jean K. Norburn, Donald L. Patrick, and Marion Danis. 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.

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

Charles E. Gessert, Nakeisha M. Curry, and Audra Robinson. Ethnicity and end-of-life care: The use of feeding tubes. *Ethnicity and Disease*, 11(1):97–106, 2001. ISSN 1049-510X.

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

Marzyeh Ghassemi, Marco AF Pimentel, Tristan Naumann, Thomas Brennan, David A Clifton, Peter Szolovits, and Mengling Feng. A multivariate timeseries modeling approach to severity of illness assessment and forecasting in icu with sparse, heterogeneous clinical data. In *AAAI*, pages 446–453, 2015.

- Monika K. Goyal, Nathan Kuppermann, Sean D. Cleary, Stephen J. Teach, and James M. Chamberlain. Racial disparities in pain management of children with appendicitis in emergency departments. *JAMA Pediatrics*, 169(11):996–1002, 2015.
- Amresh Hanchate, Andrea C. Kronman, Yinong Young-Xu, Arlene S. Ash, and Ezekiel Emanuel. 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.
- Joshua M. Hauser, Sharon F. Kleefield, Troyen A. Brennan, and Ruth L. Fischbach. 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.
- Diane E Hoffmann and Anita J Tarzian. The girl who cried pain: a bias against women in the treatment of pain. *The Journal of Law, Medicine & Ethics*, 28(s4):13–27, 2001.
- Faith P. Hopp and Sonia A. Duffy. Racial variations in end-of-life care. *Journal of the American Geriatrics Society*, 2000.
- Alistair E. W. Johnson, Tom J. Pollard, and Roger G. Mark. Reproducibility in critical care: a mortality prediction case study. In Finale Doshi-Velez, Jim Fackler, David Kale, Rajesh Ranganath, Byron Wallace, and Jenna Wiens, editors, *Proceedings of the 2nd Machine Learning for Healthcare Conference*, volume 68 of *Proceedings of Machine Learning Research*, pages 361–376, Boston, Massachusetts, 18–19 Aug 2017. PMLR. URL <http://proceedings.mlr.press/v68/johnson17a.html>.
- Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3, 2016.
- Arthur L. Whaley. Cultural mistrust: An important psychological construct for diagnosis and treatment of african americans. 32:555–562, 12 2001.
- Janet J. Lee, Ann C. Long, J. Randall Curtis, and Ruth A. Engelberg. 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>.
- Norman G. Levinsky, Wei Yu, Arlene S. Ash, Mark A. Moskowitz, Gail S. Gazelle, Olga Saynina, and Ezekiel Emanuel. Influence of age on medicare expenditures and medical care in the last year of life. 286:1349–1355, 09 2001.
- Sarah Muni, Ruth A. Engelberg, Patsy D. Treece, Danae Dotolo, and J. Randall Curtis. 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.
- Henry S Perkins, Cynthia Geppert, Adelita Gonzales, Josie D Cortez, and Helen P Hazuda. Cross-cultural similarities and differences in attitudes about advance care planning. *Journal of General Internal Medicine*, 17(1):48–57, 2002.

- William R.G. Perry, Alvin Kwok, Christina Kozycki, and Leo Anthony Celi. Disparities in end-of-life care: A perspective and review of quality. 16, 2013.
- Sean M Phelan, Diane J Burgess, Mark W Yeazel, Wendy L Hellerstedt, Joan M Griffin, and van M Ryn. Impact of weight bias and stigma on quality of care and outcomes for patients with obesity. *Obesity Reviews*, 16(4):319–326, 2015.
- Robert S. Pritchard, Elliott S. Fisher, Joan M. Teno, Sandra M. Sharp, Douglas J. Reding, William A. W Knaus, John E. Wennberg, and Joanne Lynn. Influence of patient preferences and local health system characteristics on the place of death. *Journal of the American Geriatrics Society*, 1998.
- Riker RR, Picard JT, and Fraser GL. Prospective evaluation of the sedation-agitation scale for adult critically ill patients. 1999.
- Dorothy Ruiz. Epidemiology of schizophrenia: Some diagnostic and sociocultural considerations. 138:315–326, 1983.
- C. Sessler, M. Gosnell, M.J. Grap, G. Brophy, P. O’Neal, K. Keane, E. Tesoro, , and R. Elswick. The richmond agitation–sedation scale validity and reliability in adult intensive care unit patients.
- R.J. Simon, J.L. Fleiss, B.J. Gurland, P.R. Stiller, and L. Sharpe. Depression and schizophrenia in hospitalized black and white mental patients. 28:509, 04 1973.
- Astha Singhal, Yu-Yu Tien, and Renee Y. Hsia. 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>.
- Alexander K. Smith, Ellen P. McCarthy, Elizabeth Paulk, Tracy A. Balboni, Paul K. Maciejewski, Susan D. Block, and Holly G. Prigerson. Racial and ethnic differences in advance care planning among patients with cancer: Impact of terminal illness acknowledgment, religiousness, and treatment preferences. *Journal of Clinical Oncology*, 26(25):4131–4137, 2008. doi: 10.1200/JCO.2007.14.8452. URL <https://doi.org/10.1200/JCO.2007.14.8452>. PMID: 18757326.
- Harini Suresh, Nathan Hunt, Alistair Johnson, Leo Anthony Celi, Peter Szolovits, and Marzyeh Ghassemi. Clinical intervention prediction and understanding with deep neural networks. In *Machine Learning for Healthcare Conference*, pages 322–337, 2017.
- Harriet Washington. *Medical Apartheid: The Dark History of Medical Experimentation on Black Americans from Colonial Times to the Present*. 2007. ISBN 978-0385509930.
- Hannah Wunsch, Carmen Guerra, Amber E Barnato, Derek C Angus, Guohua Li, and Walter T Linde-Zwirble. Three-year outcomes for medicare beneficiaries who survive intensive care. *Jama*, 303(9):849–856, 2010.
- Christopher J. Yarnell, Longdi Fu, Doug Manuel, Peter Tanuseputro, Theres Stukel, Ruxandra Pinto, Damon C. Scales, Andreas Laupacis, and Robert A. Fowler. Association between immigrant status and end-of-life care in ontario, canada. In *JAMA*, 2017.