Quantifying Racial Disparities in End-of-Life Care

by

William Boag

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Science in Computer Science and Engineering

at the

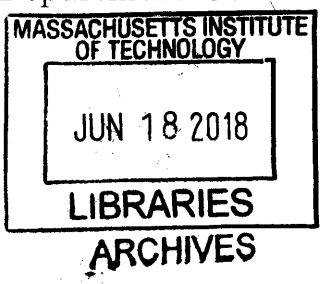
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2018

Author ..... Signature redacted .....
Department of Electrical Engineering and Computer Science
May 23, 2018

Certified by ..... Signature redacted .....
Peter Szolovits
Professor of Electrical Engineering and Computer Science
Thesis Supervisor

Accepted by ..... Signature redacted .....
Leslie A. Kolodziejski
Professor of Electrical Engineering and Computer Science
Chair, Department Committee on Graduate Students



Quantifying Racial Disparities in End-of-Life Care

by

William Boag

Submitted to the Department of Electrical Engineering and Computer Science
on May 23, 2018, in partial fulfillment of the
requirements for the degree of
Master of Science in Computer Science and Engineering

Abstract

There are established racial disparities in healthcare, particularly during end-of-life care when poor communication and historical inequities can lead to suboptimal options and outcomes for patients and their families. Previous work has suggested that medical disparities can reflect higher rates of mistrust for the healthcare system among black patients. 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. In this thesis, 'I demonstrate that racial disparities which have been reported previously are also present in two public databases. I explore the notion that one underlying cause of this disparity is due to mistrust between patient and caregivers, and develop a multiple trust metric proxies to measure such mistrust more directly. These metric demonstrate even stronger disparities in end-of-life care than race does and statistically significant higher levels of mistrust for black populations. I hope that this work will serve as a useful view for bias and fairness in clinical data, and that future work can better understand mistrust so that its underlying factors (e.g. poor communication and perceived discrimination) can be addressed.

Thesis Supervisor: Peter Szolovits
Title: Professor of Electrical Engineering and Computer Science

Acknowledgments

Since arriving at MIT two years ago, I have benefited from many supportive co-workers, peers, and friends. Most notably, I've been very lucky to have Peter Szolovits as my adviser during this time of exploration, growth, and discovery. Pete has been everything I needed in an adviser and his style allowed me to explore my own ideas and realize the kind of research I love working on!

I also owe a great deal of gratitude to the Clinical Decision-Making Group. Marzyeh Ghassemi gave me the initial idea to work on this thesis as well as generally fantastic life and work advice! Tristan Naumann and I have been working together for five years now, and his guidance and fondness for cardigans has stuck with me. Also thank you to Michele and Ziyu for their banter, and to Matthew and Elena for their exciting philosophical debates.

I must also thank Khan Academy and the MIT Open Courseware program. Six years ago, I learned Computer Science from some free online lectures. This profoundly changed the course of my life; I was so grateful to have access to world-class education for free. I certainly never expected I would be fortunate enough to call this place my home, and every day, I still feel that same excitement to be here.

To my friends that have supported me through emotional turmoil(s): Katie, Dom, Alyssa, and Amanda. Your friendship and support has helped me better understand issues I've struggled with. And to Jose for getting me into audiobooks, which was the catalyst for one of the biggest life improvements for me in the last few years.

My dog Alex saw me start this journey two years ago, but he wasn't able to make it to the end. He was the best dog in the world.

And finally, I owe so much to my parents. Their support over many, many years has been the most important gift I've ever received.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In this work, I present the following contributions:

- I demonstrate racial disparities in aggressive end-of-life (EOL) care using two public datasets.
- I present multiple proxy metrics for measuring trust during a patient's stay.
- I show that higher levels of mistrust are more associated with black patients than white patients.
- I demonstrate trust-based disparities in aggressive EOL care are even more pronounced than race-based disparities.
- I show that the proposed mistrust metrics demonstrate reasonable properties in prediction and analysis.

1.1 Biases in Health

For both outcomes and rates of incidence, there are stark racial disparities in health.

1.1.1 Worse Health Outcomes for Minorities

The life expectancy gap between the African American and white populations is about 4.5 years. Although the gap is slowly closing, it would take approximately 40 years for the African American life expectancy to "catch up" at its current rate, assuming

white life expectancy were held constant [3]. Infant mortalities are still twice as high for African Americans than for their white counterparts, regardless of socioeconomic status [12]. Both African Americans and American Native populations are more than twice as likely to have diabetes as white populations are [7]. African American men are 50% more likely to develop prostrate cancer than white men, and are twice as likely to die from it [6].

Racial disparities are also evident in patient care. Even when controlling for covariates such as age, sex, and time of treatment, African Americans (both children and adults) are less likely to receive pain medication than their white counterparts [19, 55]. Racial minorities [34], women [26], and obese patients [47] tend to have poorer treatment options available, and established longitudinal health outcomes. Doctors are more likely to diagnose African American patients with schizophrenia and other psychotic disorders, yet less likely to diagnose them with depression [54, 1, 50]

1.1.2 Biases in End-of-Life Care

Even during end-of-life (EOL) care, when all patients directly confront death, there are still observable and consistent racial variations in care [42]. During EOL, minorities are more likely to receive high-intensity, life-sustaining treatments [48, 39, 16] and have fewer advance directives [56]. 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 [9]. White patients are more likely to utilize hospice care and are less likely to disenroll in it than nonwhite patients [14, 27]. While some of these differences may be attributed to cultural preferences, many issues are the result of poor communication or unclear expectations. Family members of African American patients are more likely to cite absent or problematic communication with physicians about EOL care [24]. Similarly, in surveys, African Americans report lower rates of satisfaction with the quality of care that they received by physicians [21].

In a recent overview of racial disparities in EOL care, physicians conclude that further study is required in understanding why such apparent discrepancies by race

occur [46]. One such recommendation from one of the authors is to control for illness severity when measuring disparities in treatment selections, which has previously not been examined.

Previous work has suggested that medical disparities may reflect higher levels of mistrust for the healthcare system among black patients. 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.

1.1.3 Racial Iatrophobia

Not only are there disparities in care patterns for African Americans, but there may also be disparities in attitudes towards the health care system among African American patients.

In her 2007 book, Harriet Washington argues that the medical exploitation of African Americans by white institutions throughout American history has created "Black Iatrophobia[1]" [59]. Some contend it is irrational for African Americans to mistrust the medical community, believing that race-based medical experimentation has been stopped after the ratification of the Belmont Report following the public outrage of the Tusskgee Syphillis Study.[2] However, while the Tuskgee Study might still be the most notorious example of African American exploitation, it is far from the only example, and not the most recent.

Medical abuse in America has plagued the African American community from the beginning of US history all the way through modern times. Going back to 1801, Thomas Jefferson injected 80 of his own slaves with smallpox to prototype vaccines [29]. In the late 1840's, Dr. James Marion Sims (considered by some to be "the Father of Gynecology") surgically experimented on and mutilated his female slaves — who were unable to refuse his operations — without anesthesia [38].

As recently as 1987–1991, US scientists administered as much as five hundred times

[1] *Iatrophobia* is a Latin word that translates to *a fear of doctors*

[2] in which a group of African American men with syphilis were denied the cure for three decades in an effort to study the progression of the disease

the approved dosage of the experimental Edmonton-Zagreb vaccine against measles to African American and Hispanic babies in Los Angeles without communicating to the parents on informed consent forms that the vaccine was experimental or unlicensed [8].

Harriet Washington's hypothesis of Black Iatrophia has also been validated in the published literature. Socialized mistrust of the medical community in minority groups has been established as a factor in care differences [59]. Family members of African American patients are more likely to cite absent or problematic communication with physicians about EOL care [24]. Similarly, in surveys, African Americans report lower rates of satisfaction with the quality of care that they received by physicians [21].

Poor trust can specifically impact end-of-life care and potentially help understand racial disparities in aggressive treatments such as mechanical ventilation and vasopressors [59]. When a critical-condition patient's treatments are not working, a doctor may decide that further invasive procedures are unlikely to succeed or return the patient to a normal lifestyle. In those situations they would recommend withdrawing treatment and transitioning to comfort-based measures to ensure the patient does not suffer any further. However, when a patient or healthcare proxy does not trust the clinician's assessment – perhaps because of the suspicion that the hospital doesn't want to waste the resources – they might decline the option for palliative care and instead demand more aggressive interventions. [14, 27].

1.2 Machine Learning

The quantity of health-related data is increasing rapidly, from genetic data to electronic health records to medical images like x-rays [35, 49]. This growth has facilitated large-scale machine learning methods to guide care.

1.3 Racial Bias in Machine Learning

Biases are especially troubling in the context of machine learning applied to clinical data. For example, Black and Hispanic patients are often given less pain medication for equivalent injuries and reported pain levels [19, 55]. If this pattern is present in the training data for a model built to recommend treatment, it will learn that race correlates with pain medication dosage. Bias can be propagated in the model's future recommendations, and can also exacerbate them in a feedback loop where it reinforces unrepresentative data samples [13]. However, including information about race may be important for some clinical tasks, e.g., if there are differences in recommended care based on genetic makeup. In such a setting, quantifying bias and establishing proxy measures for medical trust is particularly important.

1.3.1 Clinical Natural Language Processing

Although there is a trend toward digitizing patient records in an increasingly structured manner, much information is still hidden in unstructured narrative text. In their primary role, electronic health record (EHR) notes facilitate patient care by recording communication among care staff. These clinical notes capture patient data that provide insight into a patient's status and course of care, such as patient history, recommended treatments, records of meetings, and more. Often, this granularity of data does not appear in equivalent detail in a structured form elsewhere in the EHR.

Clinical Natural Language Processing (NLP) uses machine learning techniques to leverage the narrative prose in doctors' notes to comprehend and better understand patient care. Early stages of the NLP pipeline include concept extraction [52, 4] and relation extraction [58] to identify basic units of interaction between the concepts in notes. Tools for basic information extraction help enable more challenging tasks, such as recommending appropriate billing codes [57], clustering patient records [43], and generating clinical paraphrases [23]. Further still, some tasks provide immediate use for patients and physicians, such as outcome prediction [17, 5, 51, 20] and question answering [53].

1.4 Quantifying Trust

Previous work has suggested that medical disparities can reflect higher levels of mistrust for the healthcare system among black patients. Blacks were suspicious of the clinical motives in advance directives and do-not-resuscitate (DNR) orders [61], and believed that the healthcare system was controlling which treatments they could receive [45]. When a patient does not believe the doctor has their best interests in mind, they 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 [9].

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 [36]. 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 [15].

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 [37].

In particular, trust surveys are not available for the datasets in this work. However, we turn to another source to estimate a patient and clinician's trust relationship: clinical notes. Throughout a patient's stay, caregivers write narrative prose to document administered care and family meetings, record patient preferences, issue reminders and warnings, and comment on the patient's quality of care. Most notes are written by nurses, though we also consider physician notes, social worker notes, and discharge summaries. In documenting their impressions of how to best understand and interact with their patients, caregivers can give clues into the level of trust in their relationship with their patient. All of these interactions help to paint a picture of the difficult relationship this patient has with their doctor. Clinical notes have

been used for prediction tasks in previous work [17] but not for quantifying mistrust.

Chapter 2

Data

2.1 Data

This work uses ICU data from two databases: Philips eICU [31, 18] and MIMIC III [32]. These datasets are two of the largest publicly available ICU databases, containing EHR records of patient demographics, admissions, treatments, and outcomes.

2.1.1 MIMIC III

The Medical Information Mart for Intensive Care (MIMIC III) v1.4 is a publicly-available dataset of ICU stays [33]. 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. For the experiments about treatment disparities, I focus on a cohort of black and white patients in end-of-life care who spent at least one day at the hospital.

In this dataset, there is no explicit flag to indicate end-of-life care. To address this issue – and overcome the issue of data scarcity – I employ two definitions of EOL: one strict and one broad. Under the strict definition, a patient is in EOL if they died in-hospital or were discharged to a hospice setting. However, some patients have hospice care indicated in their notes but are discharged to skilled nursing facilities (SNF). As a result, the more relaxed definition of EOL care also includes patients

discharged to SNF (in addition to in-hospital mortality and hospice). Throughout this thesis, I report my results on the broader EOL definition, and apply the same experiments on the stricter definition in Appendix A, which largely demonstrate the same trends but sometimes with less statistical significance due to smaller sample sizes.

Both the data extraction and modelling code are made available[1] to enable reproducibility and further study [30].

Table 2.1 displays summary statistics of the cohort by race. 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 average lengths-of-stay (p=0.222), but significantly different population ages ($p < 0.001$). Black patients tend to be significantly younger than white patients.

2.1.2 MIMIC III Treatments

The main focus for this work is measuring disparities in aggressive end-of-life procedures, so I extract treatment durations (in minutes) from MIMIC's derived mechanical ventilation (`ventdurations`) and vasopressor (`vasopressordurations`) tables.[2] Due to the noisiness of clinical measurements — for instance, when one treatment span is erroneously coded as two back-to-back smaller spans — I merge any treatment spans that occurred within 10 hours of each other.[3] If a patient had multiple spans, such as an intubation-extubation-reintubation, then I consider the patient's treatment

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

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

[3] 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.

Table 2.1: **MIMIC III Population characteristics**. Parenthetical numbers for categorical variables denote % membership. Bracketed numbers for continuous variables denote confidence intervals.

Variable	Value	Black	White	p-value
Cohort 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

duration to be the sum of the individual spans.

2.1.3 MIMIC III Chart Events

A large part of this thesis aims to better quantify the nuances of a patient's interactions with their nurses and doctors. The `chartevents` table for all MIMIC contains coded interpersonal interactions that have been documented with the patients. I extract chartevents data for all patients — not just end-of-life ones — in order to have a larger dataset from which I can derive proxies for trust.

Table 2.2 summarizes the chartevents features, 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 agitation scales (Riker-SAS and Richmond-RAS). In total, there are 620 binary features.

Table 2.2: Coded interpersonal variable 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	feamily 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?

2.1.4 MIMIC III Notes

Throughout a patient's stay, caregivers write narrative prose notes to document administered care and family meetings, record patient preferences, issue reminders and warnings, and comment on 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 [17] but not for investigating 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 [41].

We obtain the notes of any patient who had a stay of at least 12 hours in the ICU. This resulted in 48,273 admissions and over 800,000 notes.

2.2 Philips eICU

Philips Healthcare provides an eICU service to caregivers across the country enabling a team of expert physicians to remotely offer guidance and support to ICUs. Some of the data collected during the ordinary operation has been released for research as the eICU Collaborative Research Database, which covers patients who were admitted to

Table 2.3: **eICU Population characteristics**. Parenthetical numbers for categorical variables denote % membership. Bracketed numbers for continuous variables denote confidence intervals.

Variable	Value	Black	White	p-value
Cohort Size		1820	13214	- - - - -
Discharge Location	Deceased	785 (43.13)	5208 (39.41)	< 0.001
	Nursing Home	117 (6.43)	509 (3.85)	
	Skilled Nursing Facility	918 (50.44)	7497 (56.74)	
Gender	F	918 (50.44)	6477 (49.02)	0.265
	M	902 (49.56)	6737 (50.98)	
Length of stay		8.93 [4.43,16.48]	7.97 [4.28,13.93]	< 0.001
Age		66.00 [56.00,76.00]	73.00 [63.00,82.00]	< 0.001

critical care units in 2014 and 2015 [18, 31]. The full database contains records for over 160,000 unique patients totaling 200,000 ICU stays across 208 hospitals throughout the United States. Just as with MIMIC III, I use a strict EOL cohort and a broad EOL cohort. Because "Hospice" is not a selectable discharge location, the strict EOL definition only includes patients who died in-hospital. The broader EOL cohort includes those patients as well as ones who were discharged to skilled nursing facilities and nursing homes. See Appendix A for this work's experiments replicated on the small, stricter cohort.

Table 2.3 displays summary statistics of the cohort by race. A χ^2 test shows significant differences for discharge location ($p < 0.001$) but comparable distributions for gender (p=0.265). Although discharge locations do not include hospice (thus rendering the comparison to MIMIC III not perfect), black patients actually show significantly higher in-hospital mortality than white patients ($p < 0.001$), which is the opposite trend from MIMIC III. Using the Mann-Whitney test, the two populations have significantly different lengths-of-stay and ages ($p < 0.001$ each). Black patients tend to be much younger than white patients but have longer stays.

(a) How many admissions would be available for each given threshold of per-hospital compliance.

(b) How compliant the full, heterogeneous dataset would be for each given threshold of per-hospital compliance.

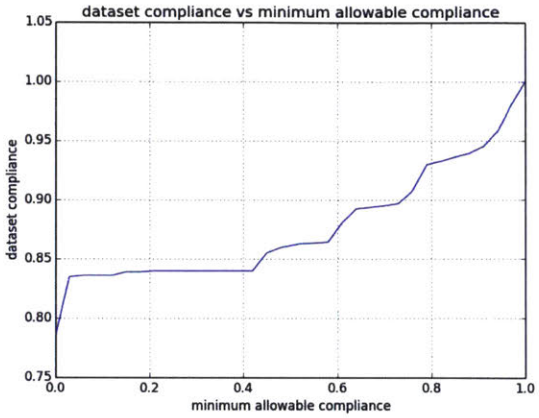


Figure 2-1: Effects on the final dataset as the minimum hospital compliance threshold is varied.

Reliable Treatment Reporting in the Database

The eICU has a heterogeneous population of hospital locations, sizes, and work cultures. The different workplace environments carry with them different levels of compliance for documenting events in the EHR. Some hospitals are very strict about recording every treatment event, whereas others might be a little more lax about failing to log some information. More lax hospital cultures would underestimate the incidence or durations of some treatments recorded in the `treatment` table of the eICU-CRD. I address this issue by filtering out hospitals if they do not maintain a reasonable level of documenting procedures which are coded elsewhere. In essence, I use a known outcome to "grade the homework" of whether a hospital diligently logs procedures in the `treatments` table.

In order to measure each hospital's reporting-compliance level, I compare the recorded `treatment` data for mechanical ventilation in the first day to the automatically-logged "oOBIntubDay1" flag in the `apachePatientResult` table. This flag indicates whether the patient actually was ventilated during the first day of the hospitalization. As a sanity check, I verified that nearly all patients who have ventilation events in the `treatment` table also have the oOBIntubDay1 flag on, which means that treatments

are only logged when they actually occur. On the other hand, there was much more variability in the other direction:

Definition: Hospital Reporting-Compliance (2.1)

When intubation truly occurred (i.e. oOBIntubDay1 is true), how often was ventilation actually recorded in the `treatment` table for a given hospital?

Using this per-hospital metric for recording-compliance, Figure 2-1a shows how the size and full-dataset compliance would vary for various thresholds of a minimum per-hospital compliance. As we lower the threshold, we allow more hospitals into the set but at the expense of a less reliable dataset. Figure 2-1b shows that we can maintain a full-dataset compliance of 95% using a threshold of .9, which would leave about 60,000 admissions (roughly half the size of the original dataset).

eICU Treatments

From the treatment dataset, I extract the timestamped events that indicate the patient is receiving aggressive treatments (**mechanical ventilation** and **vasopressors**). Because these entries are individual timepoints — rather than the start-stop durations from MIMIC III — I estimate the duration of treatment for a patient with the following heuristic: subtract the highest timestamp from the lowest timestamp. Treatment durations are measured in minutes.

This dataset does not have any available notes or chartevents, so I only use it for racial disparity analysis, in Chapter 3.

Chapter 3

Racial Disparities in End-of-Life Care

In this chapter, I demonstrate racial disparities in end-of-life care for both public ICU datasets: MIMIC III and eICU. I perform these experiments using the broader definition of EOL. For a comparison of these results on the stricter EOL cohort, see Appendix A.

I examine differences in populations for aggressive end-of-life treatments. For continuous variables (none of which are normally distributed), I determine statistical significance between population medians using the Mann-Whitney test. For categorical variables, I perform a χ^2 test. In accordance with prior work, p-values $< .05$ are considered statistically significant.

In order to show differences in treatments, every visualization of black-vs-white treatment distributions is a figure of overlayed Cumulative Distribution Functions (CDFs). It is easier to see the total effects of shifting probability masses when they are aggregated in a CDF rather than a Probability Distribution Function (PDF), because the black and white PDFs would have very similar shapes.

3.1 Aggressive End-of-Life Interventions

Following previous work [62, 42, 37], I examine differences in the duration mechanical ventilation and vasopressors that a patient receives at end-of-life.

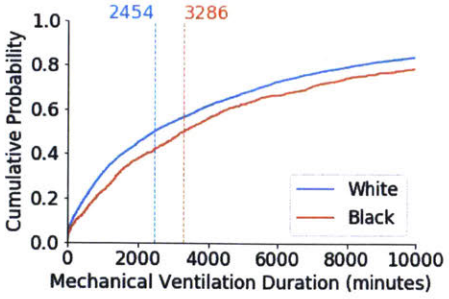
3.1.1 Mechanical Ventilation

Figure 3-1 highlights the differences of mechanical ventilation durations in white and black populations. We can see that for both datasets, black patients receive statistically significantly longer durations of ventilation than white patients.

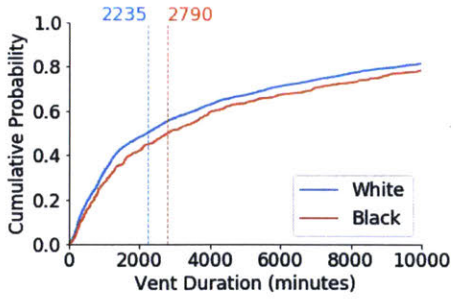
Vasopressors

Similarly, we examine racial disparities in vasopressor usage for both MIMIC III and eICU. Figure 3-2 shows that the median black patient does receive longer treatments in both datasets – 106 and 198 more minutes, respectively – the differences were not significant.

Figure 3-1: **Mechanical Ventilation:** CDF of ventilation duration by race, where dotted lines represent the median duration treatment for a population. In multiple datasets, the median black patient receives statistically significant longer ventilation durations than the median white patient.

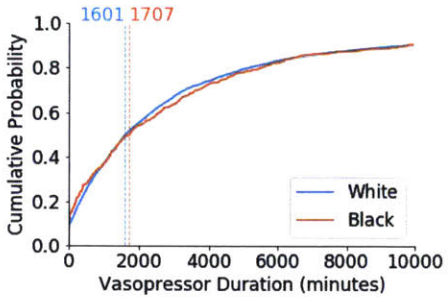


(a) *MIMIC Mechanical Ventilation*
White: 4810 patients
Black: 510 patients
p=0.005

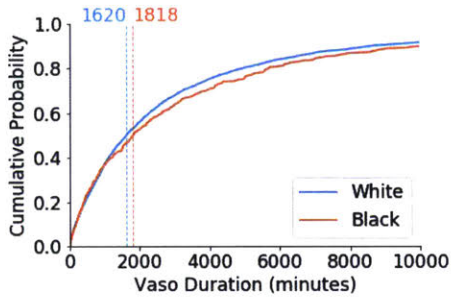


(b) *eICU Mechanical Ventilation*
White: 4911 patients
Black: 655 patients
$p < 0.001$

Figure 3-2: **Vasopressors:** In both datasets, the median black patient receives a longer duration of vasopressors than the median white patient. This trend is not statistically significant in either dataset..



(a) *MIMIC Vasopressors*
White: 4458 patients
Black: 453 patients
p=0.122



(b) *eICU Vasopressors*
White: 3479 patients
Black: 464 patients
p=0.422

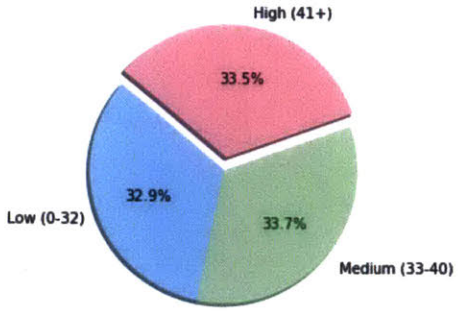
3.2 Race and Severity of Illness

One notable shortcoming of previous work is that none of the studies controlled for patient severity of illness. Because black patients have worse health [60], examining racial disparities without adjusting for illness could yield misleading results. In particular, if longer treatment durations were simply an effect of illness (e.g. sicker patients receive more treatment) then sub-stratifying patients by risk scores would help tease apart those confounding factors. For an EOL dataset, as this is, all of these patients are severely ill; however when they first arrive in the ICU, they might have a lower severity of illness that only later develops into something life threatening. By stratifying into categories of risk/severity, I compare similar-situationed-patients, rather than comparing across illness severity buckets.

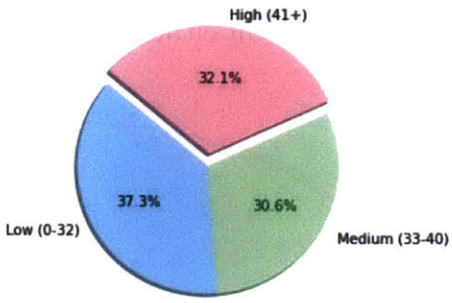
3.2.1 Racial Breakdown of Severity Scores

To test the hypothesis that black patients tend to be at higher levels of risk, I break both the white and black populations into low, medium, and high risk sub-groups. Then, I check to see whether black patients are, in fact, more likely to be high-risk than white patients. Low/Medium/High thresholds are chosen to split a given

Figure 3-3: **MIMIC OASIS:** score breakdown for white/black cohorts.

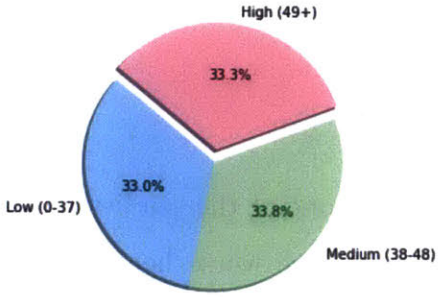


(a) White (n=9987)

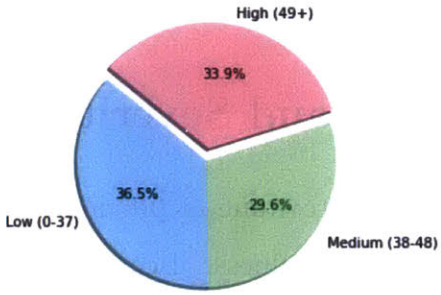


(b) Black (n=1214)

Figure 3-4: **MIMIC SAPS II:** score breakdown for white/black cohorts.

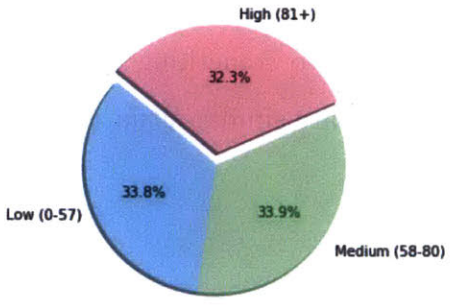


(a) White (n=9987)

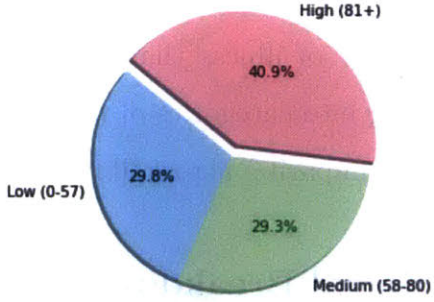


(b) Black (n=1214)

Figure 3-5: **eICU Apache:** score breakdown for white/black cohorts.



(a) White (n=13214)



(b) Black (n=1820)

dataset's combined black-and-white EOL cohort into three equal-sized tertiles. I use both the OASIS and SAPS II scores for MIMIC, and the Apache Score for eICU.

Figures 3-3, 3-4, and 3-5 show the severity makeups for the three different metrics. Because they are nearly ten times as many white patients as black patients, the thresholds to divide the full dataset into tertiles are calibrated to almost evenly split white patients into 33/33/33 splits. On the other hand, black populations are less uniformly divided, with 40% of the eICU black population being High Risk.

Surprisingly, in the MIMIC experiments, its more likely that a black patient is low-risk than medium- or high-risk. For both MIMIC SAPS II and eICU Apache, a black patient is more likely to be high-risk than a white patient is, but this trend is only very strong for eICU Apache. A χ^2 test for each of these three black-white pairs shows statistically significant ($p < 0.001$) differences in the two distributions.

3.3 Risk-Stratification

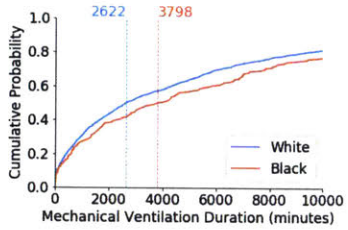
Because of the significant differences in severity between the black and white populations, a closer look at risk-stratified subgroups is warranted.

One word of caution, however, is that the populations were already small to begin with. Further stratification yields very small cohorts, so some of these experiments are better seen as investigative than conclusive. Black severity cohorts are typically 150-250 patients, which means only obvious patterns will stand out while subtler trends will fail to be significant.

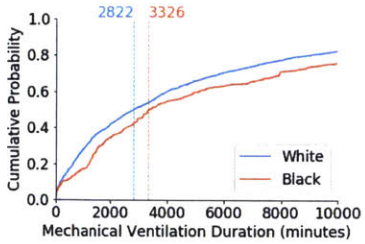
3.3.1 Mechanical Ventilation

Even when stratifying by severity, mechanical ventilation durations are still consistently higher for black patients than white patients. In the eICU dataset, shown in Figure 3-6 the gap is statistically significant at all three levels. On the other hand, we see a more interesting trend in the MIMIC dataset in Figure 3-7. Ventilation durations are virtually interchangeable at medium and high levels of severity, but have a significantly large disparity at low levels.

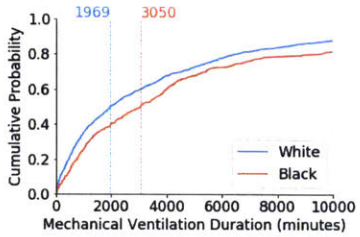
Figure 3-6: **eICU ventilation:** Black patients received statistically significantly longer median ventilation durations than white patients at every level of acuity.



(a) *Ventilation Low-Risk*
Apache Score: 0-70
White: 1702 patients
Black: 166 patients
p=0.033

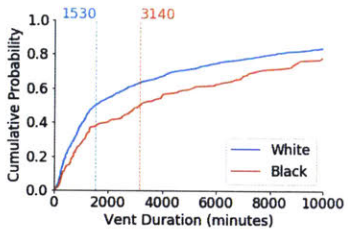


(b) *Ventilation Medium-Risk*
Apache Score: 71-98
White: 1661 patients
Black: 217 patients
p=0.004

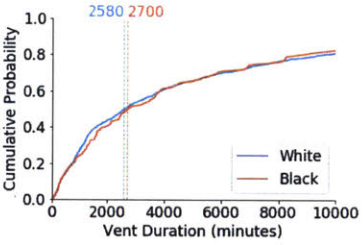


(c) *Ventilation High-Risk*
Apache Score: 99+
White: 1548 patients
Black: 272 patients
$p < 0.001$

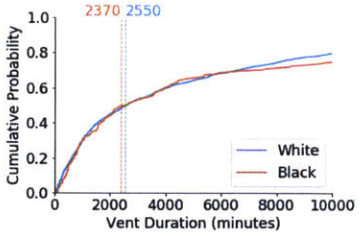
Figure 3-7: **MIMIC ventilation:** Black patients received longer median ventilation durations than white patients for low severities, but virtually the same care in higher-risk situations.



(a) *Ventilation Low-Risk*
OASIS Score: 0-36
White: 1782 patients
Black: 178 patients
$p < 0.001$



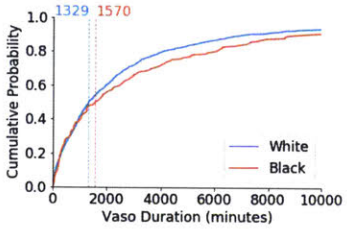
(b) *Ventilation Medium-Risk*
OASIS Score: 37-43
White: 1511 patients
Black: 153 patients
p=0.410



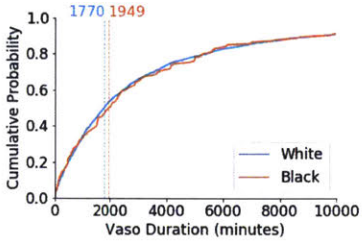
(c) *Ventilation High-Risk*
OASIS Score: 44+
White: 1517 patients
Black: 179 patients
p=0.333

3.3.2 Vasopressors

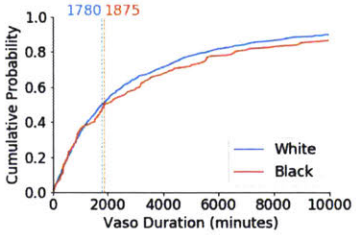
Figure 3-8: **MIMIC vasopressors:** Black patients received longer median vasopressor durations than white patients did at every level of acuity, though no cohort had a statistically significant difference.



(a) *Ventilation Low-Risk*
OASIS Score: 0-35
White: 1668 patients
Black: 169 patients
p=0.132

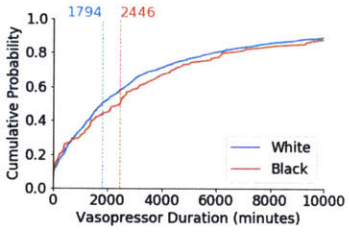


(b) *Ventilation Medium-Risk*
Apache Score: 36-43
White: 1472 patients
Black: 142 patients
p=0.459

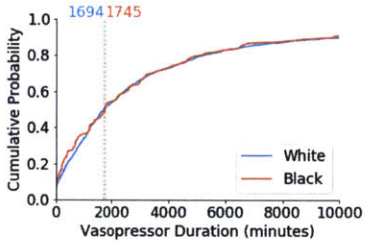


(c) *Ventilation High-Risk*
Apache Score: 44+
White: 1318 patients
Black: 142 patients
p=0.245

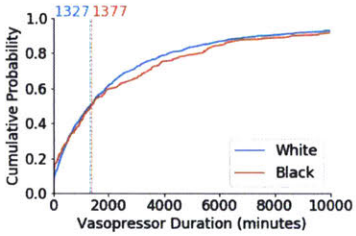
Figure 3-9: **eICU vasopressors:** Though not statistically significant (note small population sizes), black patients received longer median vasopressin durations than white patients did in the eICU at every level of acuity.



(a) *Vasopressin Low-Risk*
Apache Score: 0-74
White: 1217 patients
Black: 129 patients
p=0.223



(b) *Vasopressin Medium-Risk*
Apache Score: 75-102
White: 1159 patients
Black: 127 patients
p=0.275



(c) *Vasopressin High-Risk*
Apache Score: 103+
White: 1103 patients
Black: 208 patients
p=0.391

For vasopressors, shown in Figures 3-8 and 3-9, we see the same pattern. Just as with the non-stratified cohorts, none of these groups have significantly different

vasopressor durations. However, even though the datasets do not have the size to show statistical power, it is worth noting that all 6 of 6 groups show black patients receiving longer durations, even if only by an hour. Also, for both MIMIC III and eICU, the low risk subgroups show bigger racial disparities.

Chapter 4

Quantifying Trust in Clinical Care

As discussed previously, many believe that one of the large causes for disparities in aggressive treatment for end-of-life care comes from a socialized mistrust that non-whites have of the medical community. To investigate this claim, I want to quantify the level of trust a patient has and see what effect this has on the treatment gap. I believe that by directly modeling the mistrust in those relationships – rather than indirectly using race as a proxy – then disparities in end of life care will be even wider.

For the rest of this work, I only consider the MIMIC III dataset, because I make use of clinical notes, which are not available in the eICU database. After discussing the data available in MIMIC III, I outline three potential metrics for mistrust between a doctor and patient: likelihood for being noncompliant, likelihood for requesting an autopsy, and sentiment of documented notes.

4.1 Signs of Medical Mistrust

Though previous work has measured patient trust using surveys, such surveys are not available for the MIMIC III dataset [15]. However, there *are* both structured and unstructured parts of the data which capture the relationship between caregiver and patient.

Table 4.1: 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	feamily 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?

4.1.1 Structured Data

Introduced in Section 2.1.3, the *chartevents* table for all MIMIC contains coded interpersonal interactions that have been documented with the patients. As a reminder, Table 4.1 has been copied here, and it summarizes the chartevents features. We extract these coded variables to use as features for a supervised machine learning task. In total, we extract 620 binary features.

4.1.2 Unstructured Data

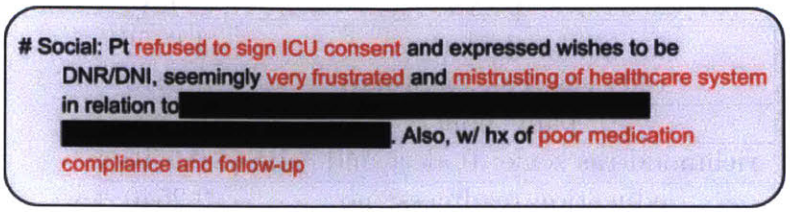
Throughout a patient's stay, caregivers write narrative prose notes to document administered care and family meetings, record patient preferences, issue reminders and warnings, and comment on the patient's quality of care. In documenting their impressions of how to best understand and interact with their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

Trust is central to a healthy doctor-patient relationship. A working relationship requires both sides to be active participants. Some indications of trust/mistrust include (but are not limited to):

(a) Does the patient listen to the doctor when asked to follow directions?

(b) Is there good communication between patient and caregiver?

Figure 4-1: An example of a nursing note documenting mistrust (in red). Situation-specific identifying information has been blacked out.



(c) How comfortable is the patient asking for assistance?

(d) How comfortable is the patient telling the doctor something personal?

(e) Do the nurses and patient show respect to each other?

(f) Is the family reluctant to donate organs?

(g) Does the family feel that an autopsy is necessary to double check the doctor's work?

Figure 4-1 shows an example of a patient who is reluctant to consent to procedures, is frustrated at the entire health system, and has poor compliance and adherence to physician instruction.

4.2 Noncompliance

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 reasonable proxy for mistrust.

To begin, I use a simple rule-based search through the notes to determine whether the patient has the phrase "noncompliant" documented somewhere in their notes (e.g. for not adhering to medical advice, regimens, follow-ups, etc). Of the 48,273 hospital admissions, I find 464 with noncompliant patients.

Although only 464 patients have documented noncompliance, I want to assign a mistrust score to every patient in the dataset. To accomplish this, I pose mistrust identification as a supervised Machine Learning problem. Using `chartevents`

Table 4.2: 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

variables as features, I predict whether a given admission has noncompliance documented in one of its notes. To accomplish this, I use an L1-regularized logistic regression model from scikit-learn [44]. Once the model is trained, I use the resulting predicted probability of noncompliance as a measure of mistrust for a new patient.

Table 4.2 shows the three most positively and most negatively informative weights used to predict a mistrust metric (Section 4.1). The features align well with 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.

For visualization, the score is normalized to zero-mean and unit variance across the corpus, which also helps make comparisons of different metrics. As shown in Figure 4-2, there is a statistically significant racial disparity in the mistrust metric. Per the Mann-Whitney test, the median black patient has a higher level of mistrust than the median white patient ($p < 0.001$). This result agrees with the extensive literature investigating differences in iatrophobia by race [59], and increases confidence that noncompliance is a reasonable prediction target for learning a linear combination of chartevents features.

4.3 Autopsy Rates

One way we can observe the mistrust held against doctors is to observe the fraction of patients who undergo an autopsy. In the last few decades, the autopsy rate has

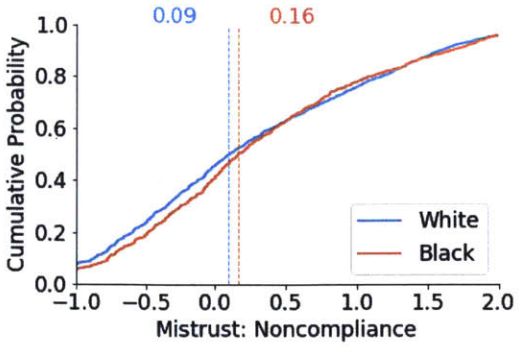


Figure 4-2: Racial disparity in noncompliance-derived mistrust metric.
White: 9923 patients
Black: 1202 patients
$p < 0.001$

declined more than 50 percent from 1972 through 2007: from 19.3 percent to 8.5 percent [28].

Autopsies are growing increasingly rarer for a variety of factors, including: rising costs of autopsies without being reimbursed by third-parties, doctor-held belief that technology has rendered autopsies uninformative, the cumbersome process for obtaining consent, and that in 1971 the Joint Commission on Accreditation of Healthcare Organizations dropped its requirement that hospitals must perform autopsies at least 20% of the time in order to be accredited [2].

However, as of 2006, the Beth Israel[1] autopsy rate is 150 postmortem examinations per year (representing approximately 20% of the deaths at the Hospital) [25]. Further, according 2006 Beth Israel Autopsy Manual states "it is the position of both this Department and the Hospital Administration that every attempt should be made to secure autopsies on every patient dying at the Beth Israel Deaconess Medical Center. Quite clearly, the value of the autopsy is greatest when it is performed properly and the results of the examination are communicated to the patient's physician effectively and promptly" [25].

One of the most obvious benefit of an autopsy is to provide 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

[1]the hospital from which MIMIC III originates.

Table 4.3: Autopsy rates by race in MIMIC III.

population	consent	decline	% consent
Asian	2	23	8.0%
White	161	505	24.2%
Other	56	102	32.9%
Black	32	51	38.6%
Hispanic	9	11	45.0%
ALL	260	692	27.3%

Table 4.4: Top-3 most positively and negatively informative chartevent features for tuning the autopsy-derived mistrust metric.

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

it, since they trust that the doctor did everything they could. Conversely, higher autopsy rates – in conjunction with other indicators – could serve as a proxy for mistrust between the family and doctor.

I use a simple rule-based parser to read through the notes for automatic consent extraction. I look for lines that mention the word "autopsy" and find phrases involving either *consent/agreed/requested* or *declined/refused/denied.* This approach does not assume implicit non-consents, and only counts yes/no for autopsies if the autopsy is explicitly mentioned in the note.

Autopsy rates by race are shown in Table 4.3. As is often the case, the white population (24.3% consent) is much larger than the others and dictates where the global average (27.3% consent) is set. Both black and hispanic populations have much higher rates at 38.5% and 45.0%, respectively. These nonwhite populations are almost twice as likely as white patients to want quality assurance of proper care. On the other hand, the asian autopsy rate is exceedingly low: 8.0%. This might also be explained by cultural norms, for instance in Japan autopsy rates are very low, in part

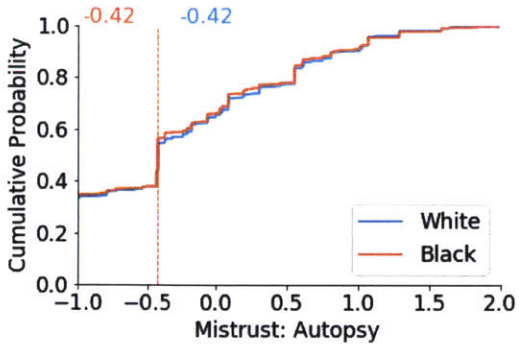


Figure 4-3: Racial disparity in autopsy-derived mistrust metric.
White: 9923 patients
Black: 1202 patients
p=0.126

because families do not want the body to be damaged by autopsy [40].

Just as with noncompliance, I build a L1-regularized logistic regression model to predict whether a patient is likely to consent to an autopsy based on their chartevents features. The top-3 most positive and most negative features are shown in Table 4.4. Again, patients with good communication, healthcare literacy, and no pain are more trusting, while patients who are restrained are less trusting.

For a given patient, their mistrust is calculated as the linear combination of weights learned from the logistic regression model. This score is then normalized to zero-mean and unit variance across the corpus, to make comparisons more interpretable. Figure 4-3 displays the distributions for autopsy-derived mistrust score by race, and there is statistically no difference. This raises possible doubts about the effectiveness of this score as a metric for trust.

4.4 Sentiment Analysis

Rather than trying to derive a score from chartevents features to predict some mistrust proxy variable, I also attempt to capture the tone of the doctor-patient relationship from the caregiver's own words. Using the Pattern software package [11], I perform sentiment analysis using the notes to gain more insight to a patient's care. Because the two previous metrics measure *mistrust* instead of *trust*, for consistency I flip the

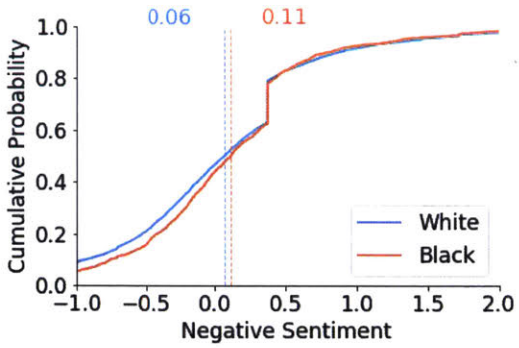


Figure 4-4: Racial disparity in (negative) sentiment.
White: 9669 patients
Black: 1173 patients
p=0.007

sign of all sentiment analysis scores, effectively measuring negative sentiment. This ensures that high scores for the Sentiment metric (i.e. strongly negative sentiment scores) would reasonably indicate high levels of mistrust.

All of a patient's notes for their hospital admission are concatenated into one document and tokenized using whitespace as a delimiter.[2] The sentiment scores of the full black-white cohort are then normalized to be zero-mean and unit-variance, which helps for comparison.

Figure 4-4 shows that both black and white patients have means greater than zero on the negative-sentiment scale, which indicates that the distribution has skew towards the majority of notes carrying negative sentiment. However, black patients received statistically significantly more negative sentiment scores from their notes (p=0.007). This simple heuristic analysis of the notes agrees with the previous work that black patients are less satisfied with their care [10, 37].

4.5 Not Just Some Severity Score Proxies

Because so many confounding factors all occur at once, it can be hard to tease out a desired signal. In particular, before showing that these mistrust metrics are effective

[2]This step is actually important because a naive application of tokenization results in even positive notes which contain "Date:[**5-1-18**]" to be tagged as negative because the tool's string-matching algorithm was identifying ":[" as negative emoticon.

at stratifying the data for even larger treatment disparities, I need to make sure that they aren't simply picking up some severity-indicating signal. Certainly, high-risk patients are treated differently than the general population. To dispel this concern, I compare the pairwise Pearson correlation coefficients between the three mistrust scores, OASIS, and SAPS II.

Table 4.5 shows that the two well-established acuity scores, OASIS and SAPS II, have a strong correlation of 0.68. On the other hand, none of the mistrust scores have much of a correlation with these metrics: the largest severity-mistrust correlation being 0.086 between Sentiment and SAPS II. Interestingly, the Autopsy 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.

4.6 Limitations

One major limitation of this work is the definition of mistrust metrics. While I use multiple metrics to try to combat the issue of not *truly* capturing trust, the issue still remains. The gold standard for measuring trust would be patient-administered surveys which directly ask about the doctor-patient relationship. In lieu of such data, I try to approximate such measurements using coded interpersonal variables and clinical notes written by caregivers. However, each of my three mistrust metrics

Table 4.5: Pairwise Pearson correlations between severity scores and mistrust score.

	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

has its shortcoming:

1. **Noncompliant.** There are numerous reasons a patient could be noncompliant that are unrelated to trust, such as lack of resources (e.g. couldn't refill medication because the patient had no car).

2. **Autopsy.** While an autopsy being performed *could* indicate mistrust, there are numerous other reasons why an autopsy could occur.

3. **Sentiment.** Not only is the sentiment analysis tool used both for general-domain tasks and providing just shallow analysis, but sentiment used in notes doesn't necessarily refer to how the caregiver feels about the patient; it could be picking up "death" as a negative word even if the patient had an unfavorable outcome despite excellent care.

In future work, it would be very useful to evaluate how well these – and other metrics – correlate with gold-standard patient surveys.

Chapter 5

Explaining Disparities with Trust

In order to assess whether trust-based factors influence the amount of treatment a patient gets during end-of-life care, I replicate the race-based treatment comparisons from Chapter 3 except with trust-based cohorts. Specifically, for a given treatment and mistrust score, I rank every patient by their scores and split them into a High Trust and a Low Trust cohort.[1]

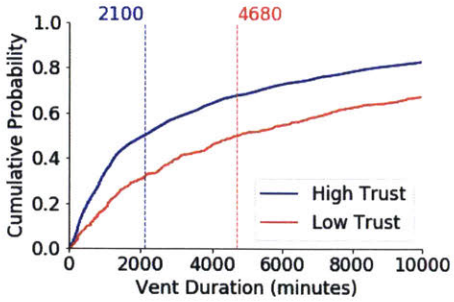
In the following three sections, I examine how well the three different mistrust scores are able to stratify cohorts for treatment disparities. The three metrics behave mostly similarly, and can be summarized with the following observations:

1. Disparities in mechanical ventilation are much starker than in vasopressor usage. This holds across all metrics, even when controlling for severity of illness.

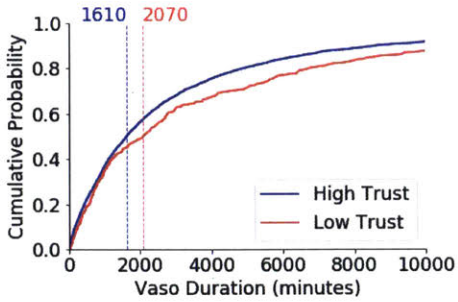
2. With few exceptions, lower risk patients have more significant disparities than high risk patients. This could be caused by higher levels of discretion and judgment being used for low-severity patients, whereas high-severity ones are subject to pre-defined protocols which promote standardized care.

[1]For each treatment, I preserve the same size difference of stratified groups. For instance, because the black group contains 510 patients for ventilation, I compare the 510 lowest trust patients against the 4811 highest trust patients. These numbers vary a very small amount because not every patient has chartevents features or notes.

Figure 5-1: **Noncompliance Cohort Disparities:** A cohort of noncompliance-derived mistrust admissions yields significant differences in both ventilation and vasopressor duration.



(a) **Mechanical Ventilation**
White: 4810 patients
Black: 510 patients
$p < 0.001$



(b) **Vasopressors**
White: 4456 patients
Black: 453 patients
p=0.001

5.1 Mistrust: Noncompliance

These experiments stratify the data based on the noncomplaince-derived mistrust score. The results, as shown below, indicate that this score is very effective at selecting high-treatment patients.

5.1.1 Aggressive End-of-Life Interventions

Figure 5-1 shows significant trust-based disparities for both ventilation and vasopressor durations. 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 5-1a: the trust-based stratification shows a 2580-minute difference between medians, in contrast to the 832-minute gap for the race split in Figure 3-1a. This threefold-increase in the treatment gap suggests that trust might be one of the contributing factors for the original racial disparity.

Figure 5-2: **Risk-Controlled Noncompliance Cohort Ventilation:** A cohort of noncompliance-derived mistrust admissions yields significant ventilation duration differences at all three levels of severity ($p < 0.001$).

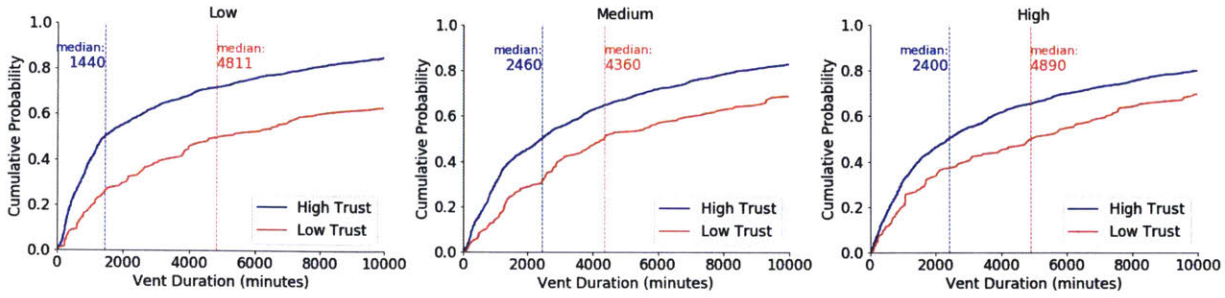
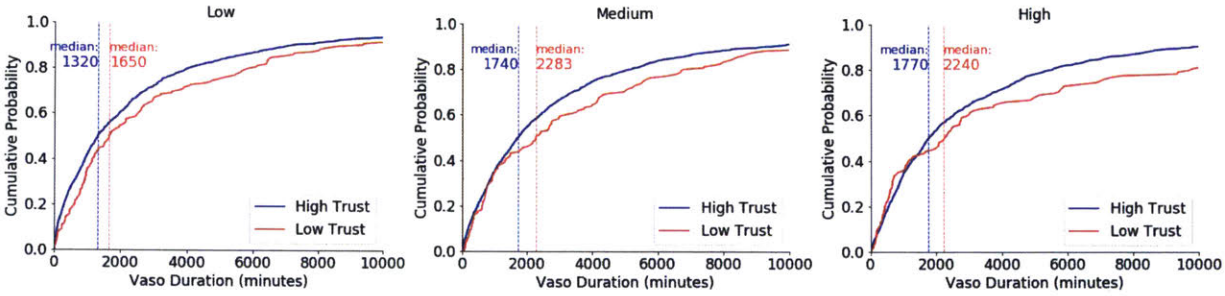


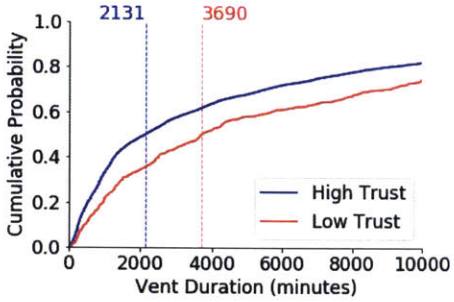
Figure 5-3: **Risk-Controlled Noncompliance Cohort Vasopressor:** A cohort of noncompliance-derived mistrust admissions yields vasopressor duration differences at all three levels of severity, though only low and medium (p=0.005 and p=0.034) are significant. High risk disparities are not significant (p=0.191).



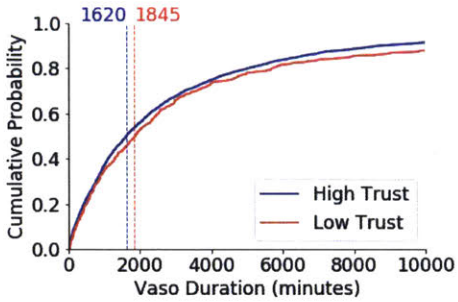
5.1.2 Risk Stratification

This metric continues to demonstrate significant treatment disparities across severities of illness. Figure 5-2 shows severity-controlled mechanical ventilation disparities, and Figure 5-3 shows the same for vasopressors. Whereas in Figure 3-7 – where controlling for risk eliminated the race-based ventilation gap for medium- and high-risk patients – we can instead see here that all three levels of severity still have significant trust-based treatment gaps. Similarly, although there were no signfificant race-based differences in vasopressor treatment at any level of risk from Figure 3-8, such disparities persist for both low- and medium-risk cohorts for trust-based cohorts.

Figure 5-4: **Autopsy Cohort Disparities:** A cohort of autopsy-derived mistrust admissions yields significant differences in ventilation, but a non-significant difference in vasopressor duration.



(a) **Mechanical Ventilation**
White: 4810 patients
Black: 510 patients
$p < 0.001$



(b) **Vasopressors**
White: 4456 patients
Black: 453 patients
p=0.059

5.2 Mistrust: Autopsy

These experiments stratify the data based on the autopsy-derived mistrust score. As shown below, this score shows a less pronounced – but still mostly significant – treatment gap than the above noncompliance-based mistrust score. This score still has wider gaps than race-based cohorts, though perhaps it is not as effective as a mistrust metric because it was fit to fewer labels ($\approx 1,000$) in Section 4.3 than the noncompliance score ($\approx 48,000$) was.

5.2.1 Aggressive End-of-Life Interventions

Figure 5-4 reflects the same conclusions as race-based stratification: mechanical ventilation has significant disparities ($p < 0.001$) whereas vasopressors do not (p=0.059). However, just as noncompliance-based 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, as shown in Figures 3-1a and 3-2a.

Figure 5-5: **Risk-Controlled Autopsy Cohort Ventilation:** A cohort of autopsy-derived mistrust admissions yields ventilation duration differences at all three levels of severity, though only low and medium ($p < 0.001$ and p=0.006) are significant. High risk disparities are not significant (p=0.170).

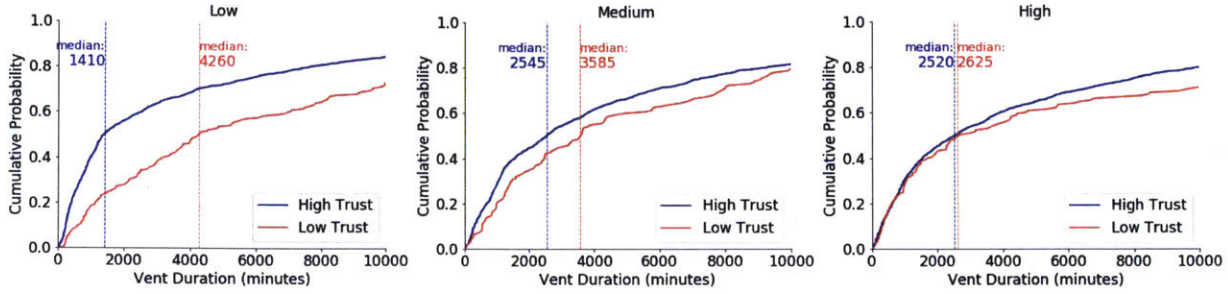
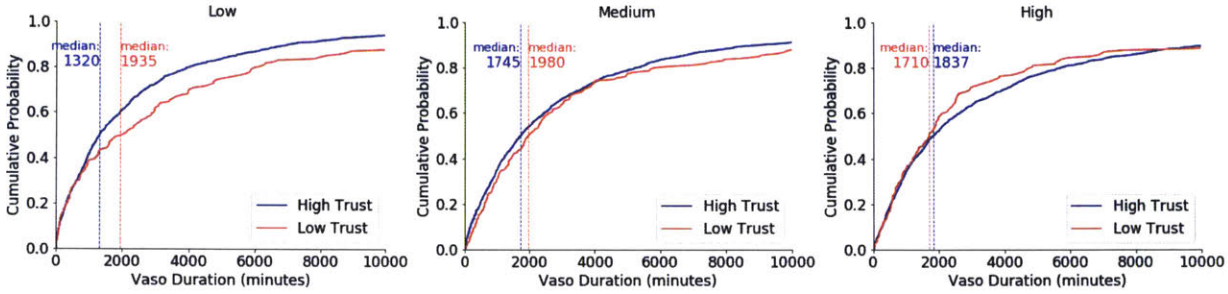


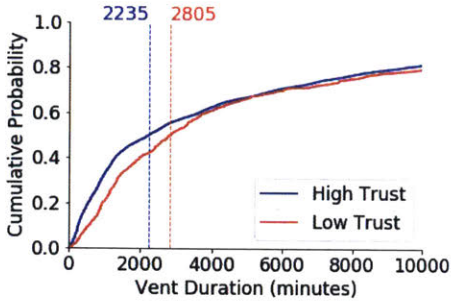
Figure 5-6: **Risk-Controlled Autopsy Cohort Vasopressor:** A cohort of autopsy-derived mistrust admissions yields significant vasopressor duration differences for low risk patients (p=0.025). Medium and high risk cohorts have little-to-no disparities (p=0.111 and p=0.156).



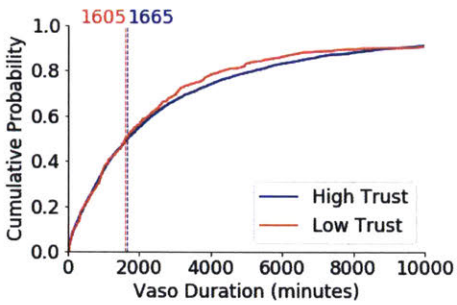
5.2.2 Risk Stratification

Just as indicated above, this metric seems to be at a halfway point between the effects of racial disparities and noncompliance-derived mistrust disparities. Figures 5-5 and 5-6 show that this score yields significant differences for both low- and medium-risk cohorts for ventilation (whereas race-based only saw this for low, and noncompliance-derived saw this for all three levels). Similarly, there was a large difference in low-severity vasopressor usage between High Trust and Low Trust populations (585 minutes), whereas the racial difference in low-severity vasopressor usage was just 241 minutes.

Figure 5-7: **Sentiment Cohort Disparities:** A cohort of negative sentiment analysis admissions yields significant differences in ventilation, but virtually no differences in vasopressor duration.



(a) **Mechanical Ventilation**
White: 4646 patients
Black: 492 patients
$p < 0.001$



(b) **Vasopressors**
White: 4284 patients
Black: 427 patients
p=0.241

5.3 Sentiment

These results show that sentiment analysis is a bit of an outlier from the other two mistrust metrics. Most of the experiments show virtually no sentiment-based treatment disparities, and the ones that are present don't follow the same trends indicated by stratifications based on race or the other two mistrust metrics. Though sentiment analysis does give a useful window into the doctor-patient relationship, the shortcomings and biases of this metric cause it to yield unintuitive results for a mistrust score, as described below.

5.3.1 Aggressive End-of-Life Interventions

As per usual, Figure 5-7 shows a significant difference for ventilation ($p < 0.001$) but not for vasopressors (p=0.241). There seems to be virtually no sentiment-based difference at all in vasopressor usage, with the negative sentiment ("Low Trust") and positive sentiment ("High Trust") medians differing by just 60 minutes. 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).

Figure 5-8: **Risk-Controlled Sentiment Cohort Ventilation:** A cohort of negative sentiment admissions yields significant ventilation duration differences for medium and high risk patients (p=0.008 and $p < 0.001$). Low risk cohorts have a disparity but it is not significant (p=0.171).

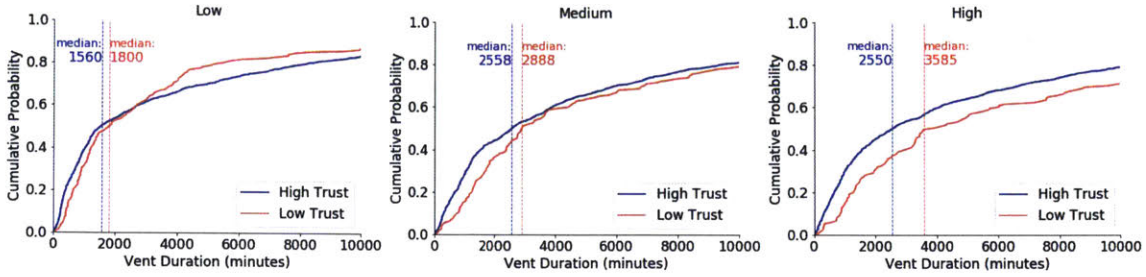
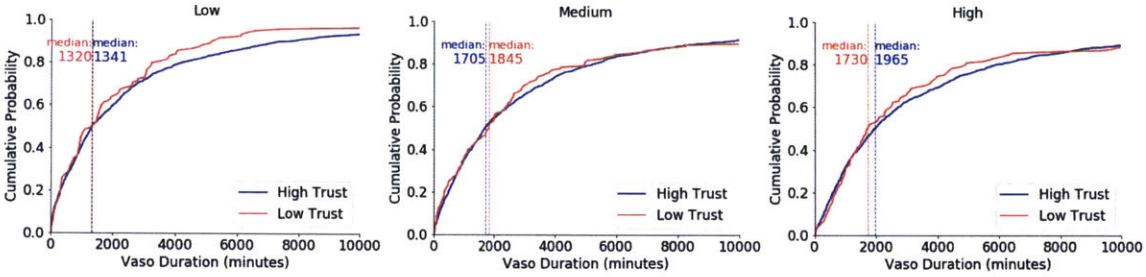


Figure 5-9: **Risk-Controlled Sentiment Cohort Vasopressor:** There were no significnt vasopressor differences at any level of severity for sentiment-based cohorts (p=0.152, p=0.282, 0.353).



5.3.2 Risk Stratification

It is not surprising that the sentiment metric did not find meaningful differences in vasopressor usage, which has proven to have more subtle disparities. However, unlike previous partitions – which demonstrated weak signals for disparities – Figure 5-9 demonstrates that there do not appear to be any sentiment-based differences at all.

On the other hand, Figure 5-8 shows something quite surprising: patients with negative sentiment analysis scores receive similar care to positively-scored patients at low-severity, but as their conditions worsen, their care diverges. This is likely the result of high-severity cases using strongly informative words such as "bloody", "declining", and even "aggressive" in their notes. These words tend to carry negative sentiment, and therefore the most negative notes tend to reflect the most grave cases. Consequently, negative sentiment patients tend to require more treatment than their positive counterparts.

Chapter 6

Evaluating Mistrust Metrics

In the previous chapter, I proposed three possible metrics to serves as proxies for mistrust. Of course, each of these scores is an approximation of the complex and potentially asymmetric doctor-patient relationship. A gold standard evaluation would examine how well a given metric correlated with a sample of patient survey responses that explicitly discuss trust. In lieu of such an analysis, I explore a few other aspects for evaluating a trust-based metric.

6.1 Sentiment as an Evaluation

Using sentiment scores as a mistrust metric yielded results that were unexpected because they were not designed to optimize relationship-based targets. However, one benefit is that its results are intuitively interpretable from reading the notes. As a result, I also use sentiment scores to evaluate the *other* metrics. I would expect that mistrustful patients have more negative notes than trustful patients, and I hypothesize that this sentiment gap is larger than a race-based sentiment gap.

Table 6.1 shows the differences in sentiment analysis scores between race, severity of illness, and trust. I observed statistically significant differences in the population means ($p < .05$) for all five 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

Table 6.1: Median sentiment analysis of cohorts stratified by race, severity, and trust.

population	**N**	**median**	p-value
White	9629	-0.064	p=0.008
Black	1164	-0.111	
Low OASIS	9629	-0.058	$p < 0.001$
High OASIS	1164	-0.163	
Low SAPS II	9629	-0.052	$p < 0.001$
High SAPS II	1164	-0.205	
Low Noncompliance Mistrust	9629	-0.054	$p < 0.001$
High Noncompliance Mistrust	1164	-0.224	
Low Autopsy Mistrust	9629	-0.033	$p < 0.001$
High Autopsy Mistrust	1164	-0.373	

most extreme negative sentiments. The median trust gaps (0.17 and 0.34) are larger than the severity gaps (.105 and .153), which are in turn larger than the race gap (.047). This further suggests that the mistrust metrics are able to tease out the cases with poor caregiver interactions and impressions.

6.1.1 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, I examine two trust-associated outcomes (*Code Status*[1] and *Whether the patient leave Against Medical Advice (AMA)*) and one more standard outcome (*in-hospital mortality*). I am 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. I take the average AUCs of 100 runs from randomly chosen 60/40 train/test splits, trained using a L1-regularized logistic regression model.

The results can be found in Table 6.2, which show that race and trust both improve outcome prediction in the tasks. Performance is variable across the tasks: no single feature is most useful for all three tasks. As is often the case, combining

[1] either *"Full Code"* or *"DNR / DNI / Comfort Measures Only"*

Table 6.2: 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

all of the features achieves the best results on each task — sometimes even with statistical significance, as for in-hospital 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).

The average classifier weights are shown in Table 6.3. 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. This agrees with previous studies which demonstrate the positive association between doctor-patient trust and favorable outcomes.

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). In general, noncompliance has the highest coefficient values of the three mistrust metrics, suggesting that it

captures some meaningful aspects of the doctor-patient relationships. On the other hand, 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 racial features add little predictive value or are dropped altogether. Age, however, is a very powerful predictor of these various outcomes, though not always as an indicator of breakdowns in the caregivers' relationship. For instance, while older patients are more likely to expire in-hospital, they are less likely to leave the hospital against medical advice. The mistrust scores — especially noncompliance-derived mistrust — is the only feature positively associated with each outcome, and consistently demonstrates predictive value.

Table 6.3: 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

Chapter 7

Conclusion

7.1 Future Work

This thesis performs an initial examination for the relationship between trust and treatment disparities. Because trust is such a broad notion, I offer initial investigations into the value of trust-based analysis but do not claim to have the best possible definition for measuring it. In addition, there is much more work to be done to understand how interpersonal relationships affect a patient's care.

7.1.1 Mistrust Metric can be Improved

This thesis presents preliminary work in modeling the mistrust between a doctor and patient, and using that mistrust to quantifiably explain racial disparities in care. As discussed earlier, I use three trust proxies in this work to demonstrate the utility of such a score, but none of the metrics fully capture all of the subtleties of the relationship.

Using clinical notes as labels and text-for-sentiment-analysis have significant shortcomings, because all of the content is generated by the caregiver. Occasionally a note might have a quote from a patient — though I do not consider that separately in this work — but that is very rare. In reality, trust is not a symmetric relationship: a patient can be mistrustful of their doctor, whereas the doctor could be oblivious

and assume all is fine between them. In order to truly measure a patient's feelings of how comfortable they are with their care, we would need a dataset which hears from them directly.

Further, although the supervised ML framework for tuning `chartevents` offers an intuitive appeal, it faces the risk of overfitting to a particular proxy target. It is possible that better-calibrated trust scores may be achieved with non-ML methods or perhaps that a ML-based method would benefit from multitask prediction of many trust-associated targets.

7.1.2 Sensitive Variable Protection

Resent work in bias and fairness for machine learning has explored the possibility of protecting from discrimination on the basis of pre-defined "sensitive variables" [22, 64, 63]. Such methods have not yet been explored in the medical domain, but this work could provide first step towards quantifying variables such as trust and communication. Perhaps it may prove useful to treat such scores as protected "sensitive variables" as well, though even if not, such scores could be incorporated into the creation of value-based care metrics for better understanding and evaluating healthcare delivery.

7.2 Conclusion

The goal of the work described in this thesis was to replicate previously observed racial disparities in healthcare treatments using a public dataset, and to investigate the role that patient mistrust plays in exacerbating those differences. This work accomplishes these two aims: first by showing racial disparities in two separate datasets and then by demonstrating that proxy metrics for trust are able to model longer aggressive durations and poor outcomes more effectively than race does.

To investigate racial differences in aggressive treatment durations, I use two datasets: MIMIC III and the Philips eICU Collaborative Research Database. In particular, I look at the duration of time that a given patient is on mechanical ventilation

and receives vasopressors during end-of-life care. I find significant differences in treatment for ventilation usage, and consistent differences for vasopressors. The result was the same for each case: black patients received higher levels of aggressive treatment during end-of-life. Previous researchers who have observed this phenomenon have suggested that these disparities are actually not driven by doctor-originating implicit bias, but rather come from mistrust that patients hold about their doctors and the healthcare system.

Even when controlling for severity (using the OASIS and Apache acuity scores), signs of these treatment gaps persist. However, for higher levels of severity, the gaps did begin to disappear. One reason for this might be that when the patient comes in at critical status, the main goal is to stabilize them before entering EOL discussions, whereas low-severity patients are earlier able to use their discretion for whether they want to pursue aggressive treatments.

In an attempt to better understand the effects that mistrust has on end-of-life care, I model mistrust with three possible metrics derived from coded interpersonal variables and clinical notes. The first two metrics used coded variables (such as pain assessment, healthcare literacy, family communication, etc.) as features in a supervised machine learning framework to predict note-derived labels of whether the patient was ever noncompliant or whether the patient consented/declined an autopsy. The third metric used standard sentiment analysis tools to determine how negative the written caregiver descriptions were of the patient's hospital stay. The noncompliance-derived and negative sentiment scores both showed statistically significantly higher levels of mistrust among black patients than white patients.

Stratifying patients by trust score instead of by race shows even larger disparities in aggressive end-of-life treatments. Once again, this trend holds when controlling for severity of illness. In addition, trust-based disparities had larger differences in sentiment from clinical notes. Finally, using the mistrust scores as features in predictive models added more value than adding race-based features — though a combination of both race and mistrust features performed the best, suggesting they captured complementary information.

Racial disparities in healthcare and machine learning are being studied at increasing levels. In addition to issues of implicit and structural bias, these studies will need to confront questions of culture and preference. Of particular import is assessing communication and informed consent in medical procedures. These goals are very difficult, and will require collaborations between data scientists, physicians, social scientists, and possibly patients themselves. It is my hope that this work can serve as preliminary steps for machine learning in trying to quantify and assess the relationship between trust and medical care.

Appendix A

Strict EOL Results

In this appendix, I repeat the experiments from Chapters 3 (Race-based Treatment Disparities) and 5 (Trust-based Treatment Disparities) using the stricter definitions of EOL. The broad definition – used throughout the thesis – includes patients discharged to skilled nursing facilities, which includes additional patients on hospice who were not labeled as such in the EHR. Doing so roughly doubles the datasize, but at the cost of introducing some patients who are not truly EOL.

Here, I demonstrate that the analyses in this thesis are robust to this definition choice. In essence, adding more patients very often does not change the comparative population differences, but does add statistical power.

Table A.1: Racial disparities in MIMIC comparing strict and broad definitions of end-of-life. Red p-values indicate statistical significance. Blue cohorts indicate the strict and broad cohorts agree on which race received higher durations of treatment.

	Strict				Broad			
cohort	**black n**	**black median**	**white median**	**p-value**	**black n**	**black median**	**white median**	**p-value**
vent	234	5825	5190	0.300	510	3286	2454	0.005
vent low-risk	76	7485	5700	0.110	178	3140	1530	< 0.001
vent medium-risk	87	5651	5340	0.448	153	2700	2580	0.410
vent high-risk	71	3825	4102	0.343	179	2550	2370	0.333
vaso	240	3146	2670	0.203	453	1707	1601	0.122
vaso low-risk	90	2969	2520	0.149	169	1570	1329	0.132
vaso medium-risk	73	3449	2943	0.386	136	1949	1770	0.459
vaso high-risk	77	3146	2585	0.475	142	1875	1780	0.154

A.1 Racial Treatment Disparities in MIMIC

Table A.1 shows that the broader EOL definition brings the cohort sizes from 234 and 240 for ventilation and vasopressors, respectively, up to 510 and 453. In doing so, two results become statistically significant while still maintaining the black-white treatment gap (i.e. for the strict vent-low cohort, there is a gap of 7485-5700=1785 minutes, whereas for the broader cohort the gap is 3140-1530=1610 minutes). The two definitions agree on which race received higher levels of treatment for 7 of the 8 comparisons, with the only disagreement being the high-risk ventilation cohort.

As is the case for each strict-vs-broad comparison, the strict population has higher levels of aggressive treatment because adding not-truly-EOL patients to the cohorts does bring down the average. However, as noted above, the effect of these "noise" patients does not seem to affect one race more than the other. As a result, using the broader group allows us to make conclusions from a larger data source without deviating from the signal of the stricter cohort.

Table A.2: Racial disparities in eICU comparing strict and broad definitions of end-of-life. Red p-values indicate statistical significance. Blue cohorts indicate the strict and broad cohorts agree on which race median received higher durations of treatment.

	Strict				Broad			
cohort	**black n**	**black median**	**white median**	**p-value**	**black n**	**black median**	**white median**	**p-value**
vent	383	3050	2429	0.005	655	2790	2235	< 0.001
vent low-risk	105	4906	3653	0.094	166	3798	2622	0.033
vent medium-risk	127	2700	1509	0.274	217	3326	2822	0.004
vent high-risk	151	2750	1542	< 0.001	272	3050	1969	< 0.001
vaso	300	1517	1478	0.486	464	1818	1620	0.422
vaso low-risk	78	2687	1058	0.116	129	2446	1794	0.223
vaso medium-risk	95	1287	1504	0.214	127	1745	1694	0.275
vaso high-risk	127	1217	1028	0.269	208	1377	1327	0.391

A.2 Racial Treatment Disparities in eICU

Just as with MIMIC, eICU is not very sensitive to the choice of EOL definition. Table A.2 shows that the two definitions again agree on 7 of 8 population comparisons. While two cohorts were already statistically significantly different for the strict definition, almost doubling our data size (105-to-166 and 127-to-217) results in two more significant racial disparities for the broader cohorts.

Table A.3: Noncompliance-derived mistrust disparities in MIMIC comparing strict and broad definitions of end-of-life. Red p-values indicate statistical significance. Blue cohorts indicate the strict and broad cohorts agree on which group median received higher durations of treatment.

	Strict				Broad			
cohort	**mistrust n**	**mistrust median**	**trust median**	**p-value**	**mistrust n**	**mistrust median**	**trust median**	**p-value**
vent	234	9540	4945	< 0.001	510	4680	2100	< 0.001
vent low-risk	83	11940	5655	< 0.001	153	4811	1440	< 0.001
vent medium-risk	82	10520	5190	< 0.001	197	4360	2460	< 0.001
vent high-risk	69	6928	3825	0.006	160	4890	2400	< 0.001
vaso	240	3765	2750	0.008	453	2070	1610	0.001
vaso low-risk	97	3533	2567	0.082	161	1650	1320	0.005
vaso medium-risk	75	4228	2954	0.018	151	2283	1740	0.034
vaso high-risk	68	3146	2701	0.233	141	2240	1770	0.191

A.3 Noncompliance-derived Treatment Disparities

In Table A.3, we can see that the choice of EOL definition had hardly any impact: they agree on 8 of 8 cohorts, and only one additional cohort – low-risk vasopressors – shows significant racial disparities after increasing the size of the dataset. Given that this cohort was almost significant with just 98 samples in the black population, using the broader EOL dataset did would not change the results or conclusions much, if at all.

Table A.4: Autopsy-derived mistrust disparities in MIMIC comparing strict and broad definitions of end-of-life. Red p-values indicate statistical significance. Blue cohorts indicate the strict and broad cohorts agree on which group median received higher durations of treatment.

	Strict				Broad			
cohort	**mistrust n**	**mistrust median**	**trust median**	**p-value**	**mistrust n**	**mistrust median**	**trust median**	**p-value**
vent	234	7480	5095	0.002	510	3690	2131	< 0.001
vent low-risk	98	8165	5666	0.002	182	4260	1410	< 0.001
vent medium-risk	73	5700	5301	0.280	190	3585	2545	0.006
vent high-risk	63	4771	3994	0.168	138	2625	2520	0.170
vaso	240	2955	2791	0.204	453	1845	1620	0.059
vaso low-risk	101	3075	2568	0.068	137	1935	1320	0.025
vaso medium-risk	73	2970	3090	0.330	169	1980	1745	0.111
vaso high-risk	66	2540	2805	0.432	147	1837	1710	0.156

A.4 Autopsy-derived Treatment Disparities

In, Table A.4 the two EOL definitions again agree on 6 of 8 population comparisons.

Two additional cohorts attain significant racial differences in the broad dataset.

Table A.5: Sentiment-derived mistrust disparities in MIMIC comparing strict and broad definitions of end-of-life. Red p-values indicate statistical significance. Blue cohorts indicate the strict and broad cohorts agree on which group median received higher durations of treatment.

	Strict				Broad			
cohort	**mistrust n**	**mistrust median**	**trust median**	**p-value**	**mistrust n**	**mistrust median**	**trust median**	**p-value**
vent	224	4335	5301	0.475	492	2805	2235	< 0.001
vent low-risk	68	4030	5890	0.260	146	1800	1560	0.171
vent medium-risk	73	4611	5625	0.425	203	2888	2558	0.008
vent high-risk	83	4337	4020	0.135	143	3585	2550	< 0.001
vaso	226	2548	2849	0.148	427	1605	1665	0.241
vaso low-risk	62	2160	2715	0.073	107	1320	1341	0.152
vaso medium-risk	84	2970	3210	0.346	143	1845	1705	0.282
vaso high-risk	80	2548	2820	0.410	147	1730	1965	0.353

A.5 Sentiment-derived Treatment Disparities

The sentiment-based trust score shows the most sensitivity to definition choice, as indicated in Table A.5. Only half of the rows agree, and two of those disputed attain significant racial disparities as large groups.

For the strict definition, the full ventilation and medium-risk ventilation patients with very positive sentiment received longer durations of treatment. However, when adding additional patients for the broader group this trend flips; the very positive sentiment patients receive significantly lower treatments than their negative sentiment counterparts. This might be attributable to the not-truly-EOL patients having notes which use fewer critical terms, thus allowing more non-critical patients to end up in the "trustful" (i.e. more positive) population.

Table A.6: Racial disparities in mistrust scores for strict and broad definitions of end-of-life. Red p-values indicate statistical significance. Blue cohorts indicate the strict and broad cohorts agree on which race median received higher durations of treatment.

	Strict				Broad			
cohort	**black n**	**black median**	**white median**	**p-value**	**black n**	**black median**	**white median**	**p-value**
noncompliance	386	0.21	0.20	0.492	1202	0.16	0.09	0.001
autopsy	386	-0.42	-0.31	0.117	1202	-0.42	-0.42	0.126
negative sentiment	370	0.226	0.213	0.111	1173	0.11	0.06	0.007

A.6 Racial Disparities in Trust

Table A.6 shows the distribution of mistrust scores for the white populations and the black populations. All but one metric are consistent across EOL definitions. We can see that for the two metrics that do agree, both increase their size threefold and attain significant racial disparities in mistrust for the broad groups.

Bibliography

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

[2] Lawrence K. Altman. Health: Hospital policy; sharp drop in autopsies stirs fears that quality of care may also fall. New York Times, 1988.

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

[4] Willie Boag, Elena Sergeeva, Saurabh Kulshreshtha, Peter Szolovits, Anna Rumshisky, and Tristan Naumann. Cliner 2.0: Accessible and accurate clinical concept extraction. In *NIPS 2017 Workshop on Machine Learning for Health Workshop*, 2017.

[5] Karla L. Caballero Barajas and Ram Akella. Dynamically modeling patient's health state from electronic medical records: A time series approach. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 69–78. ACM, 2015.

[6] CDC. Prostate cancer rates by race and ethnicity. `https://www.cdc.gov/cancer/prostate/statistics/race.htm`, 2014.

[7] Edward A. Chow, Henry Foster, Victor Gonzalez, and LaShawn McIver. The disparate impact of diabetes on racial/ethnic minority populations. *Clinical Diabetes*, 30(3):130–133, 2012.

[8] Marlene Cimons. CDC says it erred in measles study. Los Angeles Times, June 17, 1996.

[9] 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.

[10] Adolfo Gabriel Cuevas. *Exploring Four Barriers Experienced by African Americans in Healthcare: Perceived Discrimination, Medical Mistrust, Race Discordance, and Poor Communication.* PhD thesis, Portland State University. Department of Psychology, 2013.

[11] T. De Smedt and W. Daelemans. Pattern for python. In *Journal of Machine Learning Research*, volume 13, page 2031âĂŞ2035, 2012.

[12] Danielle M. Ely, Anne K. Driscoll, and T.J. Matthews. Infant mortality rates in rural and urban areas in the united states, 2014. 2017.

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

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] Ary L. Goldberger, Luis A. N. Amaral, Leon Glass, Jeffrey M. Hausdorff, Plamen Ch, Roger G. Mark, Joseph E. Mietus, George B. Moody, Chung-Kang Peng, and H. Eugene Stanley. PhysioBank, PhysioToolkit, and PhysioNet. *Circulation*, 101(23):e215–e220, June 2000.

[19] 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.

[20] Paulina Grnarova, Florian Schmidt, Stephanie Hyland, and Carsten Eickhoff. Neural document embeddings for intensive care patient mortality prediction. In *NIPS 2016 Workshop on Machine Learning for Health Workshop*, 2016.

[21] 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.

[22] Moritz Hardt, Eric Price, Nati Srebro, et al. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pages 3315–3323, 2016.

[23] Sadid A. Hasan, Bo Liu, Joey Liu, Ashequl Qadir, Kathy Lee, Vivek Datla, Aaditya Prakash, and Oladimeji Farri. Neural clinical paraphrase generation with attention. In *Proceedings of the Clinical Natural Language Processing Workshop at COLING 2016*, 2016.

[24] 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.

[25] Jonathan L. Hecht, Jeffery Joseph, Peter Ciano, William C. Quist, and Melissa Upton. The autopsy manual. The Beth Israel Deaconess Medical Center- Department of Pathology, 2006.

[26] 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.

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

[28] Donna L. Hoyert. The changing profile of autopsied deaths in the united states, 1972âĂŞ2007. `https://www.cdc.gov/nchs/data/databriefs/db67.pdf`, 2011.

[29] Thomas Jefferson. Correspondence with John Vaughn. Reprinted in Vaughn, found at Massachusetts Historical Society., November 1801.

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

[31] Alistair EW Johnson, Tom J Pollard, and Roger G Mark. eICU collaborative research database, 2017.

[32] 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.

[33] 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.

[34] Kimberly S Johnson. Racial and ethnic disparities in palliative care. *Journal of palliative medicine*, 16(11):1329–1334, 2013.

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

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

[37] 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.

[38] Barron H. Lerner. Scholars argue over legacy of surgeon who was lionized, then vilified. New York Angeles Times, October 28, 2003.

[39] 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.

[40] Shoichi Maeda, Etsuko Kamishiraki, Jay Starkey, and Noriaki Ikeda. Why are autopsy rates low in japan? views of ordinary citizens and doctors in the case of unexpected patient death and medical error. *Journal of healthcare risk management : the journal of the American Society for Healthcare Risk Management*, 33(1):18–25, 2013.

[41] Thomas H. McCoy, Victor M. Castro, Andrew Cagan, Ashlee M. Roberson, Isaac S. Kohane, and Roy H. Perlis. 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.

[42] 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.

[43] Olga Patterson and John F. Hurdle. Document clustering of clinical narratives: a systematic study of clinical sublanguages. 2011:1099–107, 01 2011.

[44] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *The Journal of Machine Learning Research*, 12:2825–2830, 2011.

[45] 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.

[46] 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.

[47] 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.

[48] Robert S. Pritchard, Elliott S. Fisher, Joan M. Teno, Sandra M. Sharp, Douglas J. Reding, illiam 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.

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

[50] Dorothy Ruiz. Epidemiology of schizophrenia: Some diagnostic and sociocultural considerations. 138:315–326, 1983.

[51] Anna Rumshisky, Marzyeh Ghassemi, Tristan Naumann, Peter Szolovits, Victor Castro, Thomas McCoy, and Roy Perlis. Predicting early psychiatric readmission with natural language processing of narrative discharge summaries. In *Translational Psychiatry*, 2016.

[52] Guergana K. Savova, James J. Masanz, Philip V. Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C. Kipper-Schuler, and Christopher G. Chute. Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513, 2010.

[53] Guergana K. Savova, Martha S. Palmer, Rodney D. Nielsen, James J. Masanz, James H. Martin, Wayne H. Ward, and Brian L. Cairns. The mipacq clinical question answering system. In *AMIA Annual Symposium proceedings*, volume 2011, pages 171–180, 2011.

[54] 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.

[55] 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.

[56] 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. PMID: 18757326.

[57] Michael Subotin and Anthony R. Davis. A system for predicting icd-10-pcs codes from electronic health records. In *Proceedings of the 2014 Workshop on Biomedical Natural Language Processing (BioNLP 2014)*, 2014.

[58] Ozlem Uzuner, Brett R. South, Shuying Shen, and Scott L. DuVall. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556, 2011.

[59] Harriet Washington. *Medical Apartheid: The Dark History of Medical Experimentation on Black Americans from Colonial Times to the Present.* 2007.

[60] David R. Williams. Race, socioeconomic status, and health the added effects of racism and discrimination. *Annals of the New York Academy of Sciences*, 896(1):173–188, 1999.

[61] 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.

[62] 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.

[63] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. Fairness constraints: Mechanisms for fair classification. *arXiv preprint arXiv:1507.05259*, 2017.

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