Deadly Stigma*

Manasvini Singh

Carnegie Mellon University

Dept of Social and Decision Sciences

msingh01@cmu.edu

DRAFT IN PROGRESS – PLEASE DO NOT CIRCULATE

Feb 2, 2026

Abstract

Stigma — shared negative beliefs about traits that violate social norms — can distort resource allocation and outcomes. Yet stigma is difficult to study in real-world settings because it is hard to measure and typically entangled with the informational content of the stigmatized trait. I address these challenges using a regression discontinuity design that exploits a potentially discrete shift in a widespread form of stigma — weight stigma — in the high-stakes setting of healthcare. Leveraging the fact that BMI categories are prominently displayed in electronic medical records, I document a discontinuous increase in in-hospital mortality at the "obese" cutoff. Evidence suggests stigma-based discrimination as a mechanism. Just-obese patients receive fewer diagnostic orders, procedures, and lab tests than almost-obese patients, consistent with reduced clinical effort. Stigmatizing language in clinical notes (identified with a validated LLM) rises at the cutoff for high-stigma providers, closely tracking the observed mortality effects. Two defining features of stigma are also present: stronger effects when norm deviation is greater, and a rise in judgmental language at the cutoff.

*I gratefully acknowledge feedback from seminar participants at Carnegie Mellon University (Behavioral Economics Group), Harvard Business School (Negotiations and Management Group), the Behavioral Science and Health Symposium at the University of Pennsylvania, Indiana University (Kelley School of Business), and the Behavioral Economics Design Initiative (BEDI) Conference on AI, Health, and Well-Being. The use of em dashes in this paper — despite ChatGPT's notoriety for overusing them — should not be taken as evidence of LLM involvement; I just happen to love them.

1 Introduction

Stigma is a shared set of negative beliefs about an individual trait that violates a social norm. Crucially, the trait is perceived to be voluntarily assumed and thus invites moral judgment and blame, both from others and the self. Stigma is a powerful and distortionary social force. On one side, it causes individuals to undertake personally costly actions to avoid being stigmatized: those eligible for welfare may decline benefits, and people with stigmatized health conditions (such as HIV or mental illness, (Smith, n.d.; Yang et al., 2021)) may avoid lifesaving care. On the other side, it leads individuals to discriminate against those seen as norm violators, even when the norm is arbitrary and the violation was not a choice. Employers, for instance, may discriminate against individuals with a history of unemployment or those who are overweight (Biewen & Steffes, 2010). In a world saturated with such norms — and thus endless opportunities to stigmatize — it is essential to understand how stigma shapes resource allocation and outcomes in real-world settings.

Yet, the rigorous study of stigma is complicated by two empirical challenges. The first challenge, familiar from the discrimination literature, is that the stigmatized trait may contain meaningful information about the individual. As a result, differences in outcomes in the data may reflect differences in the signal rather than arising from stigma per se. Identifying the effect of stigma therefore requires a research design that isolates exogenous variation in stigma while holding the true signal constant. Second, stigma is hard to measure: its defining feature — moral judgment — is rarely observable in administrative data. As a result, much of the literature relies on experimental methods, which are well suited to demonstrating the existence of stigma and isolating its psychological antecedents, but less equipped to capture its real-world dynamics. Outside of experimental settings, researchers often rely on self-reported experiences of stigma, which are prone to bias: individuals may report experiencing stigma where none exists, and may fail to recognize or report their own stigmatizing behavior.[1] Thus, I turn to healthcare — a high-stakes setting with unique institutional features — to overcome these challenges and identify the causal effects of a common and highly consequential form of stigma, namely, *weight stigma.*

Few social norms are as universal and enduring as the one concerning body size. Across cultures and throughout history, there is usually a "socially acceptable" body size, and those who

[1] Only a handful of observational studies have data rich enough to move beyond self-reports. For example, Celhay, Meyer, and Mittag (2025) link survey-based measures of welfare participation with actual welfare take-up and show that the gap is larger in high-income areas — consistent with the idea that stigma, rather than eligibility or need, depresses program participation.

deviate from it are often subject to moral judgment. For instance, across a range of contemporary settings, individuals with obesity are perceived and portrayed as lazy, clumsy, of lower ability, or burdensome to society.[2] This stigma is also well-documented in health care, where the belief that obesity is the result of personal failure (e.g., due to poor self-control over diet and exercise) can shape how obese individuals are treated by clinicians and how they engage with the system. Beyond its social importance, however, what makes this setting particularly attractive for the empirical study of stigma is the use of Body Mass Index (BMI) categories in medicine — a formal classification system that both reflects and reinforces norms around body size. Higher BMI categories, which more visibly violate these norms, carry stigmatizing labels such as "obese" and are also prominently displayed for the provider in the patient's medical record. I exploit the salience of BMI cutoffs using a Regression Discontinuity Design (RDD) and, leveraging recent advances in large language models to extract stigma from unstructured clinical text, causally link stigma to the highest-stakes outcome possible: death.

The intuition behind this research design is straightforward. The term "obese" is stigmatizing — both in health care and in broader social contexts. In a patient's medical record, certain information is persistently and saliently displayed, and BMI is one such variable. Importantly, beyond the continuous BMI value, categorical BMI distinctions (e.g., "normal", "overweight", "obese" etc) are usually made salient, for example, through explicit labeling, color-coding, or other visual cues. Given that people tend to over-rely on categories a the expense of continuous information due to cognitive constraints, I hypothesize that the formal BMI cutoffs will produce discontinuous increases in provider stigma at the obesity threshold *without a corresponding discontinuous change in the underlying signal* (i.e., health status). This, in turn, may lead to discontinuous changes in provider behavior and patient outcomes. This paper presents a body of evidence supporting this hypothesis.

For this research study, I use the Medical Information Mart for Intensive Care IV (MIMIC-IV) – a publicly available, de-identified electronic health record dataset from Beth Israel Deaconess Medical Center in Boston that covers all patients admitted to the emergency department or an intensive care unit between 2008 and 2022. MIMIC-IV contains detailed structured clinical data (e.g., diagnoses, procedures, labs, and orders) and de-identified free-text notes, enabling analysis of both clinical decisions and downstream outcomes. Unlike administrative data most commonly used in research on healthcare settings, this dataset has two variables critical for this study: patient BMI, and free-text

[2] Citations regarding the workplace, education, personal relationships, and media.

clinical notes written by the provider and care team, specifically, the hospital discharge summary. I thus limit my analytic dataset to all patients admitted to the hospital with a non-missing BMI.

I begin with the outcome of interest – death – and work backward to uncover the underlying mechanism. First, I show that, using an RDD and the Calonico et al. local non-parametric specification, in-hospital mortality increases by 40% at the obesity cutoff. I then establish the design's validity using standard diagnostics: McCrary density tests show no evidence of manipulation around the BMI threshold; baseline covariates predictive of mortality — including race, gender, age, comorbidity burden, insurance status, and predicted mortality — are balanced across the cutoff; placebo tests at alternative thresholds reveal no discontinuities; and a donut-hole RDD, which accounts for potential rounding near the cutoff, yields similarly significant results. This result is highly robust to linear and nonlinear parametric specifications, covariate inclusion, and alternative choices of kernel and clustering. (I also test for analogous mortality effects at the other BMI category cutoffs and find none.)

What could be driving the jump in mortality at the obesity cutoff? The RDD tests above reveal no underlying change in health status that could explain it. A natural explanation is weight stigma. The "obese" label may trigger moral judgment and cause providers to overfixate on a patient's body weight. This may disrupt the diagnostic process by leading providers to invest less time, effort, or clinical resources than necessary to reach an accurate diagnosis, ultimately resulting in worse patient outcomes. A wealth of qualitative evidence — both anecdotal and academic — supports this theory. Individuals with greater weight frequently report that their symptoms are discounted (often patronizingly), necessary tests are withheld, and their concerns are dismissed, sometimes with severe consequences. For example, one user on Reddit writes:

> *"I'm on the way to the doctor for the 5th time this month ... and all I keep hearing is "have you tried losing weight?".... They don't even do any tests or X-rays either. Awhile back I complained about a cough and they just said I have asthma. It wasn't until I ended up in the ER that they said it was actually my thyroid and I have to have surgery. It's like they don't take me seriously because I'm bigger."*

I examine two dimensions of the data to uncover evidence of this mechanism: (i) structured clinical decisions and (ii) unstructured clinical notes. If stigma is driving worse outcomes, we should observe some degradation in the quality of clinical decision-making for patients labeled "obese." I find

evidence consistent with this hypothesis in the first dimension I analyze — *structured* clinical data capturing provider orders, tests, labs, procedures, etc. Using the same RDD framework, I show that crossing the obesity cutoff leads to statistically significant and economically meaningful reductions in diagnostic and treatment intensity. Specifically, patients just above the cutoff receive fewer diagnostic orders in the first 24 hours of admission, undergo fewer procedures at the intensive margin, and are less likely to receive blood and urine labs, the most common labs administered during a hospital stay.

This pattern is suggestive of a stigma-based mechanism, though not, on its own, sufficient to prove it. An alternative explanation is that clinicians are not actively stigmatizing patients, but rather overemphasizing the "obese" label in their diagnostic process because they believe it conveys meaningful information about clinical risk. However, there are several logical and empirical reasons to doubt this interpretation. First, it is not obvious why an overemphasis on the "obese" label would lead to *less* clinical effort. Higher BMI is usually regarded a signal of elevated clinical risk, so one might expect providers to *increase* diagnostic intensity, not reduce it. Even if the label draws attention to weight-related conditions, this should result in a shift in diagnostic focus — not an overall reduction in effort and resource use.

Second, I conduct a placebo-style test using blood pressure (BP) categories and do not find similar results. BP classifications (e.g., "normal," "high," "extremely high") are similar to BMI categories in several key ways: they are saliently displayed in the medical record, reflect categorizations of continuous variables into risk categories, and are widely regarded by providers as clinically meaningful. However, they differ in one important respect — higher BP categories are not stigmatized the way higher BMI categories are. If the pattern observed at the obesity cutoff were driven by provider overfocus on salient clinical risk categories, we might expect similar reductions in care intensity at the high blood pressure threshold. But I find no such pattern. Crossing into a high BP category does not reduce clinical intensity or worsen outcomes. This provides suggestive evidence that there is something distinct about the "obese" label: it not only draws attention, but also appears to legitimize lower effort, reduced clinical investment, and ultimately worse care in a way that other salient clinical signals do not.

Third — and perhaps most compelling — I find direct evidence of stigma within the second dimension of data I investigate, i.e., *unstructured* clinical notes. I hypothesize that, if the "obese" label triggers moral judgment and stigma, it may be reflected in how providers write about patients. Specifically, crossing the obesity cutoff may make providers more likely to describe patients using negative

or stigmatizing language in the medical record. Fortunately, recent advances in large language models, particularly in the analysis of clinical text, make this test feasible. I focus on the discharge summary, a narrative free-text document written by the treating team that summarizes the patient's clinical course, key findings, and decision-making during the hospitalization. I apply methods developed and validated by Apakama et al. (2024), who design, refine, and validate — using physician evaluators — a zero-shot extraction prompt to identify potentially stigmatizing language in clinical documentation. This model has strong predictive performance, with a sensitivity of 95% and specificity of 86% for detecting biased languages in the clinical notes. It outputs all instances of stigma in a note, classifying each into one of four categories of biased languages (as defined by the large literature on stigma) — judgmental, stereotyping, negative labeling, and discrediting — along with a brief explanation of why it was flagged. An example of the kind of language it flags:

> *... malingering ... repeatedly attempted to manipulate the hospital system ... threatening with passive [suicidal ideation] if her demands were not met...*

For computational and feasibility reasons, I limit the analysis of clinical notes to all patients with a discharge summary whose BMI falls in the overweight or obese range, comprising approximately 39% of the full mortality sample (mortality results remain significant in this subsample). On average, using the RDD framework, I find no effect of the obesity cutoff on the overall likelihood of stigmatizing language in clinical notes. However, there is substantial variation across physicians in their use of such language. I calculate each physician's average rate of stigmatizing language across notes and divide providers into quartiles, with the first and fourth quartiles representing the least and most stigmatizing physicians, respectively. To examine heterogeneity in the treatment effect of the obesity cutoff, I implement a Differences-in-Discontinuities (Diff-in-Disc) design — identified under the same assumptions as the baseline RDD — comparing the effect of the BMI cutoff among Low-stigma providers (quartiles 1–3) and High-stigma providers (quartile 4) using Calonico, Cattaneo, and Farrell (2020)'s robust, bias-corrected inference with optimal bandwidth selection. The idea is that the cutoff may not increase stigma uniformly across providers, but may do so more for those with a greater underlying propensity to stigmatize — a form of behavioral heterogeneity commonly exploited in research (i.e., "judge designs").

Using the Diff-in-Disc design, I find two patterns consistent with a stigma-driven mortality mechanism: (1) the cutoff significantly increases the use of stigmatizing language more among High-

stigma providers relative to Low-stigma providers, and (2) the mortality effect of the cutoff is higher amongst the High-stigma providers than the Low-stigma providers. Moreover — though more suggestively — the mortality effect is larger for patients whose notes contain any stigmatizing language than for those whose notes contain none. Finally, an additional interesting result, unrelated to this LLM-based stigma measure, is that the obesity cutoff is associated with a decrease in note length, and note length is often used as a measure of provider effort.

I present further evidence in support of the stigma-driven mortality mechanism. The two defining features of stigma are that: (i) it targets individuals who violate a social norm, and (ii) it is rooted in moral judgment. I begin by testing the "norm violation" component using three separate analyses. First, I show that the effect of the obesity cutoff on stigmatizing language is concentrated among shorter individuals. The logic is that, conditional on BMI, an obese person who is shorter appears more visibly norm-violating than someone who is taller. Second, I show that the effect of the cutoff on stigma in clinical notes is larger for patients whose BMI is higher than what would be expected — or "normative" — based on a predictive model of BMI using observable characteristics such as race, gender, ethnicity, insurance status, and other demographics. The idea here is that physicians may mentally adjust body size norms based on patient characteristics, and thus stigma is more likely when a patient's weight exceeds that adjusted reference point. Third and finally, I examine how stigma in clinical notes varies with BMI categories within-patient and within-physician. Limiting the sample to patients with multiple hospital visits, I find that stigma moves in tandem (within-patient) with their BMI categorization: when the same patient moves into a higher BMI category relative to their prior admission, the likelihood of stigmatizing language in the note increases, holding constant physician-specific time-invariant characteristics. This last test is particularly illuminating because it highlights a key feature of stigma: since it is tied to a social norm, it can shift even within the same individual as they move closer to or further from that norm.

I next test the "moral judgment" component of stigma, and provide again three pieces of evidence poiting to this pathway. A distinctive feature of the model validated in Apakama et al. (2024) is that it not only identifies instances of stigma in a note, but also classifies each instance by *type* of biased language. One such category is "judgmental" language, which the model's creators explicitly describe as capturing expressions of moral judgment. First, I find that the Diff-in-Disc stigma results described above are primarily driven by this category of biased languages. Specifically, the cutoff increases the use of judgmental language significantly more among High-stigma providers

than among Low-stigma providers. This effect is not observed for other categories of descriptors. Second, I show that it is not merely the presence of judgmental language in isolation — the cutoff also increases the co-occurrence of judgmental terms and weight-related mentions in the same note, again more so for High-stigma providers than for Low-stigma providers. Third, recognizing that LLM-based classifiers may be viewed as a "black box," I conduct a simple manual search for commonly used judgmental clinical terms (e.g., "non-compliant"). While these manual flags generate more false positives than the LLM-based approach, I nonetheless find that the cutoff increases the use of such terms among High-stigma providers relative to Low-stigma providers.

Overall, the evidence of stronger effects for norm violators, combined with evidence that the cutoff increases moral judgment in clinical documentation, suggests that the obesity cutoff increases stigma — an effect concentrated among providers with a greater underlying propensity to stigmatize — and that this stigma can help explain the observed mortality results.

Contribution and Related Literature. This paper contributes to and extends two primary literatures: (i) on stigma, and (ii) on discrimination.

To begin with, it adds to a small but growing body of work in economics that studies *stigma*. This paper offers several novel contributions to that literature. First, unlike much of the existing work, which focuses on the experiences or behaviors of stigmatized individuals, this paper centers the stigmatizer — measuring stigma as it appears in the behavior and language of decision-makers. While prior research often documents how individuals alter their behavior to avoid social sanction, we rarely observe how those norms are enforced by others. The fact that the stigmatizing behavior is exhibited by healthcare providers — actors who are professionally and socially expected to act impartially and uphold norms of care — makes the setting especially striking.

Second, most studies examine the effects of stigma on intermediate outcomes such as program take-up or medication adherence. This paper instead shows that stigma can affect the final and arguably most consequential outcome: mortality.

Third, it captures stigma more directly than prior work, using written clinical documentation that reflects the provider's own words and judgments. This complements prior studies on stigma that rely on self-reports, perceptions, or inferred behavior.

Fourth, and most importantly, this setting allows for unusually clean identification. Much of

the stigma literature — both outside of[3], and within economics — struggles to separate stigma from the informational content of the stigmatized trait. For instance, the literature on unemployment stigma long suggested that long-term unemployment carried a penalty in hiring decisions. But a recent field experiment found little evidence of such effects once employer-side information was more tightly controlled. By leveraging a clinical labeling threshold that induces a discontinuous shift in how individuals are categorized — without a discontinuous change in their underlying health — this paper isolates the causal effect of stigma more cleanly than much of the existing literature.

The second literature this paper contributes to is discrimination. While the economics literature on discrimination is large and well-established, this paper expands it in two key ways. First, nearly all existing work focuses on what the sociology literature calls ascribed traits — characteristics such as race, gender, age, or disability status that are largely fixed and unchangeable. In contrast, this paper examines discrimination rooted in stigma, which is tied not to immutable traits but to perceived deviations from social norms. In doing so, it highlights a form of transient discrimination — one that can be triggered or reversed by social architecture, such as the labeling of BMI cutoffs in electronic medical records. This type of discrimination is particularly important because it is, at least in principle, more malleable. If discriminatory behavior arises from norm salience — for example, from highlighting a BMI threshold in clinical software — then it may be possible to reduce discrimination through changes in institutional design. This stands in contrast to racial or gender-based discrimination, where the underlying traits cannot be cast aside and interventions often face deeper structural constraints. More broadly, the paper underscores a double-edged feature of social norms: while norms can coordinate behavior and support cohesion, their overapplication — or misapplication in inappropriate contexts — can distort decision-making and reinforce inequality. This suggests that greater care is needed when mainstreaming norms, particularly in high-stakes settings like healthcare.

Second, this paper contributes to the discrimination literature by providing direct evidence of taste-based discrimination, where the taste in question is stigma. A standard distinction in economics separates discrimination into statistical discrimination, driven by incomplete information, and taste-based discrimination, where decision-makers are willing to incur costs to avoid associating with certain groups. Stigma-driven discrimination can be viewed as a particular form of taste-based discrimination: rather than animus in the classic Becker sense, homophily, or status threat, the relevant "taste" reflects

[3]For example, a widely cited PNAS paper interprets lower buyer interest in products from sellers in disadvantaged neighborhoods as evidence of neighborhood stigma. But if those neighborhoods are also, on average, further away or perceived as less safe, then the lower offer rate could reflect rational sorting rather than stigma.

negative moral judgments toward individuals perceived to violate a social norm. However, while a large body of work in economics documents patterns *consistent* with taste-based discrimination, it rarely captures these non-informational motives directly. Much of the evidence is necessarily indirect – showing that disparities persist when observable productivity is held constant, or that outcomes covary with proxies for prejudice or social distance. For instance, audit and correspondence studies (e.g., Bertrand and Mullainathan, 2004; Charles and Guryan, 2008) often stop at documenting disparities "consistent with" taste, without directly providing evidence of the discriminatory taste itself. In contrast, this paper provides more direct evidence of stigma-driven, taste-based discrimination by analyzing providers' language in clinical documentation. A discontinuous decline in clinical resource use *and* a discontinuous rise in biased language at the obesity cutoff reflect not only differential treatments consistent with taste-based discrimination, but also the underlying negative beliefs that motivate it – in particular, judgmental and discrediting language that statistical discrimination would not predict and that is typically unobservable in most economic settings.

2 Setting and Data

2.1 Setting

Body mass index (BMI) is a salient and routinely displayed component of the electronic medical record (EMR), and its inputs – height and weight – are routinely collected in clinical care. For example, when a patient enters a clinical setting, height and weight are typically measured at the beginning of the visit and recorded in the EMR, after which the system automatically computes BMI and assigns the patient to a BMI category that is clearly and saliently highlighted in the chart (for a generic example, see Figure A.2). Standard categories include underweight, normal weight, overweight, and obese. Some systems further subdivide obesity into classes I–III, although not all EHR interfaces display these subcategories, and it is unclear whether the system underlying my data does so.

These BMI categories create a useful setting for studying stigma because the labels signal whether – and by how much – a patient deviates from socially accepted norms of body size. For instance, "obese" is a clinical designation, but it is also a term used in everyday language to describe a socially stigmatized body size. As a result, crossing a clinical BMI cutoff assigns the patient a clinical label with a different degree of social stigma. The key empirical feature is that these body-size labels – and the stigma associated with them -=- change discretely at fixed BMI cutoffs, even though patient

underlying health risk varies smoothly with BMI. In other words, patients just below and just above a cutoff are likely similar in underlying clinical risk, yet they are assigned different labels in the medical record. This discontinuity motivates the paper's research design.

2.2 Data

I use MIMIC-IV v3.1 (Johnson et al., 2025; Johnson et al., 2023), a large deidentified electronic health record database covering all patients admitted to the emergency department or an intensive care unit at Beth Israel Deaconess Medical Center (Boston, MA) between 2008 and 2022. The database includes time stamped structured data generated in routine care – demographics, diagnoses and procedures, laboratory tests, bedside vitals, medications, and medication administration – organized in a relational format that supports transparent cohort construction and linkage across hospital modules. For instance, to measure stigmatizing language in documentation, I link a patient's structured hospital records in the module MIMIV-IV-*Hosp* (which covers 546,028 unique hospitalizations for 223,452 unique individuals) to their records in the module MIMIC-IV-*Note* (Johnson et al., n.d.), which provides deidentified free text clinical documentation (including discharge summaries and other note types) associated with hospital and emergency department encounters. While MIMIC-IV provides rich patient-level data, one downside is that its deidentification process limits cross-patient temporal comparisons. Dates are shifted consistently within each patient but not across patients, preventing alignment on a shared real-world timeline. The only calendar information available is a coarse year group indicator (e.g., 2008–2010, 2011–2013), which allows broad temporal controls but not precise time-based analyses.

I impose two sample restrictions. First, I limit the data to hospital stays, as only these encounters include discharge summaries, which are essential for identifying stigma in clinician language. Second, I restrict to stays with a non-missing BMI. As there may be multiple values of BMI during a patient's encounter, I use the first recorded value of height or weight. If both are present, I calculate BMI directly; if one is missing (usually height), I use the first BMI value reported in the structured data.[4] I assigned BMI categories using this continuous BMI variable and following standard clinical guidelines: (BMI < 18.5), normal ($18.5 \leq$ BMI < 25), overweight ($25 \leq$ BMI < 30), obese (also called class I) ($30 \leq$ BMI < 35), moderately obese (class II), ($35 \leq$ BMI < 40), and severely obese (class III)

[4] I do this because in MIMIC-IV data, height and weight are reported to two decimal place while BMI is reported to only one. In over 99% of cases, my calculated BMI matches the reported value but is more precise. Because clinical thresholds (e.g., BMI ≥ 30) rely on two-digit precision in the real-world, I use the calculated value when available. I also perform a donut-hole RDD as a robustness check against rounding artifacts.

(BMI $\geq$ 40). I drop all BMI values below 10 and above 60, as those are biologically extreme values (representing 0.7% of the data) and are likely data errors.

3 Mortality Effects of the Obesity Cutoff

Empirical Strategy. I use in-hospital mortality (i.e., deauth during the hospital visit) as my main outcome. Figure A.3 plots a histogram of BMI alongside raw in-hospital mortality rates, using bins of width 0.1. Most patients fall within the normal and overweight categories, while mortality rates begin to rise at the tails of the distribution. This pattern may reflect the underlying clinical reality documented in prior work – that both low and high BMI are associated with elevated health risk (Flegal et al., 2007; LUSKY et al., 1996) – or it may simply reflect greater noise and smaller sample sizes in the extremes of the distribution. I focus on the overweight-obese cutoff as the label "obese" is stigmatizing in a way no other BMI label is. However, I also present results using the overweight, obese II, and obese III cutoff as these represent deviations from the norm to varying degrees.

I implement a sharp regression discontinuity design that compares patients whose BMI falls just below versus just above 30 during the hospital stay. In this RD design, treatment is defined by the clinical obesity cutoff: patients with BMI $\geq$ 30 are classified as "obese" in the medical record (a potentially stigmatizing label, and thus treated), while patients with BMI $<$ 30 are classified as "overweight" (a less stigmatizing label, and thus the comparison group). Let $X_{it} \equiv \text{BMI}_i - 30$ denote the running variable for patient i in coarse year group t, and let $D_{it} = \mathbf{1}\{X_{it} \geq 0\}$ indicate being at or above the obesity cutoff. I estimate the following specification using the data-driven optimal bandwidth local-polynomial RD procedure of Calonico, Cattaneo, and Titiunik (2014) and Calonico et al. (2017):

$$Y_{it} = \gamma_0 + \tau\, \mathbf{1}\{\text{BMI}_{it} \geq 30\} + \gamma_2 f_1(X_{it}) + \gamma_3 f_2(X_{it})\, \mathbf{1}\{\text{BMI}_{it} \geq 30\} + \varepsilon_{it}, \tag{1}$$

where Y_i is an indictor for in-hospital mortality, and $f_1(\cdot)$ and $f_2(\cdot)$ flexibly capture the smooth relationship between BMI and mortality on either side of the cutoff (implemented via local polynomials within the chosen bandwidth).

The baseline specification uses a local linear fit ($p = 1$) with a triangular kernel and an MSE-optimal bandwidth (h_{mserd}), with robust, bias-corrected (RBC) standard errors clustered at the patient level. I report a range of robustness checks, including local quadratic fits ($p = 2$), estimates using

a uniform kernel, including age, sex, and race as covariates, and alternative clustering schemes. In particular, I two-way cluster by patient × year-group to accommodate the dataset's imbalanced panel structure, and by patient × clinical service (e.g., surgery, internal medicine, labor and delivery, etc) to capture within-service correlation in documentation and care. (Patient × service clustering is the primary specification for the stigma analyses in the mechanism section, since clinical notes are produced by clinical service teams.)

Results. In Fig II I depict the unadjusted relationship between raw mortality and BMI (in bins of 0.1). There is a clear jump in raw mortality at the obesity cutoff. This is confirmed by the formal RD analysis. Table 1 shows that, regardless of specification, the obesity cutoff increases mortality at the cutoff by approximately 0.4 percentage points (pp) or 40% ($p < 0.1 - p < 0.05$, depending on specification).

Results from re-estimating Equation 1 using the overweight, obese II, and obese III cutoff are presented in Table A.2. There are no mortality effects at any of these cutoffs. Not finding effects at the "overweight" cutoff is consistent with two potential considerations. First, overweight status is both more prevalent and less norm-violating, making it less likely to trigger stigma. Second, for individuals near the cutoff — that is, those included in the RDD analysis — the stigma associated with the "obese" label may plausibly be stronger than that attached to the "overweight" label, as individuals just above the obesity cutoff are more likely to display visible markers of norm violation (i.e., larger body size) than those just above the overweight cutoff. Finally, the null results at the obese II and obese III cutoffs are consistent with two possibilities: the EMR interface may not visibly differentiate these gradations, or the additional labels may not generate discrete changes in stigma beyond the initial transition into obesity.

Identification Assumptions. I implement a standard battery of RD validity tests to assess whether I can interpret this RD estimate as the causal effect of the obesity cutoff. First, I use the McCrary density test to test for a discontinuity in the distribution of BMI at the cutoff, which would suggest sorting or measurement heaping at BMI = 30. Manipulation is theoretically difficult, as BMI involves a complex calculation of a patient's height and weight, and neither the patient nor the provider can exert any pressure on the BMI value (and thus the associated category), especially in the short-run. Figure A.4 shows the data do not allow the rejection of the null of no sorting at the threshold (p = 0.16).

Second, I assess covariate balance using a summary index of baseline risk. I first estimate each patient's *predicted* mortality from a logit model that relates mortality to a rich set of predetermined covariates: age, race, sex, ethnicity, insurance status, marital status, the number of Elixhauser comorbidities, and the Elixhauser mortality index (a widely used and validated measure based on 31 high-risk comorbidities) (Elixhauser et al., 1998). Figure A.1 shows that predicted mortality closely tracks observed mortality, indicating that it captures clinically relevant risk. When I use predicted mortality as the outcome in the RD, the estimated discontinuity at BMI = 30 is close to zero and statistically insignificant, consistent with smooth underlying risk at the cutoff (see Column 1 in Table 2). In Appendix Table A.1, I also report RD balance tests for each covariate separately. Estimated discontinuities are small and statistically insignificant for all covariates except sex, where the share female rises modestly at the cutoff (significant at the 10% level). This pattern works against the main mortality result: females have lower mortality on average, so any discrete increase in the share female would, if anything, bias the estimated mortality effect downward. Consistent with this, the main mortality estimate is essentially unchanged when controlling for sex (see Table 1 Col 5), suggesting that the results are not driven by this minor imbalance.

Third, as a falsification exercise, I re-estimate the RD specification at placebo cutoffs away from the obesity cutoff, specifically, at the median values for the subsets of encounters below and above the obesity cutoff. These placebo thresholds should not trigger discrete changes in obesity labeling. Results are presented in Table 2 Columns (2) and (3), and show no effects of these placebo cutoffs. Finding no corresponding discontinuities at placebo cutoffs supports the interpretation that any effects observed at BMI = 30 reflect the clinical threshold rather than generic nonlinearities in the BMI–outcome relationship.

Finally, I estimate a donut-hole RD that drops encounters with BMI within ±0.1 of the cutoff to address potential heaping and rounding around BMI = 30. This ensures that small recording differences that could mechanically shift a patient across the threshold are not driving the results. The donut-hole estimates are presented in Table 2 Column (4) and are similar to the results from the main specification.[5]

[5]If anything, the estimate is larger in magnitude and more precisely estimated. This pattern is reassuring, because it says that even if there is some rounding-induced misclassification, it is *attenuating* the RD effect of the obesity cutoff on mortality – some patients whose true BMI during the encounter was just below 30 (and thus classified as overweight" in the EMR) appear at or above 30 in MIMIC-IV because BMI is recorded only to one decimal place. Categorizing treatment based on this rounded value mechanically assigns some untreated observations to the treated group, biasing the estimated mortality effect toward zero (since these patients do not receive the "obese" label, and the associated stigma, in practice – they are only misclassified in the research data).

4 Mechanism: Stigma

4.1 Conceptual support for mechanism.

What could be driving the jump in mortality at the obesity cutoff? The RD identification checks suggest that patients' underlying mortality risk is not changing discontinuously at the cutoff, so this cannot be explained by a discrete shift in baseline severity. One explanation – amongst others discussed after this section — is weight stigma, triggered by the label "obese" in the medical record. The obese label may cause providers to hyperfixate on a patient's weight and activate the moral judgment and blame that are hallmarks of stigma. Certainly, this is not a novel observation. There is a wealth of descriptive evidence – both anecdotal and academic – suggesting that weight stigma is common in healthcare settings, and may derail the diagnostic process and lead to worse outcomes.

Figures A.5 and A.6 present illustrative evidence from a range of Reddit subreddits (online spaces organized around a topic) – from patients and physicians, respectively – documenting how weight can shape the care process. Though it is entirely possible that weight is relevant for the underlying medical problem, patients often complain of their symptoms being taken seriously by providers because of their weight ("*what is with doctors blaming everything on weight*", Figure A.5 Panel (b)), having their symptoms ignored ("*I have so many symptoms...but nobody will diagnose me and they all say it's due to my weight*", Panel (f)), and being outright moralized against because of their weight ("*doctor made a comment assuming I eat a greasy meal everyday*", Panel (d)). One account that is particularly relevant to the stigma-based mechanism examined in this paper appears in Panel (b), where the author explicitly links her weight to adverse patient outcomes, in this case hospitalization: *"They don't do any tests or X-rays . . . It wasn't until I ended up in the ER that they said . . . I have to have surgery."*

On the other side, providers online discuss the challenges of caring for overweight and obese patients. Without denying the realities of providing care for such patients in a system where providers are overworked and underresourced, one can still identify language consistent with stigma in subreddits where physicians discuss medicine – describing patients using biased languages ("*obese patients can ruin your day*", Figure A.6 Panel (d)) and making moral judgments about lifestyle ("*I have yet to meet an obese patient who admits to eating poorly*", Panel (a)).

Finally, the prevalence of weight stigma is widely recognized by the medical workforce – the top three medical journals in the world, *JAMA*, *NEJM*, and *The Lancet*, have all published perspective

pieces explicitly validating and decrying weight stigma in healthcare (Kolata, 2016; Okwerekwu, 2016). There are also countless popular news articles such as in the New York Time) discussing how larger patients *may* receive worse care (Kolata, 2016; Okwerekwu, 2016), often drawing on descriptive studies suggesting potential pathways through which stigma could harm outcomes. For example, survey evidence suggests that obese patients may anticipate weight stigma and disengage from care, which could worsen health outcomes. Importantly for this paper, however, it does not appear that (a) weight stigma has been carefully identified via revealed provider behavior in a quantitative way, or that (b) its impacts have been causally linked to health outcomes beyond the general claim that stigma may be perceived and that perception may contribute to behaviors that harm health.

To investigate stigma as a potential mechanism, I examine two dimensions of the dataset. First, I analyze the structured data on clinical decision-making, including provider orders, diagnostic tests, labs, and procedures. If stigma is influencing care, one possible manifestation is that providers may discount patients' symptoms – attributing them to weight-related issues or deeming further diagnostic effort unwarranted. Such responses could plausibly lead to underdiagnosis or delayed treatment, offering a potential explanation for the observed increase in mortality at the obesity cutoff.This reduced clinical effort should be observable in the structured data, in the form of fewer diagnostic tests, procedures, or treatments ordered for patients just above the obesity cutoff.

Second, I turn to the unstructured data – specifically, the text of clinical notes like the discharge summary – to detect evidence of stigmatizing language using large language models. I examine whether crossing the obesity cutoff increases providers' use of biased language in clinical notes, which I interpret as evidence of stigmatizing documentation. Importantly, biased language need not reference weight explicitly and may just simply reflect a nontargeted bias against patients – consistent with prior work showing that people are rarely so explicit when discriminating against others, even in medicine. I discuss this pathway in the next subsection on the analysis of unstructured data.

This second dimension of data – unstructured notes – is central to establishing stigma as a mechanism. A reduction in clinical effort alone is not, by itself, definitive evidence of stigma; one alternative explanation is that the "obese" label simply shifts providers' diagnostic focus, leading them to mistakenly attribute symptoms to weight *without malicious intent.* In this view, the observed clinical patterns reflect misattribution of patient symptoms to a clinical signal they believe is informative, rather than stigma. I return to this possibility later in the paper, but the key empirical test is whether there is direct evidence of stigmatizing language in provider documentation. If present, such

language would distinguish stigma from benign misdiagnosis and strengthen the case that labeling alters provider perception and behavior in a harmful way.

4.2 Analysis of Structured Data

Clinical orders in the first 24 hours of arrival MIMC-IV captures all clinical orders placed for a patient. These structured records provide a detailed account of providers' diagnostic and treatment choices, including orders for laboratory tests, imaging, medications, and procedures, as well as their timing over the course of the encounter. I use these data to test whether patients just above the obesity cutoff receive less clinical effort and fewer resources in the first 24 hours after arrival. The motivation for choosing this time window is that the initial diagnostic workup often shapes the subsequent trajectory of care – early testing can trigger follow-up evaluation and treatment, whereas early inattention or underdiagnosis can generate cascades that are difficult to reverse. For example, if early signs of infection are missed, delayed recognition and treatment can allow rapid clinical deterioration, including progression to sepsis and death.

I create two variables capturing resource use within the 24 hours: total orders and diagnostic orders. Because MIMIC-IV contains hundreds of distinct order types and subtypes (e.g., laboratory orders such as CBC, BMP, troponin, lactate, and blood cultures; imaging orders such as chest X-ray, CT head, CT abdomen/pelvis, and echocardiography; cardiopulmonary diagnostics such as ECG and arterial blood gases; etc), I use an LLM (ChatGPT 5) to classify orders as diagnostic. I define diagnostic orders as those placed with the primary purpose of generating clinical information to evaluate, confirm, or rule out diagnoses, rather than directly treating the patient. This variable is thus an indicator for any order whose type is explicitly test-oriented – for instance, all "lab," "radiology," or "neurology" orders as classified as diagnostic orders, as well as several additional subtypes such as pulmonary function testing within respiratory orders; blood testing within blood bank orders; fetal monitoring within obstetric orders; and ECGs, echocardiograms, Holter monitoring, and exercise or treadmill stress tests within cardiology orders. I then split diagnostic orders into early versus total orders, defining early orders as those placed within 24 hours of hospital arrival.

In Figure III Panels (a) and (b) report RD estimates of the effect of the obesity cutoff on standardized measures of (i) total orders and (ii) diagnostic orders placed within the first 24 hours of arrival. Each panel presents the robust, bias-corrected RD estimate under five data-driven, optimal-bandwidth local-polynomial specifications. Specification (1) matches the baseline mortality analysis

(local linear, triangular kernel, patient-clustered standard errors, no covariates). Specification (2) uses a local quadratic fit ($p = 2$). Specification (3) adds predetermined covariates (race indicators, sex, private insurance, and age). Specification (4) uses two-way clustering by patient and coarse year-group. Specification (5) uses two-way clustering by patient and clinical service unit.

Point estimates are stable across specifications, although precision varies. Both panels tell a consistent story: the obesity cutoff generally yields a small, statistically insignificant reduction in total orders in the first 24 hours, but a statistically significant reduction in diagnostic orders in the first 24 hours. This pattern is consistent with reduced diagnostic workup at the threshold, which aligns with – though is not uniquely diagnostic of – a stigma-based mechanism in which the "obese" label lowers providers' effort to investigate patients' symptoms (alternative explanations are discussed later).

Volume of procedures. Another measure of clinical effort and resource use is the set of procedures performed during the hospital stay. Examples of the types of procedures that can occur in inpatient settings include interventions such as intubation and ventilation management, central or arterial line placement, dialysis/CRRT, blood transfusion, bronchoscopy, thoracentesis, chest tube placement, etc. Using the procedures recorded in MIMIC-IV, I construct two outcomes that capture the extensive and intensive margins: an indicator for whether the patient receives any procedure during their hospital stay, and the number of procedures conditional on receiving at least one. Figure III shows no discontinuity at the obesity cutoff in the extensive margin (Panel (c)). Panel (d) points to a modest reduction in the intensive margin that is estimated imprecisely and is statistically significant in some specifications. Overall, the cutoff does not affect whether a patient receives any procedure, but may result in fewer procedures among those who receive at least one. This pattern is consistent with reduced treatment intensity on the margin of additional procedures, in line with the broader evidence of lower resource use at the cutoff.

Lab tests. Because laboratory testing is among the most common diagnostic inputs in inpatient care, I next examine MIMIC-IV's laboratory tables to measure the depth of diagnostic evaluation. The lab tables record detailed information for each lab result, including the lab type (e.g., hematology, chemistry panels etc), specimen (fluid) type, the timing of collection, and the specific test performed. I focus on standard blood and urine testing, restricting to chemistry and hematology panels, which together comprise the core first-line diagnostic workup used to evaluate a wide range of inpatient presentations: chemistry panels capture metabolic status and organ function (e.g., electrolytes, glucose,

creatinine), while hematology panels capture blood-cell composition and related markers of infection, anemia, and bleeding risk (e.g., white blood cells, hemoglobin, platelets). I exclude rarer fluids such as ascites and cerebrospinal fluid, which are non-routine and often require invasive collection, and therefore reflect a different clinical margin than standard diagnostic workup.

Results are suggestive but imprecise across specifications. Overall, the estimates consistently point toward fewer of these common blood panels (both hematology and chemistry) and fewer urine tests (chemistry, but not hematology) at the cutoff. This pattern is consistent with a reduction in the breadth of diagnostic evaluation, reinforcing the broader evidence of lower diagnostic intensity around the obesity threshold.

Other measures of effort. To transparently report the full set of tests I conducted, I also examine consults, 30-day readmissions, and length of stay, and I find no statistically significant discontinuities at the cutoff for any of these outcomes.

4.3 Analysis of Unstructured Data

What would weight stigma in clinical notes look like? A purposefully extreme example would be something like: "this patient is fat, their illness is because of their weight, and their weight is their fault." In practice, however, stigma is unlikely to appear in such explicit terms – either as direct commentary about weight or as overtly stigmatizing language – because clinical settings impose strong professional and social constraints against overt disparagement (consistent with the physician norm of "do no harm"). When discriminatory motives are normatively unacceptable, research emphasizes that bias tends to surface indirectly [6] For these reasons, restricting attention to explicitly "anti-weight" language in clinical notes will likely miss much of the relevant behavior. The empirical expectation is that weight stigma manifests through more indirect and broader forms of biased documentation such as more negative affect; greater skepticism toward patient-reported symptoms; more judgment-laden descriptions of patient behavior (for example, terms like "noncompliant" or "nonadherent" may appear clinically descriptive, but often serve as shorthand for assigning blame to the patient while ignoring common structural barriers to adherence, such as cost, access, limited health literacy, etc); and lower thresholds for negative labeling in response to identical behaviors (e.g., a patient labeled "obese" may be more likely to be described as "belligerent" than an otherwise identical patient without the label

[6] See, for example, the justification–suppression model of **crandall˙justification˙2003** and the aversive racism framework of **dovidio˙aversive˙2000**; as well as literature showing that employers often refer to "fit" to justify discrimination.

exhibiting the same behavior). Thus, I specifically examine whether the obesity cutoff increases biased language in inpatient clinical documentation.

I rely on Apakama et al. (2024) to identify biased language in hospital discharge summaries. Discharge summaries are an important setting for studying biased language because they are a central and standardized component of clinical documentation, written at the conclusion of care to summarize the patient's course and are intended to be read by future providers. As such, they serve not only as a clinical summary but also as a narrative reflection of how the providers perceived and characterized the patient, making them a natural place for bias to manifest. Moreover, drafting of the discharge summary is typically undertaken by junior clinical team members – such as residents and interns – who are most directly involved in the day-to-day care of the patient. These drafts are then reviewed, edited, and/or signed off by the attending physician of record, after which the note becomes part of the permanent medical record (Kantor, Evans, & Shieh, 2015; Kind & Smith, 2008; Ng et al., 2025). In practice, this workflow means that discharge summaries largely reflect the documentation habits and cultural norms of the clinical service team responsible for the hospitalization, albeit with attending-level oversight. This makes discharge summaries a particularly natural note type for studying stigma in clinical documentation: they are authored by those most involved in patient care but also shaped by the documentation norms and standards of the clinical service unit (e.g., labor and delivery, surgery, internal medicine, etc) overseeing the hospitalization.

In this section, I first test whether the obesity cutoff increases biased language using an RD design. I do not find effects on average, but uncover important heterogeneities by provider type which mirror and thus could explain mortality patterns (using a Diff-in-Disc design). Then, I extend the analysis to test for more direct markers of weight stigma, focusing on the two defining elements emphasized in classic accounts of stigma: (i) moral judgment and (ii) norm violation, both discussed in greater details in the relevant subsections.

Measurement of Stigma. I follow (Apakama et al., 2024)'s paper in identifying instances of biased language in clinical documentation. In this paper, the authors identify biased language in clinical text using LLMs and categorize them into four conceptually distinct categories: discrediting (language that undermines patient credibility, e.g., "patient claimed"), judgmental (language that imposes moral judgment), pejorative labeling (negative labels applied to patients), and stereotyping (assumptions tied to demographics or social status). An institution-wide documentation initiative (the FAIR documen-

tation workgroup) developed these category definitions and reviewer procedures based on an extensive literature review and interviews with stakeholders (patients and providers), which are then used to structure the LLM prompt.

To detect biased language at scale, the authors use GPT-4 in a zero-shot prompting framework (no fine-tuning). They iteratively refine the prompt by reviewing model performance on a small set of sentences (100 EMR sentences balanced between biased and unbiased text), then apply the finalized prompt to a large sample of notes via the GPT-4 OpenAzure HIPAA-compliant API. The model is asked to flag whether a biased language ("an instance of bias") is present and, if so, which category it most likely belongs to. For validation, Apakama et al. (2024) compare GPT-4's classifications to independent clinician review. Two physicians reviewers independently assess whether GPT-4's bias flags and subtype assignments are correct. This validation step is especially useful for my setting because the authors includes MIMIC-IV discharge summaries to test and validate their prompt – the same note type used in my analysis. Their model had great validity: for "any bias" present, sensitivity was 95%, Specificity was 86%, positive predictive value was 84%, and NPV was 96%. Percentage agreement between GPT and human reviewers was 90%.

The full prompt is reproduced in Appendix Section A. Relative to 2024, , I make several minor modifications. First, I use OpenAI's o3 model rather than GPT-4, reflecting the availability of newer reasoning-oriented models at the time of analysis. Second, I explicitly instruct the model not to hallucinate. Third, I expand the examples of social groups that may be stereotyped against to include "obese or overweight people," to better align the prompt with the mechanism of interest. Fourth, I ask the model to focus on negative judgment (rather than judgment more generally), since positive judgments are not informative about stigmatizing documentation in this setting.[7] Due to computational constraints and feasibility, I restrict the stigma analysis to patients classified as overweight or obese, comprising approximately 39% of the full mortality sample (the main mortality results remain robust on this restricted sample).

The LLM prompt returns three outputs for each discharge summary: (1) a list of all flagged instances of biased language within the note; (2) a classification of each instance into one of the four bias types (judgmental, discrediting, labeling, or stereotyping); and (3) a brief rationale explaining why each instance was assigned to that category. I then create a variable called *Any Stigma*, which

[7] I also asked the LLM to give a quantitative measure of bias severity, and number of bias instances found, but I do not use them in this paper as they do not reveal any interesting results, likely because these are not validated in the way that the "any biased language" measure is.

is an indicator for whether the discharge summary contains at least one instance of biased language as identified by the LLM. This serves as the primary outcome for the unstructured data analysis, as it corresponds to the validated classification used in 2024, Later, I extend the analysis by examining each bias category separately as its own outcome, allowing me to assess which specific types of biased language are most affected by the obesity cutoff.

Definition excerpts and example LLM output for each bias type are presented in Table 3. As is clear, the LLM is able to pick up on clear nuanced of language that even the most sophisticated older text analysis tools would have been unable to. Moreover, as a descriptive analysis, Figure IV shows univariate correlations between the presence of Any Stigma and patient characteristics that prior literature suggests should be associated with differential documentation. Consistent with expectations, Medicaid insurance is positively correlated with stigma indicators, whereas private insurance is negatively correlated. Black patients exhibit higher rates of biased languages, and female patients exhibit lower rates relative to males, aligning with prior work suggesting that women are less likely to be explicitly stigmatized but more likely to be ignored. Being married is the strongest protective correlate against biased languages. While marital status may proxy for unobserved patient characteristics, it may also reflect the presence of a companion during the hospital stay, which could reduce provider irritation and reduce biased language; or alternatively, it may increase the credibility of patient reports and lower the prevalence of biased language, since a significant share of biased language is discrediting in nature and conveys doubt in patient reports.

The strongest positive correlates of stigma coding involve alcohol and drug use diagnoses, with roughly 30% of biased languages occurring among patients with an ICD-9/10 code for substance use, consistent with the high levels of documented bias toward these conditions. Mental illness is also a high predictor of stigma in clinical notes. To assess whether the stigma signal simply reflects illness severity rather than biased documentation, I examine correlations between biased language and three types of cancer diagnoses, which are generally regarded as sympathetic conditions; biased languages are notably less frequent in these groups. While these associations are correlational, their direction aligns with theoretical expectations and prior research on bias in clinical documentation. Surprising, having a diagnosis code of "obesity" in your medical record is not a strong predictor of having a biased language in your discharge summary.

Empirical strategy and results. I use the indicator variable "Any Stigma" as my primary outcome. I use the same Calonico, Cattaneo, and Farrell (2020) RD design using, using a linear local non-parametric data-driven optimal bandwidth-selection procedure using triangular kernel and two-way clustering of robust, bias-corrected standard errors at the patient-level and the clinical-service level (as notes are written by clinical service teams, and thus share note-writers, templates, and norms). Similar to previous sections, I also show robustness of my main results to using local quadratic specification, covariates, and an alternative two-way clustering scheme, i.e., using patient and year-group.

Panel A of Table 4 shows that the obesity cutoff does not significantly affect the average likelihood of stigma appearing in the clinical note. This null finding may reflect two possibilities: (i) the true effect of the cutoff is genuinely zero, or (ii) stigma is measured with substantial noise by the LLM classifier, and the available sample is not large enough to detect a true effect. The latter is particularly relevant given the data-hungry nature of the optimal bandwidth procedure in RD estimation. Since the resulting estimate is not a precise null and I am unable to obtain a larger sample, I turn to examining heterogeneity motivated by prior research. Specifically, a large literature on physician behavior suggests that even when average effects are muted, treatment may have larger effects among certain types of providers. I hypothesize that the obesity cutoff disproportionately increases stigma among providers predisposed to stigmatize—those whose documentation style suggests greater baseline stigmatizing tendencies.

Following the EHR meta-data literature that infers provider involvement from behavioral traces in the record, I proxy for the primary provider during a hospitalization as the provider who enters the plurality of orders for that admission. (Kahn et al., 2022; Soulakis et al., 2015). Specifically, I assign a patient's "primary provider" as the clinician who placed the greatest number of orders during the hospital encounter, based on the assumption that the provider most actively engaged in ordering is also the one most directly involved in day-to-day clinical decision-making—and therefore most likely to be central to drafting the discharge summary.[8] For each provider, I compute mean

[8]This should not be interpreted as the attending of record. MIMIC-IV does not provide a reliable attending identifier, and order entry in inpatient settings is often performed by junior physicians and other frontline clinicians. For the purposes of this paper, however, the "primary ordering provider" represents a substantively meaningful unit of analysis. These providers are typically the most involved in patient-facing care, are most likely to interact with the electronic health record (where the obesity BMI label is prominently displayed), and are therefore the clinicians most likely to be exposed to the "treatment" in this study. They are also almost certainly among the primary contributors to the discharge summary and thus at least partially responsible for any stigmatizing language it contains. Discharge summaries may also be supplemented by other members of the clinical team – typically residents, interns, or advanced practice providers– working within the same clinical service unit (e.g., internal medicine, general surgery, cardiology). The clinical service unit refers to the organizational and supervisory unit responsible for the patient's care during hospitalization, and it plays a central role in shaping documentation culture and norms. Final review and authentication of the discharge summary is then performed by the attending physician within that clinical service unit.

"Any Stigma" for their patients' discharge summaries based on the LLM classification. I use this provider-level variable to rank providers into quartiles of stigmatizing behavior. Quartile 4 comprises providers whose patients' notes most frequently contain anuy biased language. I define these providers as "High-stigma providers" and group quartiles 1–3 as "Low-stigma providers".

I then re-estimate the RD specification separately for High- and Low-stigma providers to test whether the effect of crossing the obesity cutoff on stigmatizing documentation is concentrated among providers with a higher baseline propensity to use biased language. Conceptually, this is a Difference-in-Discontinuities design (Diff-in-Disc): it compares (i) the discontinuity in stigma at the obesity cutoff among Low-stigma providers to (ii) the corresponding discontinuity among High-stigma providers, and interprets the difference between these two discontinuities as the heterogeneous effect of the cutoff by provider type. Identification requires that crossing the cutoff does not differentially shift patients into high- versus Low-stigma providers. Under these conditions, the Differences-in-Discontinuities estimate can be interpreted as the causal differential effect of obesity classification on stigmatizing documentation for high- versus Low-stigma providers.

Results are reported in Table 4 Panel B. On average, the obesity cutoff increases the probability of any stigmatizing language in discharge summaries by 4.4 percentage points ($p = 0.13$). This effect is concentrated among High-stigma providers: the discontinuity for High-stigma providers is 8 percentage points larger ($p < 0.05$) than the corresponding effect for Low-stigma providers. Table A.3 supports the identifying assumption for the Differences-in-Discontinuities design: patients just above and just below the cutoff are equally likely to be treated by high-stigma providers, and the cutoff does not generate meaningful differential sorting on observables between high- and low-stigma providers. The only notable Diff-in-Disc estimate is a small difference in Medicaid coverage (significant at 10%), but Medicaid is included as a covariate in RD robustness specifications (which doesn't affect the RD estimates) and would, if anything, attenuate the main results. This is most evident in the predicted mortality variable, which falls more at the cutoff for High-stigma than for Low-stigma providers – a pattern that works *against* the increased mortality and increased stigma findings. Taken together, these results suggest limited evidence of an average effect on stigmatizing documentation, but a meaningful increase at the cutoff among providers with a higher baseline propensity to use biased language.

I then present two pieces of evidence consistent with stigma as a mechanism for the mortality results in the previous section. First, Table 5 shows that the obesity cutoff increases in-hospital

mortality among High-stigma providers (robust, bias-corrected RD estimate = 1 pp, $p < 0.001$), and the Differences-in-Discontinuities estimate implies that the mortality effect for High-stigma providers is 0.9 percentage points larger ($p < 0.10$) than for Low-stigma providers. Second, patients whose discharge summaries contain any stigmatizing language have 0.1 percentage points higher in-hospital mortality ($p < 0.10$) than patients without such language (see Table 6. Together, these patterns are consistent with a link between the obesity label, stigmatizing documentation, and in-hospital mortality.

While biased language can be interpreted as evidence of generalized stigmatizing documentation activated at the cutoff, I further probe the stigma mechanism by examining two defining features of stigma: (i) that stigma is triggered by perceived deviation from a social norm, and (ii) that stigma is often accompanied by moral judgment.

Stigma as a response to norm deviation. I test the norm-deviation channel in two ways. First, I exploit the idea that the obesity label may be more stigmatizing when norm deviation is more visually salient. Holding BMI fixed, higher body mass is typically more apparent for shorter individuals than for taller individuals, so the same labeling threshold may generate stronger stigma responses for shorter patients. I therefore split patients in the overweight and obese BMI ranges into quartiles of height and estimate heterogeneous RD effects by height quartile. The quartile-specific RD estimates presented in Table A.4 are noisy, but the pattern is consistent with the visual-salience prediction: the largest (most positive) cutoff effect appears in the shortest height quartile, while the smallest (most negative) effect appears in the tallest quartile. Comparing the shortest and tallest quartiles, the Differences-in-Discontinuities estimate is -8.8 percentage points ($p < 0.1$), implying that the effect of crossing the cutoff is 8.8 percentage points smaller for the tallest patients than for the shortest. While imprecise, these results are consistent with the interpretation that the cutoff has the strongest impact when patients are most likely to appear as deviating from body-size norms.

Second, I examine within-patient changes in stigmatizing documentation for patients with multiple hospital encounters. The key idea is that stigma is tied to perceived norm violation rather than to fixed attributes of the person, so it should vary with the patient's proximity to the norm even when the individual is held constant. I therefore estimate a specification with patient fixed effects (to absorb time-invariant patient characteristics) and physician fixed effects (to absorb persistent differences in documentation style and team-level norms), and ask whether the same patient receives

more stigmatizing language when they move into higher BMI categories and thus face different labels in the EMR. Concretely, I regress the change in stigma since the patient's prior encounter (stigma in the current discharge summary minus stigma in the prior discharge summary) on the corresponding change in BMI category between encounters.[9] Table A.5 shows that stigma moves in tandem with BMI category: For the same patient, a one-category increase in BMI since the last hospital visit associated with an average change of 3 pp in stigma since the last visit, holding provider fixed effects constant ($p < 0.05$). This pattern supports the norm-violation feature of stigma because the same individual is documented more negatively when they move farther from the socially accepted body-size norm.

Stigma as a form of moral judgment. I can also move beyond an omnibus measure of "any biased language." The LLM does not simply flag negative language; for each flagged instance, it classifies the descriptor into one of four categories: judgmental, discrediting, labeling, or stereotyping. This is useful because moral judgment is a definitional feature of stigma, and it is directly captured by the judgmental category in the prompt. I therefore examine which specific category drives the main results. Results are presented in Table 7. The pattern indicates obesity increases biased languages across multiple categories, but the two largest and most precisely estimated discontinuity for High-stigma providers is in *judgmental* descriptors and *discrediting* descriptors. The judgemental dimension is, in particular, consistent with the moral-judgment component of stigma intensifying at the obesity cutoff.

I take two additional steps to address remaining concerns by testing whether the language changes can be explicitly tied to weight. First, I show that the discontinuity is not merely an increase in a given type of biased language overall, but an increase in such language when it appears alongside weight-related terminology (e.g., BMI, weight, habitus, obese) in the discharge summary. I implement this by searching for weight terms in the underlying clinical note and constructing an indicator equal to one when the LLM flags a given negative-descriptor category *and* the note contains at least one weight-related term. Results are reported in Table **??**. The co-occurrence results sharpen the interpretation: judgmental descriptors paired with weight mentions exhibit the largest and most precisely estimated effect. The obesity cutoff increases judgmental descriptors that co-occur with a weight term more for

[9] I estimate a first-difference model because clinicians may implicitly evaluate a patient relative to a reference point – most naturally, the patient's prior-visit BMI category – which I use as a proxy for the patient's short-run "norm." First differences therefore isolate whether stigmatizing language changes when BMI category deviates from that reference point between consecutive visits, rather than across visits that may be far apart and differ along many other dimensions. A secondary motivation is that documentation may be persistent: once stigmatizing language appears, it can carry forward in later notes, so level-based associations may partly reflect earlier labeling rather than contemporaneous BMI status.

High-stigma providers than for Low-stigma providers by 5 percentage points ($p < 0.05$). Discrediting descriptors paired with weight mentions also rise for High-stigma providers, but the Differences-in-Discontinuities estimate is smaller and only marginally significant (2.7 percentage points, $p = 0.103$).

Second, to provide a transparent check that does not rely on the LLM, I implement a simple dictionary-based measure that flags a small set of "raw" judgment-related terms using a regex search. I construct two outcomes: (i) an indicator for whether the discharge summary contains any judgmental term describing patient behavior (e.g., "difficult," "rude," "uncooperative"), and (ii) an indicator for whether such a judgmental term appears *and* the note also contains a weight-related term. Table 9 reports results consistent with the stigma hypothesis for both measures: the obesity cutoff increases the use of raw judgmental language by 8 pp ($p < 0.05$, Column 1), and it increases the co-occurrence of judgmental language with weight mentions by 4.8 pp ($p < 0.05$, Column 2), more for High-stigma providers than for Low-stigma providers.

5 Conclusion

This paper uses a regression discontinuity design at the BMI 30 obesity threshold to isolate the causal relationship between stigma – due to the prominently displayed "obese" label in most electronic medical records – and resource allocation and outcomes in the high-stakes of inpatient healthcare. I first find that crossing into the "obese" category generates a discontinuous increase in in-hospital mortality. Consistent with a mechanism of stigma-based discrimination, patients just above the cutoff receive a less intensive clinical workup – fewer diagnostic orders in the first 24 hours, fewer procedures on the intensive margin, and fewer common laboratory panels, consistent with reduced effort and resource use for individuals with the "obese" label. In the unstructured text of discharge summaries, a physician-validated LLM detects a discontinuous increase in biased documentation at the cutoff that is concentrated among providers with higher baseline propensities to stigmatize, and the language most affected is that categorized as "judgmental" and "discrediting". Additional evidence is consistent with stigma's defining features: stronger effects when norm deviation is more salient and co-occurrence between judgmental language and weight-related mentions.

The results highlight a policy-relevant and potentially tractable margin: the design of provider-facing clinical interfaces and classification schemes. If discrete labels amplify stigma without commensurate clinical benefit, health systems should reconsider how prominently BMI categories – especially

the "obese" label – are displayed, whether thresholds are necessary for routine workflows, and whether alternative presentations (e.g., continuous measures, less stigmatizing terminology, or decision-support that emphasizes evidence-based risk factors rather than categorical labels) can preserve clinical usefulness while reducing harm. More broadly, the findings suggest that stigma can operate through subtle channels – reduced diagnostic effort and biased documentation – even in settings governed by strong professional norms. Interventions that shift documentation norms and redesign default information displays may therefore meaningfully reduce inequities in care and improve patient outcomes.

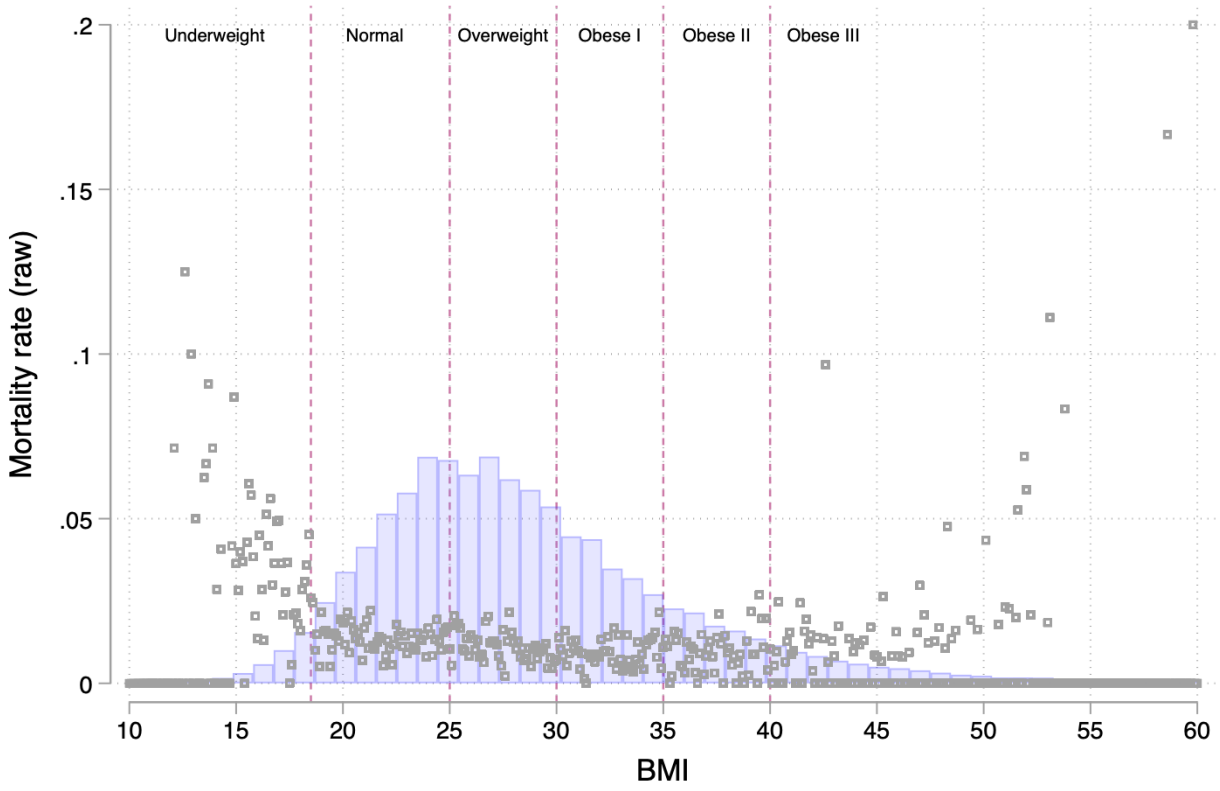


FIGURE I
Distribution of BMI and Raw Mortality

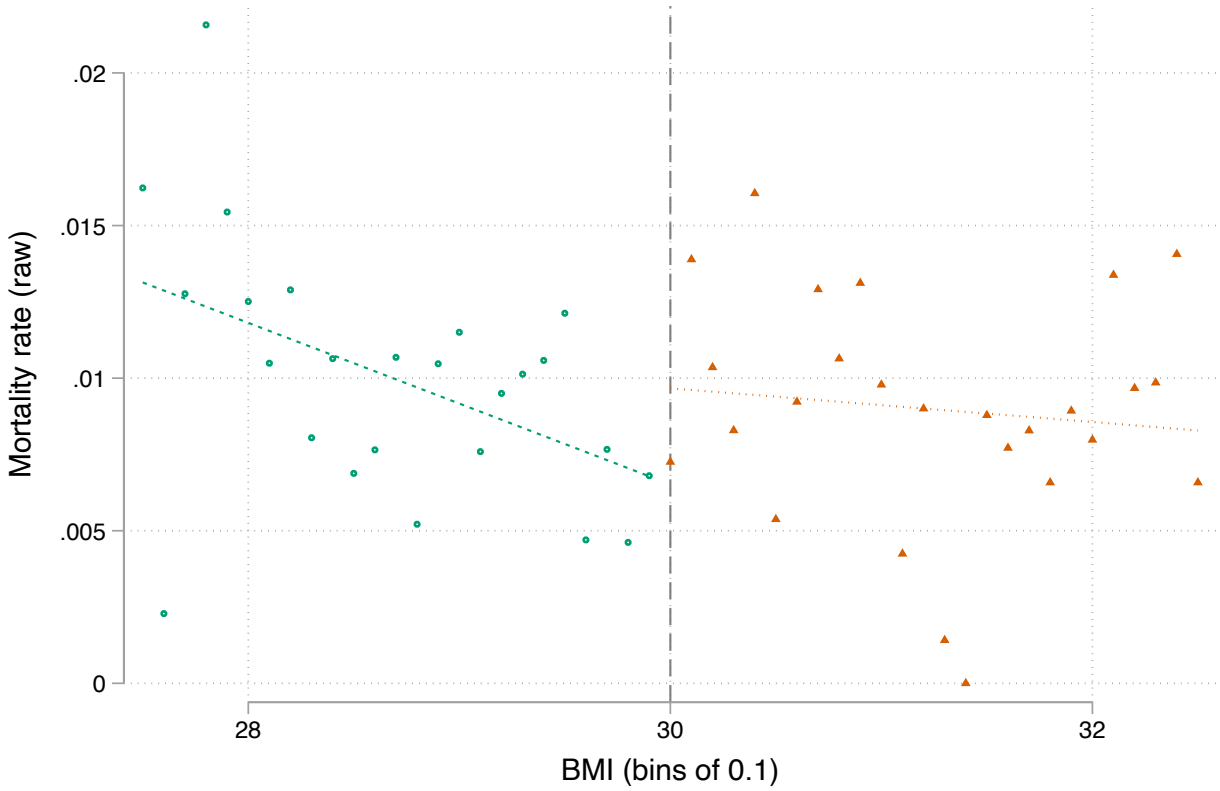
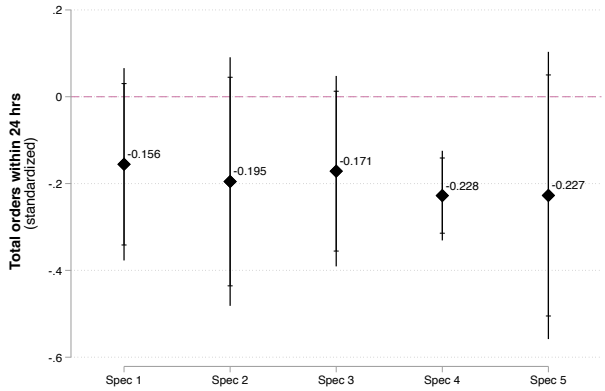
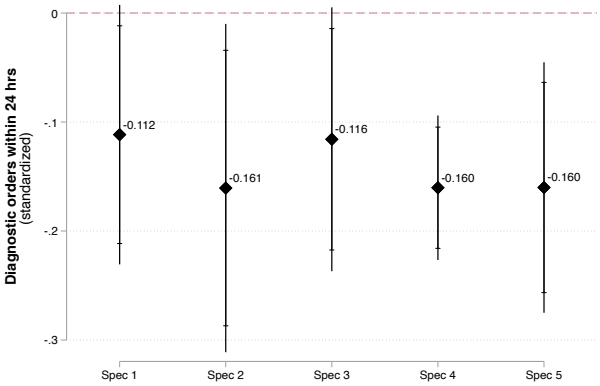


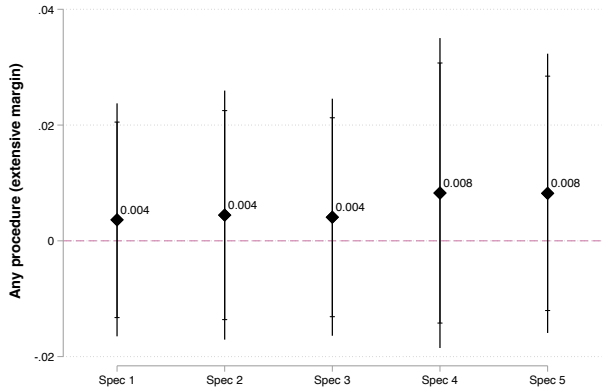
FIGURE II
Mortality and BMI at the Obesity Cutoff



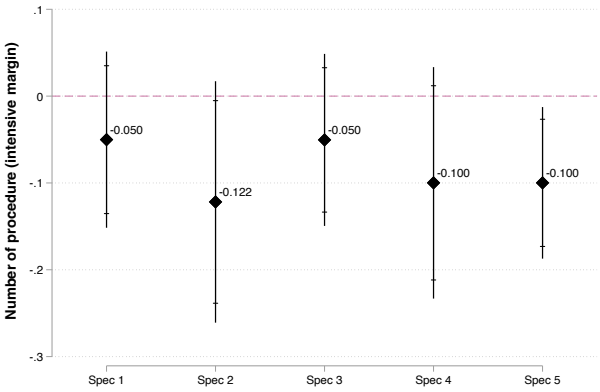
Panel (a): Total orders within first 24 hrs



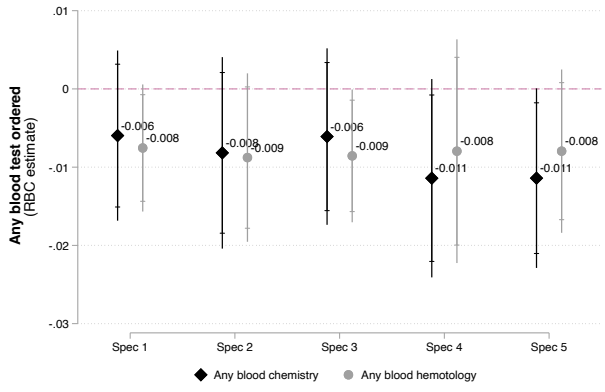
Panel (b): Diagnostic orders within first 24 hrs



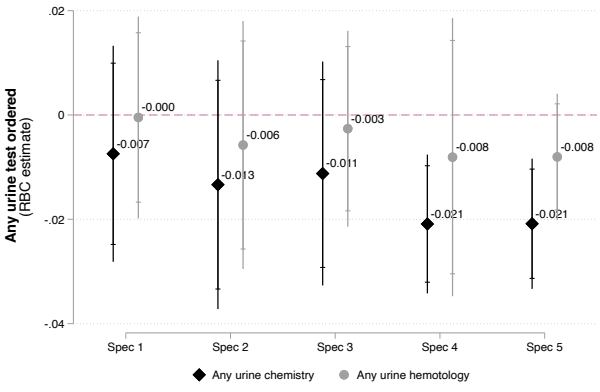
Panel (c): Any procedure



Panel (d): # of procedures (conditional on > 0)



Panel (e): Any blood test



Panel (f): Any urine test

FIGURE III Analysis of Structured Data

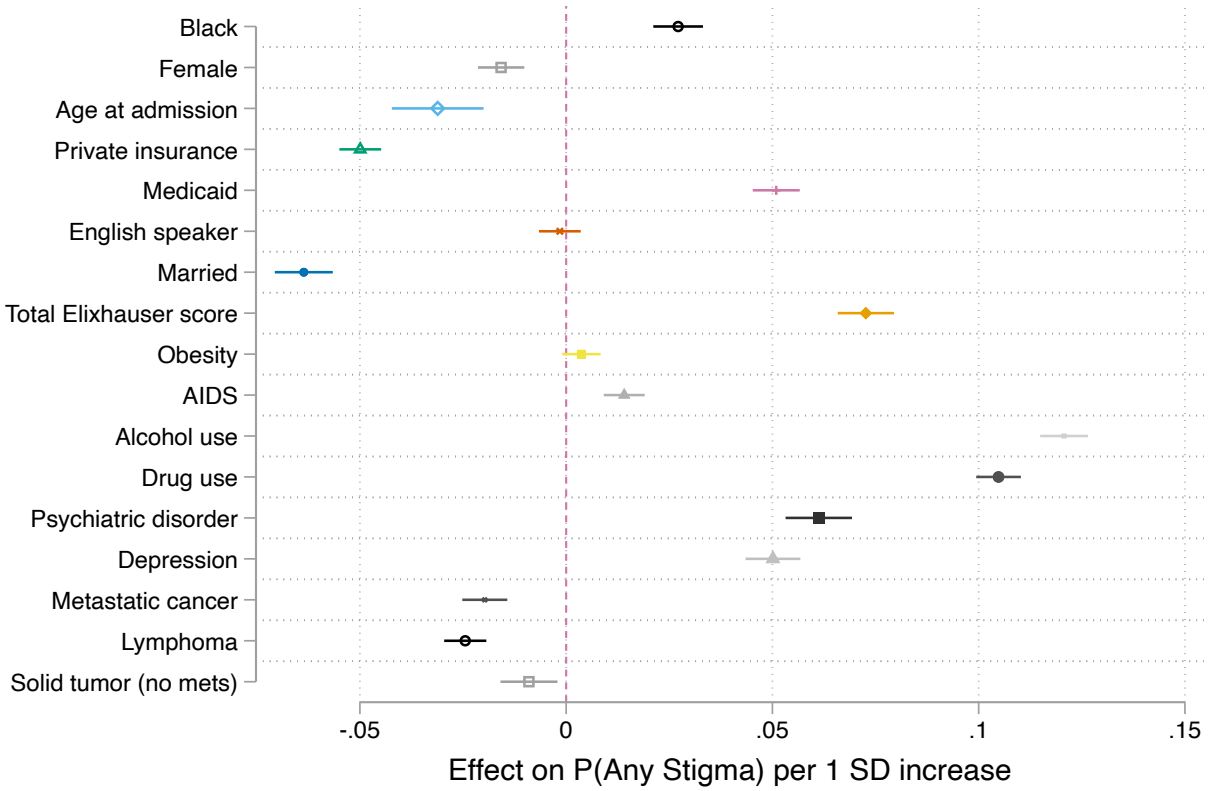


FIGURE IV
Correlations of Any Stigma with Patient Characteristics

	Dep Var: In-Hospital Mortality						
Model	(1)	(2)	(3)	(4)	(5)	(6)	(7)
RD estimate (RBC)	0.004* (0.002)	0.005** (0.002)	0.004* (0.002)	0.004* (0.002)	0.004* (0.002)	0.005** (0.002)	0.005** (0.002)
P-value (RBC)	0.071	0.024	0.058	0.097	0.069	0.012	0.038
Observations left	109,501	109,501	109,501	109,501	109,501	109,501	109,501
Observations right	58,863	58,863	58,863	58,863	58,863	58,863	58,863
Effective obs left	29,142	21,126	48,161	41,252	29,103	29,144	29,144
Effective obs right	21,613	17,106	30,129	27,488	21,602	21,613	21,613
Polynomial order (p)	1	1	2	2	1	1	1
Kernel	Tri	Uni	Tri	Uni	Tri	Tri	Tri
Bandwidth (h) (L/R)	2.969	2.201	4.645	4.069	2.959	2.970	2.970
Cluster vars	P	P	P	P	P	P,Y	P,S
Covariates	No	No	No	No	Yes	No	No
Mean, left of threshold	0.010	0.010	0.011	0.011	0.010	0.010	0.010

TABLE 1 RD robustness to polynomial order, kernel, covariates, and clustering

Outcome	**Pred. Mortality**	**Obs. Mortality**		
	(1)	(2)	(3)	(4)
RD estimate (RBC)	-0.000 (0.001)	0.000 (0.002)	0.002 (0.002)	0.005** (0.002)
P-value (RBC)	0.921	0.892	0.433	0.027
Mean, left of threshold	0.011	0.013	0.009	0.010
Cutoff	30	24.80	34.50	30
Kernel	Triangular	Triangular	Triangular	Triangular
Polynomial order (p)	1	1	1	1
Cluster variables	patient	patient	patient	patient
Bandwidth (L/R)	4.401	2.745	5.318	2.754
Observations left	106,283	54,987	138,915	109,060
Observations right	57,497	113,377	29,449	57,898
Effective obs left	44,638	28,307	36,466	26,529
Effective obs right	28,424	31,494	17,202	19,375

TABLE 2 RD placebo and donut checks

Bias type	Definition (Excerpt) and Example LLM Output
Discrediting	`Language that unjustifiably questions the legitimacy of a patient's self-reported symptoms, their concerns, their experiences with their illness, and their reports about adherence to their medication ... pay attention when adverbs such as reportedly, allegedly, and supposedly, and/or similar words like ''[the patient] claims'', are used to modify the validity of what is being presented.` **Example LLM Output:** **(1) Bias term:** *these symptoms were likely psychogenic* **Bias type:** Discrediting **Reason:** Dismisses the patient-reported side-effects as psychogenic without presenting objective evidence, questioning the legitimacy of the patient's concerns. **(2) Bias term:** *she ... had "diarrhea" with abdominal pain* **Bias type:** Discrediting **Reason:** Quotation marks can suggest doubt about the legitimacy of the symptom, subtly undermining the patient's self-report. **(3) Bias term:** *She was lying in bed comfortably, checking her facebook page. She immediately started to complain of pain the moment I walked in* **Bias type:** Discrediting **Reason:** Implies the patient's pain is not genuine and only performed for the clinician's benefit, casting unjustified doubt on her reported symptoms. **(4) Bias term:** *after eating a "poisoned taco", found to have HCV* **Bias type:** Discrediting **Reason:** The use of quotation marks around "poisoned taco" suggests skepticism and subtly mocks the patient-reported etiology, questioning her account without offering supporting evidence. **(5) Bias term:** *Reports prior fever though no fever in OSH documentation.* **Bias type:** Discrediting **Reason:** Contrasts the patient's self-report with chart data in a way that casts doubt on the accuracy of the patient's history without providing further clinical evaluation, implying unwarranted skepticism.
Judgmental	`Language that conveys moral judgment or reflects personal values onto the patient. Identify language as judgmental only when it implies moral failing or character flaws unrelated to the clinical context.` **Example LLM Output:** **(1) Bias term:** *to attain certain secondary benefits (i.e., expedite or force medical professionals to order x-rays for other purposes)* **Bias type:** Judgmental **Reason:** Suggests the patient's prior self-harm was manipulative for personal gain, conveying moral judgment about motives rather than neutrally describing clinical history. **(2) Bias term:** *The patient was supposed to restart his diuretic ... but did not do so.* **Bias type:** Judgmental **Reason:** Subtly assigns blame to the patient for not restarting the medication, implying a failure of responsibility without additional context.

TABLE 3 Examples of LLM output by bias type.

Bias type	Definition (Excerpt) and Example LLM Output
Judgmental	*(continued)* **Example LLM Output (continued):** **(3) Bias term:** *has bad willpower so finds he needs to come to hospital* **Bias type:** Judgmental **Reason:** Frames difficulty adhering to dietary recommendations as a character flaw ("bad willpower") rather than a clinical or contextual challenge. **(4) Bias term:** *blaming others for his problems with difficulty taking responsibility for his own care and actions* **Bias type:** Judgmental **Reason:** Attributes moral failing (refusing responsibility) to the patient's character rather than documenting observable behavior in neutral terms. **(5) Bias term:** *Despite h.o former substance abuse and statements suggesting wanting sobriety, he continued to ask for certain medications that he formerly abused.* **Bias type:** Judgmental **Reason:** Frames medication requests as morally inconsistent with sobriety, implying wrongdoing rather than neutrally documenting the request and context.
Pejorative Labeling	`Language that unfairly reduces a patient to their medical condition ... Pay specific attention to when a patient is referred to as a negative adjective (for example ... patient is belligerent, agitated, drug-seeking behavior, and/or aggressive).` **Example LLM Output:** **(1) Bias term:** *Confused – always.* **Bias type:** Pejorative Labeling **Reason:** Adding "always" labels the patient with a persistent negative trait, potentially depersonalizing and stigmatizing the individual. **(2) Bias term:** *hypersexual behavior* **Bias type:** Pejorative Labeling **Reason:** Attaches a stigmatizing behavioral label to the patient, potentially reinforcing stigma. **(3) Bias term:** *found to be crazy* **Bias type:** Pejorative Labeling **Reason:** Uses a pejorative label ("crazy") to describe the patient's presentation, reducing the patient to a stigmatizing stereotype of mental illness. **(4) Bias term:** *patient is a poor historian and unable to provide a reliable history* **Bias type:** Pejorative Labeling **Reason:** Labels the patient as a "poor historian," reducing the patient to a perceived deficiency and potentially conveying blame rather than neutrally describing limitations in history-taking.

TABLE 3 Examples of LLM output by bias type (continued).

Bias type	Definition (Excerpt) and Example LLM Output
Pejorative Labeling	*(continued)*
	(5) Bias term: *Pt spoke in gibberish* **Bias type:** Pejorative Labeling **Reason:** Uses a pejorative descriptor ("gibberish") that negatively labels the patient's speech rather than clinically characterizing it.
Stereotyping	`Presumptions about a patient based on demographic groups and without individual evidence ... (pay specific attention to marginalized groups like racial/ethnic minorities, women, elderly, overweight or obese persons, patients dealing with poverty, non-English speakers, patients dealing with homelessness, and patients who identify as LGBTQIA.)` **Example LLM Output:** **(1) Bias term:** *Obese ___ woman with corn-rolls in her hair* **Bias type:** Stereotyping **Reason:** Calling attention to body habitus and hairstyle (often racialized) is not clinically relevant in a mental-status exam and can perpetuate weight and cultural stereotypes. **(2) Bias term:** *overweight caucasian F dressed in hospital gown holding cake, puddings in lap; ... eats cake, pudding, sandwich and other foods very rapidly throughout interview* **Bias type:** Stereotyping **Reason:** Highlights weight, race, and eating behavior in a way that reinforces stereotypes about overeating in overweight individuals without clear clinical necessity. **(3) Bias term:** *skin color appropriate for ethnicity* **Bias type:** Stereotyping **Reason:** Invokes a presumed "appropriate" skin color based on ethnicity, applying a generalized expectation about a demographic group rather than an individual-specific clinical observation. **(4) Bias term:** *protective factors against self-harm include ... female gender, race* **Bias type:** Stereotyping **Reason:** Uses gender and race generically as protective factors, applying population-level generalizations to the individual rather than documenting patient-specific protective factors. **(5) Bias term:** *Tattoo of joker on leg as well as an ankle bracelet* **Bias type:** Stereotyping **Reason:** Including an ankle bracelet (often associated with criminal monitoring) is not clinically relevant and can foster negative stereotypes about the patient's character.

TABLE 3 Examples of LLM output by bias type (continued).

	Dep Var: Any Stigma in Notes							
	estimate	SE (RBC)	p-value	total N	Eff. N (L)	Eff. N (R)	h	mean (L)
Panel A: Baseline RD estimate								
Full sample	-0.019	0.015	0.204	65,788	9,309	8,322	1.311	0.247
Panel B: Heterogeneous RD estimates								
Low-stigma providers	-0.036*	0.019	0.058	49,594	8,640	7,728	1.657	0.188
High-stigma providers	0.044	0.029	0.128	16,194	1,968	1,772	1.194	0.425
Diff-in-Disc estimate	**0.0800****	**0.034**	**0.019**	**65,788**				

TABLE 4 RD heterogeneous effects by provider stigma group

	Dep Var: In-hospital mortality						
	estimate	SE (RBC)	p-value	total N	Eff. N (L)	Eff. N (R)	h
RD estimate							
Low-stigma providers	0.003	0.003	0.347	127,784	27,780	19,267	3.570
High-stigma providers	0.013***	0.003	0.000	38,946	6,776	4,825	3.075
Diff-in-Disc estimate	**0.0104***	**0.006**	**0.067**	**166,730**			

TABLE 5 RD heterogeneous effects on in-hospital mortality by provider stigma group

	Dep Var: In-hospital mortality						
	estimate	SE (RBC)	p-value	total N	Eff. N (L)	Eff. N (R)	h
RD estimate							
No stigma in notes	0.005**	0.003	0.005	49,228	7,020	6,210	1.312
Any stigma in notes	0.006	0.004	0.886	16,560	2,491	2,262	1.437
Diff-in-Disc estimate	**0.0009***	**0.000**	**0.057**	**65,788**			

TABLE 6 RD heterogeneous effects on in-hospital mortality by stigma in notes

	estimate	SE (RBC)	p-value	total N	Eff. N (L)	Eff. N (R)	h
Outcome: Judgemental							
RD estimate							
Low-stigma providers	-0.019	0.012	0.108	49,594	6,482	5,995	1.253
High-stigma providers	0.058**	0.025	0.023	16,194	1,629	1,489	0.993
Diff-in-Disc estimate	**0.0766*** **	**0.020**	**0.000**	**65,788**			
Outcome: Stereotyping							
RD estimate							
Low-stigma providers	-0.000	0.002	0.957	49,594	8,566	7,708	1.643
High-stigma providers	0.009*	0.005	0.087	16,194	3,020	2,554	1.700
Diff-in-Disc estimate	**0.0090****	**0.005**	**0.046**	**65,788**			
Outcome: Pejorative Labeling							
RD estimate							
Low-stigma providers	-0.024*	0.012	0.058	49,594	10,211	8,470	1.856
High-stigma providers	0.004	0.032	0.893	16,194	2,258	1,991	1.314
Diff-in-Disc estimate	**0.0279**	**0.031**	**0.373**	**65,788**			
Outcome: Discrediting							
RD estimate							
Low-stigma providers	-0.019	0.013	0.152	49,594	7,584	6,776	1.439
High-stigma providers	0.057***	0.015	0.000	16,194	1,769	1,624	1.073
Diff-in-Disc estimate	**0.0759****	**0.024**	**0.002**	**65,788**			

TABLE 7 RDx heterogeneous effects by provider stigma group – stigma components

	estimate	SE (RBC)	p-value	total N	Eff. N (L)	Eff. N (R)	h
Outcome: Judgemental (weight mention)							
RD estimate							
Low-stigma providers	-0.003	0.013	0.832	49,594	8,091	7,177	1.530
High-stigma providers	0.067**	0.029	0.019	16,194	1,748	1,609	1.036
Diff-in-Disc estimate	**0.0700*** **	**0.021**	**0.001**	**65,788**			
Outcome: Stereotyping (weight mention)							
RD estimate							
Low-stigma providers	0.001	0.001	0.631	49,594	8,045	7,158	1.515
High-stigma providers	-0.004	0.009	0.609	16,194	3,095	2,597	1.771
Diff-in-Disc estimate	**-0.0051**	**0.008**	**0.515**	**65,788**			
Outcome: Pejorative Labeling (weight mention)							
RD estimate							
Low-stigma providers	-0.011	0.012	0.380	49,594	8,177	7,269	1.567
High-stigma providers	-0.006	0.022	0.793	16,194	2,274	2,013	1.337
Diff-in-Disc estimate	**0.0052**	**0.021**	**0.807**	**65,788**			
Outcome: Discrediting (weight mention)							
RD estimate							
Low-stigma providers	0.003	0.006	0.544	49,594	10,791	8,904	1.934
High-stigma providers	0.031**	0.013	0.017	16,194	2,054	1,867	1.206
Diff-in-Disc estimate	**0.0272**	**0.017**	**0.107**	**65,788**			

TABLE 8 RD heterogeneous effects by provider stigma group – stigma components with weight mentions

	estimate	SE (RBC)	p-value	total N	Eff. N (L)	Eff. N (R)	h
Outcome: Judgemental words (raw)							
RD estimate							
Low-stigma providers	-0.028	0.021	0.182	49,594	8,494	7,594	1.604
High-stigma providers	0.052**	0.017	0.003	16,194	3,064	2,566	1.731
Diff-in-Disc estimate	**0.0802****	**0.032**	**0.012**	**65,788**			
Outcome: Judgemental words (raw) + weight mention							
RD estimate							
Low-stigma providers	-0.018	0.015	0.217	49,594	8,164	7,202	1.553
High-stigma providers	0.029**	0.013	0.029	16,194	3,267	2,694	1.843
Diff-in-Disc estimate	**0.0475****	**0.017**	**0.004**	**65,788**			

TABLE 9 RD heterogeneous effects by provider stigma group – judgement outcomes

References

Apakama, D. U., Nguyen, K.-A.-N., Hyppolite, D., Soffer, S., Mudrik, A., Ling, E., Moses, A., Temnycky, I., Glasser, A., Anderson, R., Parchure, P., Woullard, E., Edalati, M., Chan, L., Kronk, C., Freeman, R., Kia, A., Timsina, P., Levin, M. A., … Nadkarni, G. N. (2024, October). Identifying and Characterizing Bias at Scale in Clinical Notes Using Large Language Models [Pages: 2024.10.24.24316073]. https://doi.org/10.1101/2024.10.24.24316073

Bertrand, M., & Mullainathan, S. (2004). Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *American economic review*, *94*(4), 991–1013.

Biewen, M., & Steffes, S. (2010). Unemployment Persistence: Is There Evidence for Stigma Effects? *Economics Letters*, *106*(3), 188–190. https://doi.org/10.1016/j.econlet.2009.11.016

Calonico, S., Cattaneo, M. D., & Farrell, M. H. (2020). Optimal bandwidth choice for robust bias-corrected inference in regression discontinuity designs. *The Econometrics Journal*, *23*(2), 192–210.

Calonico, S., Cattaneo, M. D., Farrell, M. H., & Titiunik, R. (2017). Rdrobust: Software for regression-discontinuity designs [Publisher: SAGE Publications Sage CA: Los Angeles, CA]. *The Stata Journal*, *17*(2), 372–404.

Calonico, S., Cattaneo, M. D., & Titiunik, R. (2014). Robust Nonparametric Confidence Intervals for Regression-Discontinuity Designs: Robust Nonparametric Confidence Intervals. *Econometrica*, *82*(6), 2295–2326. https://doi.org/10.3982/ECTA11757

Celhay, P., Meyer, B. D., & Mittag, N. (2025). Stigma in Welfare Programs. *The Review of Economics and Statistics*, 1–37. https://doi.org/10.1162/rest_a_01556

Charles, K. K., & Guryan, J. (2008). Prejudice and Wages: An Empirical Assessment of Becker's The Economics of Discrimination. *Journal of Political Economy*, *116*(5), 773–809. https://doi.org/10.1086/593073

Elixhauser, A., Steiner, C., Harris, D. R., & Coffey, R. M. (1998). Comorbidity measures for use with administrative data. *Medical care*, 8–27.

Flegal, K. M., Graubard, B. I., Williamson, D. F., & Gail, M. H. (2007). Cause-Specific Excess Deaths Associated With Underweight, Overweight, and Obesity. *JAMA*, *298*(17), 2028–2037. https://doi.org/10.1001/jama.298.17.2028

Johnson, A., Pollard, T., a-chahin, Blundell, J., Gow, B., erinhong, Schubert, M., Dang, K., Paris, N., shu98, JackieMe, Carlson, E., Barros, A., Goi, W., Bulgarelli, L., Hao, S., alexmbennett2, Porschen, C., Hegselmann, S., … van den Berg, T. (2025, April). MIT-LCP/mimic-code: V3.0.1. https://doi.org/10.5281/ZENODO.821871

Johnson, A., Pollard, T., Horng, S., Celi, L. A., & Mark, R. (n.d.). MIMIC-IV-Note: Deidentified free-text clinical notes. https://doi.org/10.13026/1N74-NE17

Johnson, A. E. W., Bulgarelli, L., Shen, L., Gayles, A., Shammout, A., Horng, S., Pollard, T. J., Hao, S., Moody, B., Gow, B., Lehman, L.-w. H., Celi, L. A., & Mark, R. G. (2023). MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, *10*(1), 1. https://doi.org/10.1038/s41597-022-01899-x

Kahn, J. M., Minturn, J. S., Riman, K. A., Bukowski, L. A., & Davis, B. S. (2022). Characterizing intensive care unit rounding teams using meta-data from the electronic health record. *Journal of critical care*, *72*, 154143. https://doi.org/10.1016/j.jcrc.2022.154143

Kantor, M. A., Evans, K. H., & Shieh, L. (2015). Pending Studies at Hospital Discharge: A Pre-post Analysis of an Electronic Medical Record Tool to Improve Communication at Hospital Discharge. *Journal of General Internal Medicine*, *30*(3), 312–318. https://doi.org/10.1007/s11606-014-3064-x

Kind, A. J. H., & Smith, M. A. (2008). Documentation of Mandated Discharge Summary Components in Transitions from Acute to Subacute Care. In K. Henriksen, J. B. Battles, M. A. Keyes, & M. L. Grady (Eds.), *Advances in Patient Safety: New Directions and Alternative Approaches (Vol. 2: Culture and Redesign)*. Agency for Healthcare Research; Quality (US). Retrieved January 12, 2026, from http://www.ncbi.nlm.nih.gov/books/NBK43715/

Kolata, G. (2016). Why Do Obese Patients Get Worse Care? Many Doctors Don't See Past the Fat. *The New York Times*. Retrieved January 10, 2026, from https://www.nytimes.com/2016/09/26/health/obese-patients-health-care.html

LUSKY, A., BARELL, V., LUBIN, F., KAPLAN, G., LAYANI, V., SHOHAT, Z., LEV, B., & WIENER, M. (1996). Relationship between Morbidity and Extreme Values of Body Mass Index in Adolescents. *International Journal of Epidemiology*, *25*(4), 829–834. https://doi.org/10.1093/ije/25.4.829

Ng, I. K. S., Tung, D., Seet, T., Yow, K. S., Chan, K. L. E., Teo, D. B., & Chua, C. E. (2025). How to write a good discharge summary: A primer for junior physicians. *Postgraduate Medical Journal*, *101*(1198), 764–772. https://doi.org/10.1093/postmj/qgaf020

Okwerekwu, J. A. (2016, June). In treating obese patients, too often doctors can't see past weight. Retrieved January 10, 2026, from https://www.statnews.com/2016/06/03/weight-obese-doctors-patients/

Smith, E. C. (n.d.). Stigma and Social Cover: A Mental Health Care Experiment in Refugee Networks. *Working Paper*.

Soulakis, N. D., Carson, M. B., Lee, Y. J., Schneider, D. H., Skeehan, C. T., & Scholtens, D. M. (2015). Visualizing collaborative electronic health record usage for hospitalized patients with heart failure. *Journal of the American Medical Informatics Association*, *22*(2), 299–311. https://doi.org/10.1093/jamia/ocu017

Yang, D., Allen, I., James, Mahumane, A., Riddell, I., James, & Yu, H. (2021, April). Knowledge, Stigma, and HIV Testing: An Analysis of a Widespread HIV/AIDS Program. https://doi.org/10.3386/w28716

Online Appendix

A LLM prompt

You are a highly esteemed hospitalist physician recognized for your expertise in identifying and addressing biased language in inpatient clinical documentation.

Your mission is to scrutinize hospital discharge summaries from the electronic medical record, focusing specifically on detecting language that may contribute to patient stigmatization. For this task, we categorize "stigmatizing language" into four distinct types: Discrediting, Judgmental, Stigmatizing/Labeling, and Stereotyping.

Below are the definitions and guidelines for each category:

Discrediting: Language that unjustifiably questions the legitimacy of a patient's self-reported symptoms, their concerns, their experiences with their illness, and their reports about adherence to their medication. Tag language as discrediting only if it shows an unjustified pattern of skepticism not supported by clinical evidence or dismisses patient reports without proper investigation. Specifically pay attention when adverbs such as reportedly, allegedly, and supposedly, and/or similar words like "[the patient] claims", are used to modify the validity of what is being presented.

Judgmental: Language that conveys moral judgment or reflects personal values onto the patient. Identify language as judgmental only when it implies moral failing or character flaws unrelated to the clinical context. Specifically include the phrase "non-compliance" when given without reason, or "toxic habits", or "no toxic habits" or any other phrasing that implies the patient has failed in some manner. Ignore terms that are factual and have medical implications like "active smoker", "toxic appearing", or "ill appearing." This should only capture negative judgments; in other words, language that portrays the patient in a positive light or is neutral should not be categorized as judgmental.

Stigmatizing/Labeling: Language that unfairly reduces a patient to their medical condition can be stigmatizing. Pay specific attention to when a patient is referred to as a negative adjective (for example, but not limited to, patient is belligerent, agitated, drug-seeking behavior, and/or aggressive). Consider language stigmatizing when it unnecessarily labels patients with negative adjectives, contributing to stigma or negatively depersonalizing the individual. Any references to medical history, age, and clinically relevant behaviors are not biased when they are stated factually and without negative adjectives that could imply judgment.

Stereotyping: Presumptions about a patient based on demographic groups and without individual evidence. Mark language as stereotyping when it applies generalized traits or behaviors to a broad

group rather than assessing the individual patient (pay specific attention to marginalized groups like racial/ethnic minorities, women, elderly, overweight or obese persons, patients dealing with poverty, non-English speakers, patients dealing with homelessness, and patients who identify as LGBTQIA. For the purposes of this category only, ignore bias mentioned in past medical history, exam findings about mental status and alertness (as these tend to have clinical value). Also ignore when a patient is referred to as their age and gender.

Review each discharge summary provided.

Identify and categorize any instance of language that falls into the above categories in that discharge summary. Do not make up instances of stigmatizing language.

For each identified instance in the given discharge summary, explain why it has been categorized as such, referencing the guidelines provided. Bias in clinical notes can fall into multiple categories, but please try to find the most pertinent fit. For instance, "drug seeking behavior" can be considered discrediting, judgmental, Stigmatizing/Labeling, and Stereotyping depending on the use, but it most falls in line with Stigmatizing/Labeling for this review as it usually refers to the patient as an adjective.

You have to provide 2 scores for each discharge summary. The first score, $stigma_intensity$, should reflect the average intensity of the stigmatizing language used in the discharge summary, and can range from 0 to 10. This score is equal to 0 if there are no instances of stigmatizing language identified. If instances are identified, a $stigma_intensity$ score of 1 means that the language is mildly stigmatizing on average, while a score of 10 indicates that the language is extremely stigmatizing on average.

The second score, $stigma_frequency$, should reflect the number of instances of stigmatizing language in the discharge summary, and can range from 0 to n where 0 means no instances, and n means the total number of instances.

Objective: Your goal is to enhance the awareness and understanding of how language can influence patient care and perception in clinical settings. This exercise will contribute to developing best practices for creating unbiased and respectful clinical documentation without stigma. Keep your responses brief and to-the-point.

Expected JSON Format for Response: { "instances": ["bias term": "phrase taken verbatim from the discharge summary which suggests the presence of any of the four types of bias (make sure exact phrase is in the discharge summary, and if there are more than 10 instances, only return the 10 instances with the strongest instances of bias) ", "bias type": "type of bias - Discrediting,

Judgmental, Stigmatizing/Labeling, Stereotyping ", "reason": "short explanation why the phrase is considered biased ", }], "stigma_intensity": "0-10 integer score reflecting the average intensity of the stigmatizing language in the note across *all* biased phrases you found ", "stigma_frequency": "0-n integer score reflecting the *total* number of instances of stigmatizing language in the note (even if you only returned the top 10 instances before) " If there are no bias terms in the discharge summary please return: "stigma_intensity": "0", "instances": "None", "stigma_frequency": "0"

Here is the discharge note: ""

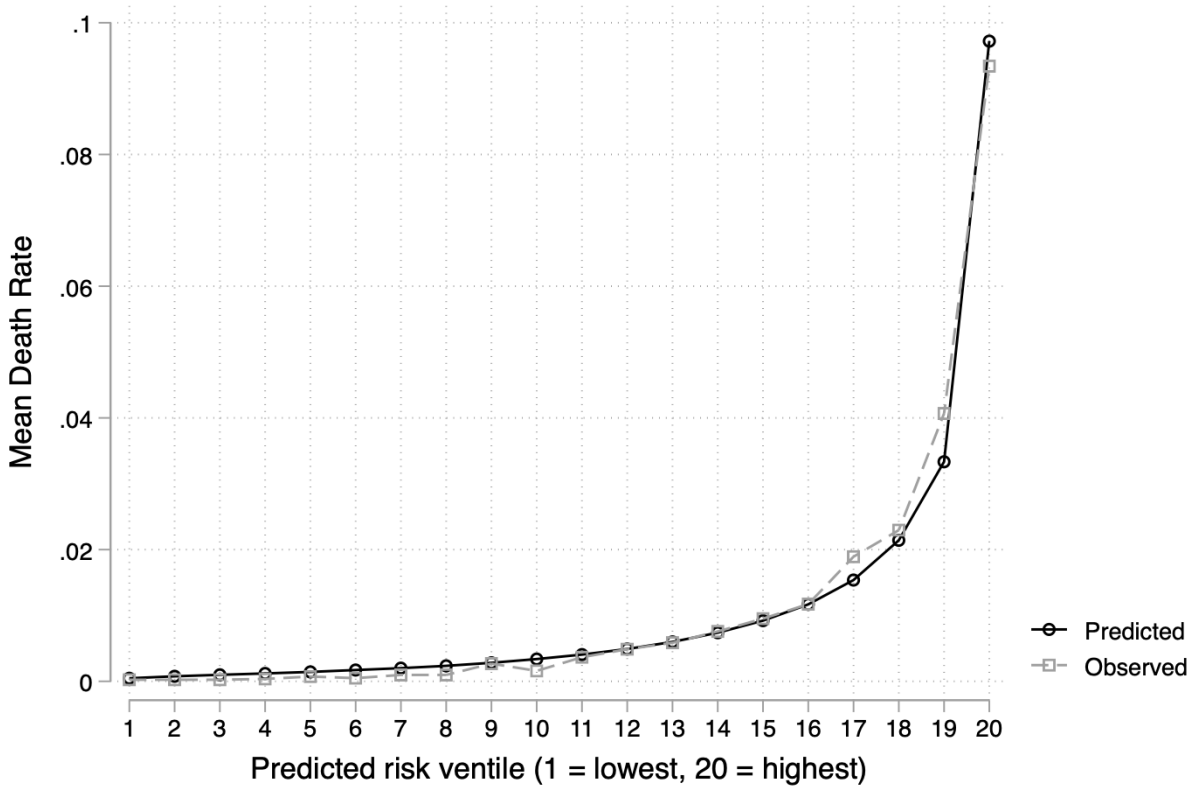


FIGURE A.1
Predicted Mortality and Observed Mortality

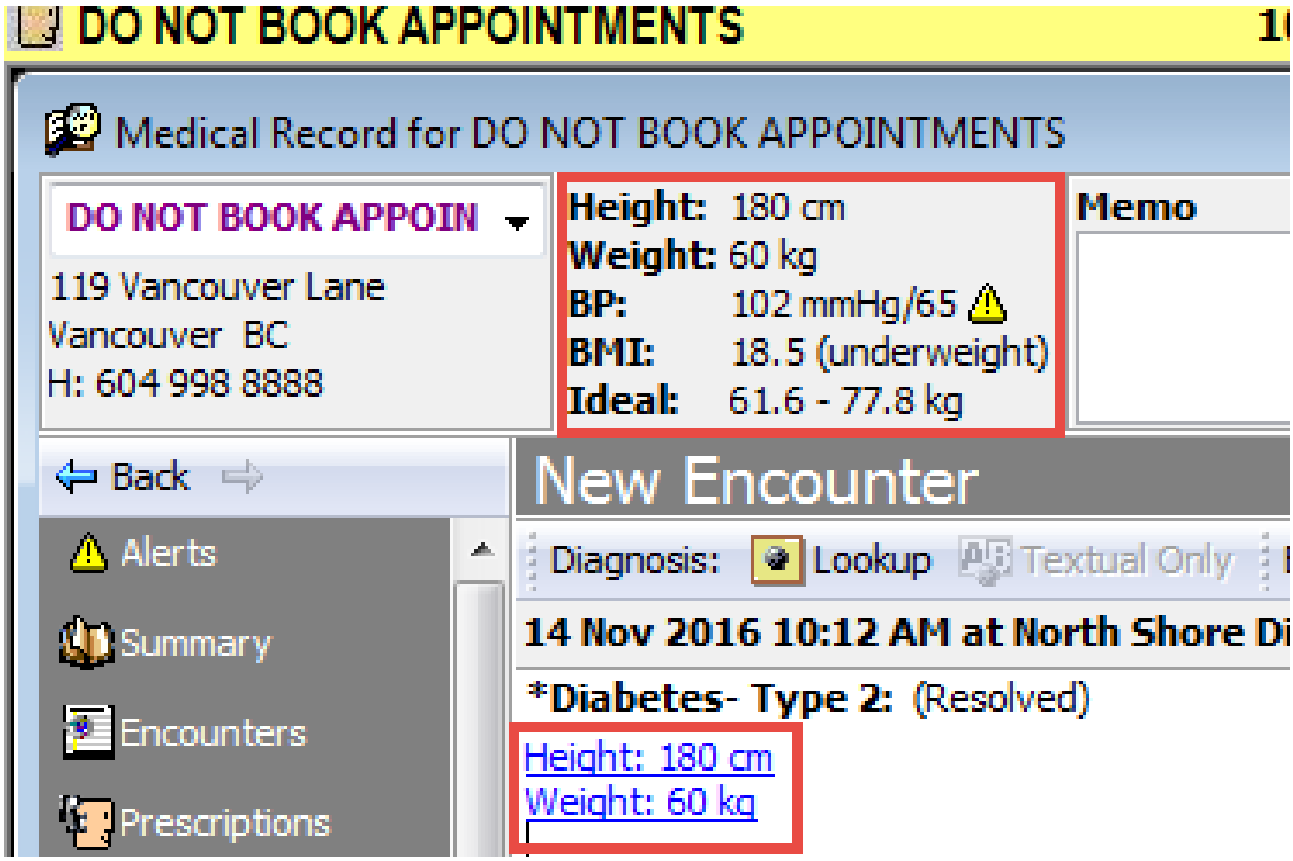


FIGURE A.2
Example of an Electronic Medical Record

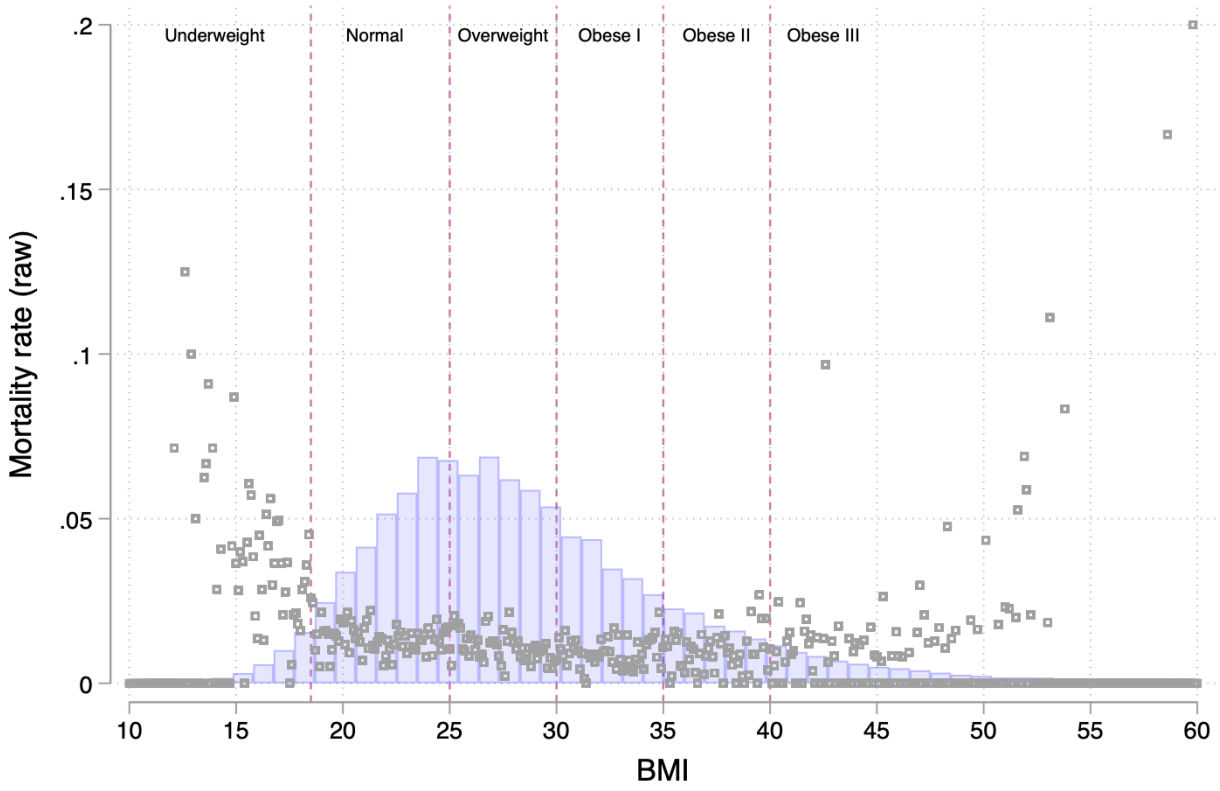


FIGURE A.3
Distribution of BMI and Raw Mortality

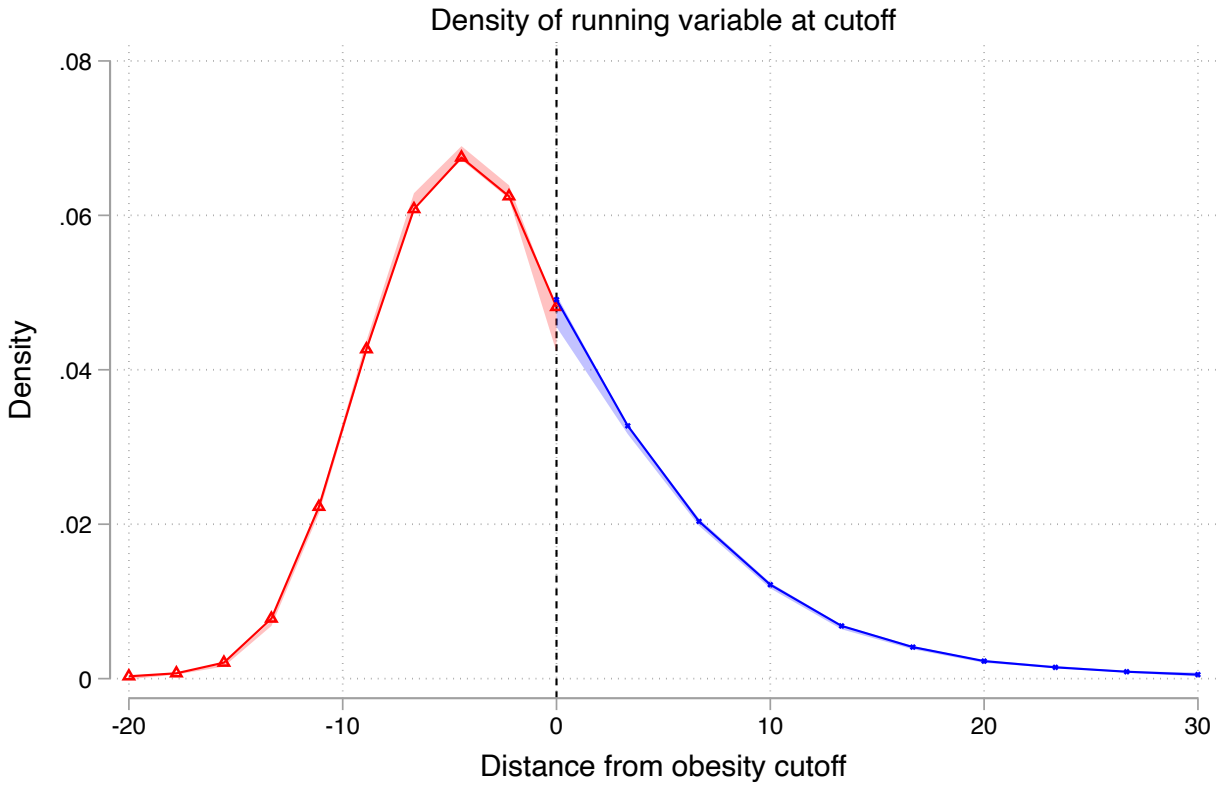
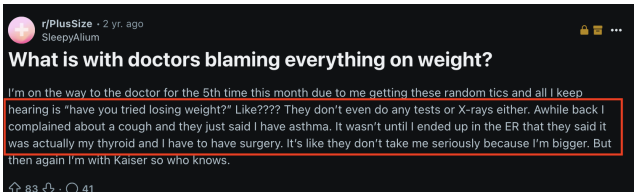


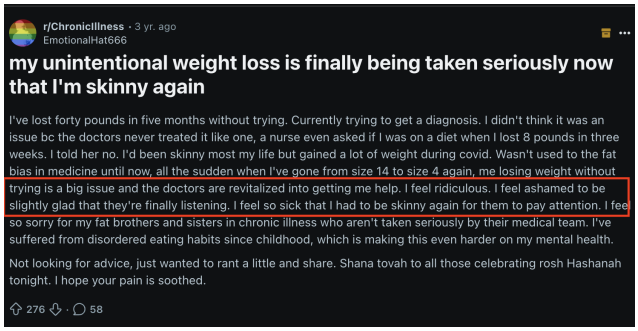
FIGURE A.4
McCrary Test



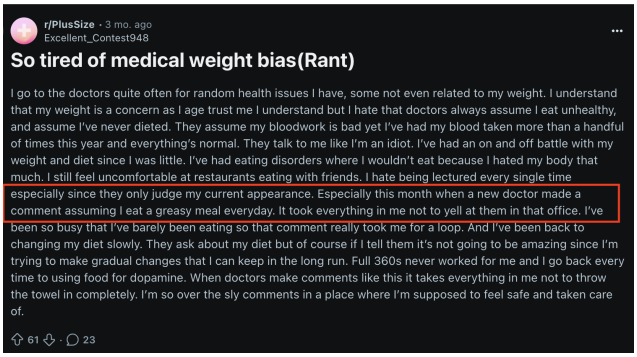
Panel (a)



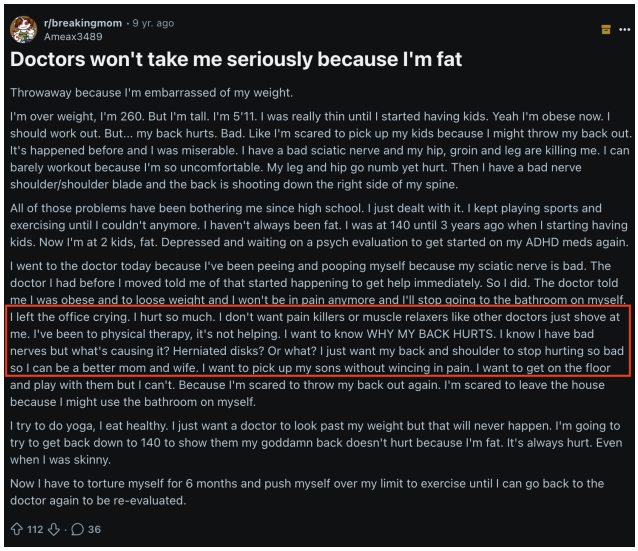
Panel (b)



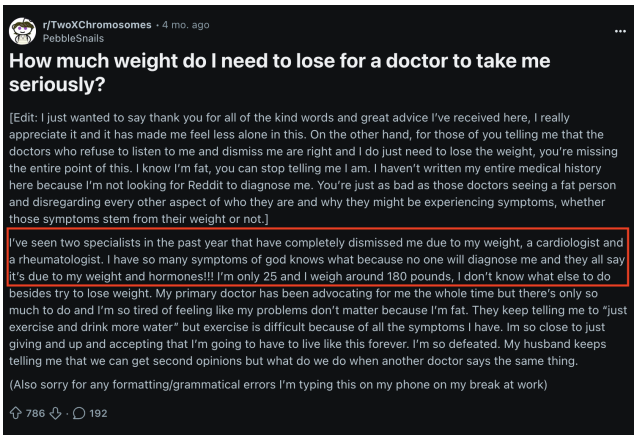
Panel (c)



Panel (d)

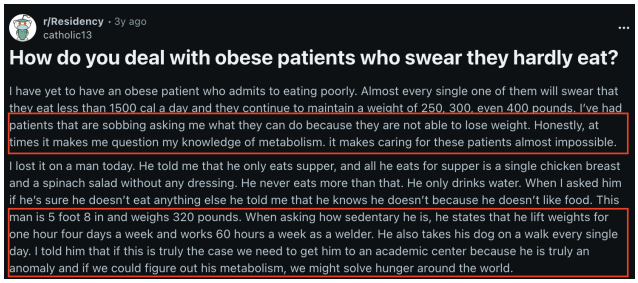


Panel (e)

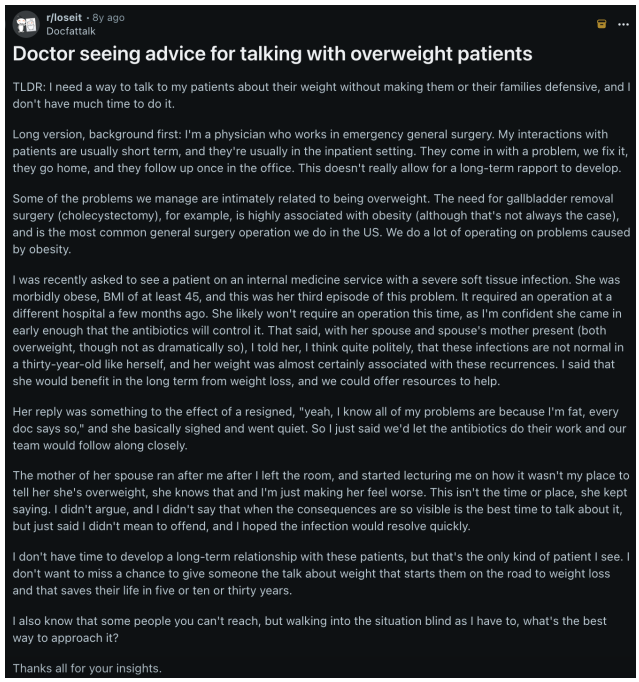


Panel (f)

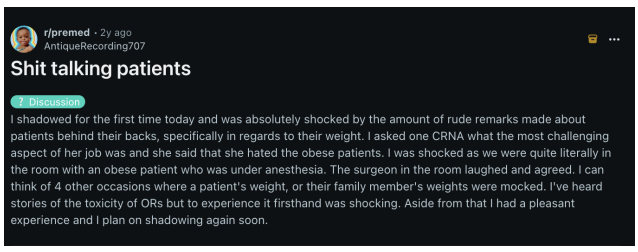
FIGURE A.5 Anecdotal evidence by patients consistent with weight stigma



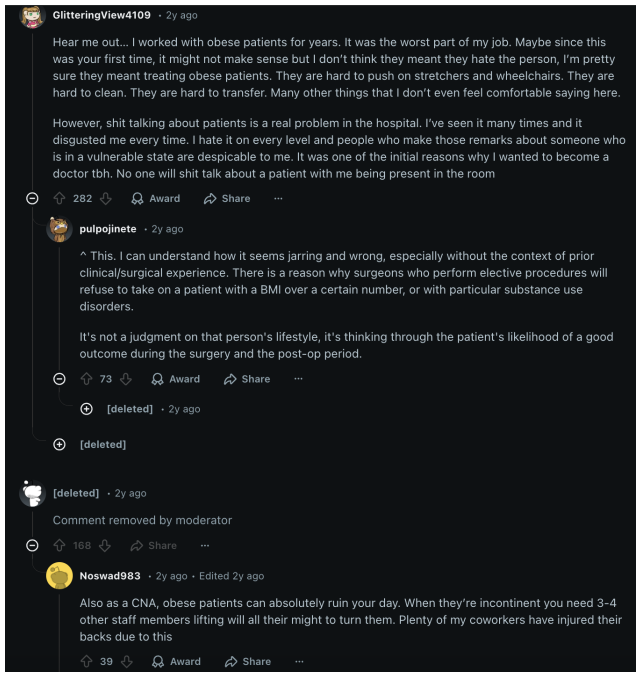
Panel (a)



Panel (b)



Panel (c)



Panel (d)

FIGURE A.6 Anecdotal evidence by patients consistent with weight stigma

Outcome	RD estimate	p-value	Mean (left)	Bandwidth h	N (L/R)	Eff. N (L/R)
Female	0.024* (0.014)	0.079	0.47	2.728	109,501 / 58,863	26,726 / 20,259
Age at admission	0.136 (0.422)	0.746	61.71	3.879	109,501 / 58,863	39,250 / 26,367
Black	-0.004 (0.009)	0.621	0.15	3.519	109,501 / 58,863	35,789 / 24,795
Medicaid	0.003 (0.009)	0.707	0.17	4.935	108,318 / 58,478	51,060 / 31,118
Medicare	0.002 (0.011)	0.831	0.49	4.722	108,318 / 58,478	48,472 / 30,351
Private insurance	-0.008 (0.011)	0.449	0.32	4.398	108,318 / 58,478	45,006 / 28,678
Married	0.003 (0.012)	0.789	0.50	4.028	107,379 / 57,882	40,313 / 26,933
Hispanic	-0.002 (0.006)	0.665	0.06	5.029	109,501 / 58,863	53,068 / 31,873
Num of Elixhauser comorbidities	0.059 (0.048)	0.215	3.08	5.261	109,471 / 58,838	55,373 / 32,759
Elixhauser mortality index	0.190 (0.215)	0.377	7.38	3.592	109,471 / 58,838	36,261 / 24,953

TABLE A.1 Covariate balance at the cutoff

	Dep Var: In-Hospital Mortality		
	(1)	(2)	(3)
RD estimate (RBC)	-0.002 (0.002)	-0.000 (0.002)	-0.004 (0.003)
P-value (RBC)	0.337	0.905	0.314
Mean, left of threshold	0.013	0.009	0.009
Cutoff BMI	25.0	35.0	40.0
Cutoff label	Overweight	Obese II	Obese III
Kernel	Triangular	Triangular	Triangular
Polynomial order (p)	1	1	1
Cluster variables	patient	patient	patient
Bandwidth (L/R)	2.686	5.179	4.871
Observations left	56,604	141,014	156,451
Observations right	111,760	27,350	11,913
Effective obs left	27,491	32,751	14,735
Effective obs right	30,904	15,805	6,674

TABLE A.2 RD estimates at alternative BMI cutoffs

	Low-stigma			High-stigma			Diff-in-Disc		
Characteristics	Estimate	SE(RBC)	P-value	Estimate	SE(RBC)	P-value	Estimate	SE(RBC)	P-value
Female	0.026	0.017	0.122	0.044*	0.023	0.059	0.018	0.026	0.496
Age at admission	-0.064	0.488	0.896	0.389	0.913	0.670	0.453	0.977	0.643
Black	-0.003	0.008	0.728	-0.011	0.012	0.341	-0.009	0.014	0.548
Medicaid	0.014	0.008	0.106	-0.008	0.014	0.575	-0.021*	0.012	0.067
Medicare	0.005	0.013	0.710	-0.006	0.032	0.854	-0.011	0.036	0.765
Private insurance	-0.017**	0.008	0.041	0.007	0.021	0.736	0.024	0.022	0.288
Married	0.003	0.019	0.860	-0.009	0.017	0.605	-0.012	0.030	0.686
Hispanic	0.001	0.004	0.858	-0.003	0.006	0.663	-0.003	0.007	0.622
Num of Elixhauser comorbidities	0.078	0.094	0.406	0.021	0.111	0.849	-0.057	0.191	0.766
Elixhauser mortality index	0.170	0.288	0.555	0.445	0.412	0.280	0.275	0.647	0.671
Predicted Mortality	0.001	0.001	0.359	-0.002**	0.001	0.042	-0.003**	0.001	0.042

TABLE A.3 Covariate balance at the cutoff by provider type

	estimate	SE (RBC)	p-value	total N	Eff. N (L)	Eff. N (R)	h
Outcome: Any Stigma in Notes							
RD estimate by height quartile							
Height Q1 (lowest)	0.023	0.028	0.410	17,038	3,150	2,875	1.759
Height Q2	-0.031	0.052	0.548	15,657	2,809	2,403	1.681
Height Q3	0.001	0.040	0.971	15,635	2,231	1,942	1.362
Height Q4 (highest)	-0.065	0.046	0.162	15,888	2,306	2,039	1.302
Diff-in-Disc (Q4 − Q1)	**-0.0877***	**0.046**	**0.058**	**65,788**			

TABLE A.4 Heterogeneous RD cutoff effects by height quartile (Diff-in-Disc: highest vs lowest quartile)

	Dep. var.: $\Delta Stigma_{t-(t-1)}$
	(1)
$\Delta BMI_{t-(t-1)}$	0.030**
	(0.011)
p-value	0.008
Observations	18,322
Patient FE	Yes
Provider FE	Yes
Clustered SE	patient, provider

TABLE A.5 First-difference regression of stigma changes on BMI-category changes