Deadly Stigma*

Manasvini Singh

Carnegie Mellon University

msingh01@cmu.edu

(DRAFT IN PROGRESS)

March 18, 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 discrete shift in a widespread form of stigma — weight stigma — in the high-stakes setting of healthcare. Leveraging the prominent categorization and display of BMI in electronic medical records, I document a discontinuous increase in in-hospital mortality at the "obese" cutoff, without any associated changes in patient health. 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 using a physician-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 violation 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 University (the Economics Dept and the Negotiations and Management Group at HBS), the CHIBE 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 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 a trait perceived to violate a social norm. Crucially, the trait is seen as voluntarily assumed, inviting moral judgment and blame from both others and the self (Bénabou & Tirole, 2006, 2026; Lindbeck, Nyberg, & Weibull, 1999; Link & Phelan, 2001; Moffitt, 1983). Stigma is a powerful and distortionary social force. On one side, it causes individuals to undertake costly actions to avoid being stigmatized: those eligible for welfare may decline benefits (Besley & Coate, 1992; Celhay, Meyer, & Mittag, 2025), and people with stigmatized health conditions such as HIV or mental illness (Smith, 2025; Yang et al., 2021) may avoid lifesaving care. On the other side, it leads others to discriminate against perceived norm violators, even when the norm is arbitrary and the violation was not a choice — as when employers penalize individuals with a history of unemployment (Biewen & Steffes, 2010). In a world saturated with such norms, understanding how stigma shapes resource allocation and outcomes in real-world settings is essential.

Yet rigorous study of stigma faces two empirical challenges. The first, familiar from the discrimination literature, is that the stigmatized trait may embed outcome-relevant information, so observed gaps in outcomes can reflect signals observable to decision-makers but not to researchers. Identifying stigma's effect therefore requires a design that isolates exogenous variation in stigma while holding the true signal constant (Bertrand & Mullainathan, 2004; Lang & Lehmann, 2012; Phelps, 1972). The second challenge is measurement: stigma's defining feature — moral judgment — is rarely observable in administrative data. Much of the literature thus relies on experimental methods (e.g., Friedrichsen, König, and Schmacker, 2018; Oberholzer-Gee, 2008), which are well suited to demonstrating stigma's existence and isolating its psychological antecedents but less equipped to capture its real-world dynamics. Outside of experiments, researchers often rely on self-reported experiences of stigma (e.g., Cawley, 2004a; Pescosolido et al., 2021), which are prone to bias: individuals may report stigma where none exists and may fail to recognize their own stigmatizing behavior.[1] 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: *weight stigma.*

Few social norms are as universal and enduring as the one concerning body size. Across cultures and throughout history, there is typically a "socially acceptable" body size, and those who deviate from it face moral judgment — perceived and portrayed as lazy, clumsy, of lower ability, or burdensome to society.[2] This stigma is qualitatively well-documented in healthcare, where the belief that obesity

[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 take-up and show that the gap is larger in high-income areas — consistent with stigma, rather than eligibility or need, depressing program participation.

[2] Weight stigma is rooted in moralized social norms around an "ideology of blame" (Crandall, 1994). This is vividly illustrated by contemporary discourse around GLP-1 drugs and the rise of "Ozempic shaming" (Kolata, 2023): GLP-1 users face social censure for supposedly lacking willpower and "taking the easy way out," leading many to go to great lengths to conceal that they are taking weight loss drugs. Weight stigma also manifests across a range of other settings. In labor markets, obesity is associated with a wage penalty for women (Averett & Korenman, 1996; Cawley, 2004b) and substantially fewer interview callbacks for otherwise identical applicants (Rooth, 2009). In education, overweight and obese students receive harsher grades even controlling for test performance and rich covariates (Sabia, 2007). The bias

reflects personal failure — poor self-control over diet and exercise — shapes how clinicians treat obese individuals and how those individuals engage with the system (Hagan, 2021; Rubin, 2019a; Treasure & Ambwani, 2021). Beyond its social importance, what makes this setting particularly attractive for studying stigma empirically 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 prominently displayed 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 healthcare and in broader social contexts. In a patient's medical record, BMI is persistently and saliently displayed, and categorical distinctions (e.g., "normal," "overweight," "obese") are typically made salient through explicit labeling, color-coding, or other visual cues. Given that people tend to over-rely on categories at the expense of continuous information due to cognitive constraints (Fisher & Keil, 2018; McShane & Gal, 2017; Samet & Zureich, 2025), I hypothesize that formal BMI cutoffs produce discontinuous increases in provider stigma at the obesity threshold *without a corresponding discontinuous change in underlying 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. A simple conceptual framework formalizes the stigma channel and generates five testable predictions, which the empirical analysis tests in sequence.

I use MIMIC-IV, a de-identified electronic health record database covering all emergency department and ICU admissions at Beth Israel Deaconess Medical Center (Boston) from 2008 to 2022 (Johnson et al., 2023). Unlike typical administrative data, the database contains both patient BMI and free-text clinical notes — specifically, hospital discharge summaries — enabling me to link clinical labeling to provider language and patient outcomes. I limit the analytic sample to all hospital admissions with a non-missing BMI.

I begin with the outcome of interest – death – and work backward to uncover the mechanism. Using a regression discontinuity design with the local nonparametric estimator (Calonico, Cattaneo, & Farrell, 2020; Calonico et al., 2017), in-hospital mortality rises by approximately 0.4 percentage points — or roughly 40% — at the obesity cutoff. Standard RD diagnostics support validity: there is no evidence of manipulation in the BMI density, predetermined covariates and predicted mortality are balanced, placebo cutoffs show no discontinuities, and donut-hole estimates that address rounding are similar. The result is robust to alternative functional forms, covariate adjustment, kernels, clustering,

extends to interpersonal relationships (Blodorn et al., 2016), and media content analyses document frequent weight-based disparagement (Himes & Thompson, 2007). Cross-cultural evidence shows the bias reverses where fatness signals status, underscoring the role of social norms in driving weight stigma (Macchi, 2023).

and OLS specification, and no analogous mortality effects appear at other BMI cutoffs. Moreover, the mortality effect is concentrated among high-severity patients — those in the top quartile of the Elixhauser mortality index — while low-severity patients show no detectable effect. This is consistent with a mechanism in which the label reduces clinical effort: the mortality cost of any given effort reduction is largest for the sickest patients.

What could explain the mortality jump at the obesity cutoff? The RDD validity tests show no discontinuous change in underlying patient health that would account for it. One explanation is discrimination by providers due to weight stigma: the "obese" label may activate moral judgment by healthcare providers, distorting the diagnostic process—for example, by leading them to discount symptoms, incorrectly attribute them to patient weight, or reduce diagnostic effort—and ultimately worsen outcomes. A large qualitative literature is consistent with this channel: higher-weight patients frequently report that symptoms are discounted, tests are delayed or withheld, and concerns are dismissed, sometimes with serious consequences. I provide illustrative accounts from both patients and providers later in the paper.

However, before presenting evidence in support of the stigma mechanism, I consider and rule out three potential alternative explanations. First, changes in BMI-based clinical guidelines cannot account for the result either: no known protocol should *increase* mortality at the cutoff; clinical guidelines are not BMI-based, they are usually weight-based (e.g., medication dosing) and so do not map cleanly onto BMI thresholds; and the few guideline-triggered interventions that exist at the "obese" cutoff (e.g., obesity counseling, weight-loss medication) are outpatient measures designed to improve outcomes. Second, the "obese" label is also unlikely to harm patients by introducing additional clinical complexity, since every BMI range already carries a categorical label, and no analogous mortality discontinuity appears at the normal–overweight cutoff (BMI= 25), where a patient first receives a weight-related designation.

Third, the mortality effect may reflect diagnostic distraction rather than stigma: the salient "obese" label could draw provider attention toward weight and away from appropriate care, without requiring moral judgment. To distinguish these channels, I examine blood pressure classification thresholds. BP is an ideal comparison because it shares every feature that makes BMI labels salient to providers — a continuous biometric discretized into clinical stages, prominently displayed in the same EMR, governed by the same type of guideline-based cutoffs — but carries no social stigma. An RDD at each major BP cutoff yields no mortality increase; in fact, at the Stage I hypertension cutoff, mortality is *reduced*, consistent with a large prior literature showing that non-stigmatized clinical thresholds increase resource use and weakly improve outcomes by focusing clinical attention around the cutoff (Alalouf, Miller, & Wherry, 2024; Almond et al., 2010; Barreca et al., 2011; Bharadwaj, Løken, & Neilson, 2013; Chen et al., 2019; Fukuma et al., 2020; Geiger, Clapp, & Cohen, 2021; Petersen et al., 2020; Toshiaki et al., 2021). The obesity cutoff is the only clinical threshold linked to

increased mortality. What distinguishes it is that obesity is stigmatized.

In the mechanism section, I examine two dimensions of the MIMIC-IV data to explicitly probe the hypothesis of stigma-driven discrimination: (i) structured measures of clinical decision-making and (ii) unstructured clinical notes. If stigma worsens outcomes, it should be reflected in degraded diagnostic or treatment intensity for patients labeled "obese." Consistent with this hypothesis, the structured data show economically meaningful reductions in care intensity at the cutoff: patients just above the obesity threshold receive fewer diagnostic orders in the first 24 hours of hospital arrival, undergo fewer procedures on the intensive margin, and are less likely to receive common blood laboratory panels. The pattern of reduced clinical effort is difficult to reconcile with non-stigma explanations. If providers were simply over-weighting the BMI signal without stigma, higher perceived risk should, if anything, increase diagnostic intensity — as a large prior literature on clinical thresholds has shown. And if the label merely shifted attention toward weight-related conditions, one would expect a reallocation of diagnostic focus, not a broad reduction in orders and testing.

However, the most direct evidence that the observed increase in mortality and decrease in clinical resource use at the cutoff are driven by stigma comes from unstructured clinical notes. If the "obese" label triggers moral judgment and blame, it could be reflected in how providers describe patients in the medical record. I focus on discharge summaries, narrative documents written by the treating team to summarize the hospitalization, key findings, and clinical reasoning. To measure biased language at scale, I apply the physician-validated, zero-shot LLM approach of Apakama et al. (2024), which detects biased language with high performance (sensitivity 95%, specificity 86%). The prompt extracts all flagged instances in a note, classifies each into one of four categories – Discrediting, Judgmental, Stereotyping, and Pejorative Labeling – and provides a brief rationale for each flag. For example, the LLM identifies the following phrase in a patient's discharge summary as an instance of Discrediting language: "*She was lying in bed comfortably, checking her facebook page. She immediately started to complain of pain the moment I walked in*", and outputs the following reason for doing so: "*Implies the patient's pain is not genuine and only performed for the clinician's benefit, casting unjustified doubt on her reported symptoms*".

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, providers differ substantially in their baseline propensity to use stigmatizing language. I classify providers by their average rate of stigmatizing descriptors, defining "High-stigma" providers as those in the top quartile. I then estimate a heterogeneous RD — a cross-sectional Differences-in-Discontinuities (Diff-in-Disc) design — using the procedure of Calonico et al. (2025), which yields a robust, bias-corrected difference between the effect of the obesity discontinuity

among High-stigma providers and the effect among Low-stigma providers.

Using this Diff-in-Disc design, I find two patterns consistent with a stigma-driven mortality mechanism: (1) the obesity cutoff increases the use of stigmatizing language significantly more among High-stigma providers relative to Low-stigma providers, and (2) the mortality effect of the cutoff is significantly higher amongst the High-stigma providers than the Low-stigma providers (in fact, the effect of the obesity cutoff on mortality is almost entirely concentrated amongst these providers). This is consistent with an explanation that weight stigma, unobservable to the researcher, is manifesting in both stigmatizing language, and lower quality clinical care that is leading to worse outcomes for the patient.

I present further evidence in support of the stigma-driven mortality mechanism by testing two defining features of stigma: (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 of stigma using two analyses. First, I show that the effect of the obesity cutoff on stigmatizing language is larger for shorter patients than for taller patients. The intuition is straightforward: two patients with the same BMI will differ in physical appearance—the shorter patient will appear larger, more visibly violating norms around body size, and thus the "obese" label should trigger more stigma for shorter patients than it would for taller ones. Second, I examine within-patient variation in stigmatizing language across consecutive hospitalizations, using each patient's prior visit as a natural within-person reference point—or "norm"—for the physician. Restricting to patients with multiple encounters and including both patient and physician fixed effects, I find that stigmatizing language increases when the same patient moves into a higher BMI category relative to their last admission. This pattern aligns with a core feature of stigma: it is norm-based and can shift even within an individual as they move closer to or further from their norm.

I next test the "moral judgment" component of stigma, providing three pieces of evidence. A distinctive feature of the model validated in Apakama et al. (2024) is that it not only identifies instances of stigma in clinical notes 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 results described above are largest for this category. The cutoff increases the use of judgmental language significantly more among High-stigma providers than among Low-stigma providers. (Discrediting and stereotyping language also increase.) Second, the cutoff increases not merely the presence of judgmental language in isolation but also the co-occurrence of judgmental terms and weight-related mentions (e.g., "fat," "obese," "overweight," "heavyset") within the same note, again more so for High-stigma providers. Notably, when the outcome is the co-occurrence of bias type and weight mentions, the only category that responds to the obesity cutoff is judgmental language. 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—both in isolation and jointly with weight-related words—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. This paper contributes to and extends two primary literatures: (i) on discrimination, and (ii) on stigma.

This paper advances the economics literature on *discrimination* in two ways. First, it provides direct evidence of taste-based discrimination. Economics typically distinguishes statistical discrimination, arising from incomplete information, from taste-based discrimination, where decision-makers incur costs to avoid associating with certain groups. Stigma-driven discrimination is a form of taste-based discrimination in which the relevant "taste" reflects not animus, homophily, or status threat, but negative moral judgments toward individuals perceived to violate a social norm. Yet most economic evidence for taste-based discrimination is indirect – showing that disparities persist after controlling for observable productivity, or that outcomes covary with proxies for prejudice (e.g., Bertrand and Mullainathan, 2004; Charles and Guryan, 2008). This paper goes further. A discontinuous increase in stigmatizing language at the obesity cutoff, concentrated among high-stigma providers, reveals not only differential treatment but the beliefs motivating it, which are observed in decision-makers' own words. The combination of reduced clinical effort and biased documentation at the same threshold provides joint evidence of discriminatory intent *and* discriminatory action in the same setting, closing a noticeable gap in the literature.

Second, it identifies a form of *transient discrimination* — discrimination triggered not by an immutable characteristic but salient signals that emphasize individual traits that deviate from social norms. The canonical discrimination literature focuses on ascribed traits — race, gender, age, disability — that individuals cannot shed. The discrimination documented here is different in kind: if it is activated by how an institution displays information, then it is also theoretically reversible through institutional redesign. That is, if the "obese" label in an electronic medical record triggers discriminatory behavior, then modifying the interface can, in principle, eliminate it. This distinction has first-order policy implications, opening a class of interventions — redesigning decision environments rather than changing attitudes to reduce discrimination — the viability of which the broader literature has not yet considered in depth.

This paper also makes four contributions to the *stigma* literature. First, it provides what is,

to my knowledge, the first credible evidence tracing the causal linkage between stigma and mortality. By exploiting a clinical labeling threshold that discretely changes categorization without changing underlying health, the design resolves the identification problem that has long plagued this literature: separating the causal effect of stigma from the informational content of the stigmatized trait. Despite a large body of theoretical and empirical work across economics, sociology, and public health, this fundamental confound has remained largely unaddressed. Second, the outcome is mortality — not program take-up, medication adherence, or self-reported wellbeing, but whether the patient survives. Most prior empirical work on stigma examines intermediate outcomes, leaving unanswered the question of whether stigma ultimately affects the outcomes that matter most for welfare. Third, this paper observes stigma as it is produced, not as it is perceived. Rather than relying on surveys, vignettes, or inferred behavioral responses, it captures stigmatizing language in the written clinical documentation of decision-makers — providers professionally bound to treat patients impartially. This shifts the empirical focus from the experiences of the stigmatized to the actions of the stigmatizer, a perspective that the literature has called for but not achieved thus far. Fourth and finally, by establishing that stigma carries a tangible, quantifiable mortality cost, this paper implies that interventions which reduce the prevalence of stigmatized conditions generate welfare gains that current cost-benefit frameworks do not capture. This is particularly relevant for GLP-1 receptor agonists, the class of anti-obesity medications whose clinical valuations are based primarily on weight loss, cardiovascular risk reduction, and metabolic outcomes. To the extent that reducing obesity also reduces exposure to the "obese" label — and therefore to the stigma-driven discrimination documented here — the welfare gains from these therapies are larger than existing valuations reflect.

2 Setting, Framework, 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.1). 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 – changes 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 Conceptual Framework

The setting described above has three features that lend themselves to a simple model: providers choose how much effort to exert on each patient, patients receive a discrete label at the obesity cutoff, and that label — unlike labels at other clinical thresholds — carries social stigma. I now formalize these features to generate testable predictions that guide the empirical analysis.

Consider a provider j treating patient i. The patient has underlying severity θ_i, which varies continuously with BMI, and receives a discrete label $L_i = \mathbf{1}\{\text{BMI}_i \geq 30\}$ displayed in the electronic medical record. The provider chooses clinical effort $e \geq 0$.

Patient mortality is $m(\theta_i, e)$, with $m_\theta > 0$ (sicker patients are more likely to die), $m_e < 0$ (effort reduces mortality), and $m_{e\theta} < 0$ (effort is more productive for sicker patients, so underinvestment in effort is most consequential for the most severe cases).

The provider maximizes

$$V_j = -m(\theta_i, e) - c(e) - \sigma_j \cdot L_i \cdot \psi(e) \tag{1}$$

The first term captures the provider's concern for patient outcomes — whether through professional norms, altruism, or reputational incentives. The second is the direct cost of effort, with $c' > 0$ and $c'' > 0$. The third term is the stigma channel: when the "obese" label is present ($L_i = 1$) and the provider has positive stigma propensity ($\sigma_j > 0$), exerting effort on the patient generates additional psychological disutility. This is a Beckerian taste-based discrimination channel: interacting more intensively with a stigmatized patient is more aversive (Becker, 2010), not because of preferences for homophily or animus (as is usually the case in such models), but because of moral considerations that microfound stigma. The function $\psi(e)$ satisfies $\psi' > 0$ and $\psi'' \geq 0$. The parameter $\sigma_j \geq 0$ indexes provider j's propensity to stigmatize, which varies across providers but is fixed within a provider. (This model does not attempt to speak to supply-side determinants of σ_j, or why it varies from provider-to-provider.

The first-order condition is:

$$-m_e(\theta_i, e^*) = c'(e^*) + \sigma_j \cdot L_i \cdot \psi'(e^*) \tag{2}$$

The left-hand side is the marginal benefit of effort: the reduction in patient mortality from one

additional unit of effort. The right-hand side is the marginal cost: the direct cost of effort plus, when the label is active and the provider stigmatizes, the marginal stigma disutility. I present the model's testable predictions in the order they are tested in the paper, beginning with the reduced-form consequence most directly observable in the data — mortality — before turning to the underlying channels.

*Prediction 1: Mortality rises at the cutoff..*When L_i switches from 0 to 1 at BMI = 30, the marginal cost of effort increases for any provider with $\sigma_j > 0$, reducing equilibrium effort e^*. Since $m_e < 0$, lower effort increases mortality. The model therefore predicts a discontinuous increase in mortality at the obesity threshold.

Prediction 2: Mortality effects are concentrated among high-severity patients.. Since $m_{e\theta} < 0$, the mortality cost of any given effort reduction is larger for sicker patients. The model therefore predicts that the mortality effect of the obesity label should be concentrated among high-severity patients: those for whom effort is most productive suffer the greatest harm when stigma reduces it.

*Prediction 3: Clinical effort falls at the cutoff..*The mortality effect operates through reduced effort: the "obese" label raises the marginal cost of effort via the stigma term, so patients just above the cutoff should receive fewer diagnostic tests, procedures, and other clinical resources than patients just below.

*Prediction 4: Effects are concentrated among High-stigma providers.*The effort reduction at the cutoff is proportional to σ_j. Providers with a higher propensity to stigmatize should exhibit larger increases in mortality at the cutoff, while providers with low σ_j should show little or no response to the label.

*Prediction 5: Effects are amplified when norm deviation is more salient.*If σ_j responds to the perceived salience of norm violation — for instance, if obesity is more visually apparent for some patients than others at the same BMI — then the stigma channel should be stronger when the deviation from body-size norms is more conspicuous.

2.3 Data

Testing these predictions requires data on patient mortality, clinical effort, provider behavior, and the content of clinical documentation. I use MIMIC-IV v3.1 (Johnson et al., 2025; Johnson et al., 2023), a large deidentified electronic health record database that contains all four, 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 MIMIC-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 three 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.[3] I assign BMI categories using this continuous BMI variable and following standard clinical guidelines: underweight (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) (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. My final exclusion criteria is to remove any encounters with missing data for primary provider (0.9% of the data). The final sample for the mortality analysis comprises of 166,730 patients. Summary statistics for this sample, such as demographics, insurance status, BMI categories, and chronic conditions, are presented in Table 1.

3 Mortality Effects of the Obesity Cutoff

3.1 Empirical Strategy.

I use in-hospital mortality (i.e., death during the hospital visit) as my main outcome. Figure I 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

[3] 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 $\geq$ 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.

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 ≥ 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 $X_i \equiv \text{BMI}_i - 30$ denote the running variable for patient i, and let $D_i = \mathbf{1}\{X_i \geq 0\}$ 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):

$$Y_i = \gamma_0 + \tau\, D_i + \gamma_l\, f_l(X_i) + \gamma_r\, f_r(X_i) \times D_i + \varepsilon_i, \tag{3}$$

where Y_i is an indicator for in-hospital mortality, and $f_l(\cdot)$ and $f_r(\cdot)$ flexibly capture the smooth relationship between BMI and mortality on either side of the cutoff (implemented via local polynomials within the chosen bandwidth). τ is the RD estimate of the effect of the obesity cutoff on the outcome.

Table 2 probes the sensitivity of the main result to polynomial order, kernel shape, covariate adjustment, clustering, and estimation method. The baseline specification (Column 1) uses a local linear fit ($p = 1$) with a triangular kernel, MSE-optimal bandwidth (h_{mserd}), and robust, bias-corrected (RBC) standard errors clustered at the patient level. Columns 2–4 vary polynomial order and kernel shape: Column 2 switches to a uniform kernel, Column 3 uses a local quadratic ($p = 2$) with a triangular kernel, and Column 4 combines the quadratic with a uniform kernel. Column 5 returns to the baseline specification but adds predetermined covariates (age, sex, and race). Column 6 uses two-way clustering by patient and hospital service — the organizational unit responsible for the patient's care during the hospitalization (e.g., internal medicine, general surgery, labor and delivery) — to capture within-service correlation in documentation and care practices, which is the primary clustering scheme for the stigma analyses in the mechanism section since clinical notes are produced by service teams. Finally, Column 7 estimates the discontinuity parametrically via OLS using a linear interaction model within a ± 2 BMI window around the cutoff, with patient-clustered standard errors.[4]

3.2 Results.

In Figure 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. Consistent with Prediction 1, Table 2 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).

[4]The OLS specification estimates $Y_i = \alpha + \tau\, D_i + \beta_1 \tilde{X}_i + \beta_2 \tilde{X}_i \times D_i + \varepsilon_i$ for $|\tilde{X}_i| \leq 2$, where $\tilde{X}_i = \text{BMI}_i - 30$.

Results from re-estimating Equation 3 using the overweight, obese II, and obese III cutoff are presented in Table A.3. 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.

3.3 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.2 shows the data do not allow the rejection of the null of no sorting at the threshold (p = 0.15).

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 co-morbidities, and the Elixhauser mortality index (a widely used and validated measure based on 31 high-risk comorbidities) (Elixhauser et al., 1998). Figure IIIa 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 A.1). In Appendix Table A.2, 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 2 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 A.1 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 p-value for the McCrary Density test for this distribution is now p=0.35.) The donut-hole estimates are presented in Table A.1 Column (4) and are similar to the results from the main specification.

3.4 Heterogeneity by patient severity

Prediction 2 of the conceptual framework states that, because $m_{e\theta} < 0$, the mortality consequences of reduced effort should be concentrated among high-severity patients. I test this using a heterogeneous RD design — or, equivalently, a cross-sectional Differences-in-Discontinuities (Diff-in-Disc) design — that compares the RD effect of the obesity cutoff across patient subgroups. I use the procedure of Calonico et al. (2025), which extends the Calonico, Cattaneo, and Titiunik (2014) framework to allow for heterogeneous treatment effects along a pre-specified covariate. The procedure estimates separate local-polynomial RD regressions within each subgroup and constructs a robust, bias-corrected test of the difference between them, with valid inference that accounts for the data-driven bandwidth selection in each subgroup. This approach is preferable to ad hoc alternatives (such as estimating group-specific RDs and comparing coefficients via post-estimation tests) because it jointly optimizes bandwidth choice and bias correction across subgroups.

I split patients by their Elixhauser mortality index — a validated comorbidity-based predictor of in-hospital death constructed from 31 diagnostic categories (Elixhauser et al., 1998). I define high-severity patients as those in the top quartile of the index and low-severity patients as those in the bottom three quartiles. Table 3 reports the results across five specifications that mirror the main mortality robustness analysis: (1) the baseline local linear RD with patient-clustered standard errors; (2) a local quadratic polynomial; (3) the baseline specification with covariates (age, race, sex, private insurance); (4) two-way clustering by patient and hospital service; and (5) a parametric OLS estimate within a ±2 BMI window around the cutoff.

The mortality effect of the obesity cutoff is driven entirely by high-severity patients: the RD estimate for these patients is large and statistically significant across all specifications, while the estimate for low-severity patients is close to zero. The Diff-in-Disc estimate — the difference in the mortality effect between high- and low-severity patients — is consistently positive, confirming that the

mortality penalty of the "obese" label falls disproportionately on the sickest patients. This pattern is consistent with $m_{e\theta} < 0$: when stigma-induced effort reductions are applied uniformly, the patients for whom effort matters most suffer the greatest harm.

3.5 Ruling out alternative explanations

The conceptual framework in Section 2.2 predicts that the obesity cutoff increases mortality (Prediction 1) and that this effect is concentrated among the sickest patients (Prediction 2). Before testing the remaining predictions, I consider three alternative explanations for the mortality result that do not involve stigma.

Could changes in BMI-based clinical guidelines cause this jump in mortality?. This is an appealing alternative explanation, but it is unlikely to be driving the result for several reasons. First, and most logically, there is no known protocol that should *increase* mortality – that too discontinuously – at the cutoff. Guidelines are designed to improve outcomes, not worsen them, so any guideline-based posited mechanism would need to explain why crossing the obesity threshold leads to worse outcomes. Second, clinical guidelines are not usually BMI-based, rather, they are most commonly weight-based (e.g., medication doses are based on ideal or actual weight, Erstad and Barletta, 2021), and even then do not meaningfully change until higher BMI categories (e.g., BMI ≥ 40). Since weight does not map linearly into BMI, any changes in weight-based guidelines would not show up so cleanly at the obesity cutoff. Third, treatment guidelines that *are* triggered at BMI 30—obesity counseling, weight-loss medication initiation, and specialist referrals— both improve outcomes and fall within the purview of outpatient and primary care, and thus cannot logically point toward an increased in-hospital mortality mechanism. Overall, the explanation that clinical guidelines change at the obesity cutoff which then leads to patient death is implausible.

Could the "obese" label increase complexity in clinical decision-making, thereby harming patient outcomes?. This alternative explanation seems unlikely for three reasons. First, the "obese" label applies to all patients with $BMI \geq 30$, not just those near the cutoff; if the label introduced harmful complexity, outcomes should worsen across the entire obese category, not discontinuously at the threshold. Second, every BMI range carries a categorical label—"underweight," "normal," "overweight," "obese"—so crossing the cutoff does not introduce a label where none existed; it simply replaces one label with another. Third, if the concern is that "obese" conveys more clinically actionable information than "overweight", thereby increasing cognitive load, the same logic should produce a mortality discontinuity at the normal-to-overweight cutoff (BMI = 25), where a patient transitions from a label carrying no weight-related clinical salience to one that does, but there is no such mortality discontinuity.

Could label salience alone, without stigma, explain the mortality effect?. The model attributes the mortality effect to stigma: the $\sigma_j \cdot L_i \cdot \psi(e)$ term raises the psychic cost of effort for stigmatized patients,

increasing mortality (Prediction 1) by reducing effort (Prediction 3). An alternative interpretation preserves the role of the label's salience, but not the label's stigma. If the salient "obese" classification simply makes weight salient — mistakenly drawing provider attention toward obesity as a potential cause of illness, and away from appropriate care — then mortality could rise through diagnostic distraction rather than effort withdrawal. These two channels are observationally distinct in the model, and I exploit two differences between them.

First, the channels generate opposing predictions for total clinical effort. The stigma model predicts that effort falls at the cutoff (Prediction 3), because $\sigma_j > 0$ raises the marginal psychic cost of caring for stigmatized patients. The distraction channel, by contrast, implies a reallocation of effort across tasks with no necessary decline in total quantity. I will show in the subsequent Section **??**, that structured measures of clinical effort — diagnostic tests, procedures, and labs — fall discretely at the cutoff, consistent with the stigma channel and difficult to reconcile with distraction.

Second, under the distraction alternative, the label need not be stigmatized to affect provider behavior. This amounts to adding a salience term to the provider's objective, $V_j = -m(\theta_i, e) - c(e) - \sigma_j \cdot L_i \cdot \psi(e) - L_i \cdot \delta(e)$, where $\delta(e)$ captures any salience cost of a categorical label regardless of stigma. At the obesity cutoff, both the stigma and distraction channels are active (because $\sigma_j > 0$); isolating the distraction channel requires a threshold where $\sigma_j = 0$, so that only $L_i \cdot \delta(e)$ operates. Blood pressure classification provides an ideal such test.

Like BMI, BP is a continuous biometric measured and recorded for every patient at admission; like BMI, it is categorized into discrete clinical stages defined by guideline-based cutoffs; like BMI, it is prominently displayed in the same EMR interface (see Figure A.1); like BMI, it is a chronic risk factor associated with cardiovascular and metabolic disease; and like BMI, its classification thresholds are the subject of longstanding clinical guidelines that providers are trained to be attuned to. The two measures also share similar prevalence in my sample: 34% of patients have uncomplicated hypertension compared with 31% who are overweight, and 20% have complicated hypertension compared with 19% who are obese (Table 1). Crossing a BP threshold thus sets $L_i = 1$, generating the same type of salient categorical label, through the same EMR display, for the same patient population, but with $\sigma_j = 0$, since hypertension carries no social stigma. If the salience channel $\delta(e)$ meaningfully contributes to mortality, BP thresholds should produce harmful effects. If instead the mortality effect at the obesity cutoff is driven entirely by $\psi(e)$ operating through $\sigma_j > 0$, these thresholds should have no harmful effect on mortality.

I estimate a binding-score RDD at each major BP classification threshold, where the running variable is defined by the patient's first systolic and diastolic blood pressure readings during the hospitalization.[5] Because hypertension classification guidelines have changed over the past decade

[5]When the treatment threshold is defined jointly by two running variables — here, systolic and diastolic blood pressure — the standard univariate RDD cannot be applied directly. I follow Choi and Lee (2018;), Kämpfen and Mosca (2024) and construct a *binding score*: for each cutoff, I standardize the distance of each BP dimension from its respective

— and I cannot confirm exactly which categories were displayed in the EMR at the time of each encounter — I estimate the RD at every plausible cutoff that may have been shown: Prehypertension (systolic $\geq$ 120 or diastolic $\geq$ 80; Chobanian et al. 2003), Stage 1 Hypertension (systolic $\geq$ 130 or diastolic $\geq$ 80), Stage 2 Hypertension (systolic $\geq$ 140 or diastolic $\geq$ 90; Whelton et al. 2018), and Hypertensive Crisis (systolic $\geq$ 180 or diastolic $\geq$ 120). Results are presented in Table A.11 and rule out the distraction channel: not only are all estimates negatively signed — the opposite direction from the BMI results — the effect at the Stage 1 threshold on mortality is statistically significant and negative.

This effect of the hypertension cutoff replicates prior research in outpatient settings, also using RDD designs, that hypertension thresholds improve patient outcomes, largely by raising awareness and increasing clinical resource use around hypertension cutoffs (Chen et al., 2019; Dai et al., 2022). Moreover, this pattern is also consistent with a large body of prior research showing that clinical thresholds — operating as salience devices that concentrate clinician attention — increase resource use and weakly improve patient outcomes. The *Very Low Birth Weight* classification at 1,500 grams triggered higher spending and weakly improved short- and long-term outcomes (Almond et al., 2010; Barreca et al., 2011; Bharadwaj, Løken, & Neilson, 2013). The *Advanced Maternal Age* cutoff at 35 sharply reduced adverse birth outcomes (Geiger, Clapp, & Cohen, 2021). The HbA1c 6.5% diabetes diagnosis threshold increased spending and improved outcomes (Alalouf, Miller, & Wherry, 2024; Petersen et al., 2020; Toshiaki et al., 2021), and similar patterns appear at the cancer screening age cutoffs (Srikanth & Erin, 2016). In every such case, $\sigma_j = 0$: these conditions are not socially stigmatized, and their thresholds improve (or leave unchanged patient outcomes.

The effect of the obesity cutoff stands in stark contrast. Where every other clinical threshold in prior work weakly improves outcomes and resource use, crossing the obesity threshold *increases* mortality. What distinguishes the obesity threshold is that $\sigma_j > 0$: obesity is stigmatized in ways these other conditions are not, activating the $\psi(e)$ channel that is absent at non-stigmatized thresholds. The stigmatizing nature of the obesity threshold is well-documented in prior research using similar RDD designs: crossing the "overweight" and "obese" BMI cutoffs in RDD designs has been shown to cause significant declines in subjective well-being (Zeng & Yu, 2019) and substantial labor market earnings losses (Levasseur, 2019).

Having established the mortality result and ruled out the standard alternative explanations, I now turn to testing the remaining predictions of the framework: that clinical effort falls at the cutoff (Prediction 3), that effects are concentrated among High-stigma providers (Prediction 4), and that effects are amplified when norm deviation is more salient (Prediction 5).

threshold and take the maximum. The resulting scalar equals zero exactly when the patient's binding constraint first reaches the classification boundary, and is positive when the patient exceeds the threshold on at least one dimension. This reduces the two-dimensional frontier to a single running variable, enabling standard local polynomial RD estimation at the binding point.

4 Mechanism: Stigma

4.1 Weight stigma in healthcare

The conceptual framework predicts that the obesity label increases mortality by activating provider stigma, reducing clinical effort. Before turning to the quantitative tests, I briefly review the descriptive evidence that weight stigma is prevalent in healthcare — providing context for the formal predictions. There is a wealth of 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.4 and A.5 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 not being taken seriously by providers because of their weight ("*what is with doctors blaming everything on weight*", Figure A.4 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 language ("*obese patients can ruin your day*", Figure A.5 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, *JAMA*, *NEJM*, and *The Lancet*, have all published perspective pieces explicitly validating and decrying weight stigma in healthcare (Hagan, 2021; Rubin, 2019b; Treasure & Ambwani, 2021). There are also countless popular news articles such as in the New York Times 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 (Phelan et al., 2015; Puhl & Heuer, 2009). 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 by patients and that perception may contribute to behaviors that harm health.

I now test the framework's remaining predictions using two dimensions of the MIMIC-IV data. First, I analyze the structured data on clinical decision-making — provider orders, diagnostic tests, labs, and procedures — to test whether clinical effort falls at the cutoff (Prediction 3). If stigma is influencing care, providers may discount patients' symptoms, attributing them to weight-related issues or deeming further diagnostic effort unwarranted, resulting in fewer diagnostic tests, procedures, or treatments for patients just above the cutoff.

Second, I turn to the unstructured data — specifically, the text of discharge summaries — to test whether stigmatizing language increases at the cutoff and whether this effect is concentrated among High-stigma providers (Prediction 4). I also test whether effects are amplified when norm deviation is more salient (Prediction 5). I detect stigmatizing language using large language models, following the physician-validated approach of Apakama et al. (2024). Importantly, biased language need not reference weight explicitly and may 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.

4.2 Analysis of Structured Data

4.2.1 Variable Construction

*Clinical orders in the first 24 hours of arrival*MIMIC-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 (GPT 5.2) 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.

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: total number of procedures a patient receives during their hospital stay; and since 25% of patients have no procedures listed on their record, the number of procedures conditional on receiving at least one (i.e., the intensive margin).

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, specimen type, the timing of collection, and the specific test performed. I focus on blood chemistry and blood hematology panels, which together comprise the core first-line diagnostic workup in inpatient settings. Blood chemistry panels capture metabolic status and organ function (e.g., electrolytes, glucose, creatinine, liver enzymes), while blood hematology panels capture blood-cell composition and related markers of infection, anemia, and bleeding risk (e.g., white blood cells, hemoglobin, platelets). The choice to order blood labs reflects a provider's judgment that the patient's presentation warrants a standard diagnostic workup, making them a natural margin on which to detect reductions in routine diagnostic effort. For each panel type, I construct an indicator for whether the patient received any such panel during the hospitalization.

4.2.2 Empirical strategy and results

Figure IV reports RD estimates of the effect of the obesity cutoff on these structured measures of clinical effort described above: total orders within 24 hrs of hospital arrival; diagnostic orders within 24 hrs of hospital arrival; total number of procedures; total number of procedures conditional on having at least one; any blood chemistry panel; and any blood hematology panel. For each dependent variable, the figure 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, sex, insurance, and age). Specification (4) uses two-way clustering by patient and hospital service. Specification (5) uses OLS.

The broader pattern in Figure IV points to a reduction in clinical resource use for patients just above the obesity cutoff relative to those just below. The clearest result is for diagnostic orders placed within 24 hours of arrival, which show a statistically significant reduction across all five specifications. The remaining outcomes reinforce this pattern: the number of procedures conditional on receiving at least one, and the probability of receiving blood chemistry and blood hematology panels, all show negative point estimates, with at least one specification reaching significance at the 10% or 5% level. (The one exception is the total number of procedures, where the point estimate is negative and statistically insignificant).

Taken together, these results are consistent with Prediction 3: the obesity cutoff reduces the intensity of the clinical workup. This pattern supports a stigma-based mechanism in which the "obese" label reduces providers' diagnostic effort, though it is not sufficient on its own to establish stigma as the channel. The direction of the effects does, however, help narrow the set of plausible explanations. As discussed in Section 3.5, the distraction alternative predicts that providers overweight the salient label in their clinical reasoning. But higher BMI is generally regarded as a signal of elevated clinical risk, so if providers were responding to the informational content of the label, one would expect diagnostic intensity to *increase* at the cutoff, consistent with the pattern documented in prior RDD studies at other clinical thresholds. Moreover, if the label redirected clinical attention toward weight-related conditions, the expected result would be a reallocation of diagnostic effort, not such a consistent reduction in orders, procedures, and laboratory testing. The structured data therefore point away from a pure salience or distraction mechanism: the obesity cutoff leads providers to do *less*, not more. I now turn to unstructured clinical documentation to test Predictions 4 and 5 and provide more direct evidence of stigma.[6]

4.3 Analysis of Unstructured Data

In this section, I test whether the obesity cutoff increases the prevalence of stigmatizing language in free-text clinical notes – which along with the reduced effort result documented prior, may provide more direct evidence that the "obese" label activates the stigma channel posited in the conceptual model. (Note that that this analysis is *not* designed to examine whether stigmatizing language *causes* mortality.) In the framework, the "obese" label triggers stigma among providers with $\sigma_j > 0$, which raises the psychological cost of effort and degrades care. If this channel is operative, it should leave traces in how providers document the patient. The analysis tests Prediction 4 (effects concentrated

[6]To transparently report the full set of tests I conducted, I also examine the effect of the cutoff on total consults, 30-day readmissions, and length of stay, and I find no statistically significant discontinuities at the cutoff for any of these outcomes.

among High-stigma providers) and Prediction 5 (effects amplified when norm deviation is more salient), with the latter discussed in greater detail in the relevant subsections.

I focus on hospital discharge summaries, which are well-suited for detecting stigmatizing language for two reasons. First, they are lengthy narrative documents — averaging 1,674 words in my data — written at the conclusion of care to summarize the patient's entire hospital course, giving providers ample scope to reveal how they perceived and characterized the patient. Second, they are drafted by the junior physicians and hospital service team members most directly involved in day-to-day care, then reviewed and signed by the attending physician of record (Kantor, Evans, & Shieh, 2015; Kind & Smith, 2008; Ng et al., 2025), meaning they reflect the documentation habits, cultural norms, and implicit attitudes of the provider and hospital service team responsible for the hospitalization.

What would weight stigma look like in clinical notes? A large literature documents the prevalence of biased language in clinical documentation (Healy, Richard, & Kidia, 2022; Himmelstein, Bates, & Zhou, 2022; Park et al., 2021; Sun et al., 2022), but stigma is unlikely to appear as explicit or overtly discriminatory commentary about weight. Clinical settings impose strong professional constraints against overt disparagement, and when discriminatory motives are normatively unacceptable, bias tends to surface indirectly.[7] 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, Wilder et al., 2021); 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 exhibiting the same behavior).

4.3.1 Variable Construction.

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).

The category definitions and review procedures were developed by an institution-wide documentation initiative (the FAIR documentation workgroup) through an extensive literature review

[7] See, for example, the justification–suppression model of Crandall and Eshleman (2003) and the aversive racism framework of Dovidio and Gaertner (2000); as well as literature showing that employers often refer to "fit" to justify discrimination.

and interviews with patients and providers; these definitions structure the LLM prompt. Apakama et al., 2024 use GPT-4 in a zero-shot framework (no fine-tuning) to detect biased language at scale. The authors iteratively refined the prompt on a balanced set of 100 EMR sentences, then deployed it via the GPT-4 OpenAzure HIPAA-compliant API. For each note, the model flags whether biased language is present and classifies each instance by category. To validate the classifier, two physician reviewers independently assessed GPT-4's bias flags and subtype assignments. Crucially, the testing and validation sample in Apakama et al. (2024) includes MIMIC-IV discharge summaries — the same dataset, note type, and clinical setting used in my analysis — meaning the classifier has been validated directly on the data source I study. The classifier performs well: for "any bias" present, sensitivity is 95%, specificity 86%, positive predictive value 84%, and negative predictive value 96%, with 90% agreement between model and human reviewers. The full prompt is reproduced in Appendix Section A.

Relative to 2024, I make four modifications to the prompt. 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 subject to stereotyping to include "obese or overweight people," to better align the prompt with the mechanism of interest. Fourth, I restrict the judgmental category to negative judgment, since positive judgments are not informative about stigmatizing documentation.[8] Due to computational constraints, 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 in this subsample (RD estimate $= 0.006$, SE $= 0.003$, $p < 0.05$).

For each discharge summary, the prompt returns: (1) all flagged instances of biased language; (2) a classification of each instance into one of the four bias types (*Judgmental*, *Discrediting*, *Pejorative Labeling*, or *Stereotyping*); and (3) a brief rationale for each classification. From this output, I construct the primary outcome variable, *Any Stigma in Note*, an indicator equal to one if the discharge summary contains at least one flagged instance. This binary measure corresponds to the validated classification in 2024, I later disaggregate by bias category to assess which types of language are most responsive to the obesity cutoff.

Definition excerpts and example LLM output for each bias type are presented in Table 4. As a descriptive validation of the classifier, Figure V shows univariate correlations between *Any Stigma* and patient characteristics that prior literature suggests should predict differential documentation. The correlations align with expectations: biased language is more prevalent among Black patients, Medicaid recipients, and patients with substance use or mental illness diagnoses — conditions that carry well-documented stigma in clinical settings. The strongest positive correlates are alcohol and

[8]The prompt also returns a quantitative severity measure and a count of bias instances, but I do not use these outputs as they are not validated in 2024,

drug use diagnoses, with roughly 30% of biased language occurring among patients with an ICD-9/10 code for substance use. Conversely, biased language is less prevalent among privately insured patients and amongst patients with cancer diagnoses (conditions generally regarded as sympathetic rather than stigmatized). Interestingly, being married is the most protective against stigmatizing documentation, which is consistent with two channels: having a spouse present may corroborate patient-reported symptoms, reducing the discrediting language that arises when providers doubt unaccompanied patients' accounts; and having a companion to assist with daily needs during the hospitalization may reduce the patient's reliance on providers, lowering the friction that can prompt negative documentation. Female patients also exhibit lower rates of explicitly biased language, consistent with prior work suggesting that bias toward women is more likely to manifest as dismissal or inattention than as overtly stigmatizing documentation. Notably, having an ICD diagnosis code for obesity is not itself a strong predictor of biased language, suggesting that the classifier is capturing documentation style rather than mechanically flagging notes that mention weight-related clinical terminology.

Before turning to the RD estimates, I examine descriptively whether provider attention to weight in clinical notes varies with BMI. I construct a simple dictionary-based indicator that flags a note if it contains any weight-related term, identified via regular expression.[9] I apply this dictionary to two sources: the raw clinical note and the LLM output. In the raw text, the dictionary captures any mention of weight — clinical or otherwise — and is therefore a broad measure of whether providers attend to body size at all. In the LLM output, a weight-related term appears only if the classifier flagged it as an instance of biased language or cited it in its reasoning, providing a narrower measure of whether that attention is stigmatizing. Panel (a) of Figure VI plots the share of notes containing at least one weight-related word in the raw text; Panel (b) applies the same dictionary to the LLM output. Both panels show that weight-related language increases with BMI — suggesting that increasing body size is something providers notice and document regardless of the label. The substantially lower rate in the LLM output than in the raw notes is consistent with the classifier filtering out benign clinical mentions and retaining instances where provider attention to weight takes a stigmatizing form.

4.3.2 Empirical strategy and results.

The primary outcome is the indicator *Any Stigma in Note*. I estimate the effect of the obesity cutoff using the Calonico, Cattaneo, and Farrell (2020) local polynomial RD estimator with a triangular kernel, MSE-optimal bandwidth, and robust, bias-corrected standard errors clustered two ways by patient and hospital service, because discharge summaries are produced by service teams that share note-writers, templates, and documentation norms. I also show robustness to a local quadratic specification, covariate adjustment, an alternative two-way clustering scheme by primary physician

[9] The dictionary includes stems and expressions such as *obese/obesity, overweight, underweight, fat, heavyset, habitus, BMI, body mass index, wt, weight,* and *weigh*, as well as numeric weight expressions (e.g., "80 kg," "175 lbs") and BMI unit notation (kg/m^2). The indicator requires no language model.

and hospital service, and OLS regression.

Panel A of Table 5 shows that the obesity cutoff does not significantly affect the average likelihood of stigmatizing language in a discharge summary. This is not surprising: the conceptual framework does not predict a uniform increase in stigma across all providers. The stigma channel operates through σ_j, which varies across providers — only those with $\sigma_j > 0$ should respond to the label. If a substantial share of providers have low or zero stigma propensity, the average effect will be attenuated even if the cutoff strongly increases stigma among high-σ_j providers. Measurement noise in the LLM classifier compounds this concern, as does the data-intensive nature of optimal bandwidth selection in RD estimation. The conceptual framework motivates a direct test of Prediction 4: the effect should be concentrated among providers with a higher baseline propensity to stigmatize.

To classify providers by stigma propensity, I first identify each patient's primary provider as the clinician who placed the plurality of orders during the hospitalization, following the EHR metadata literature that infers provider involvement from behavioral traces in the record (Kahn et al., 2022; Soulakis et al., 2015).[10] I then compute each provider's mean *Any Stigma* rate across all their patients' discharge summaries and rank providers into quartiles. I define providers in the top quartile as "High-stigma providers" and group the remaining three quartiles as "Low-stigma providers."

I then re-estimate the RD specification separately for High- and Low-stigma providers, using the same Diff-in-Disc framework introduced in Section 3.4, 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. Identification requires that crossing the cutoff does not differentially shift patients into High- versus Low-stigma providers. Figure A.3 supports this assumption: patients just above and just below the cutoff are equally likely to be treated by High-stigma providers, and Table A.4 shows no meaningful differential sorting on observables between High- and Low-stigma providers. The only notable imbalance is a small difference in Medicaid coverage (significant at 10%), but insurance status is included as a covariate in robustness specifications and would, if anything, attenuate the main results. The predicted mortality variable falls more at the cutoff for High-stigma than for Low-stigma providers — a pattern that works *against* the findings reported below.

Results are shown in Figure VII. Panel A presents the stigma results: the obesity cutoff increases the probability of stigmatizing language in discharge summaries by 4.4 percentage points among High-stigma providers ($p = 0.13$), while reducing it by 3.5 percentage points among Low-stigma providers

[10]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 hospital service. Final review and authentication of the discharge summary is then performed by the attending physician.

($p = 0.06$). The Diff-in-Disc estimate — the differential RD effect across provider types — is 8 percentage points ($p < 0.05$), indicating that the obesity label triggers a significantly larger increase in stigmatizing language among providers with higher baseline propensity for such language. Panel B presents the corresponding mortality results: the mortality effect of the obesity cutoff is driven almost entirely by High-stigma providers, with crossing the cutoff increasing in-hospital mortality by 1.3 percentage points among High-stigma providers ($p < 0.001$) and having no detectable effect among Low-stigma providers. The Diff-in-Disc estimate implies that the mortality discontinuity is 1 percentage point larger for High-stigma providers ($p < 0.001$).

Taken together, these results confirm Prediction 4: the providers whose language responds most to the "obese" label are also the providers whose patients experience the largest mortality increase.[11] Table 5, Panel B reports the full details of the heterogeneous RD analysis for stigma, and Table 6 reports the corresponding estimates for mortality. Tables A.5 and A.6 confirm that both sets of results are robust to alternative polynomial order, covariate adjustment, clustering, and OLS specification.[12]

Having established that the obesity cutoff increases both stigmatizing language and mortality among High-stigma providers, I next probe two defining features of stigma more directly: (i) that it is triggered by perceived deviation from a social norm, and (ii) that it is accompanied by moral judgment.

*Stigma as a response to norm deviation.*Prediction 5 states that the stigma channel should be stronger when deviation from body-size norms is more salient. I test this in two ways. First, holding BMI fixed, shorter patients appear physically larger than taller patients, making norm deviation more visually apparent. The obesity label should therefore trigger more stigmatizing language among shorter patients. I split patients into quartiles of height and estimate heterogeneous RD effects on stigmatizing language by height quartile. The quartile-specific RD estimates presented in Table A.8 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 to the tallest quartile, the Diff-in-Disc estimate implies that the effect of crossing the cutoff on stigmatizing language is 8.8 percentage points larger for the shortest

[11]These results do not identify a causal effect of stigmatizing language on mortality (an empirical endeavor that would face severe confounding). Rather, I estimate the RD effect of the obesity cutoff on each outcome separately, within each provider subgroup. The pattern is consistent with an underlying mechanism in which stigma, triggered by the "obese" label, manifests both in clinical language and, through reduced clinical effort, in patient outcomes.

[12]To ensure that these results are not mechanically driven by how I classify provider type, I re-estimate the Diff-in-Disc using two alternative definitions. First, I exclude patients with an obese ICD diagnosis code when calculating a provider's stigmatizing language rate, since patients above the obesity cutoff are more likely to receive such a code, which could inflate measured stigma rates for providers who treat more patients near the threshold. Second, I construct a "leave-one-out" measure, where for each patient I calculate the provider's mean stigma rate across all other patients, excluding the focal patient. This addresses the concern that including the focal patient's own stigma value in the provider-level measure could generate a mechanical relationship between provider classification and the patient-level outcome. Results, presented in Table A.7, show that the Diff-in-Disc estimate remains positive and statistically significant for both stigmatizing language and mortality. Neither concern has a clear theoretical basis in the RD setting: both would require High-stigma providers to differentially treat patients just above versus just below the cutoff, which Figure A.3 rules out.

patients than for the tallest ($p < 0.1$). While imprecise, these results are consistent with Prediction 5: the cutoff has the strongest impact when patients are most likely to appear as deviating from body-size norms.

Second, I examine within-patient variation in stigmatizing documentation across consecutive hospitalizations. Each patient's prior-visit BMI category serves as a natural reference point — a within-person "norm" — against which the physician implicitly evaluates the patient. If stigma responds to perceived norm deviation, the same patient should be documented more negatively when they move into a higher BMI category relative to their last visit. I restrict to patients with multiple encounters and estimate a first-difference specification, regressing the change in stigmatizing language between consecutive visits on the corresponding change in BMI category. Patient fixed effects ensure that the identifying variation comes from within-patient changes in BMI category rather than cross-patient differences in body size, and physician fixed effects ensure I account for provider-specific documentation style.[13] Table A.9 shows that stigma moves in tandem with BMI category: for the same patient, a one-category increase in BMI since the last hospital visit is associated with a 3 percentage point increase in stigmatizing language ($p < 0.05$). This pattern further supports Prediction 5: 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. Moral judgment is a defining feature of stigma, and the LLM classifier allows a direct test: because each flagged instance is classified into one of four categories — *Judgmental*, *Discrediting*, *Pejorative Labeling*, or *Stereotyping* — I can examine whether the obesity cutoff specifically increases language expressing moral judgment, not just generic negativity. Results are presented in Table 7. The obesity cutoff increases biased language across multiple categories, but the two largest and most precisely estimated Diff-in-Disc effects are for *Judgmental* and *Discrediting* descriptors. The judgmental dimension is, in particular, consistent with the moral-judgment component of stigma intensifying at the obesity cutoff.

The results above show that the obesity cutoff increases judgmental and discrediting language, but this language need not be directed at the patient's weight. A sharper test asks whether the cutoff increases stigmatizing language that co-occurs with weight-related terminology (e.g., BMI, weight, habitus, obese) in the same note — directly linking the label to weight-specific moral judgment. I construct an indicator equal to one when the LLM flags a given bias category *and* the note contains at least one weight-related term. Results are reported in Table 8. 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

[13] 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.

weight term more for 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, $p = 0.103$).

Non-LLM analysis of notes.. The LLM classifier used in this paper is validated against independent physician review, and the prompt is reproduced in full in the Appendix. Nonetheless, for readers who prefer fully transparent measures, I complement the LLM results with simple text-based analyses that require no language model. Table A.10 reports RD estimates for word count, character count, and average word length. Because these measures are inexpensive to compute, they can be estimated on the full sample of patients with a discharge summary and a recorded BMI, rather than restricting to the overweight and obese subsample. On average, notes have 1,674 words, 11,096 characters, and an average word length of 5.39 characters. The estimates show that crossing the obesity cutoff significantly reduces both word count and character count both by 1.7%, consistent with a reduction in clinical effort. Interpretation should be cautious, however, because these aggregate measures are not designed to detect stigma and there is no established mapping from note length to biased language or discriminatory behavior.

As a final complement, I implement a simple dictionary-based measure that flags a small set of judgmental terms in the discharge summary using a regex search, requiring no language model. I construct two outcomes: (i) an indicator for whether the discharge summary contains any judgmental term describing patient behavior,[14] 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

[14] The dictionary includes: *difficult*, *demanding*, *argumentative*, *combative*, *belligerent*, *hostile*, *aggressive*, *threatening*, *rude*, *abusive*, *verbally abusive*, *non-cooperative*, and *uncooperative*.

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 study's biggest limitation is that it draws on a single hospital. The extent to which the findings generalize is an open question, though there is an inherent tradeoff: multi-hospital datasets rarely provide the combination of continuous BMI, free-text clinical notes, and sufficiently de-identified documentation needed to measure stigma at scale through an LLM. The richness of the data that makes identification possible here is precisely what most multi-site administrative datasets lack. That said, weight stigma is among the most thoroughly documented and universally experienced forms of social stigma — reported across cultures, clinical settings, and provider types — making it difficult to argue that the mechanisms documented here are unique to a single institution.

The results highlight a policy-relevant and potentially tractable margin: the design of provider-facing clinical interfaces and classification schemes. Unlike discrimination rooted in immutable traits such as race or gender, the discrimination documented here is, to some extent, transient — triggered by a categorical label embedded in clinical software, and therefore reversible through changes in institutional design. 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 can preserve clinical usefulness while reducing harm.

Finally, these findings have implications for the economic valuation of anti-obesity therapies. The meteoric rise of GLP-1 receptor agonists — drugs such as semaglutide and tirzepatide that produce substantial weight loss — has generated enormous market valuations based primarily on projected reductions in obesity-related disease. Yet the results of this paper suggest an additional, unpriced benefit: by moving patients below clinical obesity thresholds, these drugs may also reduce exposure to weight stigma in healthcare, improving the quality of clinical attention and diagnostic effort such patients receive. If the "obese" label itself contributes to excess mortality through the channels documented here, then current valuations of GLP-1 therapies, which do not account for the elimination of stigma-driven harm, likely understate their true social value.

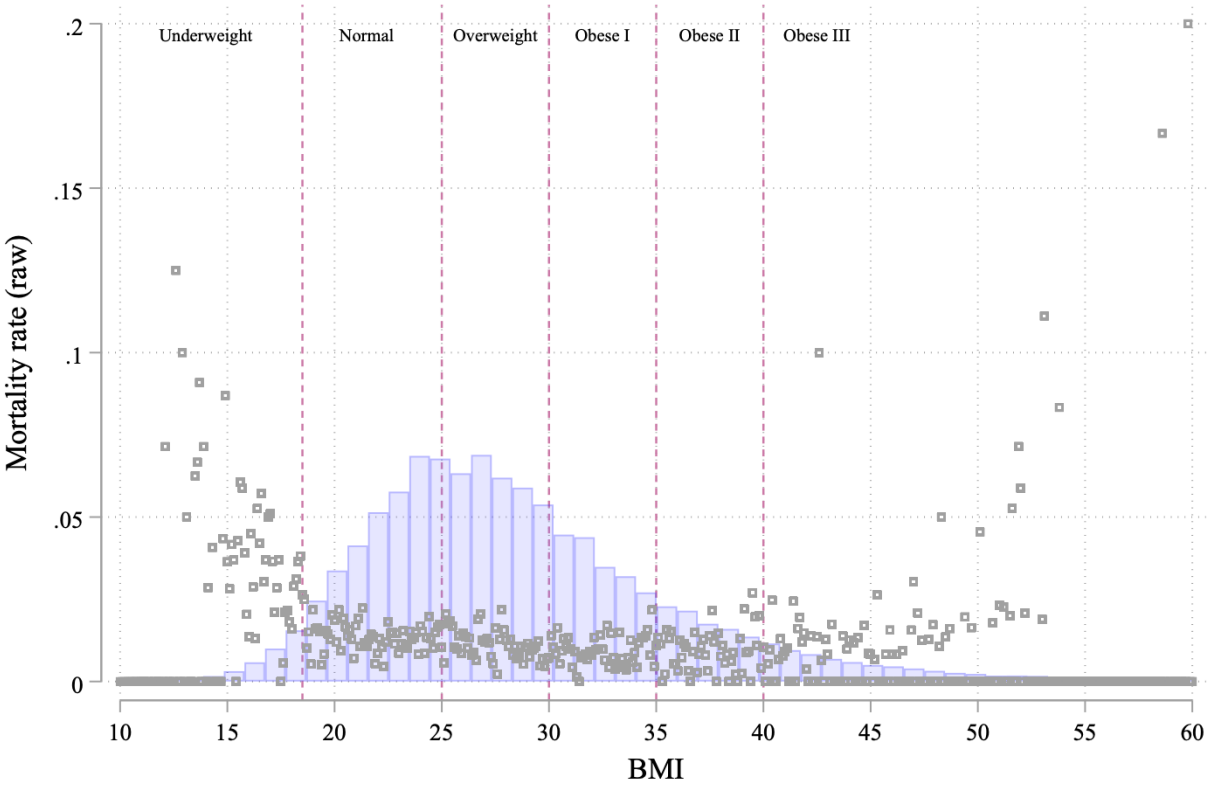


FIGURE I
Distribution of BMI and Raw Mortality

Notes: Bars show the frequency of hospital encounters by BMI bin (width = 1). Scatter points show mean in-hospital mortality within bins of width 0.5. Dashed vertical lines mark standard BMI category cutoffs: underweight ($<$ 18.5), normal (18.5–25), overweight (25–30), obese I (30–35), obese II (35–40), and obese III ($\geq$ 40).

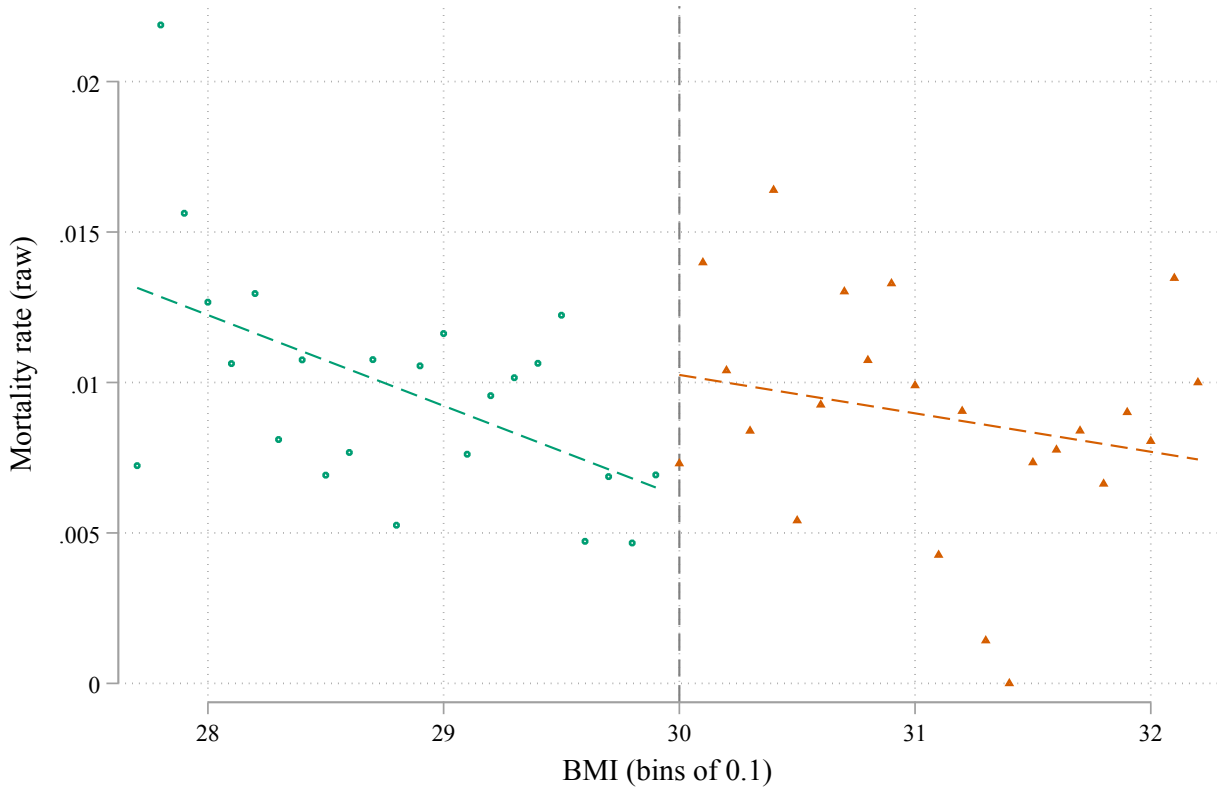
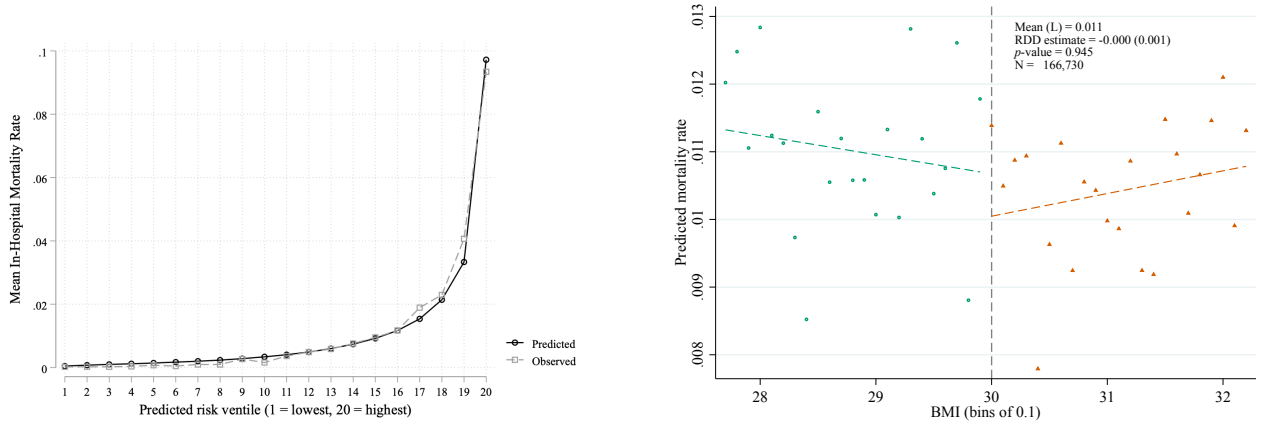


FIGURE II

Mortality and BMI at the Obesity Cutoff

Notes: Each point represents raw mean in-hospital mortality within a BMI bin of width 0.45, sized by the number of observations in each bin. Solid lines are local linear fits estimated separately on each side of the obesity cutoff (BMI = 30). The running variable is BMI − 30. The reported RBC estimate, standard error, and p-value are from the Calonico, Cattaneo, and Titiunik (2014) procedure with MSE-optimal bandwidth and patient-clustered standard errors.



Panel (a): Predicted vs. Observed Mortality

Panel (b): Predicted Mortality RDD

FIGURE III

RDD Estimate of Obesity Cutoff on *Predicted* Mortality

Notes: Panel (a) plots the mean of observed in-hospital mortality and predicted mortality across ventiles of predicted mortality. Panel (b) plots the RD in predicted mortality at the obesity cutoff (BMI = 30). Predicted mortality is generated from a logit model of observed mortality on baseline covariates (age, sex, race, insurance, marital status, Elixhauser comorbidities and mortality index). In Panel (b), hollow circles are bin means (bin width = 0.45), sized by the number of observations in each bin; solid lines are local linear fits estimated separately on each side of the cutoff. The reported RBC estimate, standard error, and p-value are from the Calonico, Cattaneo, and Titiunik (2014) procedure with MSE-optimal bandwidth and patient-clustered standard errors.

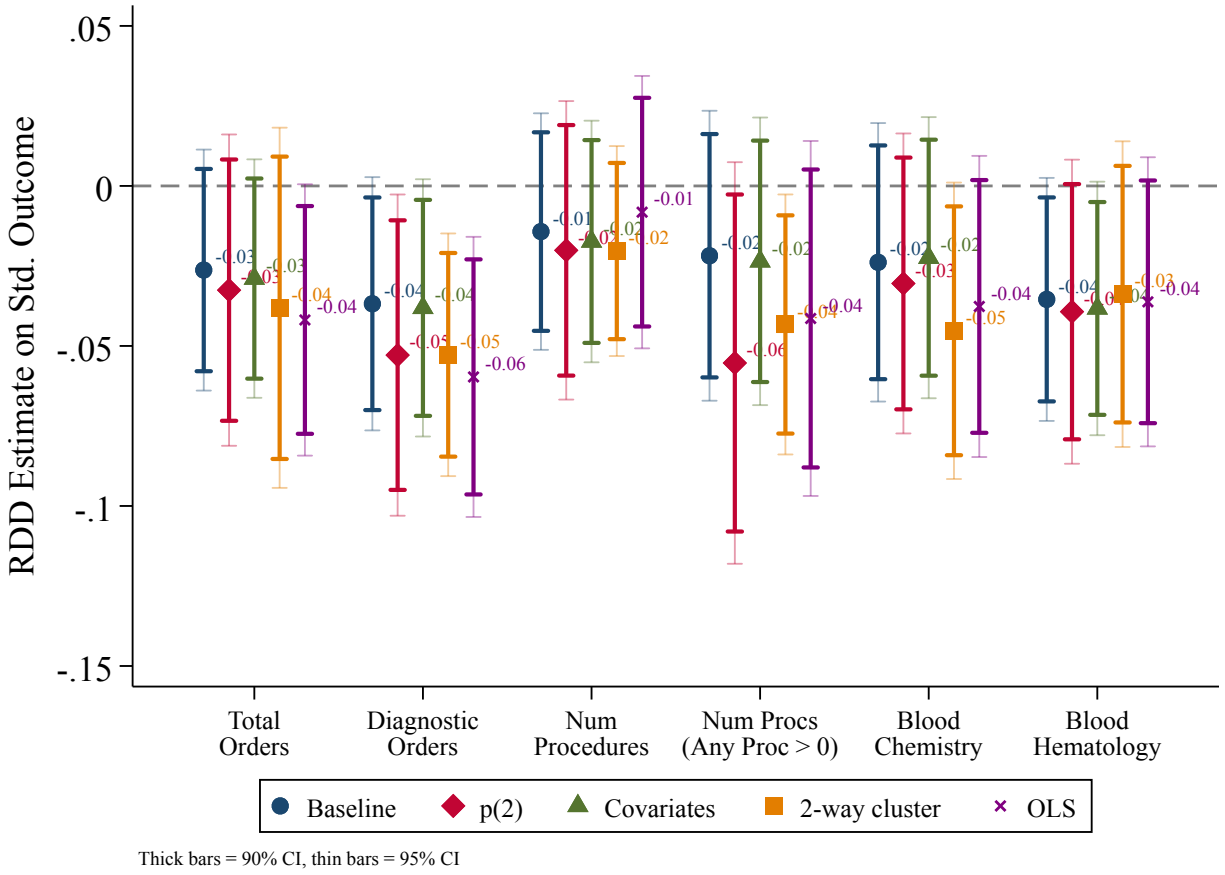


FIGURE IV RDD Estimates Across Structured Clinical Outcomes

Notes: Each group on the x-axis represents a different clinical outcome. Points show bias-corrected RD estimates at the obesity cutoff (BMI = 30) under five specifications: (1) baseline local linear with patient-clustered standard errors; (2) local quadratic; (3) local linear with covariates (age, sex, race, insurance); (4) two-way clustering by patient and hospital service; (5) OLS within ±2 BMI units. Thick bars denote 90% confidence intervals; thin bars denote 95% confidence intervals. "Num Procs (Any Proc > 0)" conditions on patients who received at least one procedure. All outcomes are standardized.

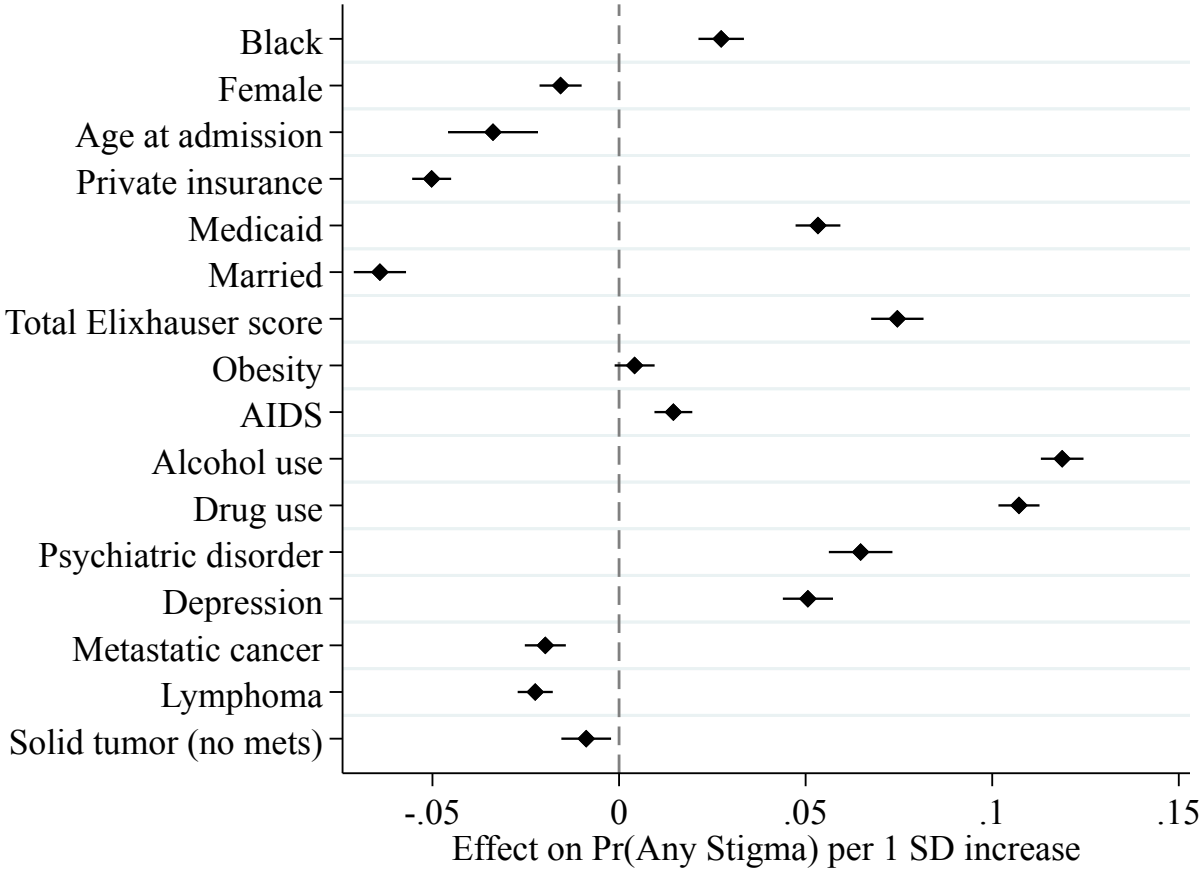
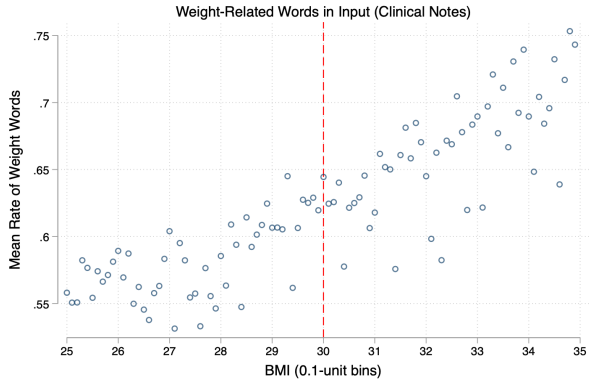


FIGURE V

Correlates of Stigmatizing Language in Discharge Summaries

Notes: Each estimate shows the univariate correlation between the indicated patient characteristic on the Y axis and an indicator for any stigmatizing language in the discharge summary, as classified by the LLM described in Section 4.3. Sample restricted to hospital admissions with a discharge summary and BMI in the overweight or obese range.

(a) Input (Clinical Notes)



(b) Output (LLM)

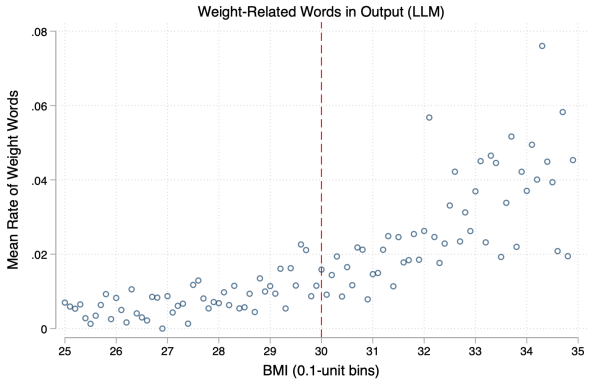
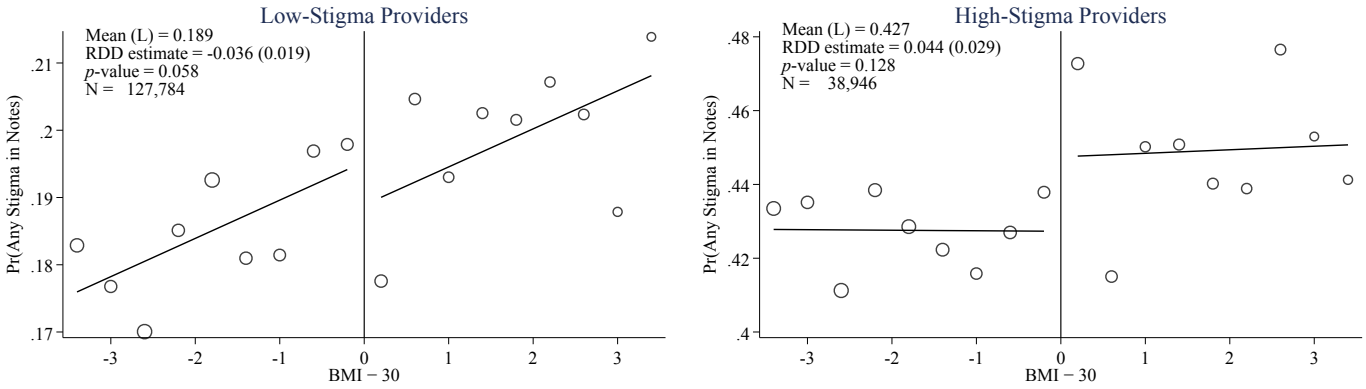


FIGURE VI

Mean rate of weight-related words around the obesity cutoff

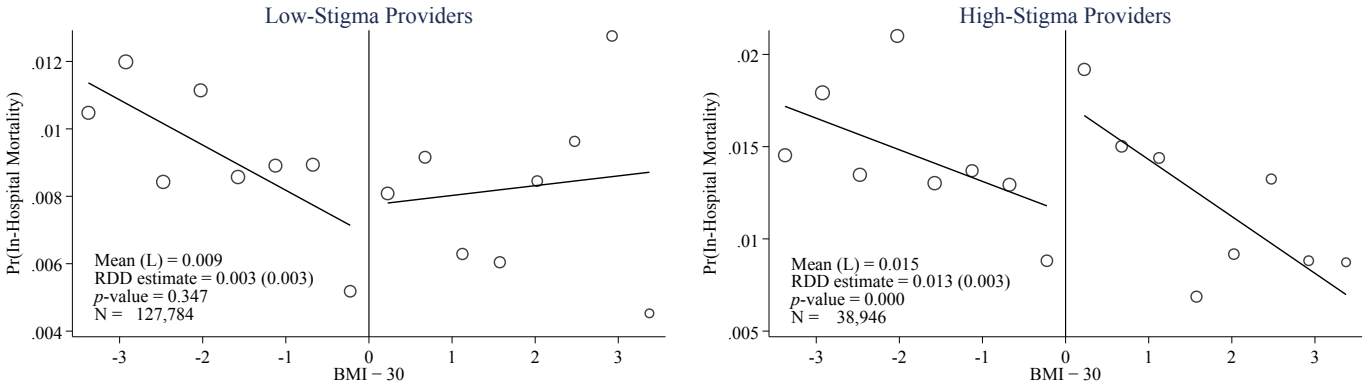
Notes: Each point plots the share of discharge summaries containing at least one weight-related word within 0.1-unit BMI bins, for BMI between 25 and 35. Weight-related words are identified by a regular expression matching terms such as *obese/obesity, overweight, fat, heavyset, habitus, underweight, BMI, body mass index, weight*, and numeric expressions of weight in kg or lbs. Panel (a) searches the raw clinical text (input); Panel (b) searches the LLM-extracted stigma output. The vertical dashed line marks the obesity cutoff (BMI = 30).

Panel A: Any Stigma in Notes



Diff-in-Disc estimate (High − Low) = 0.0800 (0.034), $p = 0.019$**

Panel B: In-Hospital Mortality



Diff-in-Disc estimate (High − Low) = 0.0104* (0.006), $p = 0.067$

FIGURE VII

Stigma and Mortality at the Obesity Cutoff by Provider Type

Notes: Each panel plots raw bin means for Low-stigma (left) and High-stigma (right) providers. Panel A shows stigmatizing language prevalence; Panel B shows in-hospital mortality. Circle size is proportional to the number of observations in each bin. Solid lines are local linear fits estimated separately on each side of the obesity cutoff (BMI = 30). RD estimates are robust, bias-corrected estimates from Calonico et al. (2025). Diff-in-Disc estimates are shown in red.

	Mean	SD
Demographics		
Age at admission	60.7	18.4
Female	0.523	0.499
Black	0.158	0.364
Hispanic	0.054	0.226
Married	0.456	0.498
Insurance		
Medicaid	0.188	0.390
Medicare	0.472	0.499
Private insurance	0.316	0.465
BMI category		
Underweight (<18.5)	0.034	0.180
Normal (18.5–24.9)	0.302	0.459
Overweight (25.0–29.9)	0.314	0.464
Obese I (30.0–34.9)	0.187	0.390
Obese II (35.0–39.9)	0.092	0.289
Obese III (≥40.0)	0.071	0.257
Elixhauser comorbidities		
Number of Elixhauser comorbidities	3.3	2.3
Elixhauser mortality index	7.48	9.04
Congestive heart failure	0.178	0.382
Cardiac arrhythmias	0.233	0.423
Valvular disease	0.094	0.291
Pulmonary circulation disorders	0.054	0.226
Peripheral vascular disorders	0.076	0.265
Hypertension, uncomplicated	0.349	0.477
Hypertension, complicated	0.201	0.401
Paralysis	0.014	0.117
Other neurological disorders	0.087	0.281
Chronic pulmonary disease	0.202	0.401
Diabetes, uncomplicated	0.124	0.330
Diabetes, complicated	0.138	0.345
Hypothyroidism	0.134	0.341
Renal failure	0.186	0.389
Liver disease	0.093	0.291
Peptic ulcer disease	0.009	0.097
AIDS/HIV	0.007	0.083
Lymphoma	0.046	0.209
Metastatic cancer	0.067	0.250
Solid tumor w/o metastasis	0.061	0.240
Rheumatoid arthritis	0.039	0.193
Coagulopathy	0.094	0.292
Obesity (diagnosis)	0.127	0.333
Weight loss	0.072	0.258
Fluid/electrolyte disorders	0.207	0.405
Blood loss anemia	0.009	0.092
Deficiency anemia	0.054	0.226
Alcohol abuse	0.066	0.248
Drug abuse	0.049	0.217
Psychoses	0.024	0.154
Depression	0.204	0.403
Outcome		
In-hospital mortality	0.012	0.108
Observations	166,730	

TABLE 1 Summary Statistics

Notes: Elixhauser comorbidities and mortality index follow Elixhauser et al. (1998).

	Dep Var: In-Hospital Mortality						
Model	(1)	(2)	(3)	(4)	(5)	(6)	(7)
RD estimate (RBC)	0.004** (0.002)	0.006** (0.002)	0.005** (0.002)	0.004* (0.002)	0.004** (0.002)	0.006** (0.002)	0.004** (0.002)
Observations left	108,342	108,342	108,342	108,342	108,342	108,342	18,946
Observations right	58,388	58,388	58,388	58,388	58,388	58,388	15,686
Effective obs left	27,988	20,270	47,403	40,884	27,918	27,989	18,946
Effective obs right	20,883	16,634	29,817	27,259	20,809	20,883	15,686
Polynomial order (p)	1	1	2	2	1	1	1
Kernel	Tri	Uni	Tri	Uni	Tri	Tri	—
Bandwidth (h) (L/R)	2.899	2.178	4.601	4.064	2.885	2.900	±2
Cluster vars	P	P	P	P	P	P,H	P
Covariates	No	No	No	No	Yes	No	No
Estimator	RD	RD	RD	RD	RD	RD	OLS
Mean, left of threshold	0.010	0.010	0.011	0.011	0.010	0.010	0.009

TABLE 2 RD robustness to polynomial order, kernel, covariates, and clustering

Notes: Each column reports an estimate of the effect of the obesity cutoff (BMI = 30) on in-hospital mortality, with standard errors in parentheses. Columns (1)–(6) use the robust, bias-corrected (RBC) local polynomial RD estimator of Calonico, Cattaneo, and Titiunik (2014, 2017) with MSE-optimal bandwidth; Column (7) estimates the discontinuity via OLS within a ±2 BMI window. Tri = triangular kernel; Uni = uniform kernel. P = patient, H = hospital service. Covariates in Column (5) include age, sex, and race indicators. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.001$.

	Dep Var: In-Hospital Mortality				
	(1)	(2)	(3)	(4)	(5)
RD estimate					
Low-severity patients	0.001	0.001	0.001	0.001	-0.000
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
High-severity patients	0.020*	0.018*	0.020**	0.020**	0.018**
	(0.010)	(0.010)	(0.010)	(0.006)	(0.008)
Diff-in-Disc estimate	**0.0187***	**0.0174***	**0.0190***	**0.0187****	**0.0183****
	(0.010)	**(0.010)**	**(0.010)**	**(0.006)**	**(0.008)**
Polynomial	Linear	Quadratic	Linear	Linear	Linear
Covariates	No	No	Yes	No	No
Clustering	P	P	P	P, H	P
Method	RD	RD	RD	RD	OLS
Bandwidth	MSE-opt.	MSE-opt.	MSE-opt.	MSE-opt.	±2
Total N	166,730	166,730	166,730	166,730	34,632

TABLE 3 Effect of Obesity Cutoff on In-Hospital Mortality by Patient Severity

Notes: Each column reports an estimate of the effect of the obesity cutoff (BMI = 30) on in-hospital mortality, with standard errors in parentheses, separately for low- and high-severity patients. Patient severity is defined by the Elixhauser mortality index (Elixhauser et al., 1998): high-severity patients are those in the top quartile. The Diff-in-Disc estimate is the difference in the RD effect between high- and low-severity patients (High − Low). Columns (1)–(4) use the robust, bias-corrected (RBC) local polynomial RD estimator of Calonico, Cattaneo, and Titiunik (2014, 2017) with MSE-optimal bandwidth; Column (5) estimates the discontinuity via OLS within a ±2 BMI window. P = patient, H = hospital service. Covariates in Column (3) include age, race, sex, and insurance indicators. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.001$.

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 4 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 4 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 4 Examples of LLM output by bias type

Notes: This table presents the prompt definition (excerpt, in `monospace`) and representative LLM-flagged instances (drawn from MIMIC-IV discharge summaries) for each of the four categories of biased language: discrediting, judgmental, pejorative labeling, and stereotyping. For each flagged instance, the table reports the bias term (verbatim from the note), the assigned bias type, and the LLM's reasoning. Patient identifiers have been removed by the MIMIC-IV de-identification pipeline.

	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 5 Effect of Obesity Cutoff on Stigmatizing Language by Provider Type

Notes: The outcome is an indicator for any stigmatizing language in the discharge summary. Each row reports the robust, bias-corrected (RBC) RD estimate and standard error at the obesity cutoff (BMI = 30) using the Calonico, Cattaneo, and Titiunik (2014, 2017) procedure with MSE-optimal bandwidth. Low-stigma providers (quartiles 1–3) and High-stigma providers (quartile 4) are defined by provider-level mean rates of stigmatizing language. The Diff-in-Disc estimate is the difference in the RD effect between High- and Low-stigma providers (High − Low). All specifications use a local linear polynomial, triangular kernel, and two-way clustering by patient and hospital service. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.001$.

	Dep Var: In-hospital mortality							
	estimate	SE (RBC)	p-value	total N	Eff. N (L)	Eff. N (R)	h	mean (L)
RD estimate								
Low-stigma providers	0.003	0.003	0.347	127,784	27,780	19,267	3.570	0.012
High-stigma providers	0.013***	0.003	0.000	38,946	6,776	4,825	3.075	0.016
Diff-in-Disc estimate	**0.0104***	**0.006**	**0.067**	**166,730**				

TABLE 6 Effect of Obesity Cutoff on In-Hospital Mortality by Provider Type

Notes: The outcome is in-hospital mortality. Each row reports the robust, bias-corrected (RBC) RD estimate and standard error at the obesity cutoff (BMI = 30) using the Calonico, Cattaneo, and Titiunik (2014, 2017) procedure with MSE-optimal bandwidth. Low-stigma providers (quartiles 1–3) and High-stigma providers (quartile 4) are defined by provider-level mean rates of stigmatizing language. The Diff-in-Disc estimate is the difference in the RD effect between High- and Low-stigma providers (High − Low). All specifications use a local linear polynomial, triangular kernel, and two-way clustering by patient and hospital service. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.001$.

	estimate	SE (RBC)	p-value	total N	Eff. N (L)	Eff. N (R)	h	mean (L)
Outcome: Judgemental								
RD estimate								
Low-stigma providers	-0.019	0.012	0.108	49,594	6,482	5,995	1.253	0.071
High-stigma providers	0.058**	0.025	0.023	16,194	1,629	1,489	0.993	0.199
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	0.004
High-stigma providers	0.009*	0.005	0.087	16,194	3,020	2,554	1.700	0.018
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	0.123
High-stigma providers	0.004	0.032	0.893	16,194	2,258	1,991	1.314	0.284
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	0.042
High-stigma providers	0.057***	0.015	0.000	16,194	1,769	1,624	1.073	0.102
Diff-in-Disc estimate	**0.0759****	**0.024**	**0.002**	**65,788**				

TABLE 7 Effect of Obesity Cutoff on Stigmatizing Language by Bias Category

Notes: The outcome in each panel is an indicator for whether the discharge summary contains any instance of a given bias category: judgmental, stereotyping, pejorative labeling, or discrediting. Each row reports the robust, bias-corrected (RBC) RD estimate and standard error at the obesity cutoff (BMI = 30) using the Calonico, Cattaneo, and Titiunik (2014, 2017) procedure with MSE-optimal bandwidth. Low-stigma providers (quartiles 1–3) and High-stigma providers (quartile 4) are defined by provider-level mean rates of stigmatizing language. The Diff-in-Disc estimate is the difference in the RD effect between High- and Low-stigma providers (High − Low). All specifications use a local linear polynomial, triangular kernel, and two-way clustering by patient and hospital service. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.001$.

	estimate	SE (RBC)	p-value	total N	Eff. N (L)	Eff. N (R)	h	mean (L)
Outcome: Judgemental (+ weight mention)								
RD estimate								
Low-stigma providers	-0.003	0.013	0.832	49,594	8,091	7,177	1.530	0.042
High-stigma providers	0.067**	0.029	0.019	16,194	1,748	1,609	1.036	0.116
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	0.002
High-stigma providers	-0.004	0.009	0.609	16,194	3,095	2,597	1.771	0.009
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	0.070
High-stigma providers	-0.006	0.022	0.793	16,194	2,274	2,013	1.337	0.157
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	0.023
High-stigma providers	0.031**	0.013	0.017	16,194	2,054	1,867	1.206	0.057
Diff-in-Disc estimate	**0.0272**	**0.017**	**0.107**	**65,788**				

TABLE 8 Effect of Obesity Cutoff on Stigmatizing Language Co-occurring with Weight Mentions

Notes: The outcome in each panel is an indicator equal to one when the discharge summary contains any instance of a given bias category (judgmental, stereotyping, pejorative labeling, or discrediting) *and* the note also contains at least one weight-related term (e.g., BMI, weight, habitus, obese). Each row reports the robust, bias-corrected (RBC) RD estimate and standard error at the obesity cutoff (BMI = 30) using the Calonico, Cattaneo, and Titiunik (2014, 2017) procedure with MSE-optimal bandwidth. Low-stigma providers (quartiles 1–3) and High-stigma providers (quartile 4) are defined by provider-level mean rates of stigmatizing language. The Diff-in-Disc estimate is the difference in the RD effect between High- and Low-stigma providers (High − Low). All specifications use a local linear polynomial, triangular kernel, and two-way clustering by patient and hospital service. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.001$.

	estimate	SE (RBC)	p-value	total N	Eff. N (L)	Eff. N (R)	h	mean (L)
Outcome: Judgemental words (raw)								
RD estimate								
Low-stigma providers	-0.028	0.021	0.182	49,594	8,494	7,594	1.604	0.155
High-stigma providers	0.052**	0.017	0.003	16,194	3,064	2,566	1.731	0.252
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	0.090
High-stigma providers	0.029**	0.013	0.029	16,194	3,267	2,694	1.843	0.144
Diff-in-Disc estimate	**0.0475****	**0.017**	**0.004**	**65,788**				

TABLE 9 Effect of Obesity Cutoff on Dictionary-Based Judgmental Language

Notes: The outcome in the first panel is an indicator for any dictionary-based judgmental word (e.g., "difficult," "combative," "uncooperative") in the discharge summary; in the second panel the outcome additionally requires the note to contain at least one weight-related term. Each row reports the robust, bias-corrected (RBC) RD estimate and standard error at the obesity cutoff (BMI = 30) using the Calonico, Cattaneo, and Titiunik (2014, 2017) procedure with MSE-optimal bandwidth. Low-stigma providers (quartiles 1–3) and High-stigma providers (quartile 4) are defined by provider-level mean rates of stigmatizing language. The Diff-in-Disc estimate is the difference in the RD effect between High- and Low-stigma providers (High − Low). All specifications use a local linear polynomial, triangular kernel, and two-way clustering by patient and hospital service. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.001$.

References

Alalouf, M., Miller, S., & Wherry, L. R. (2024). What Difference Does a Diagnosis Make?: Evidence from Marginal Patients. *American Journal of Health Economics*, *10*(1), 97–131. https://doi.org/10.1086/724415

Almond, D., Doyle, J. J., Kowalski, A. E., & Williams, H. (2010). ESTIMATING MARGINAL RETURNS TO MEDICAL CARE: EVIDENCE FROM AT-RISK NEWBORNS. *The quarterly journal of economics*, *125*(2), 591–634. https://doi.org/10.1162/qjec.2010.125.2.591

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

Averett, S., & Korenman, S. (1996). The Economic Reality of the Beauty Myth. *Journal of Human Resources*, *31*(2), 304–330. Retrieved February 24, 2026, from https://ideas.repec.org//a/uwp/jhriss/v31y1996i2p304-330.html

Barreca, A. I., Guldi, M., Lindo, J. M., & Waddell, G. R. (2011). Saving Babies? Revisiting the effect of very low birth weight classification*. *The Quarterly Journal of Economics*, *126*(4), 2117–2123. https://doi.org/10.1093/qje/qjr042

Becker, G. S. (2010). *The economics of discrimination*. University of Chicago press.

Bénabou, R., & Tirole, J. (2006). Incentives and Prosocial Behavior. *American Economic Review*, *96*(5), 1652–1678. https://doi.org/10.1257/aer.96.5.1652

Bénabou, R., & Tirole, J. (2026). Laws and Norms. *Journal of Political Economy*, *134*(2), 731–772. https://doi.org/10.1086/738343

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.

Besley, T., & Coate, S. (1992). Understanding welfare stigma: Taxpayer resentment and statistical discrimination. *Journal of Public Economics*, *48*(2), 165–183. https://doi.org/10.1016/0047-2727(92)90025-B

Bharadwaj, P., Løken, K. V., & Neilson, C. (2013). Early Life Health Interventions and Academic Achievement. *American Economic Review*, *103*(5), 1862–1891. https://doi.org/10.1257/aer.103.5.1862

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

Blodorn, A., Major, B., Hunger, J., & Miller, C. (2016). Unpacking the psychological weight of weight stigma: A rejection-expectation pathway. *Journal of experimental social psychology*, *63*, 69–76. https://doi.org/10.1016/j.jesp.2015.12.003

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., Palomba, F., & Titiunik, R. (2025, July). Rdhte: Conditional Average Treatment Effects in RD Designs [arXiv:2507.01128 [econ]]. https://doi.org/10.48550/arXiv.2507.01128

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

Cawley, J. (2004a). The Impact of Obesity on Wages. *The Journal of Human Resources*, *39*(2), 451–474. https://doi.org/10.2307/3559022

Cawley, J. (2004b). The Impact of Obesity on Wages. *The Journal of Human Resources*, *39*(2), 451–474. https://doi.org/10.2307/3559022

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

Chen, S., Sudharsanan, N., Huang, F., Liu, Y., Geldsetzer, P., & Bärnighausen, T. (2019). Impact of community based screening for hypertension on blood pressure after two years: Regression discontinuity analysis in a national cohort of older adults in China. *BMJ*, *366*, l4064. https://doi.org/10.1136/bmj.l4064

Chobanian, A. V., Bakris, G. L., Black, H. R., Cushman, W. C., Green, L. A., Izzo, J., Joseph L., Jones, D. W., Materson, B. J., Oparil, S., Wright, J., Jackson T., Roccella, E. J., & and the National High Blood Pressure Education Program Coordinating Committee. (2003). The Seventh Report of the Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood PressureThe JNC 7 Report. *JAMA*, *289*(19), 2560–2571. https://doi.org/10.1001/jama.289.19.2560

Choi, J.-y., & Lee, M.-j. (2018). Regression Discontinuity with Multiple Running Variables Allowing Partial Effects. *Political Analysis*, *26*(3), 258–274. https://doi.org/10.1017/pan.2018.13

Crandall, C. S. (1994). Prejudice against fat people: Ideology and self-interest. *Journal of Personality and Social Psychology*, *66*(5), 882–894. https://doi.org/10.1037//0022-3514.66.5.882

Crandall, C. S., & Eshleman, A. (2003). A justification-suppression model of the expression and experience of prejudice. *Psychological Bulletin*, *129*(3), 414–446. https://doi.org/10.1037/0033-2909.129.3.414

Dai, T., Jiang, S., Liu, X., & Sun, A. (2022). The effects of a hypertension diagnosis on health behaviors: A two-dimensional regression discontinuity analysis [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.10 *Health Economics*, *31*(4), 574–596. https://doi.org/10.1002/hec.4466

Dovidio, J. F., & Gaertner, S. L. (2000). Aversive racism and selection decisions: 1989 and 1999. *Psychological Science*, *11*(4), 315–319. https://doi.org/10.1111/1467-9280.00262

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

Erstad, B. L., & Barletta, J. F. (2021). Drug dosing in the critically ill obese patient: A focus on medications for hemodynamic support and prophylaxis. *Critical Care*, *25*, 77. https://doi.org/10.1186/s13054-021-03495-8

Fisher, M., & Keil, F. C. (2018). The Binary Bias: A Systematic Distortion in the Integration of Information. *Psychological Science*, *29*(11), 1846–1858. https://doi.org/10.1177/0956797618792256

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

Friedrichsen, J., König, T., & Schmacker, R. (2018). Social image concerns and welfare take-up. *Journal of Public Economics*, *168*, 174–192. https://doi.org/10.1016/j.jpubeco.2018.10.008

Fukuma, S., Iizuka, T., Ikenoue, T., & Tsugawa, Y. (2020). Association of the National Health Guidance Intervention for Obesity and Cardiovascular Risks With Health Outcomes Among Japanese Men. *JAMA internal medicine*, *180*(12), 1630–1637. https://doi.org/10.1001/jamainternmed.2020.4334

Geiger, C. K., Clapp, M. A., & Cohen, J. L. (2021). Association of Prenatal Care Services, Maternal Morbidity, and Perinatal Mortality With the Advanced Maternal Age Cutoff of 35 Years. *JAMA Health Forum*, *2*(12), e214044. https://doi.org/10.1001/jamahealthforum.2021.4044

Hagan, S. (2021). E66.01 and Our Culture of Shame [_eprint: https://www.nejm.org/doi/pdf/10.1056/NEJMp2111 *New England Journal of Medicine*, *385*(25), 2307–2309. https://doi.org/10.1056/NEJMp2111506

Healy, M., Richard, A., & Kidia, K. (2022). How to Reduce Stigma and Bias in Clinical Communication: A Narrative Review. *Journal of General Internal Medicine*, *37*(10), 2533–2540. https://doi.org/10.1007/s11606-022-07609-y

Himes, S. M., & Thompson, J. K. (2007). Fat stigmatization in television shows and movies: A content analysis. *Obesity*, *15*(3), 712–718. https://doi.org/10.1038/oby.2007.635

Himmelstein, G., Bates, D., & Zhou, L. (2022). Examination of Stigmatizing Language in the Electronic Health Record. *JAMA Network Open*, *5*(1), e2144967. https://doi.org/10.1001/jamanetworkopen.2021.44967

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

Kämpfen, F., & Mosca, I. (2024, February). Heterogeneous Effects of Blood Pressure Screening. https://doi.org/10.2139/ssrn.4709700

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

Kolata, G. (2023). New Obesity Drugs Come With a Side Effect of Shaming. *The New York Times*. Retrieved March 7, 2026, from https://www.nytimes.com/2023/06/14/health/obesity-drugs-wegovy-ozempic.html

Lang, K., & Lehmann, J.-Y. K. (2012). Racial Discrimination in the Labor Market: Theory and Empirics. *Journal of Economic Literature*, *50*(4), 959–1006. https://doi.org/10.1257/jel.50.4.959

Levasseur, P. (2019). Implementing a regression discontinuity design to explore the heterogeneous effects of obesity on labour income: The case of Mexico. *Journal of Public Health*, *27*(1), 89–101. https://doi.org/10.1007/s10389-018-0925-5

Lindbeck, A., Nyberg, S., & Weibull, J. W. (1999). Social Norms and Economic Incentives in the Welfare State*. *The Quarterly Journal of Economics*, *114*(1), 1–35. https://doi.org/10.1162/003355399555936

Link, B. G., & Phelan, J. C. (2001). Conceptualizing Stigma. *Annual Review of Sociology*, *27*(Volume 27, 2001), 363–385. https://doi.org/10.1146/annurev.soc.27.1.363

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

Macchi, E. (2023). Worth Your Weight: Experimental Evidence on the Benefits of Obesity in Low-Income Countries. *American Economic Review*, *113*(9), 2287–2322. https://doi.org/10.1257/aer.20211879

McShane, B. B., & Gal, D. (2017). Statistical Significance and the Dichotomization of Evidence [_eprint: https://doi.org/10.1080/01621459.2017.1289846]. *Journal of the American Statistical Association*, *112*(519), 885–895. https://doi.org/10.1080/01621459.2017.1289846

Moffitt, R. (1983). An Economic Model of Welfare Stigma. *The American Economic Review*, *73*(5), 1023–1035. Retrieved September 15, 2025, from https://www.jstor.org/stable/1814669

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

Oberholzer-Gee, F. (2008). Nonemployment stigma as rational herding: A field experiment. *Journal of Economic Behavior & Organization*, *65*(1), 30–40. https://doi.org/10.1016/j.jebo.2004.05.008

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/

Park, J., Saha, S., Chee, B., Taylor, J., & Beach, M. C. (2021). Physician Use of Stigmatizing Language in Patient Medical Records. *JAMA Network Open*, *4*(7), e2117052. https://doi.org/10.1001/jamanetworkopen.2021.17052

Pescosolido, B. A., Halpern-Manners, A., Luo, L., & Perry, B. (2021). Trends in Public Stigma of Mental Illness in the US, 1996-2018. *JAMA Network Open*, *4*(12), e2140202. https://doi.org/10.1001/jamanetworkopen.2021.40202

Petersen, I., Nicolaisen, S. K., Ricciardi, F., Sharma, M., Thomsen, R. W., Baio, G., & Pedersen, L. (2020). Impact of Being Eligible for Type 2 Diabetes Treatment on All-Cause Mortality

and Cardiovascular Events: Regression Discontinuity Design Study. *Clinical Epidemiology*, *12*, 569–577. https://doi.org/10.2147/CLEP.S251704

Phelan, S., Burgess, D., Yeazel, M., Hellerstedt, W., Griffin, J., & van Ryn, M. (2015). Impact of weight bias and stigma on quality of care and outcomes for patients with obesity. *Obesity Reviews*, *16*(4), 319–326. https://doi.org/10.1111/obr.12266

Phelps, E. S. (1972). The Statistical Theory of Racism and Sexism. *The American Economic Review*, *62*(4), 659–661. Retrieved February 24, 2026, from https://www.jstor.org/stable/1806107

Puhl, R. M., & Heuer, C. A. (2009). The stigma of obesity: A review and update. *Obesity*, *17*(5), 941.

Rooth, D.-O. (2009). Obesity, Attractiveness, and Differential Treatment in Hiring: A Field Experiment. *The Journal of Human Resources*, *44*(3), 710–735. Retrieved February 24, 2026, from https://www.jstor.org/stable/20648914

Rubin, R. (2019a). Addressing medicine's bias against patients who are overweight. *Jama*, *321*(10), 925–927.

Rubin, R. (2019b). Addressing Medicine's Bias Against Patients Who Are Overweight. *JAMA*, *321*(10), 925–927. https://doi.org/10.1001/jama.2019.0048

Sabia, J. J. (2007). The Effect of Body Weight on Adolescent Academic Performance. *Southern Economic Journal*, *73*(4), 871–900. Retrieved February 24, 2026, from https://ideas.repec.org//a/wly/soecon/v73y2007i4p871-900.html

Samet, J., & Zureich, J. (2025). The Effects of Performance Targets on Employee Learning. Retrieved February 24, 2026, from https://www.dropbox.com/scl/fi/qtcw6jrddg2j5ucmmagsz/Samet-Zureich2026.pdf?dl=0&rlkey=0l677rv5oxybj27dk01lvenuy

Smith, E. C. (2025). Stigma and Social Cover: A Mental Health Care Experiment in Refugee Networks. *Working Paper*. https://www.dropbox.com/scl/fi/t6blamvn6ulv1fa50pbdw/JMP_EmmaSmith.pdf?rlkey=0nreo50w25z8n5j3ikko5hi9i&e=1&dl=0

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

Srikanth, K., & Erin, S. (2016). How Effective is Population-Based Cancer Screening? Regression Discontinuity Estimates from the US Guideline Screening Initiation Ages. *Forum for Health Economics & Policy*, *19*(1), 87–139. Retrieved March 4, 2026, from https://ideas.repec.org//a/bpj/fhecpo/v19y2016i1p87-139n2.html

Sun, M., Oliwa, T., Peek, M. E., & Tung, E. L. (2022). Negative Patient Descriptors: Documenting Racial Bias In The Electronic Health Record. *Health Affairs*, *41*(2), 203–211. https://doi.org/10.1377/hlthaff.2021.01423

Toshiaki, I., Katsuhiko, N., Brian, C., & Karen, E. (2021). False alarm? Estimating the marginal value of health signals [toshiaki_false_2021]. *Journal of Public Economics*, *195*, 104368. https://doi.org/10.1016/j.jpubeco.2021.104368

Treasure, J., & Ambwani, S. (2021). Addressing weight stigma and anti-obesity rhetoric in policy changes to prevent eating disorders. *The Lancet*, *398*(10294), 7–8. https://doi.org/10.1016/S0140-6736(21)01109-0

Whelton, P. K., Carey, R. M., Aronow, W. S., Casey, D. E., Collins, K. J., Dennison Himmelfarb, C., DePalma, S. M., Gidding, S., Jamerson, K. A., Jones, D. W., MacLaughlin, E. J., Muntner, P., Ovbiagele, B., Smith, S. C., Spencer, C. C., Stafford, R. S., Taler, S. J., Thomas, R. J., Williams, K. A., … Wright, J. T. (2018). 2017 ACC/AHA/AAPA/ABC/ACPM/AGS/APhA/ASH/ASPC Guideline for the Prevention, Detection, Evaluation, and Management of High Blood Pressure in Adults: Executive Summary: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. *Hypertension*, *71*(6), 1269–1324. https://doi.org/10.1161/HYP.0000000000000066

Wilder, M. E., Kulie, P., Jensen, C., Levett, P., Blanchard, J., Dominguez, L. W., Portela, M., Srivastava, A., Li, Y., & McCarthy, M. L. (2021). The Impact of Social Determinants of Health on Medication Adherence: A Systematic Review and Meta-analysis. *Journal of General Internal Medicine*, *36*(5), 1359–1370. https://doi.org/10.1007/s11606-020-06447-0

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

Zeng, Q., & Yu, X. (2019). Overweight and obesity standards and subjective well-being: Evidence from China. *Economics & Human Biology*, *33*, 144–148. https://doi.org/10.1016/j.ehb.2019.02.006

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: ""

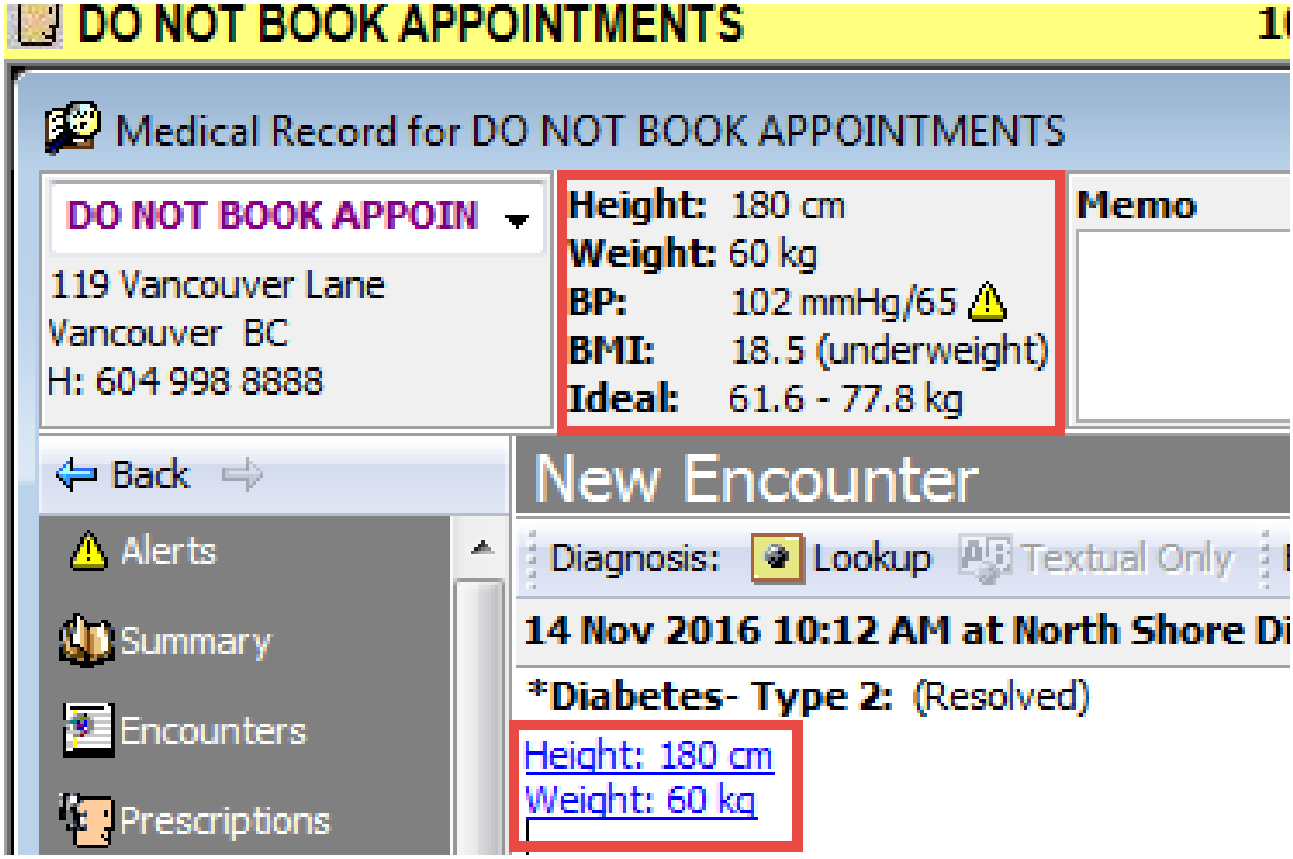


FIGURE A.1
Example of BMI Display in an Electronic Medical Record

Notes: Screenshot of a generic electronic medical record (EMR) interface illustrating how BMI and BMI category are saliently displayed for the provider. The continuous BMI value and categorical label (e.g., "underweight") are typically highlighted through color-coding or other visual cues. This example is illustrative and does not depict the specific EMR system used in the MIMIC-IV data.

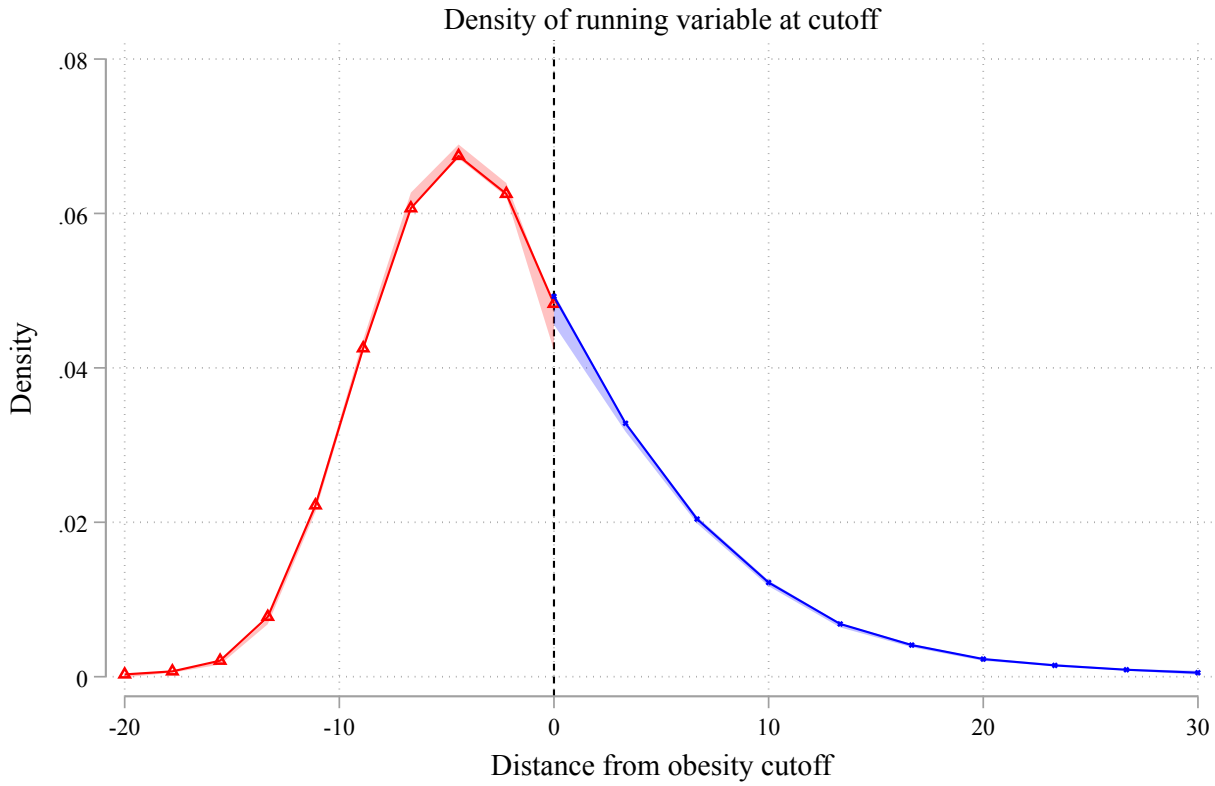


FIGURE A.2
McCrary Density Test at the Obesity Cutoff

Notes: Density of BMI estimated separately on each side of the obesity cutoff (BMI = 30) using the **cattaneo˙simple˙2020** local polynomial density estimator. The null hypothesis is no discontinuity in the density at the cutoff. Failure to reject ($p = 0.16$) indicates no evidence of sorting or manipulation around the threshold.

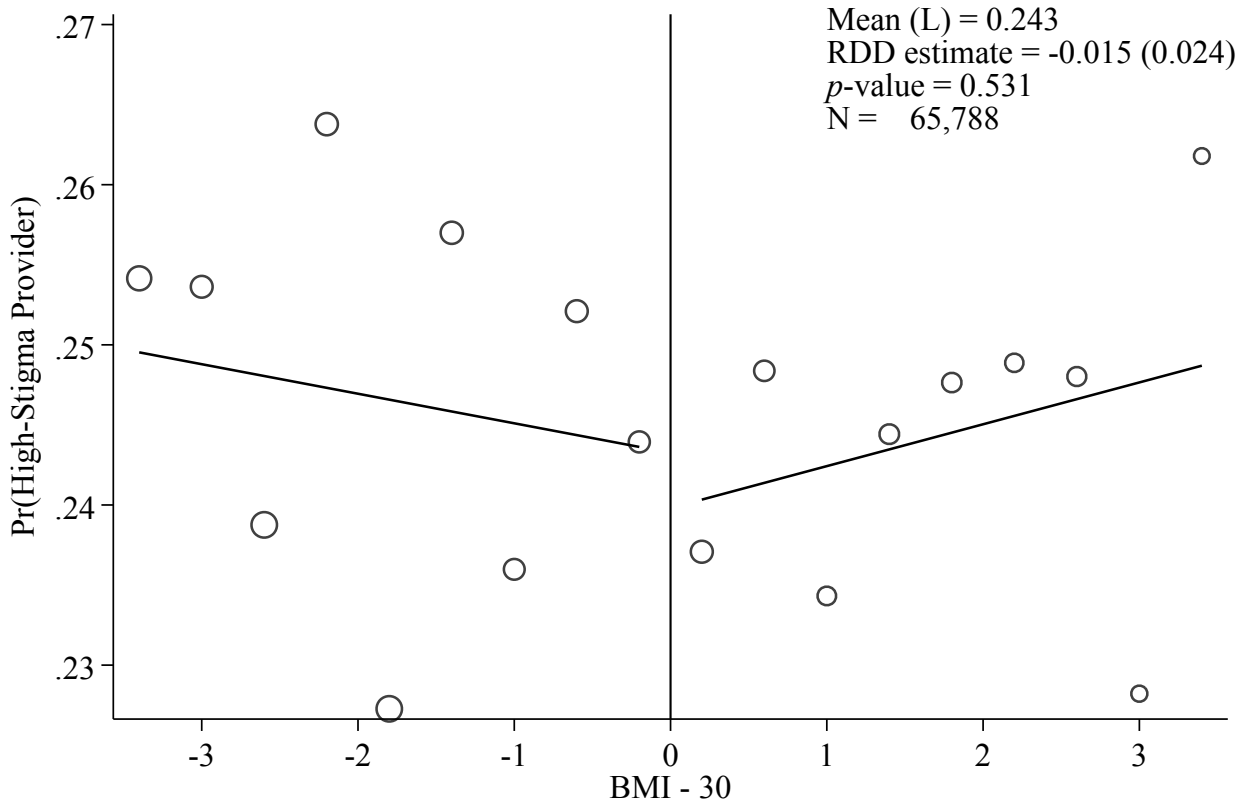
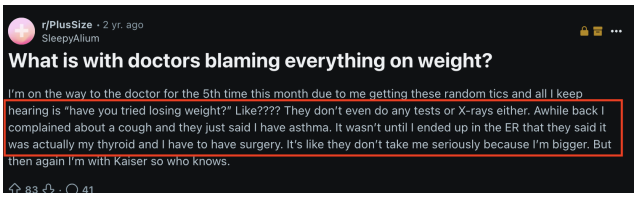


FIGURE A.3
RD estimate of provider type at the obesity cutoff

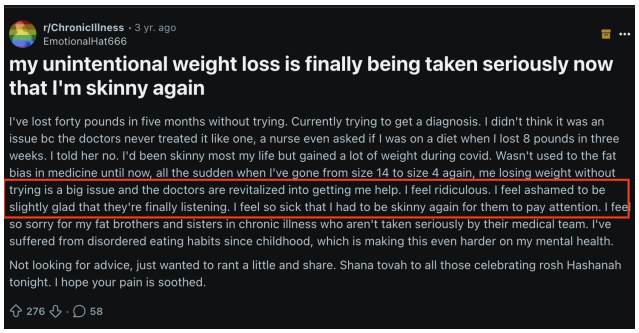
Notes: The figure plots the probability of being treated by a High-stigma provider (top quartile of mean stigmatizing language rate) against BMI relative to the obesity cutoff (BMI = 30). Each circle represents a bin mean, sized proportionally to the number of observations. Solid lines are local linear fits estimated separately on each side of the cutoff. The text box reports the robust, bias-corrected (RBC) RD estimate, standard error, p-value, the mean of the dependent variable to the left of the cutoff, and the sample size. All specifications use a local linear polynomial, triangular kernel, MSE-optimal bandwidth, and two-way clustering by patient and hospital service.



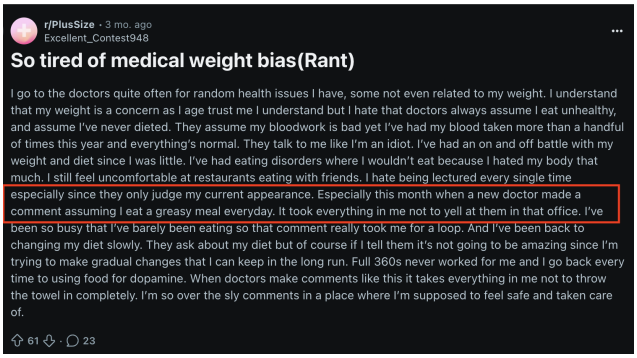
Panel (a)



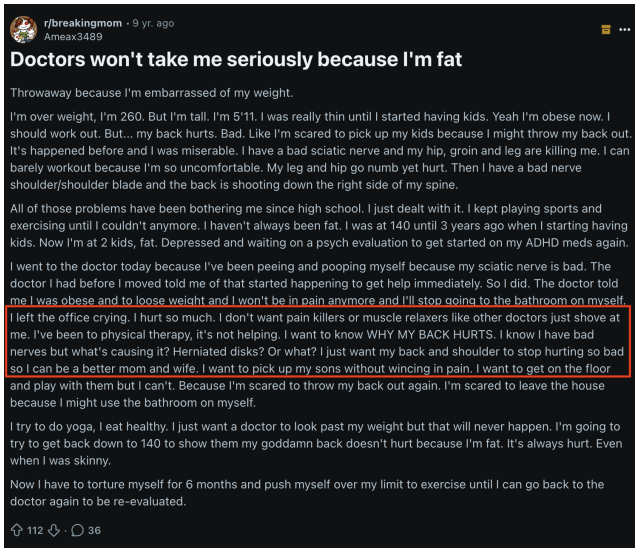
Panel (b)



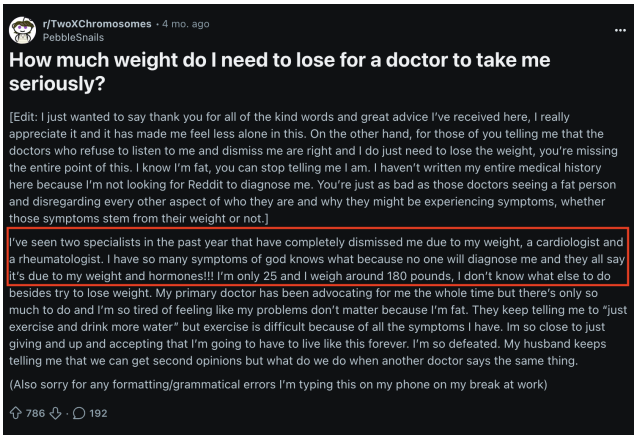
Panel (c)



Panel (d)



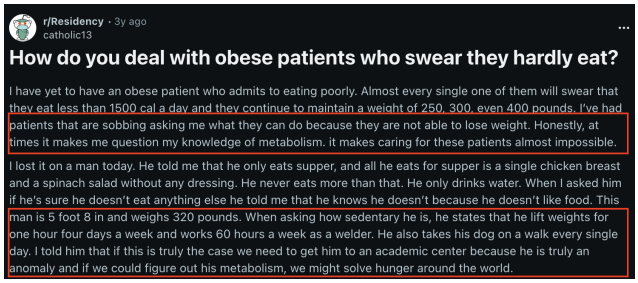
Panel (e)



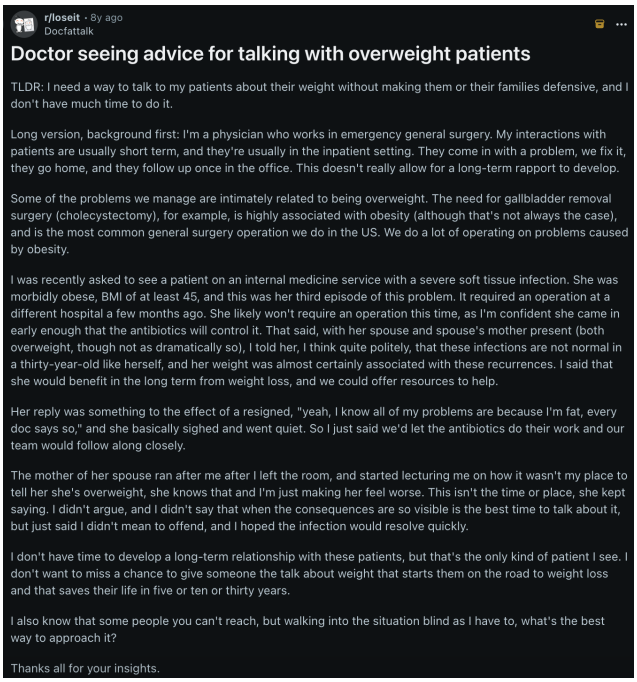
Panel (f)

FIGURE A.4 Anecdotal Evidence from Patients Consistent with Weight Stigma

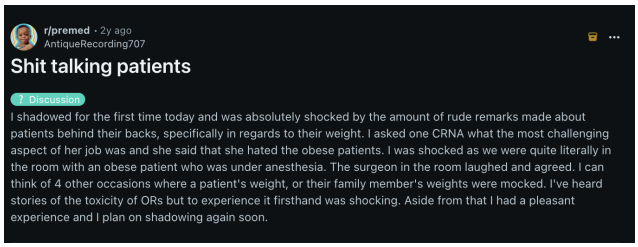
Notes: Screenshots of posts from Reddit subreddits in which patients describe experiences consistent with weight stigma in healthcare settings, including symptom dismissal, withheld diagnostic testing, and moralizing comments from providers. Posts have been cropped and identifying information redacted.



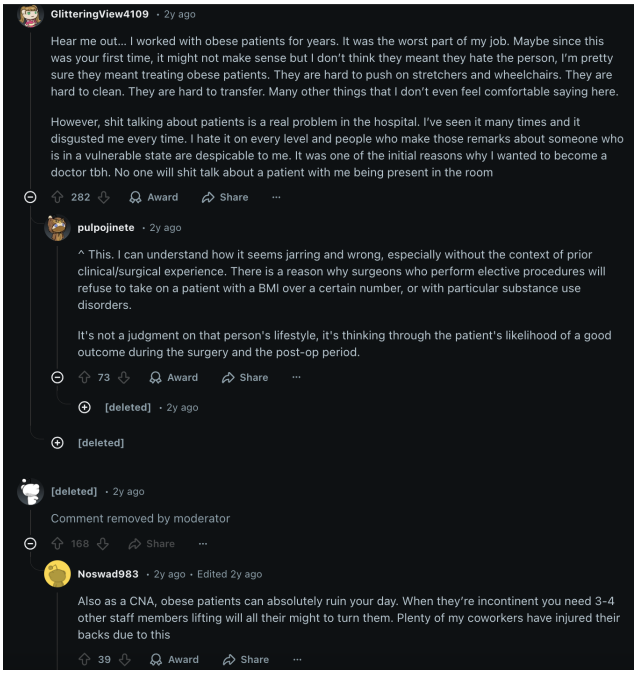
Panel (a)



Panel (b)



Panel (c)



Panel (d)

FIGURE A.5 Anecdotal Evidence from Providers Consistent with Weight Stigma

Notes: Screenshots of posts from Reddit subreddits in which physicians discuss caring for overweight and obese patients. Posts include language consistent with weight stigma, including biased descriptions of patients and moral judgments about lifestyle. Posts have been cropped and identifying information redacted.

Outcome	**Pred. Mortality**	**Obs. Mortality**		
	(1)	(2)	(3)	(4)
RD estimate (RBC)	-0.000 (0.001)	0.001 (0.002)	0.002 (0.002)	0.006** (0.002)
P-value (RBC)	0.945	0.825	0.433	0.017
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	2.739	5.321	2.684
Observations left	105,253	54,322	137,512	107,909
Observations right	57,056	112,408	29,218	57,430
Effective obs left	44,284	27,959	36,172	25,330
Effective obs right	28,204	31,175	17,080	18,758

TABLE A.1 RD placebo and donut checks

Notes: Each column reports the robust, bias-corrected (RBC) RD estimate, standard error (in parentheses), and RBC p-value using the Calonico, Cattaneo, and Titiunik (2014, 2017) procedure. All specifications use a triangular kernel, local linear polynomial, and patient-clustered standard errors. Col (1): outcome is predicted mortality from a logit of observed mortality on baseline covariates (age, sex, race, insurance, marital status, Elixhauser comorbidities and mortality index); cutoff at BMI = 30. Cols (2)–(3): outcome is observed in-hospital mortality; cutoffs are placebo thresholds at the within-group median BMI below and above 30, respectively. Col (4): donut-hole RD dropping observations within ±0.1 of the obesity cutoff. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.001$.

Outcome	RD estimate	p-value	Mean (left)	Bandwidth h	N (L/R)	Eff. N (L/R)
Female	0.025* (0.014)	0.073	0.47	2.659	108,342 / 58,388	25,504 / 19,673
Age at admission	0.149 (0.420)	0.722	61.86	3.906	108,342 / 58,388	39,481 / 26,523
Black	-0.005 (0.009)	0.578	0.15	3.501	108,342 / 58,388	35,368 / 24,539
Medicaid	0.002 (0.009)	0.812	0.17	4.565	107,255 / 58,032	46,302 / 29,325
Medicare	0.003 (0.011)	0.817	0.49	4.870	107,255 / 58,032	49,233 / 30,581
Private insurance	-0.007 (0.011)	0.533	0.32	4.672	107,255 / 58,032	47,457 / 29,731
Married	0.005 (0.012)	0.698	0.50	4.003	106,252 / 57,411	39,829 / 26,612
Hispanic	-0.003 (0.006)	0.620	0.06	4.844	108,342 / 58,388	49,575 / 30,688
Num of Elixhauser comorbidities	0.064 (0.047)	0.173	3.09	5.731	108,323 / 58,365	60,125 / 34,389
Elixhauser mortality index	0.195 (0.214)	0.363	7.46	3.622	108,323 / 58,365	36,641 / 25,035

TABLE A.2 Covariate balance at the obesity cutoff

Notes: Each row reports the robust, bias-corrected (RBC) RD estimate, standard error, and p-value for a pre-determined covariate using the Calonico, Cattaneo, and Titiunik (2014, 2017) procedure at the obesity cutoff (BMI = 30). Mean (left) is the sample mean just below the cutoff. N (L/R) is the total number of observations on each side of the cutoff; Eff. N (L/R) is the effective number of observations within the MSE-optimal bandwidth on each side; h is the MSE-optimal bandwidth. All specifications use a local linear polynomial, triangular kernel, and patient-clustered standard errors. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.001$.

	Dep Var: In-Hospital Mortality		
	(1)	(2)	(3)
RD estimate (RBC)	-0.002 (0.002)	-0.000 (0.003)	-0.004 (0.003)
P-value (RBC)	0.372	0.889	0.276
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)	2.691	5.158	4.860
Observations left	55,922	139,589	154,916
Observations right	110,808	27,141	11,814
Effective obs left	27,344	32,302	14,622
Effective obs right	30,629	15,647	6,609

TABLE A.3 RD estimates of in-hospital mortality at alternative BMI cutoffs

Notes: Each column reports the robust, bias-corrected (RBC) RD estimate, standard error, and p-value using the Calonico, Cattaneo, and Titiunik (2014, 2017) procedure at an alternative BMI cutoff where no "obese" label is assigned: (1) BMI = 25 (overweight), (2) BMI = 35 (obese class II), and (3) BMI = 40 (obese class III). The dependent variable is in-hospital mortality. Mean (left) is the sample mean to the left of the cutoff. All specifications use a local linear polynomial, triangular kernel, MSE-optimal bandwidth, and patient-clustered standard errors. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.001$.

	Low-stigma			High-stigma			Diff-in-D		
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.4 Diff-in-Disc covariate balance by provider stigma group

Notes: Each row reports the robust, bias-corrected (RBC) RD estimate, standard error, and p-value for a pre-determined covariate using the Calonico, Cattaneo, and Titiunik (2014, 2017) procedure at the obesity cutoff (BMI = 30). The first three columns report RD estimates among Low-stigma providers (quartiles 1–3); the next three among High-stigma providers (quartile 4); the final three report the Diff-in-Disc estimate (High − Low). All specifications use a local linear polynomial, triangular kernel, and patient-clustered standard errors. Predicted mortality is generated from a logit of observed mortality on baseline covariates. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.001$.

	Dep Var: Any Stigma in Notes				
	(1)	(2)	(3)	(4)	(5)
RD estimate					
Low-stigma providers	-0.036*	-0.032	-0.038*	-0.036*	-0.023*
	(0.019)	(0.027)	(0.020)	(0.019)	(0.014)
High-stigma providers	0.044	0.043*	0.044	0.044	0.019**
	(0.029)	(0.022)	(0.041)	(0.028)	(0.009)
Diff-in-Disc estimate	**0.0800****	**0.0749****	**0.0823***	**0.0800****	**0.0418****
	(0.034)	**(0.038)**	**(0.047)**	**(0.033)**	**(0.017)**
Polynomial	Linear	Quadratic	Linear	Linear	Linear
Covariates	No	No	Yes	No	No
Clustering	Pat., Svc.	Pat., Svc.	Pat., Svc.	Prov., Svc.	Pat., Svc.
Method	RD	RD	RD	RD	OLS
Bandwidth	MSE-opt.	MSE-opt.	MSE-opt.	MSE-opt.	±1.5
Total N	166,730	166,730	166,730	166,730	20,018

TABLE A.5 Robustness of the Diff-in-Disc effect on stigmatizing language

Notes: Each column reports the robust, bias-corrected (RBC) RD estimate and standard error (in parentheses) for Low- and High-stigma providers separately, as well as the Diff-in-Disc estimate (High − Low), using variations of the baseline specification. The dependent variable is an indicator for any stigmatizing language in the discharge summary. Column (1) is the baseline: local linear polynomial, triangular kernel, MSE-optimal bandwidth, and two-way clustering by patient and hospital service. Column (2) uses a local quadratic polynomial. Column (3) adds covariates (age, race, sex, private insurance) to the local linear specification. Column (4) clusters by provider and hospital service instead of patient and hospital service. Column (5) estimates the RD parametrically via OLS using a linear interaction model on a ±1.5 BMI window around the cutoff. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.001$.

	Dep Var: In-hospital mortality				
	(1)	(2)	(3)	(4)	(5)
RD estimate					
Low-stigma providers	0.003	0.003	0.003	0.003	-0.000
	(0.003)	(0.003)	(0.003)	(0.003)	(0.002)
High-stigma providers	0.013***	0.014***	0.013***	0.013***	0.008***
	(0.003)	(0.004)	(0.003)	(0.002)	(0.001)
Diff-in-Disc estimate	**0.0104***	**0.0113***	**0.0106***	**0.0104****	**0.0083*****
	(0.006)	**(0.007)**	**(0.006)**	**(0.005)**	**(0.002)**
Polynomial	Linear	Quadratic	Linear	Linear	Linear
Covariates	No	No	Yes	No	No
Clustering	Pat., Svc.	Pat., Svc.	Pat., Svc.	Prov., Svc.	Pat., Svc.
Method	RD	RD	RD	RD	OLS
Bandwidth	MSE-opt.	MSE-opt.	MSE-opt.	MSE-opt.	±3.0
Total N	166,730	166,730	166,730	166,730	51,630

TABLE A.6 Robustness of the Diff-in-Disc effect on in-hospital mortality

Notes: Each column reports the robust, bias-corrected (RBC) RD estimate and standard error (in parentheses) for Low- and High-stigma providers separately, as well as the Diff-in-Disc estimate (High − Low), using variations of the baseline specification. The dependent variable is in-hospital mortality. Column (1) is the baseline: local linear polynomial, triangular kernel, MSE-optimal bandwidth, and two-way clustering by patient and hospital service. Column (2) uses a local quadratic polynomial. Column (3) adds covariates (age, race, sex, private insurance) to the local linear specification. Column (4) clusters by provider and hospital service instead of patient and hospital service. Column (5) estimates the RD parametrically via OLS using a linear interaction model on a ±3 BMI window around the cutoff. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.001$.

	Excl. Obese ICD		Leave-One-Out	
	Any Stigma in Notes	Mortality	Any Stigma in Notes	Mortality
RD estimate				
Low-stigma providers	-0.030	0.003	-0.035*	0.003
	(0.019)	(0.003)	(0.019)	(0.003)
High-stigma providers	0.033*	0.012***	0.039	0.013***
	(0.019)	(0.003)	(0.029)	(0.003)
Diff-in-Disc estimate	**0.0632****	**0.0087****	**0.0743****	**0.0106****
	(0.026)	**(0.004)**	**(0.034)**	**(0.005)**
Total N	65,788	166,730	65,788	166,730

TABLE A.7 Robustness of Diff-in-Disc estimates to alternative provider classification methods

Notes: Each column reports the robust, bias-corrected (RBC) RD estimate, standard error (in parentheses), and Diff-in-Disc estimate for Low- and High-stigma providers using two alternative methods of classifying provider type. Columns (1)–(2) define provider type after excluding patients with an obese ICD diagnosis code from the calculation of each provider's mean stigmatizing language rate. Columns (3)–(4) use a leave-one-out measure, where each patient's provider stigma rate is calculated across all other patients of that provider, excluding the focal patient. In columns (1) and (3) the dependent variable is an indicator for any stigmatizing language in the discharge summary; in columns (2) and (4) it is in-hospital mortality. The Diff-in-Disc estimate is the difference in the RD effect between High- and Low-stigma providers. All specifications use a local linear polynomial, triangular kernel, MSE-optimal bandwidth, and two-way clustering by patient and hospital service. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.001$.

	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.8 Heterogeneous RD effects on stigmatizing language by height quartile

Notes: Each row reports the robust, bias-corrected (RBC) RD estimate, standard error, and p-value using the Calonico, Cattaneo, and Titiunik (2014, 2017) procedure at the obesity cutoff (BMI = 30). The outcome is an indicator for any stigmatizing language in the discharge summary. The sample is split into quartiles of patient height (Q1 = shortest, Q4 = tallest). The Diff-in-Disc estimate is the difference in the RD effect between the highest and lowest height quartiles (i.e., Q4 − Q1). All specifications use a local linear polynomial, triangular kernel, and patient-clustered standard errors. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.001$.

	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.9 First-difference: changes in stigmatizing language on changes in BMI

Notes: The table reports a first-difference regression of $\Delta Stigma_{t-(t-1)}$ on $\Delta BMI_{t-(t-1)}$, where each observation is a within-patient change across consecutive hospital encounters. The dependent variable is the change in an indicator for any stigmatizing language in the discharge summary; the independent variable is the change in BMI. Standard errors (in parentheses) are two-way clustered by patient and provider. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.001$.

Outcome	estimate	SE (RBC)	p-value	total N	Eff. N (L)	Eff. N (R)	h	mean (sample)
Word count	-28.092**	10.523	0.008	133,318	34,771	21,896	4.272	1674.57
Character count	-186.768**	68.711	0.007	133,318	33,585	21,599	4.197	11096.16
Average word length	-0.002	0.005	0.644	133,318	34,357	21,818	4.208	5.39

TABLE A.10 Effect of the obesity cutoff on discharge-note length

Notes: Each row reports the robust, bias-corrected (RBC) RD estimate, standard error, and p-value for a text-based measure of discharge-summary length using the Calonico, Cattaneo, and Titiunik (2014, 2017) procedure at the obesity cutoff (BMI = 30). Outcomes are word count, character count, and average word length. Mean (sample) is the sample average of the outcome. Unlike the main specifications, the sample includes all patients with a discharge summary and a recorded BMI (not restricted to the overweight and obese range). All specifications use a local linear polynomial, triangular kernel, and two-way clustering by patient and hospital service. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.001$.

	Dep Var: In-Hospital Mortality			
	(1)	(2)	(3)	(4)
RD estimate (RBC)	-0.0051 (0.0059)	-0.0144** (0.0054)	-0.0024 (0.0051)	-0.0060 (0.0068)
P-value (RBC)	0.392	0.008	0.631	0.377
Mean, left of threshold	0.024	0.019	0.015	0.008
Cutoff information				
BP category	Prehypertension	Stage 1	Stage 2	Crisis
Systolic / Diastolic threshold	120/80	130/80	140/90	180/120
Classification	JNC-7	AHA/ACC	AHA/ACC	AHA/ACC
RD details				
Observations left	11,336	15,619	22,860	31,474
Observations right	20,792	16,509	9,268	654
Effective obs left	5,643	6,782	7,824	1,699
Effective obs right	8,111	7,948	5,934	502
Polynomial order (p)	1	1	1	1
Kernel	Tri	Tri	Tri	Tri
Bandwidth (h) (L/R)	0.536	0.565	0.675	0.996
Cluster vars	P	P	P	P

TABLE A.11 Placebo RD: In-hospital mortality at hypertension blood pressure cutoffs

Notes: Each column reports the robust, bias-corrected (RBC) RD estimate of the effect of crossing a hypertension threshold on in-hospital mortality. The running variable is a binding score: the maximum of the standardized distances of systolic and diastolic blood pressure from their respective cutoffs, so that the score equals zero when the binding constraint first reaches the threshold. Column (1) uses the JNC-7 "Prehypertension" cutoff (systolic $\geq$120 *or* diastolic $\geq$80 mmHg). Columns (2)–(4) use the AHA/ACC 2017 classification: Stage 1 (130/80), Stage 2 (140/90), and Hypertensive Crisis (180/120). All thresholds use OR logic (either dimension binding), which is the correct clinical definition for Stage 1 and above. Blood pressure is measured as the first recorded value during the hospitalization. All specifications use a local linear polynomial, triangular kernel, and patient-clustered standard errors. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.001$.