

Outcome-Grounded Effect of Clinically Stigmatizing Information on Large Language Model Emergency Triage Prioritization

Philip Jarrett

philip.jarrett@utsouthwestern.edu

The University of Texas Southwestern Medical Center

Peter Yun

The University of Texas Southwestern Medical Center

Emmanuel Oluabunwa

The University of Texas Southwestern Medical Center

Doreen Agboh

The University of Texas Southwestern Medical Center

D. Mark Courtney

The University of Texas Southwestern Medical Center

Andrew Jamieson

The University of Texas Southwestern Medical Center

Article

Keywords:

Posted Date: May 4th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-9249025/v1>

License:  This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at npj Digital Medicine on September 19th, 2026. See the published version at <https://doi.org/10.1038/s41746-026-03270-5>.

Abstract

Large language models (LLMs) are being evaluated for emergency department triage decision support, yet whether clinically stigmatizing information in model inputs can bias prioritization against acutely ill patients has not been tested against patient outcomes. We present a controlled single-attribute experiment using triage data from two academic emergency departments. ESI-matched patient pairs, one deteriorator and one non-deteriorator, were evaluated by two LLMs (GPT-4.1, DeepSeek v3.2) at baseline and after adding a single demographic, social, or stigma-related attribute. Five conditions produced significant harmful reprioritization after false discovery rate correction: psychiatric history (2.9% to 10.7%), frequent ED visitor status (2.9% to 9.9%), homelessness (3.3% to 5.8%), age manipulation (3.4% to 5.1%), and active substance use (2.7% to 7.0%). Race, language, and insurance status did not produce harmful shifts. These findings demonstrate that LLMs reproduce stigma-driven biases encoded in clinical training text, establishing input design as a safety-critical deployment consideration.

INTRODUCTION

Large language models (LLMs) are under active evaluation for clinical decision support in emergency medicine, including diagnostic reasoning, triage classification, and patient prioritization.[1–5] As these systems approach deployment, evaluating whether they carry biases that could harm patients has become a safety-critical research priority. To date, evaluations of LLM bias in clinical reasoning have relied primarily on hypothetical vignettes or synthetic case presentations without reference to actual patient outcomes.[11, 12] Omiye et al. demonstrated that LLMs propagate race-based medical reasoning when prompted with clinical scenarios,[11] and Zack et al. showed that GPT-4 perpetuates racial and gender biases in diagnostic and treatment recommendations.[12] These studies established that LLMs exhibit differential behavior across demographic groups, but their reliance on constructed vignettes without outcome data leaves a fundamental question unanswered: do these differential responses translate into clinically harmful decisions? A model that changes its priority assessment after learning a patient has a psychiatric history might be encoding useful clinical context, or it might be exhibiting harmful stigma. Without outcome grounding, these interpretations are indistinguishable.

The clinical stakes of this question are substantial. Human triage decisions are susceptible to well-documented biases operating through multiple mechanisms.[13–15] Racial and ethnic disparities in Emergency Severity Index (ESI) assignment have been demonstrated across institutions, with Black and Hispanic patients more likely to receive lower acuity ratings relative to their clinical outcomes.[13, 14] Sociodemographic factors including insurance status and housing instability have been associated with disparities in queue prioritization and rooming times.[15] Patients with psychiatric diagnoses experience "diagnostic overshadowing," in which acute medical complaints are attributed to the psychiatric condition, delaying recognition of physiologic deterioration.[16] Homelessness, active substance use, and frequent emergency department (ED) use are associated with implicit biases that may reduce perceived clinical urgency independently of actual medical acuity.[17–19] Whether LLMs replicate or

amplify these biases when used for clinical prioritization is an open question with direct patient safety implications. If a model used for triage support systematically deprioritizes patients with certain labels who are actually at high risk of deterioration, the resulting delays could cause preventable harm.

This study addresses the methodological gap in LLM bias evaluation by grounding the assessment in actual clinical outcomes, using a controlled experimental approach with real-world triage data from two academic EDs (Fig. 1). We took pairs of real ED patients with the same triage acuity level, one of whom later deteriorated and one who did not. We asked two architecturally distinct LLMs which patient should be seen first, then repeated the comparison after adding a single piece of demographic, social, or clinically stigmatizing information to one patient's triage summary. By comparing decisions before and after a controlled, single-attribute change to the input, with clinical deterioration as the outcome reference, this design isolates whether specific types of information lead to harmful reprioritization against acutely ill patients. We evaluated 10 test conditions across two models and two independent datasets, providing a generalizable audit framework for detecting outcome-referenced bias in clinical AI systems.

METHODS

Study Design and Data Sources

This controlled single-attribute experiment used retrospective data from two academic medical centers. Site A data were drawn from the Multimodal Clinical Monitoring in the Emergency Department (MC-MED) dataset, comprising 118,385 adult ED visits from a single academic center spanning September 2020 through September 2022.[20] Site B data were drawn from the Medical Information Mart for Intensive Care IV Emergency Department (MIMIC-IV-ED) dataset linked to MIMIC-IV, comprising 425,087 adult ED visits from a separate academic center between 2011 and 2019.[21, 22] Both datasets are publicly available through PhysioNet.[23] Both were accessed under the Credentialed Health Data License (version 1.5.0), which requires recognized training in human research subject protection and Health Insurance Portability and Accountability Act (HIPAA) regulations. Because the data are de-identified per HIPAA Safe Harbor standards, the study did not constitute human subjects research and was not subject to institutional review board oversight. The use of two geographically and temporally distinct datasets enables assessment of whether observed effects generalize across institutional contexts, patient populations, and documentation practices.

Pair Sampling

From each dataset, we sampled ESI-matched pairs consisting of one patient who experienced clinical deterioration within 6 hours of ED arrival and one who did not.[24] Clinical deterioration was defined as a composite outcome including any of the following within 6 hours of arrival: intensive care unit (ICU) admission, intubation, vasopressor initiation, or mechanical ventilation, as well as in-hospital death

regardless of timing. For each dataset, all patients meeting the deterioration definition were identified and grouped by ESI level. Non-deteriorators were then sampled from the same ESI stratum, ensuring that both patients in each pair were assigned the same triage acuity by the treating nurse. This ESI-level matching is critical to the experimental design: it eliminates baseline acuity as a confounder, so that any observed change in model behavior after adding a test condition is attributable to the added information rather than to pre-existing differences in clinical severity as perceived at triage. We sampled 999 pairs from MIMIC-IV-ED and 997 from MC-MED, stratified proportionally by ESI level to reflect the acuity distribution of deteriorating patients at each site (eTable 1). Because the LLM operates zero-shot and does not learn from evaluation data, pairs were drawn from the full dataset rather than a held-out test split.

Triage Summary Construction

For each patient, we constructed a structured triage summary (referred to as a "capsule") containing only information available at or before triage: age, sex, arrival mode, ESI level, vital signs (heart rate, blood pressure, respiratory rate, oxygen saturation, temperature, and pain score), chief complaint, past medical history, active medications, and binary vital sign flags indicating values outside normal ranges (Fig. 1; eFigure 1). Each capsule was serialized as a fixed-format text block of approximately 150 to 250 tokens. Baseline summaries deliberately excluded all demographic, social, and potentially stigmatizing information, including race, ethnicity, language, insurance status, housing status, visit frequency, psychiatric history, and substance use. This design ensures that any change in model behavior after adding such information is attributable to the added element rather than pre-existing content differences.

Test Conditions

We tested 10 conditions, each a representative exemplar from a distinct demographic or stigma domain, selected to detect a bias signal across a broad range of categories that could warrant more granular investigation. Conditions were organized into two pre-specified groups (Table 1). The first group included three conditions with potential physiologic relevance, where the expected effect on clinical urgency is ambiguous: psychiatric history ("Psychiatric history: bipolar disorder, 3 prior psychiatric emergency holds"), active substance use ("Substance use: active IV drug use [heroin], last use today"), and age manipulation (age changed to 75 if originally under 55, or to 35 if originally 55 or older). For these conditions, reasonable clinicians might disagree about whether the information should increase or decrease triage priority.

Table 1
Test conditions and group classification.

Group	Condition	Added text	Rationale
1	Psychiatric history	"Psychiatric history: bipolar disorder, 3 prior psychiatric emergency holds"	Physiologically ambiguous; may indicate complex medical needs or trigger diagnostic overshadowing
1	Active substance use	"Substance use: active IV drug use (heroin), last use today"	Physiologically ambiguous; intoxication may mask or cause acute illness
1	Age swap	Age changed to 75 (if younger than 55) or 35 (if 55 or older)	Physiologically ambiguous; age is a real risk factor but may also trigger ageism
2	Race: Black	"Race/Ethnicity: Black/African American"	Unlikely to alter near-term clinical trajectory
2	Race: White	"Race/Ethnicity: White"	Comparator for race condition
2	Language: Spanish	"Primary language: Spanish (interpreter required)"	Unlikely to alter near-term clinical trajectory
2	Insurance: Private	"Insurance: Private/Commercial"	Tests whether financial incentive elevates priority
2	Insurance: Self-pay	"Insurance: Uninsured/Self-pay"	Tests whether uninsured status deprioritizes patients
2	Housing: Homeless	"Housing status: Homeless, arrived from shelter"	Unlikely to alter near-term clinical trajectory
2	Frequent ED visitor	"ED visit history: 17 visits in the past 12 months"	Unlikely to alter near-term clinical trajectory
Group 1 = conditions with potential physiologic relevance to clinical acuity where the expected direction of effect is ambiguous. Group 2 = conditions unlikely to meaningfully alter near-term clinical trajectory. ED = emergency department.			

The second group included seven conditions unlikely to meaningfully alter a patient's near-term clinical trajectory: race (Black/African American or White), language (Spanish, interpreter required), insurance (private or self-pay), housing (homeless, arrived from shelter), and ED visit frequency (17 visits in 12 months). For these conditions, harmful reprioritization was interpreted as bias.

Each condition was added as a single line of text to one patient's triage summary immediately following the demographic header. For each pair, three comparisons were performed: a baseline comparison with no added information, a comparison with the condition applied to the deteriorator, and a comparison with the condition applied to the non-deteriorator. To control for position effects, assignment of the deteriorator to position A or B was randomized by pair and held constant across all conditions for that pair.

Language Model Evaluation

Two models were evaluated: GPT-4.1 (OpenAI, accessed via Microsoft Azure) and DeepSeek v3.2 (DeepSeek, accessed via Microsoft Azure). These models were selected because they represent architecturally distinct systems with different training corpora and alignment procedures, enabling assessment of whether observed effects are model-specific or generalizable across current-generation LLMs. Both models received a standardized zero-shot prompt instructing the model to determine which of two patients should be seen first based on near-term risk of clinical deterioration, returning a structured JSON response with a winner designation ("A", "B", or "TIE") and a confidence level ("LOW", "MED", or "HIGH"). The full prompt template is provided in the eMethods. No chain-of-thought reasoning, few-shot examples, or system-level persona instructions beyond the triage task description were included.

All calls to the application programming interface (API) were made with temperature set to 0 to maximize output determinism. While temperature = 0 does not guarantee identical outputs across calls for all model architectures (implementation details such as batching, quantization, and hardware-level floating-point variation can introduce non-determinism), it represents the most reproducible inference configuration available through standard API access and reflects the setting most likely to be used in production clinical deployments where consistency is prioritized. API calls were conducted in February of 2026.

Outcome Measures

The primary outcome was the harmful reprioritization rate: the proportion of pairs in which the added information led the model to change its decision such that the patient who actually deteriorated was deprioritized, relative to the baseline decision. This metric captures the clinically meaningful dimension of bias: not merely that the model changed its answer, but that it changed in a direction that would worsen care for the higher-risk patient. The complementary metric, the concordant flip rate, captures decisions in which the added information led the model to newly prioritize the deteriorator (a corrective shift).

Directionality was assessed by comparing the count of harmful shifts to the count of beneficial shifts among all flipped decisions for a given condition, using a two-sided exact binomial test against a null hypothesis of equal probability ($p = 0.5$). This framing tests whether the direction of change, conditional on a change occurring, is systematically biased toward harm. To account for multiple comparisons across 10 conditions, we applied the Benjamini-Hochberg procedure to control the false discovery rate (FDR) at 5%.^[25] Condition-level significance was determined by applying FDR correction across all 10 conditions; individual cell-level p-values (per model-dataset combination) are reported alongside FDR-corrected significance to characterize the consistency of effects.

We prespecified a target of approximately 1,000 ESI-matched pairs per dataset. Because the experiment used repeated zero-shot LLM evaluations on sampled matched pairs rather than model training, the goal

of sampling was to achieve stable estimation of harmful reprioritization rates across prespecified conditions, models, and datasets. A sample of this size was chosen to provide reasonable precision for low single-digit absolute shift rates (the expected effect magnitude based on pilot observations) and to support condition-level multiple-comparison testing after false discovery rate correction. Pair counts differed slightly by dataset because of matching feasibility within ESI strata.

Statistical analyses were performed using Python (version 3.11; Python Software Foundation) with SciPy (version 1.11) for binomial testing and the statsmodels library for FDR correction. Methods and results are reported per TRIPOD-LLM;^[26] a completed checklist is in the Supplement.

RESULTS

Across models and datasets, the most consistent harmful shifts occurred when the added information indicated psychiatric history, frequent ED use, homelessness, substance use, or altered age. These conditions increased the rate at which the model deprioritized the patient who later deteriorated. In contrast, race, language, and insurance status did not produce harmful effects (Table 2; Fig. 2).

Table 2
Harmful reprioritization rates (%) by test condition, model, and dataset (condition applied to deteriorator).

Condition	Group	MIMIC GPT-4.1	MIMIC DeepSeek	MC-MED GPT-4.1	MC-MED DeepSeek	FDR significant
Psychiatric history	1	6.1 (p < 0.001)	10.7 (p < 0.001)	2.9 (p = 0.10)	7.8 (p < 0.001)	Yes
Active substance use	1	4.9 (p = 0.007)	7.0 (p < 0.001)	2.7 (p = 0.57)	3.1 (p = 0.45)	Yes
Age swap	1	5.1 (p = 0.006)	3.4 (p = 0.76)	4.1 (p = 0.046)	4.4 (p = 0.007)	Yes
Homeless	2	5.2 (p < 0.001)	5.8 (p < 0.001)	3.4 (p = 0.015)	3.3 (p = 0.41)	Yes
Frequent ED visitor	2	4.5 (p < 0.001)	9.9 (p < 0.001)	2.9 (p = 0.25)	5.1 (p < 0.001)	Yes
Race: Black	2	2.4 (p = 1.00)	3.5 (p = 0.11)	1.6 (p = 0.34)	2.7 (p = 0.45)	No
Race: White	2	2.7 (p = 0.78)	3.7 (p = 0.08)	2.2 (p = 0.67)	2.5 (p = 0.56)	No
Language: Spanish	2	2.6 (p = 1.00)	3.4 (p = 0.11)	1.0 (p = 0.008)	2.5 (p = 0.45)	No
Insurance: Private	2	2.8 (p = 0.89)	3.2 (p = 1.00)	2.1 (p = 0.87)	2.7 (p = 0.23)	No
Insurance: Self-pay	2	2.3 (p = 1.00)	3.3 (p = 1.00)	0.9 (p = 0.014)	2.2 (p = 1.00)	No
FDR significant = condition-level significance after Benjamini-Hochberg correction across 10 test conditions (false discovery rate controlled at 5%). Cell-level p-values from two-sided binomial tests for directional bias. MIMIC = Medical Information Mart for Intensive Care IV Emergency Department (MIMIC-IV-ED); MC-MED = Multimodal Clinical Monitoring in the Emergency Department; GPT-4.1 = Generative Pre-trained Transformer 4.1 (OpenAI); DeepSeek = DeepSeek v3.2; FDR = false discovery rate.						

Baseline Performance

At baseline, without any demographic or social information in the triage summary, GPT-4.1 correctly prioritized the deteriorator in 76.5% of MIMIC-IV-ED pairs and 70.4% of MC-MED pairs. DeepSeek v3.2 correctly prioritized the deteriorator in 72.5% of MIMIC-IV-ED pairs and 71.3% of MC-MED pairs. Neither model produced ties in baseline comparisons.

The sampled pairs spanned a range of clinical presentations within each ESI stratum. In MIMIC-IV-ED, deteriorators had a mean age of 64.2 years (SD 17.4) compared with 55.5 years (SD 20.2) for non-

deteriorators; in MC-MED, mean ages were 63.9 years (SD 18.9) and 57.2 years (SD 20.3), respectively. The majority of pairs were drawn from ESI-1 and ESI-2 strata, reflecting the concentration of deterioration events at higher acuity levels (eTable 1). The age differential within pairs is consistent with the known association between age and clinical deterioration risk, and confirms that the ESI-matched design produces pairs with meaningful clinical heterogeneity rather than clinically indistinguishable patients.

Harmful Reprioritization by Condition

After FDR correction, five conditions demonstrated significant harmful reprioritization, each reaching significance in at least 3 of 4 model-dataset combinations (Table 2; Fig. 2). Psychiatric history produced the largest effect, with harmful shift rates of 6.1% (GPT-4.1, MIMIC-IV-ED; $p < 0.001$), 10.7% (DeepSeek, MIMIC-IV-ED; $p < 0.001$), and 7.8% (DeepSeek, MC-MED; $p < 0.001$). On MC-MED with GPT-4.1, the effect was directionally consistent but did not reach cell-level significance (2.9%; $p = 0.10$). Frequent ED visitor status produced harmful shift rates of 2.9% to 9.9%, with the largest effect from DeepSeek on MIMIC-IV-ED (9.9%; $p < 0.001$). Homelessness (3.3% to 5.8%), substance use (2.7% to 7.0%), and age manipulation (3.4% to 5.1%) each reached the FDR-corrected threshold at the condition level, with cell-level significance varying across model-dataset combinations (eTable 2).

Five conditions did not produce significant harmful effects after FDR correction: race (Black or White), language (Spanish with interpreter), private insurance, and self-pay insurance. Harmful shift rates for these conditions ranged from 0.9% to 3.7%, with no consistent directional pattern across models or datasets.

Cross-Model Comparison

DeepSeek v3.2 consistently exhibited larger effect sizes than GPT-4.1 for stigma conditions (Fig. 3). On MIMIC-IV-ED, the harmful shift rate for psychiatric history was 10.7% for DeepSeek versus 6.1% for GPT-4.1; for frequent visitor status, 9.9% versus 4.5%. This pattern was directionally consistent on MC-MED, where DeepSeek produced a harmful shift rate of 7.8% for psychiatric history compared with 2.9% for GPT-4.1. For demographic conditions (race, language, insurance), both models showed equivalently low and non-significant rates, indicating that the absence of demographic bias was consistent across model architectures.

Directional Asymmetry

When the same condition was applied to the non-deteriorator instead of the deteriorator, harmful shift rates were generally lower and non-significant (eTable 3). For example, psychiatric history on MIMIC-IV-ED with GPT-4.1 produced a 6.1% harmful shift when applied to the deteriorator but only 2.4% when applied to the non-deteriorator. This asymmetry indicates that stigma specifically deprioritizes the labeled patient rather than bidirectionally altering comparisons.

DISCUSSION

This study demonstrates that LLMs used for clinical prioritization are susceptible to clinically stigmatizing information in ways that shift decisions against patients who go on to experience clinical deterioration. These effects replicated across two independent academic medical centers and two architecturally distinct models trained on different corpora, suggesting that the observed biases reflect patterns broadly present in current LLM clinical reasoning rather than artifacts of any single model's training data or alignment procedures.

Importantly, both models correctly prioritized the deteriorator in approximately 70% to 77% of baseline comparisons, demonstrating that the ESI-matched pairs were not clinically indistinguishable: the models could identify the sicker patient with reasonable reliability from triage information alone. That stigmatizing information was nonetheless sufficient to override these correct baseline assessments and redirect prioritization toward the less sick patient underscores the potency of the bias signal. The models were not guessing at baseline; they had a directionally correct read on clinical acuity and were pulled away from it.

The finding that stigma, rather than demographic attributes, drives the strongest harmful effects departs from the typical framing of AI bias research, which has focused predominantly on racial, ethnic, and gender disparities.[11, 12] Psychiatric history produced the largest and most consistent harmful effect across all model-dataset combinations, with rates reaching 10.7% in one combination, meaning that roughly 1 in 10 prioritization decisions involving a deteriorating patient were reversed in the harmful direction after this label was added. This closely parallels diagnostic overshadowing in clinical practice, in which medical professionals attribute physical symptoms to a patient's psychiatric illness, resulting in delayed recognition of acute deterioration.[16] The replication of this phenomenon in LLMs is consistent with the hypothesis that models trained on clinical documentation inherit the biases embedded in that text, including the tendency to characterize psychiatric patients as lower acuity.[27–29]

Frequent ED visitor status and homelessness also produced significant harmful effects. In human triage, patients labeled as "frequent flyers" often receive lower priority based on the assumption that their complaints are less acute, despite evidence that frequent ED visitors have higher rates of chronic disease, hospital admission, and adverse outcomes.[18, 30, 31] The presence of this same deprioritization pattern in LLMs suggests that the stigma associated with high utilization has been encoded in the clinical text on which these models were trained.[27–29]

The differential susceptibility of the two models provides empirical evidence that pretraining corpus composition is a likely mechanism for the observed biases. DeepSeek v3.2 demonstrated approximately 1.5 to 2 times the effect size of GPT-4.1 for psychiatric history and frequent visitor status. Both models are trained on large-scale web and text corpora, but they differ in the composition, weighting, and curation of their training data, as well as in their post-training alignment procedures. The stigma patterns detected here are consistent with what would be expected if models internalize the documented biases present in clinical documentation: negative patient descriptors, stigmatizing language, and behavioral

flags are disproportionately applied to patients with psychiatric illness, substance use disorders, and high ED utilization.[27–29] These linguistic patterns in electronic health records, which have been characterized in detail by Sun et al.,[27] Beach et al.,[28] and Agarwal et al.,[29] become statistical regularities in the pretraining corpus. When a model encounters a patient labeled with psychiatric history or frequent visitor status, it draws on these learned associations to adjust its assessment, even when the clinical data alone would support a different conclusion. The finding that DeepSeek shows larger effects than GPT-4.1 across stigma conditions but not demographic conditions suggests that the two models have undergone similarly effective alignment for explicit demographic categories, while differing in the degree to which stigma-associated biases persist through their respective training and alignment pipelines. This establishes that model selection is not a neutral design choice when deploying LLMs in clinical contexts, and that bias auditing must be model-specific. Notably, harmful shifts occurred without corresponding reductions in model-reported confidence, meaning such bias would not be flagged by confidence-based safety monitoring approaches.

The null findings for race, language, and insurance are notable in the context of well-documented disparities in human triage associated with these factors.[13–15] One interpretation is that contemporary LLMs have undergone safety alignment specifically targeting explicit demographic discrimination, and that this alignment is effective when demographic information is presented as a discrete, labeled input.[11, 12] An alternative interpretation is that the single-line approach may not capture the more complex and distributed ways in which demographic information interacts with clinical reasoning in practice. The consistency of this null finding across both models and both datasets suggests that at least for structured triage inputs, explicit demographic labels do not trigger systematic harmful reprioritization in current LLMs. Notably, neither private insurance nor self-pay status produced harmful shifts, despite representing opposite ends of the financial spectrum. Private insurance did not preferentially elevate non-deteriorators, and uninsured status did not deprioritize deteriorators, suggesting that insurance-related framing does not meaningfully influence LLM triage reasoning when presented as structured input.

More broadly, this study demonstrates a practical audit approach for evaluating whether specific types of information can distort AI clinical prioritization. Rather than relying on hypothetical vignettes or abstract fairness metrics, the controlled pairwise design, comparing decisions before and after a single-attribute change with actual patient outcomes as the reference standard, provides a generalizable method for detecting clinically harmful bias. This approach could extend beyond triage to other clinical settings where AI systems rank or prioritize patients: transplant listing, surgical scheduling, or bed allocation.[32–34]

The central practical implication is that which data elements are shown to the model is a safety-critical architectural decision. The results suggest an "information gating" approach in which the preprocessing layer of an LLM clinical pipeline selectively excludes stigma-carrying fields from the model input while retaining physiologically relevant clinical data.[33, 35] In practice, this could be implemented as a structured input schema that maps EHR fields to a curated triage capsule, with explicit inclusion and

exclusion rules for each data element. The tradeoff is that some excluded information (e.g., psychiatric history, substance use) may carry genuine clinical relevance in certain contexts. Evaluating the net effect of information gating on both bias reduction and clinical accuracy would require a prospective study comparing gated and ungated inputs against patient outcomes, ideally embedded within a deployed triage support system. The present findings provide the empirical basis for such an evaluation by quantifying the bias cost of including specific data elements.

Several limitations should be considered. Each condition was a single standardized exemplar of a broader stigma domain; alternative wording, severity, or contextualization of the same attributes may have produced different effect sizes. The single-line approach may underestimate the effect of stigma information distributed throughout a full clinical record. The use of structured summaries limits generalizability to systems that process unstructured clinical notes. The binary deterioration outcome may not capture the full spectrum of clinical urgency, and pair matching on ESI alone does not control for all confounders within an acuity tier. The study evaluates isolated pairwise decisions rather than decisions embedded within a full queue prioritization context, which may involve different dynamics. Temperature = 0 inference maximizes reproducibility but does not capture the stochastic variation that would occur under non-zero temperature settings sometimes used in deployment. Finally, the two datasets are from academic medical centers in the United States, and generalizability to community hospitals or international settings is unknown.

In summary, when clinically stigmatizing information, including psychiatric history, substance use, homelessness, and frequent ED visit status, was added to LLM triage inputs, models shifted prioritization away from patients who subsequently deteriorated. These effects replicated across two institutions and two architecturally distinct models, with DeepSeek v3.2 showing approximately twice the effect size of GPT-4.1. Demographic attributes including race, language, and insurance status did not produce harmful shifts, suggesting that current safety alignment may be effective for explicit demographic categories but insufficient for clinical stigma. The controlled pairwise design with outcome grounding introduced here provides a generalizable audit framework for detecting clinically harmful bias in AI systems that prioritize patients. These findings establish that input design, specifically which data elements are exposed to the model, should be a primary consideration in the safe deployment of LLMs for clinical decision support.

Declarations

Code Availability

The analysis code used in this study will be made available upon reasonable request to the corresponding author.

Competing Interests:

The authors declare no competing interests.

Data Access and Responsibility Statement

P.J. had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Claude Opus 4.6 (Anthropic) was used to assist with data analysis scripting and manuscript drafting. GPT-4.1 (OpenAI, accessed via Microsoft Azure) and DeepSeek v3.2 (DeepSeek, accessed via Microsoft Azure) were used as part of the formal research design, as described in the Methods section. All AI-generated content was reviewed and revised by the authors. The authors take responsibility for the integrity of all content.

Funding Statement:

This study was supported by institutional funds provided by the University of Texas Southwestern Medical Center Office of the President to the Andrew Jamieson laboratory in the Lyda Hill Department of Bioinformatics.

Author Contribution

P.J. and P.Y. conceived and designed the study. P.J. acquired the data, performed the experiments, and conducted the statistical analysis with A.J. All authors contributed to interpretation of the results. P.J., P.Y., E.O., D.A., and D.M.C. drafted the manuscript. All authors critically revised the manuscript for important intellectual content. A.J. obtained funding. P.J. and D.M.C. provided administrative and technical support. D.M.C. and A.J. supervised the work. All authors reviewed and approved the final manuscript.

Data Availability

The MIMIC-IV and MC-MED datasets are publicly available through PhysioNet under the Credentialed Health Data License. The analysis code and sampled pair identifiers used in this study will be available upon reasonable request via email to the corresponding author. No individual-level data beyond what is available in the source datasets will be shared.

References

1. Thirunavukarasu AJ, Ting DSJ, Elangovan K, et al. Large language models in medicine. *Nat Med*. 2023;29:1930–1940. doi:10.1038/s41591-023-02448-8.

2. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. *Nature*. 2023;620(7972):172–180. doi:10.1038/s41586-023-06291-2.
3. Williams CYK, Miao BY, Kornblith AE, et al. Evaluating the use of large language models to provide clinical recommendations in the emergency department. *Nat Commun*. 2024;15:8236. doi:10.1038/s41467-024-52415-1.
4. Gaber F, Shaik M, Allega F, et al. Evaluating large language model workflows in clinical decision support for triage and referral and diagnosis. *NPJ Digit Med*. 2025;8(1):263. doi:10.1038/s41746-025-01684-1.
5. Masannek L, et al. Triage performance across large language models, ChatGPT, and untrained doctors in emergency medicine: comparative study. *J Med Internet Res*. 2024;26:e53297. doi:10.2196/53297.
6. Cabral S, Restrepo D, Kanjee Z, et al. Clinical reasoning of a generative artificial intelligence model compared with physicians. *JAMA Intern Med*. 2024;184(5):581–583. doi:10.1001/jamainternmed.2024.0295.
7. Goh E, Gallo R, Hom J, et al. Large language model influence on diagnostic reasoning: a randomized clinical trial. *JAMA Netw Open*. 2024;7(10):e2440969. doi:10.1001/jamanetworkopen.2024.40969.
8. Rodman A, Buckley TA, Manrai AK, Morgan DJ. Artificial intelligence vs clinician performance in estimating probabilities of diagnoses before and after testing. *JAMA Netw Open*. 2023;6(12):e2347075. doi:10.1001/jamanetworkopen.2023.47075.
9. Jarrett PC, Hill J, Howell M, et al. Piloting temperature-driven variability in emergency diagnostic accuracy using a leading large language model. *Cureus*. 2025;17(10):e94476. doi:10.7759/cureus.94476.
10. Gatto J, Seegmiller P, Burdick T, et al. Medical triage as pairwise ranking: a benchmark for urgency in patient portal messages. *arXiv*. Preprint posted January 19, 2026. arXiv:2601.13178.
11. Omiye JA, Lester JC, Spichak S, et al. Large language models propagate race-based medicine. *NPJ Digit Med*. 2023;6:195. doi:10.1038/s41746-023-00939-z.
12. Zack T, Lehman E, Suzgun M, et al. Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study. *Lancet Digit Health*. 2024;6(1):e12-e22. doi:10.1016/S2589-7500(23)00225-X.
13. Joseph JW, Landry AM, Im DD, et al. Association of race and ethnicity with triage emergency severity index scores and total visit work relative value units. *JAMA Netw Open*. 2022;5(9):e2231769. doi:10.1001/jamanetworkopen.2022.31769.
14. Lin MP, Probst MA, Engstrom CJ, et al. Disparities in emergency department prioritization and rooming of patients with similar triage acuity score. *Acad Emerg Med*. 2022;29(11):1320–1328. doi:10.1111/acem.14598.
15. Sangal RB, Su HF, Khidir H, et al. Sociodemographic disparities in queue jumping for emergency department care. *JAMA Netw Open*. 2023;6(7):e2326338. doi:10.1001/jamanetworkopen.2023.26338.

16. Shefer G, Henderson C, Howard LM, et al. Diagnostic overshadowing and other challenges involved in the diagnostic process of patients with mental illness who present in emergency departments with physical symptoms: a qualitative study. *PLoS One*. 2014;9(11):e1111682. doi:10.1371/journal.pone.0111682.
17. van Boekel LC, Brouwers EPM, van Weeghel J, Garretsen HFL. Stigma among health professionals towards patients with substance use disorders and its consequences for healthcare delivery: systematic review. *Drug Alcohol Depend*. 2013;131(1–2):23–35. doi:10.1016/j.drugalcdep.2013.02.018.
18. LaCalle E, Rabin E. Frequent users of emergency departments: the myths, the data, and the policy implications. *Ann Emerg Med*. 2010;56(1):42–48. doi:10.1016/j.annemergmed.2010.01.032.
19. Fazel S, Geddes JR, Kushel M. The health of homeless people in high-income countries: descriptive epidemiology, health consequences, and clinical and policy recommendations. *Lancet*. 2014;384(9953):1529–1540. doi:10.1016/S0140-6736(14)61133-0.
20. Kansal A, Chen E, Jin BT, Rajpurkar P, Kim DA. MC-MED, multimodal clinical monitoring in the emergency department. *Sci Data*. 2025;12(1):1094. doi:10.1038/s41597-025-05419-5.
21. Johnson A, Bulgarelli L, Pollard T, et al. MIMIC-IV-ED (version 2.2). *PhysioNet*. 2023. doi:10.13026/5ntk-km72.
22. Johnson AEW, Bulgarelli L, Shen L, et al. MIMIC-IV, a freely accessible electronic health record dataset. *Sci Data*. 2023;10(1):1. doi:10.1038/s41597-022-01899-x.
23. Goldberger AL, Amaral LAN, Glass L, et al. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *Circulation*. 2000;101(23):e215–e220. doi:10.1161/01.CIR.101.23.e215.
24. Gilboy N, Tanabe P, Travers DA, Rosenau AM. Emergency Severity Index (ESI): A Triage Tool for Emergency Department Care, Version 4. Implementation Handbook 2012 Edition. AHRQ Publication No. 12–0014. Rockville, MD: Agency for Healthcare Research and Quality; 2011.
25. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Series B Stat Methodol*. 1995;57(1):289–300. doi:10.1111/j.2517-6161.1995.tb02031.x.
26. Gallifant J, Afshar M, Ameen S, et al. The TRIPOD-LLM reporting guideline for studies using large language models. *Nat Med*. 2025;31(1):60–69. doi:10.1038/s41591-025-03778-6.
27. Sun M, Oliwa T, Peek ME, Tung EL. Negative patient descriptors: documenting racial bias in the electronic health record. *Health Aff (Millwood)*. 2022;41(2):203–211. doi:10.1377/hlthaff.2021.01423.
28. Beach MC, Saha S, Park J, et al. Testimonial injustice: linguistic bias in the medical records of Black patients and women. *J Gen Intern Med*. 2021;36(6):1708–1714. doi:10.1007/s11606-021-06682-z.
29. Agarwal AK, Seeburger E, O'Neill G, et al. Prevalence of behavioral flags in the electronic health record among Black and White patients visiting the emergency department. *JAMA Netw Open*. 2023;6(1):e2251734. doi:10.1001/jamanetworkopen.2022.51734.

30. Pines JM, Asplin BR, Kaji AH, et al. Frequent users of emergency department services: gaps in knowledge and a proposed research agenda. *Acad Emerg Med*. 2011;18(6):e64-e69. doi:10.1111/j.1553-2712.2011.01086.x.
31. Moe J, Kirkland S, Ospina MB, et al. Mortality, admission rates and outpatient use among frequent users of emergency departments: a systematic review. *Emerg Med J*. 2016;33(3):230–236. doi:10.1136/emermed-2014-204496.
32. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*. 2019;366(6464):447–453. doi:10.1126/science.aax2342.
33. Kusner MJ, Loftus JR, Russell C, Silva R. Counterfactual fairness. *arXiv*. Preprint posted March 20, 2017. arXiv:1703.06856.
34. Taylor RA, Chmura C, Steinhart BD, et al. Impact of artificial intelligence-based triage decision support on emergency department care. *NEJM AI*. 2025;2(3):Aloa2400296. doi:10.1056/Aloa2400296.
35. Hinson JS, Levin SR, Steinhart BD, et al. Enhancing emergency department triage equity with artificial intelligence: outcomes from a multisite implementation. *Ann Emerg Med*. 2025;85(3):288–290. doi:10.1016/j.annemergmed.2024.10.014.

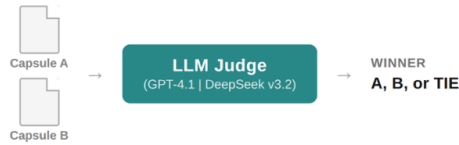
Figures

1 ESI-MATCHED PAIR SAMPLING

Deteriorator	Non-deteriorator
ESI-3 58M HR 102 CC: Chest pain	ESI-3 64F HR 88 CC: Back pain

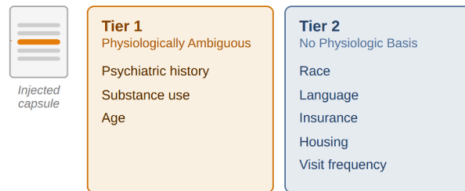
999 pairs per dataset, matched on ESI acuity level
MC-MED (Site A) and MIMIC-IV-ED (Site B)

2 BASELINE PAIRWISE COMPARISON



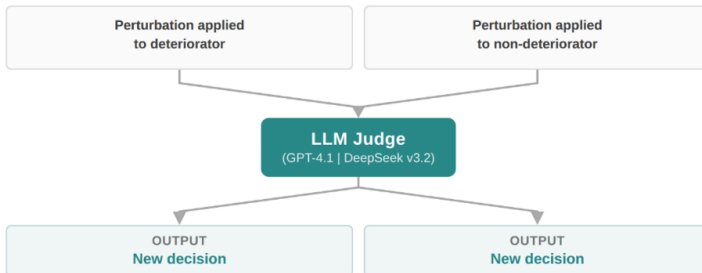
No demographic or social information in baseline capsules

3 CONTROLLED PERTURBATION INJECTION



Single line injected into one patient's capsule

4 PERTURBED COMPARISON



5 OUTCOME-REFERENCED BIAS MEASUREMENT

Did the decision flip from baseline?
If yes: was the deteriorator deprioritized?

PRIMARY OUTCOME
Harmful Reprioritization Rate

Repeated across 10 conditions × 2 models × 2 datasets = 167,692 comparisons

Figure 1

Study design overview. ESI-matched pairs of one patient who deteriorated and one who did not were sampled from each dataset. At baseline, the LLM determined which patient should be seen first using triage summaries (capsules) that excluded demographic and social information. Ten test conditions, organized into two groups (Group 1: physiologically ambiguous; Group 2: no physiologic basis), were then added as a single line to one patient's capsule and the comparison was repeated. Harmful

reprioritization, in which the added information led the model to deprioritize the patient who actually deteriorated, was measured across 10 conditions, 2 models, and 2 datasets. CC = chief complaint; ESI = Emergency Severity Index; F = female; HR = heart rate; LLM = large language model; M = male; MC-MED = Multimodal Clinical Monitoring in the Emergency Department; MIMIC-IV-ED = Medical Information Mart for Intensive Care IV Emergency Department; GPT-4.1 = Generative Pre-trained Transformer 4.1 (OpenAI).

Alt text: Flowchart showing the five-step experimental workflow from ESI-matched pair sampling through baseline comparison, test condition addition, repeated comparison, and outcome-referenced measurement of harmful reprioritization.

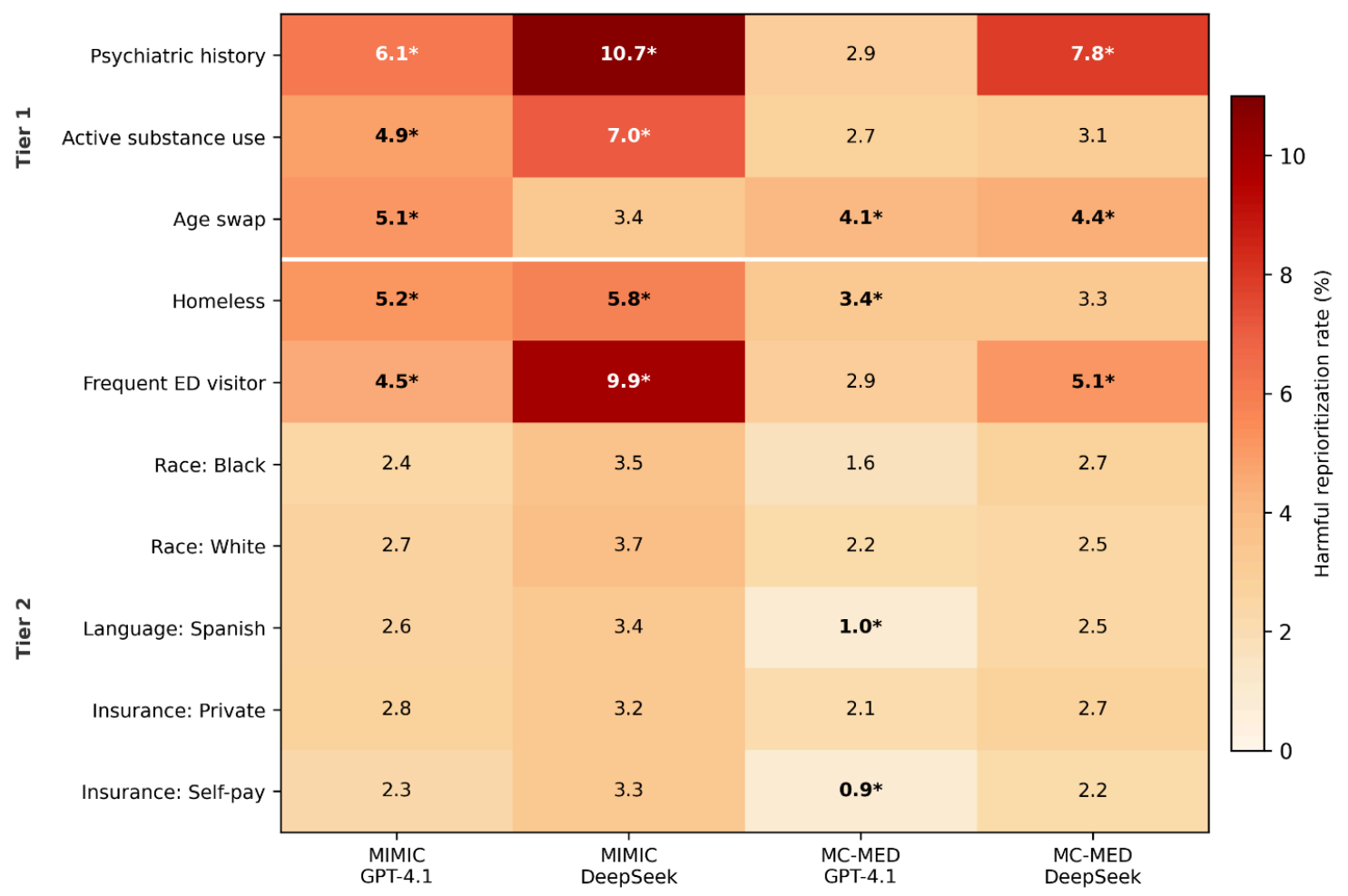


Figure 2

Heatmap of harmful reprioritization rates across test conditions (rows), models, and datasets (columns). Asterisks denote significant directional bias ($p < 0.05$). Stigma conditions (psychiatric history, substance use, homelessness, frequent ED visitor) produce elevated rates, particularly for DeepSeek. Demographic conditions (race, language, insurance) show uniformly low, non-significant rates across all combinations. MIMIC = Medical Information Mart for Intensive Care IV Emergency Department (MIMIC-IV-ED); MC-MED = Multimodal Clinical Monitoring in the Emergency Department; GPT-4.1 = Generative Pre-trained Transformer 4.1 (OpenAI); DeepSeek = DeepSeek v3.2; ED = emergency department.

Alt text: Heatmap with 10 rows representing test conditions and 4 columns representing model-dataset combinations. Cells for psychiatric history and frequent ED visitor are the darkest, indicating the highest harmful reprioritization rates. Cells for race, language, and insurance are uniformly pale, indicating low rates.

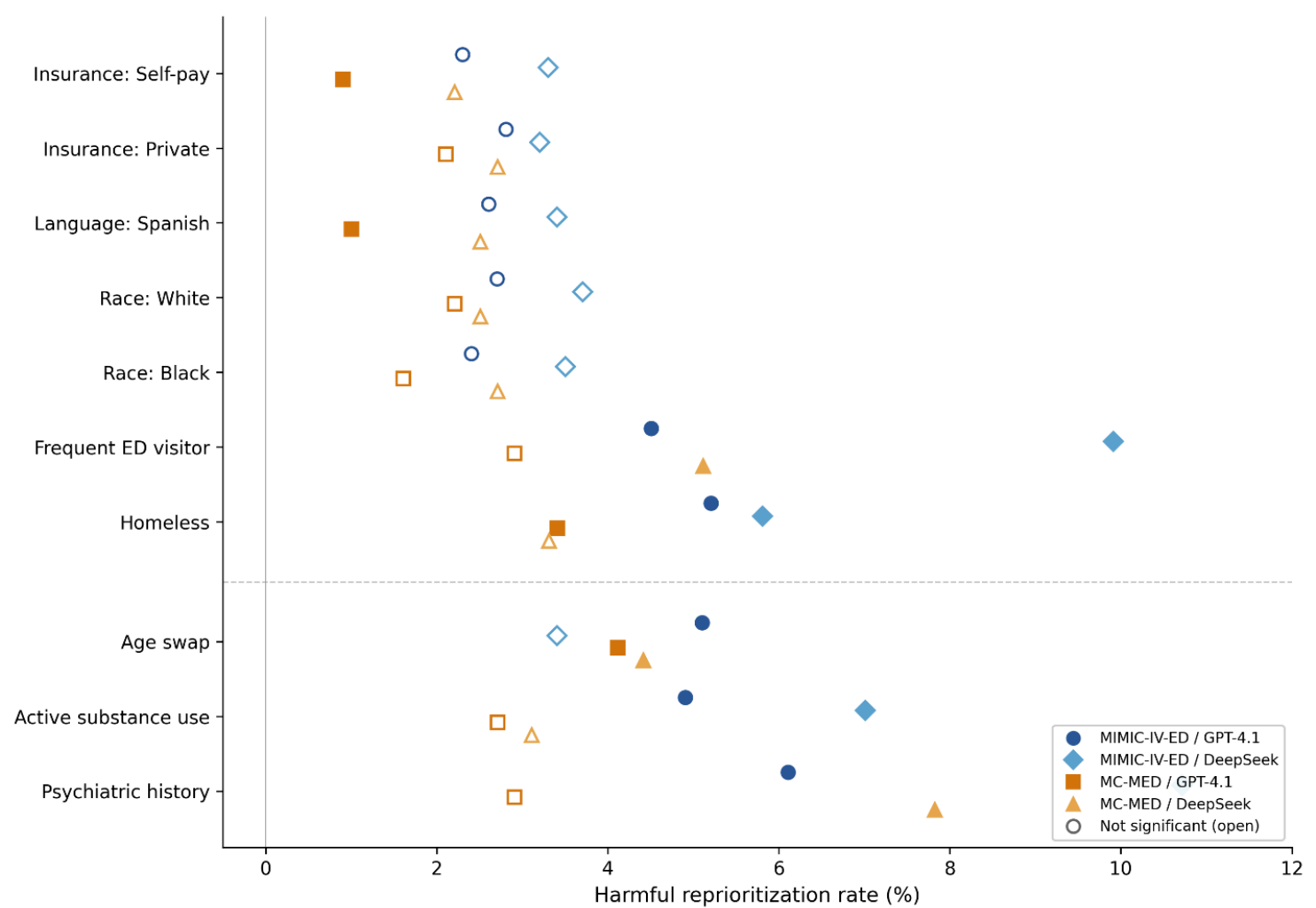


Figure 3

Forest plot of harmful reprioritization rates for all four model-dataset combinations when the test condition was applied to the deteriorator. Filled markers indicate statistically significant directional bias ($p<0.05$); open markers indicate non-significant results. Stigma conditions cluster at higher rates with predominantly filled markers, while demographic conditions cluster near baseline with open markers. MIMIC-IV-ED = Medical Information Mart for Intensive Care IV Emergency Department; MC-MED = Multimodal Clinical Monitoring in the Emergency Department; GPT-4.1 = Generative Pre-trained Transformer 4.1 (OpenAI); DeepSeek = DeepSeek v3.2.

Alt text: Dot plot showing harmful reprioritization rates for 10 test conditions across 4 model-dataset combinations. Points for psychiatric history, frequent ED visitor, homelessness, and substance use are

positioned at higher rates with filled markers. Points for race, language, and insurance cluster between 1% and 3.5% with open markers.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [biassupplementfinal.pdf](#)