

An Open, Reproducible Implementation of the CMS Hospital-Wide Readmission Measure on MIMIC-IV

Sai Sharan Kasa

sk.sharankasa@gmail.com

Independent Researcher <https://orcid.org/0009-0002-7425-4516>

Research Article

Keywords: hospital readmission, CMS Hospital-Wide Readmission, MIMIC-IV, clinical prediction models, model calibration, reproducibility, health informatics

Posted Date: October 22nd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10319770/v2>

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

[Read Full License](#)

Additional Declarations: The authors declare potential competing interests as follows: The author is employed by Amazon. This work was conducted independently, on personal time and with personal resources, using publicly available credentialed data; the employer had no role in the study and provided no funding, data, or resources.

An Open, Reproducible Implementation of the CMS Hospital-Wide Readmission Measure on MIMIC-IV

Author: Sai Sharan Kasa (independent researcher) **Correspondence:**
sk.sharankasa@gmail.com · **ORCID:** 0009-0002-7425-4516 **Code:**
<https://github.com/sharankasa/readmission-risk-prediction>

Abstract

Thirty-day hospital readmission is one of the most studied prediction targets in clinical machine learning, yet the overwhelming majority of MIMIC-based studies predict a *naive* label: any readmission within 30 days. The measure Medicare actually uses to grade hospitals on all-condition readmission, the CMS Hospital-Wide Readmission (HWR) measure, is materially different. It counts only *unplanned* readmissions, applies a published planned-readmission algorithm, stratifies patients into specialty cohorts, and removes a defined set of ineligible admissions. We present an open, versioned, end-to-end implementation of the HWR measure on MIMIC-IV, and quantify how much the target shifts when it is defined faithfully: the 30-day event rate falls from **17.6% (naive) to 15.9% (faithful)** on the same index admissions ($n = 192,879$). We release the cohort logic, the encoded planned-readmission code lists (CMS Planned Readmission Algorithm v4.0), a Quan-Elixhauser comorbidity encoder, and a leakage-guarded feature specification, together with a calibrated baseline model (ROC-AUC 0.730, well-calibrated, decision-curve positive) evaluated for subgroup fairness and temporal drift. In an ablation we add discharge-summary embeddings (Amazon Titan) to the structured baseline and find no meaningful uplift (+0.001 AUC), a leakage-controlled negative result that is itself informative. The contribution is not the classifier; it is a reusable, reproducible reference implementation of a clinically meaningful label that other researchers can adopt instead of the naive one.

1. Background and motivation

Reducing avoidable readmissions is a long-standing policy goal. The Hospital Readmissions Reduction Program attaches Medicare payment penalties to a set of condition-specific readmission measures (acute myocardial infarction, heart failure, pneumonia, COPD, elective hip/knee replacement, and CABG). Separately, the Hospital-Wide Readmission (HWR) measure, developed by the Yale/CMS team (CORE), is the broad all-condition instrument Medicare uses for public reporting and hospital star ratings; it shares the same planned-readmission machinery as the penalty measures. The measure is deliberately specific about what counts as a readmission event, and that specificity is exactly what most machine-learning studies discard.

Three design choices distinguish the HWR measure from a naive “any 30-day readmission” label:

1. **Planned readmissions are not events.** A scheduled chemotherapy cycle, a staged procedure, or a planned transplant admission is a *planned* return and does not count against the hospital. CMS defines this with an explicit algorithm (a list of always-planned procedures and diagnoses, plus a set of acute diagnoses that disqualify an otherwise-planned admission).
2. **Patients are stratified into specialty cohorts.** Each index admission is assigned to one of five mutually exclusive cohorts (medicine, surgery/gynecology, cardiorespiratory, cardiovascular, neurology) based on the discharge service and procedures, because baseline risk differs sharply across them.
3. **A defined set of admissions is ineligible.** Admissions with in-hospital death, discharge against medical advice, certain transfers, and other categories are removed from the index pool.

A model trained on the naive label is answering a different, easier, and less policy-relevant question than the one hospitals are measured on. Our aim is to make the faithful target as easy to adopt as the naive one, by releasing the implementation rather than only describing it.

Contributions.

- An open, reproducible HWR cohort builder on MIMIC-IV: index-admission selection, the planned readmission algorithm (v4.0), specialty-cohort assignment, and the standard exclusions, expressed as auditable SQL with an accompanying CONSORT diagram.
 - A quantification of the label shift: **17.6% naive vs 15.9% faithful** on identical admissions, with the reclassified admissions attributable to the planned-readmission logic.
 - Reusable reference assets: the encoded planned-readmission code lists, a 31-flag Quan-Elixhauser comorbidity encoder, and a point-in-time, leakage-guarded feature specification.
 - A calibrated baseline (discrimination, calibration, decision-curve analysis, SHAP, subgroup fairness, temporal drift) that others can use as a reference point, not a leaderboard entry.
 - A leakage-controlled ablation showing that discharge-note embeddings add no meaningful signal over the structured baseline, a negative result reported straight.
-

2. Data

We use **MIMIC-IV v3.1** (the hosp and icu modules) and **MIMIC-IV-Note v2.2** (discharge summaries), both distributed by PhysioNet under credentialed access and a Data Use Agreement. No row-level patient data is redistributed; the repository releases only code, aggregate results, and public reference code lists. Access requires independent PhysioNet credentialing.

We restrict the index pool to the **ICD-10 era** so that diagnosis- and procedure-based logic (the planned-readmission algorithm, comorbidity encoding, specialty assignment) is applied against a single, internally consistent coding system rather than mixing ICD-9 and ICD-10. This yields **192,879 index admissions**.

3. Methods

3.1 Cohort construction and the CONSORT waterfall

Beginning from all hospitalizations in MIMIC-IV (546,028), we first restrict to the ICD-10 era, which removes 291,651 pre-ICD-10 admissions and leaves 254,377. We then apply the HWR exclusions, each recorded as the first failing reason in an audit column so the full waterfall regenerates from a single query: the measure’s separately-reported cohort exclusions for medical treatment of cancer (16,897), psychiatric care (13,962), and rehabilitation (20); maternity (12,024); death during the index stay (6,301) and discharge as deceased (121); hospice (3,558); death within 30 days of discharge (3,144); admissions not assignable to one of the five HWR specialty cohorts (2,490); discharge against medical advice (1,656); and transfer to another acute facility at discharge (1,325). The result is a final cohort of **192,879 index admissions across 99,285 unique patients**; the full reduction is shown in Fig. 1. Each surviving admission is assigned to exactly one of the five specialty cohorts.

Two of these exclusions concern death rather than diagnosis and warrant a note on fidelity. The claims-based measure requires beneficiaries to be discharged alive and to retain at least 30 days of post-discharge enrollment for the 30-day outcome to be observable [2]; because MIMIC-IV has no enrollment concept, we operationalize the latter as survival through the 30-day post-discharge window, which is why the 3,144 admissions with a death within 30 days of discharge are removed rather than counted as non-readmissions. This mirrors the measure’s observability requirement, treating post-discharge death as a competing risk resolved by exclusion exactly as the specification does, rather than departing from it.

The distribution across specialty cohorts is reported in Table 1; the medicine cohort is the largest (106,148 admissions) and carries the highest observed 30-day readmission rate (17.6%).

Table 1. Specialty-cohort distribution and faithful 30-day unplanned readmission rate (n = 192,879).

Specialty cohort

- | |
|--------------------|
| Medicine |
| Surgery/Gynecology |
| Cardiovascular |
| Neurology |
| Cardiorespiratory |

All cohorts

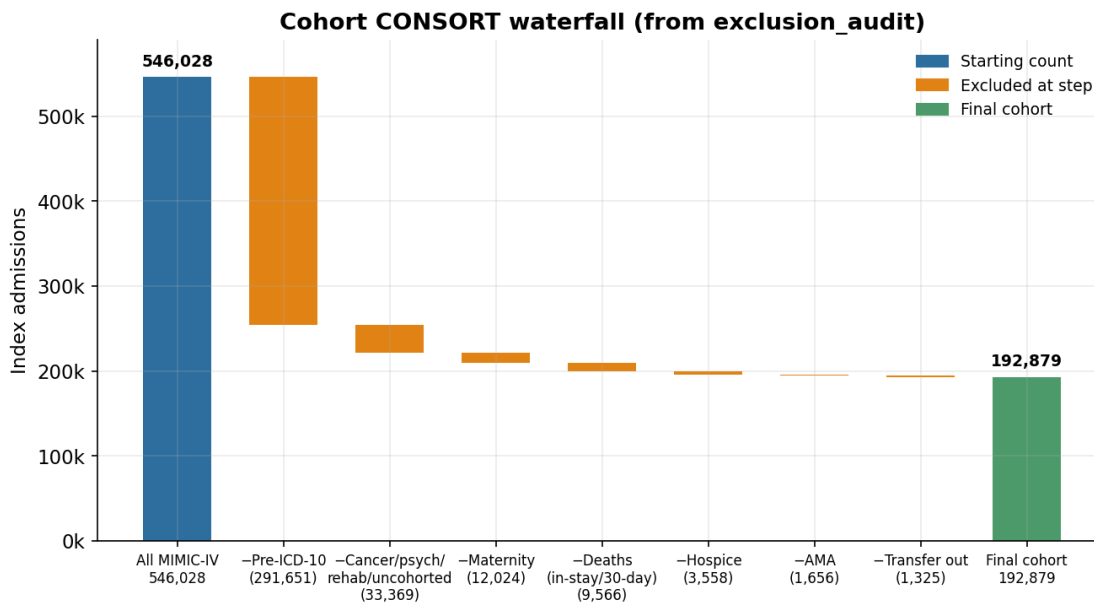


Figure 1. Cohort CONSORT waterfall, regenerated from the `exclusion_audit` column: 546,028 MIMIC-IV hospitalizations reduced to the 192,879 final index admissions.

3.2 Label definition: faithful vs naive

For each index admission we look forward 30 days from discharge for a subsequent inpatient admission. Under the **naive** definition, any such admission is a positive. Under the **faithful** definition, an admission is a positive only if it is *unplanned*: we apply the CMS Planned Readmission Algorithm v4.0, classifying a candidate readmission as planned when it contains an always-planned procedure or diagnosis and is not overridden by an acute, disqualifying diagnosis. Planned readmissions are set to non-events (they remain in the denominator as index stays but do not count as outcomes).

On the identical 192,879 index admissions, the positive rate is **17.6% under the naive definition and 15.9% (prevalence 0.1588) under the faithful definition**. The gap is the direct, quantified cost of using the wrong label, and the reclassified admissions are exactly those the planned- readmission logic identifies.

3.3 Feature engineering

Features are computed strictly from information available at or before discharge from the index stay (point-in-time correctness), with explicit leakage guards: nothing from the outcome window, and the ED prelude of the index encounter is included so that short-stay ED labs are not lost. The specification produces a **252-column** wide matrix across eight feature groups:

1. **Administrative / demographic / prior-utilization** (age, sex, admission characteristics, prior admissions, cumulative prior length of stay, days since last discharge).

2. **Laboratory** (a 23-analyte panel, emitted long and pivoted; present for 84.6% of admissions).
3. **Vitals from outpatient measurements (OMR)** (BMI, blood pressure, weight).
4. **ICU vitals** (chartevents-derived, present for the 16.8% of admissions with an ICU stay).
5. **Medications.**
6. **Procedures.**
7. **Specialty cohort** (the five HWR service lines).
8. **Comorbidity** (31 Quan-Elixhauser flags plus a count; hypertension 59%, diabetes 29%, heart failure 20%, renal 20%; mean 3.4 comorbidities per admission).

The comorbidity encoder (a released, versioned ICD-10 to Quan-Elixhauser map) and the labs panel are provided as reference assets so the feature build is reproducible without hand-curation.

3.4 Model

The baseline is a **HistGradientBoostingClassifier** (scikit-learn) with categorical features ordinal-encoded and passed to the model's native categorical support. Probabilities are calibrated with **isotonic regression**. All evaluation uses **5-fold GroupKFold grouped on subject_id**, so that no patient contributes admissions to both training and test folds. Model choice and the calibration decision are recorded as Architecture Decision Records in the repository.

3.5 Evaluation

We report discrimination (ROC-AUC, PR-AUC against the 15.9% prevalence), calibration (Brier score; reliability curve), and clinical net benefit via **decision-curve analysis**. We interpret the model with **SHAP**, run a single-feature **leakage tripwire** (no individual feature should reach near-perfect AUC), evaluate **subgroup fairness** across race, insurance, and specialty cohort, and assess **temporal drift** with the Population Stability Index (PSI) across admission-year windows.

3.6 GenAI ablation

To test whether unstructured clinical text adds signal over the structured baseline, we embed each index admission's discharge summary with **Amazon Titan Text Embeddings v2** (1024 dimensions), reduce the embeddings with unsupervised PCA (1024 to 50 components; no label, hence no leakage), and merge them into the feature matrix for the 99,314 admissions that have a discharge summary. We re-run the identical GroupKFold, calibrated pipeline with and without the note features, on both the full cohort and the note-having subset, and apply the same single-feature leakage tripwire to the note dimensions.

4. Results

4.1 The label shift (primary resource result)

On 192,879 index admissions, faithful HWR labeling yields a 30-day unplanned readmission rate of **15.9%**, versus **17.6%** under naive labeling (Fig. 2). Everything downstream (class balance, calibration targets, subgroup rates) inherits this shift, which is why adopting the faithful label matters even before any modeling.

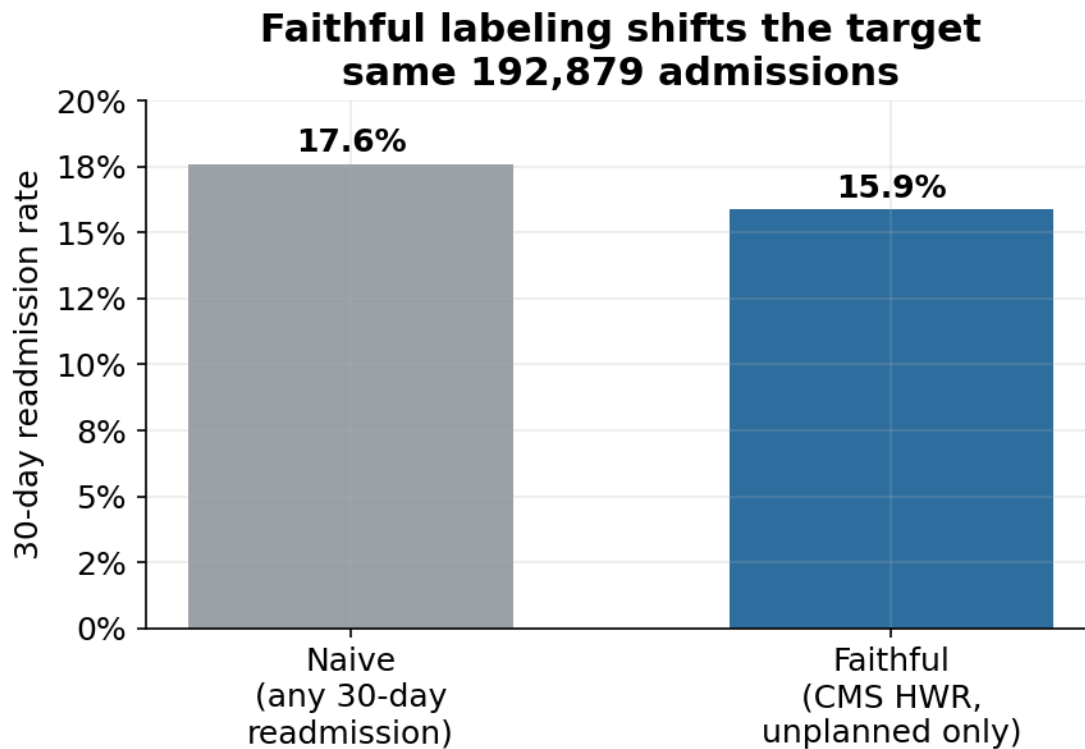


Figure 2. Faithful CMS HWR labeling versus a naive “any 30-day readmission” label on the same 192,879 index admissions: 15.9% versus 17.6%.

4.2 Discrimination and calibration

The calibrated baseline achieves **ROC-AUC 0.7305 (5-fold mean 0.7305, SD 0.0024)** and **PR-AUC 0.3726** at a prevalence of 0.159. The model is already well-calibrated before adjustment (**Brier 0.1186 raw vs 0.1187 after isotonic calibration**), consistent with gradient-boosted trees on a large, well-populated feature matrix; isotonic calibration is retained for guaranteed reliability without harming the score, and the reliability curve confirms good calibration (Fig. 3). Decision-curve analysis shows positive net benefit across the clinically relevant threshold range (Fig. 4).

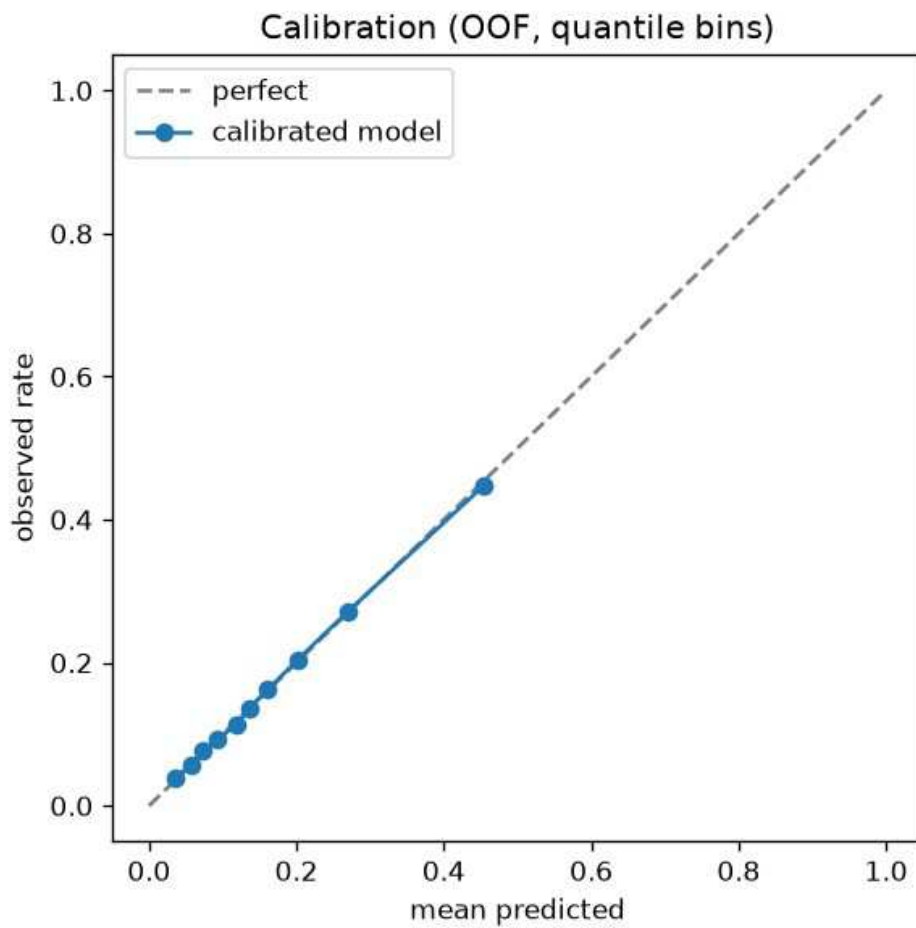


Figure 3. Calibration (reliability) curve for the isotonic-calibrated model.

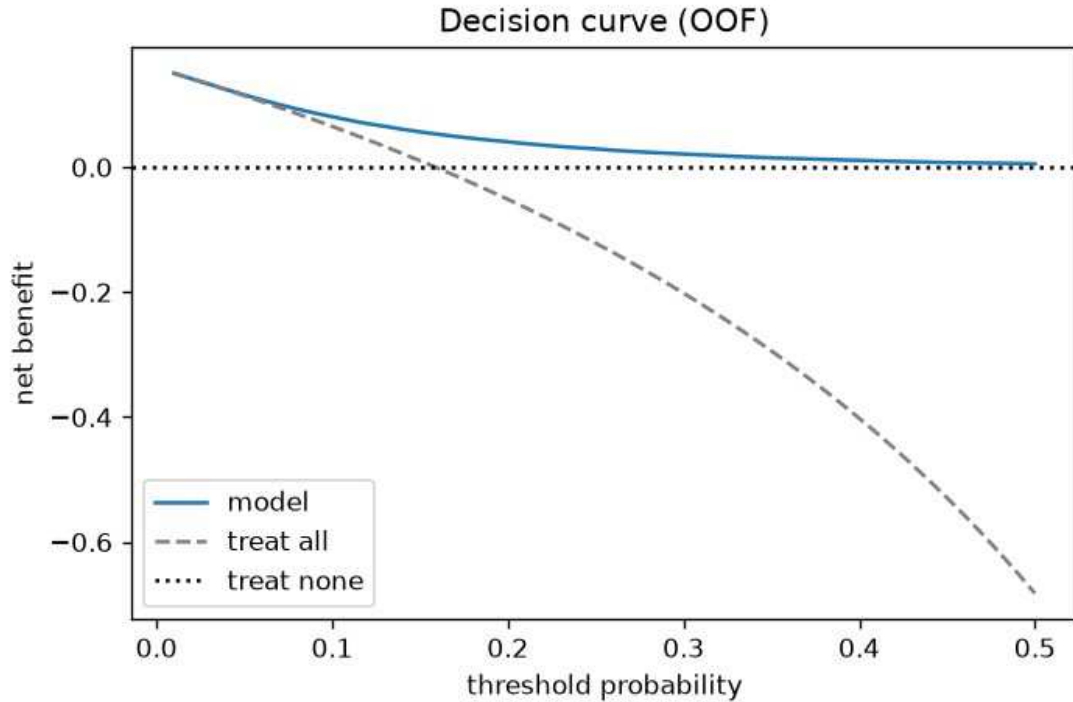


Figure 4. Decision-curve analysis: net benefit across the clinically relevant threshold range.

4.3 What drives the prediction

SHAP attributes the model largely to **prior-utilization and acuity** features (Fig. 5): the top drivers are cumulative prior length of stay, days since last discharge, and prior admission count, followed by acuity indicators, anticoagulant exposure, and red-cell distribution width (RDW). Several **demographic and administrative attributes also rank highly**, notably sex (the second-strongest feature), marital status, and insurance; we examine whether the model depends on these protected attributes, and the fairness implications, in Section 4.4. Labs and vitals contribute at the margin; note text adds essentially nothing on top (Section 4.5). This concentration of signal in prior-utilization and a handful of demographic features is the mechanistic explanation for the null GenAI result and a caution for anyone expecting rich text or physiology to dominate.

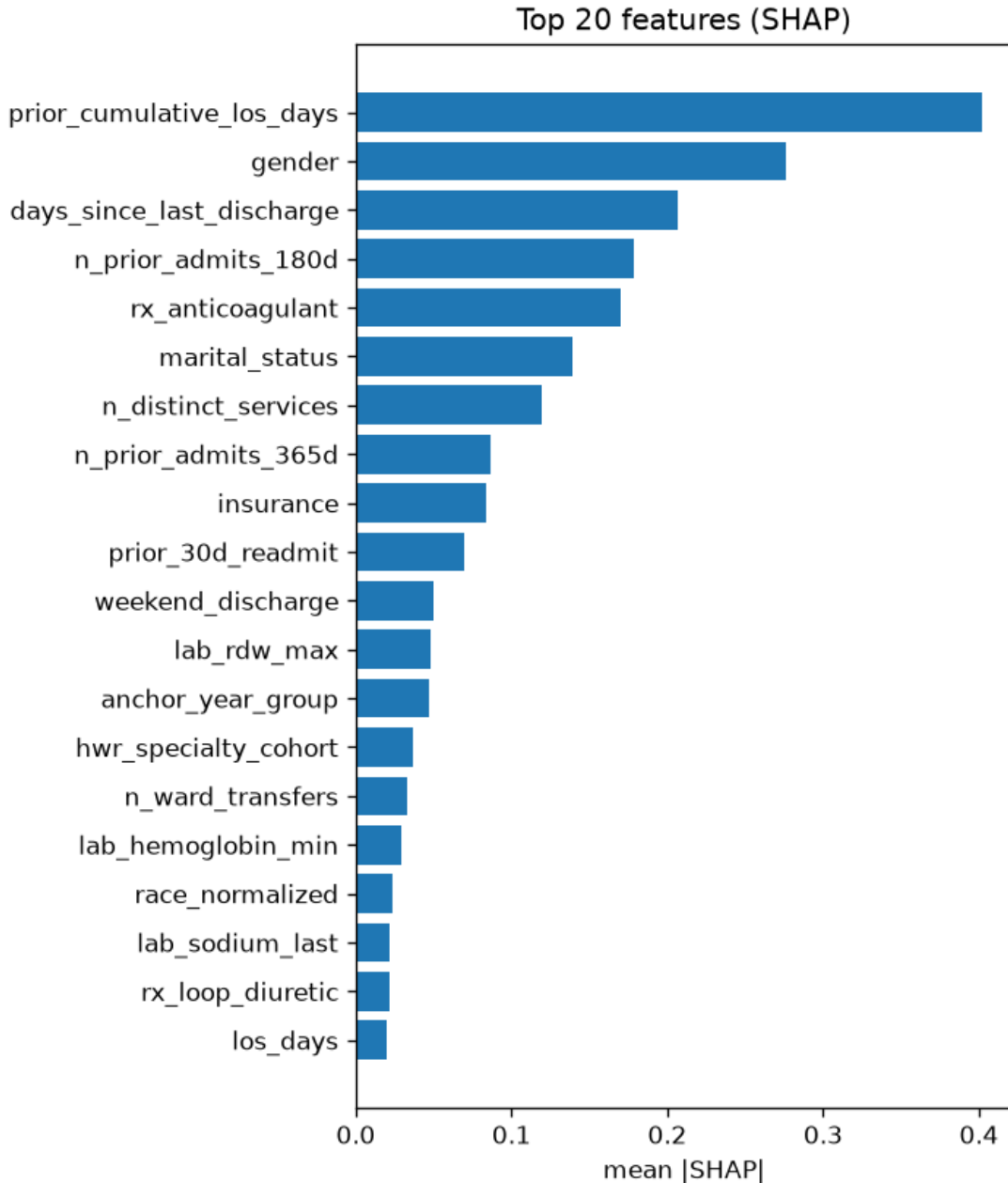


Figure 5. Global feature importance as mean absolute SHAP value per feature (average contribution magnitude to the prediction; direction and causation are not shown). Prior-utilization features lead, with sex, marital status, and insurance also among the top features.

4.4 Subgroup fairness and temporal drift

Discrimination is stable across demographic and clinical subgroups (Table 2): ROC-AUC ranges from about 0.69 to 0.75 across sex, race, insurance, and specialty-cohort strata. The only departures are in two small or degenerate strata, the $n = 88$ “no charge” insurance group and the “unknown” race group (whose 4.0% observed rate reflects missing race coding rather than a clinical subgroup). Several administrative and demographic attributes

(sex, marital status, insurance) rank highly in the SHAP analysis (Section 4.3); because a model that leans on protected attributes raises fairness concerns, a protected-attribute in/out ablation (retraining with and without those attributes to quantify the performance cost of removing them) is a natural next step, which we flag as a priority for follow-up rather than claim here.

Table 2. Subgroup performance: count, observed faithful 30-day readmission rate, and ROC-AUC.

Axis
Sex
Sex
Race
Race
Race
Race
Race
Race
Insurance
Insurance
Insurance
Insurance
Insurance
Specialty
Specialty
Specialty
Specialty
Specialty

Temporal PSI comparing the 2020-2022 window against earlier admissions flags **11 features with significant drift and 4 with moderate drift**, and a **prediction-distribution PSI of 0.19**, which crosses our retraining threshold, unsurprising given the COVID-era disruption to admission patterns. The monitor and its retraining trigger are released with the code.

4.5 GenAI ablation: a leakage-controlled null

Adding discharge-summary embeddings to the structured baseline produced **no meaningful uplift**. On the full cohort, ROC-AUC moved from **0.7304 to 0.7312 (+0.0008)**; on the note-having subset (n = 99,314), from **0.7122 to 0.7143 (+0.0021)**. Both deltas sit within fold-to-fold noise (SD 0.0024). (The uplift comparison is an independent re-run of the pipeline, so the structured baseline reads 0.7304 here versus 0.7305 in Section 4.2, a

difference well within that noise.) No note dimension tripped the leakage tripwire. The interpretation is not that clinical text is worthless in general, but that *for this target, against a strong calibrated structured baseline built largely on prior-utilization signal, discharge-summary embeddings are redundant*. Reported straight, this is a useful counterweight to the common assumption that adding an LLM embedding will move a clinical baseline.

5. Discussion

The recurring problem this work addresses is not modeling capacity but **construct validity**: a large literature optimizes a label that is not the one hospitals are accountable for. By releasing a faithful, auditable implementation, we lower the cost of doing the harder, correct thing to roughly the cost of doing the easy, wrong thing. The 1.7-percentage-point label shift is modest in absolute terms but changes the object of study, and the reusable assets (planned-readmission code lists, comorbidity encoder, feature spec) remove the main friction that pushes researchers toward the naive label in the first place.

The negative GenAI result and the SHAP analysis together tell a coherent story: prior utilization and acuity carry the signal, and additional modalities (labs, ICU vitals, note text) add little on top. That is worth knowing before investing in expensive text or multimodal pipelines for this task.

6. Limitations

- **Single-center, retrospective.** MIMIC-IV is one academic medical center (BIDMC). We make no external-validity or deployment claims; this is a methods and reproducibility contribution.
 - **Fidelity, not certification.** We implement the published HWR logic faithfully, but this is not a CMS-certified calculation and is not risk-standardized for hospital comparison; it adapts the measure's event definition to a research cohort.
 - **ICD-10 restriction.** Limiting to the ICD-10 era keeps coding consistent but reduces the usable sample and excludes earlier admissions.
 - **Note coverage.** Only about half of index admissions have a discharge summary, so the note ablation's full-cohort result partly reflects coverage, which we address by also reporting the note-having subset.
 - **Modest discrimination.** ROC-AUC around 0.73 is typical for all-condition readmission and reflects the intrinsic difficulty of the target, not a tuning deficit.
-

7. Reproducibility, ethics, and data availability

All code is released at the repository above under an open license, parameterized so no account-specific identifiers are embedded: the cohort SQL (index selection, the CMS Planned Readmission Algorithm v4.0, specialty cohorts, and exclusions), the feature-engineering SQL, the encoded reference code lists, the point-in-time feature specification,

model training, serving, deployment, drift monitoring, and the GenAI ablation. **No patient-level data is included**; the repository gitignores all data and derived patient artifacts and ships only code, aggregate results, and public CMS/AHRQ/Yale reference code lists. Reproduction requires the reader's own PhysioNet credentialing for MIMIC-IV and MIMIC-IV-Note and their own cloud resources; a runbook is provided. This use, deriving and publishing aggregate results and code from credentialed data, is the intended use under the PhysioNet Data Use Agreement. Per journal norms, the data-availability statement points to PhysioNet credentialed access rather than a direct data deposit.

Ethics. MIMIC-IV and MIMIC-IV-Note were collected under Institutional Review Board approval at the Beth Israel Deaconess Medical Center and MIT, with a waiver of informed consent; this work is a secondary analysis of that de-identified data and required no additional IRB review.

This work was conducted independently, on personal infrastructure and personal time, using only public credentialed data, with no employer resources, data, or affiliation.

References

1. Horwitz LI, Partovian C, Lin Z, et al. Development and use of an administrative claims measure for profiling hospital-wide performance on 30-day unplanned readmission. *Ann Intern Med.* 2014;161(10 Suppl):S66-S75. doi:10.7326/M13-3000.
2. Yale New Haven Health Services Corporation / Center for Outcomes Research and Evaluation (YNHHSC/CORE). Hospital-Wide All-Cause Unplanned Readmission Measure: Updates and Specifications Report, incorporating the Planned Readmission Algorithm version 4.0. Centers for Medicare & Medicaid Services (QualityNet), 2021.
3. Johnson A, Bulgarelli L, Pollard T, Gow B, Moody B, Horng S, Celi LA, Mark R. MIMIC-IV (version 3.1). PhysioNet; 2024. doi:10.13026/kpb9-mt58. Companion paper: Johnson AEW, Bulgarelli L, Shen L, et al. MIMIC-IV, a freely accessible electronic health record dataset. *Sci Data.* 2023;10:1. doi:10.1038/s41597-022-01899-x.
4. Johnson A, Pollard T, Horng S, Celi LA, Mark R. MIMIC-IV-Note: Deidentified free-text clinical notes (version 2.2). PhysioNet; 2023. doi:10.13026/1n74-ne17.
5. Quan H, Sundararajan V, Halfon P, et al. Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data. *Med Care.* 2005;43(11):1130-1139. PMID 16224307.
6. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems (NeurIPS)*; 2017.
7. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Making.* 2006;26(6):565-574.
8. Amazon Web Services. Amazon Titan Text Embeddings models. Amazon Bedrock User Guide; 2024.
9. 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. (PhysioNet resource citation.)