Recalibrating Risk: A CMS-Aligned Pipeline for 30-Day Heart Failure Readmissions

Molly Maskrey
mollymaskrey@gmail.com

June 28, 2025

Abstract

Background: Heart failure (HF) readmissions represent a significant clinical and economic burden, with 30-day readmission rates serving as a key quality metric. However, systematic changes in ICD-10 coding practices between 2016-2017 have created methodological challenges for longitudinal studies using administrative data.

Objective: To develop and validate machine learning models for predicting 30-day heart failure readmissions while addressing the impact of ICD-10 coding migration on case identification across multiple years of data.

Methods: We analyzed 3,695,822 heart failure-related hospitalizations from the Healthcare Cost and Utilization Project (HCUP) Nationwide Readmissions Database (2016-2022). A hybrid methodology was employed: I50.x primary diagnosis codes for 2016 data, and I11.x/I13.x primary with I50.x secondary diagnosis codes for 2017-2022 data to account for systematic coding shifts. Multiple machine learning approaches were evaluated, including XGBoost and stacked ensemble methods, with Synthetic Minority Oversampling Technique (SMOTE) for class imbalance handling.

Results: The hybrid coding methodology identified a substantial shift in heart failure coding patterns post-2016, with I50.x primary diagnoses declining while I11.x/I13.x codes with secondary I50.x increased dramatically. Our optimized XGBoost model achieved an AUC of 0.596 with decision threshold analysis revealing clinically actionable operating points: 99.4% sensitivity at default threshold, 72.7% sensitivity with 40.4% specificity at optimized threshold (0.80), and 90.8% sensitivity with 19.9% positive predictive value at liberal threshold (0.75). Key predictive features included chronic kidney disease (16.5% importance), chronic obstructive pulmonary disease (12.1%), and ventilator support (11.5%).

Conclusions: Systematic ICD-10 coding changes significantly impact heart failure case identification in administrative data. The hybrid methodology presented addresses this bias while maintaining model performance. Clinical features related to comorbidity burden and procedural intensity emerge as primary drivers of readmission risk. Decision threshold optimization enables practical deployment balancing sensitivity and resource utilization.

Keywords: Heart failure, readmission prediction, machine learning, ICD-10 coding, administrative data, SMOTE, decision thresholds

1 Introduction

Heart failure (HF) affects over 6 million adults in the United States and represents one of the leading causes of hospitalization among Medicare beneficiaries(5). The 30-day readmission rate for heart failure patients has become a critical quality metric, directly impacting hospital reimbursement through the Centers for Medicare and Medicaid Services (CMS) Hospital Readmissions Reduction Program(6). Despite significant clinical and policy attention, 30-day readmission rates for heart failure remain persistently high, ranging from 20-25% nationally(7).

The complexity of heart failure as a clinical syndrome, combined with its association with multiple comorbidities and social determinants of health, has made accurate prediction of readmission risk a challenging yet crucial endeavor(8). Traditional risk prediction models have

shown modest performance, with most achieving area under the curve (AUC) values between 0.55-0.65(9). Recent advances in machine learning techniques, particularly ensemble methods and techniques for handling class imbalance, offer potential improvements in predictive accuracy.

However, a critical methodological challenge has emerged in longitudinal studies of heart failure using administrative data: systematic changes in ICD-10 coding practices. Beginning in 2017, clinical coding guidelines were modified regarding the relationship between hypertension and heart disease, eliminating the requirement for explicit documentation of causal relationships(10). This change led to a systematic shift where heart failure cases that would previously be coded with I50.x as the primary diagnosis began appearing with hypertensive heart disease codes (I11.x) or hypertensive heart and chronic kidney disease codes (I13.x) as primary diagnoses, with I50.x relegated to secondary positions.

1.1 The ICD-10 Coding Migration Problem

Our preliminary analysis of the Healthcare Cost and Utilization Project (HCUP) Nationwide Readmissions Database revealed a dramatic shift in heart failure coding patterns between 2016 and subsequent years. Traditional approaches that rely solely on I50.x primary diagnosis codes would systematically undercount heart failure cases post-2016, introducing significant bias in longitudinal studies and potentially compromising the validity of predictive models trained on multi-year datasets.

This coding migration phenomenon has several implications:

1. **Case Identification Bias:** Studies using traditional I50.x primary diagnosis criteria may miss a substantial proportion of heart failure cases post-2016
2. **Temporal Inconsistency:** Longitudinal analyses may incorrectly interpret coding changes as epidemiological trends
3. **Model Training Bias:** Machine learning models trained on multi-year data without accounting for coding changes may learn spurious temporal patterns rather than clinical relationships

1.2 Study Objectives

This study addresses these methodological challenges while developing robust machine learning approaches for heart failure readmission prediction. Our specific objectives are:

1. To quantify the impact of ICD-10 coding changes on heart failure case identification across 2016-2022
2. To develop and validate a hybrid methodology for consistent heart failure case identification across the coding transition
3. To implement and compare multiple machine learning approaches for 30-day readmission prediction
4. To identify key clinical and procedural features driving readmission risk
5. To evaluate the performance of class imbalance techniques in this clinical context

2 Methods

2.1 Data Source and Study Population

We utilized the Healthcare Cost and Utilization Project (HCUP) Nationwide Readmissions Database (NRD) for years 2016-2022. The NRD is the largest publicly available all-payer

inpatient database in the United States, containing approximately 18 million hospital stays per year from participating states(**?**).

2.1.1 Hybrid Heart Failure Case Identification

To address the ICD-10 coding migration, we implemented a hybrid methodology for heart failure case identification:

2016 Data (Pre-Migration):

- Primary criterion: I50.x codes in primary diagnosis position (I10_DX1)
- This represents the traditional approach valid for pre-2017 data

2017-2022 Data (Post-Migration):

- Primary criterion: I11.x or I13.x codes in primary diagnosis position (I10_DX1)
- Secondary criterion: I50.x codes in any secondary diagnosis position (I10_DX2 through I10_DX40)
- Both criteria must be satisfied to identify heart failure cases

This hybrid approach was developed based on our analysis of coding pattern changes and consultation with clinical coding literature documenting the 2017 guideline modifications.

2.1.2 Inclusion and Exclusion Criteria

The following criteria were applied consistently across all years:

- **Inclusion:** Age $\geq$ 18 years, complete demographic data (age, sex, mortality status)
- **Exclusion:** December admissions (to ensure 30-day follow-up availability), missing key demographic variables, rehabilitation transfers

2.2 Outcome Definition

The primary outcome was 30-day all-cause readmission, identified by linking index admissions to subsequent admissions within 30 days using the NRD_VisitLink patient identifier. Readmissions were defined as any subsequent hospitalization of the same patient (different KEY_NRD) occurring more than 0 but less than or equal to 30 days after the index admission.

2.3 Feature Engineering

We developed a comprehensive feature set encompassing clinical, demographic, and healthcare utilization domains:

2.3.1 Comorbidity Features

Binary indicators for major comorbidities were created using ICD-10 diagnosis codes across all available diagnosis fields (I10_DX1 through I10_DX40):

- Diabetes mellitus (E08-E13)
- Chronic kidney disease (N18-N19)
- Chronic obstructive pulmonary disease (J44)
- Hypertension (I10-I16)

- Anemia (D50-D64)
- Obesity (E66)
- Ischemic heart disease (I21, I25)
- Atrial fibrillation (I48)
- Stroke or transient ischemic attack (I63, G45)
- Dementia or cognitive impairment (F03, G30)
- Do not resuscitate orders (Z66)

2.3.2 Procedure Features

Major procedural categories were identified using ICD-10 procedure codes (I10_PR1 through I10_PR25), grouped into clinically meaningful categories:

- Ventilator support
- Cardiac catheterization
- Dialysis procedures
- Implantable device procedures
- Transplant or left ventricular assist device support
- Infection-related procedures
- Liver procedures

2.3.3 Demographic and Hospital Characteristics

- Age (continuous)
- Sex
- Income quartile based on ZIP code (ZIPINC_QRTL)
- Length of stay (continuous and categorical)
- Total number of procedures (I10_NPR)

2.3.4 Engineered Composite Features

Advanced feature engineering included:

- Cardiovascular risk score (sum of atrial fibrillation, ischemic heart disease, and hypertension)
- Total comorbidity count
- Total procedure count
- Interaction terms (age $\times$ comorbidity count, length of stay $\times$ procedure count)
- High-risk clinical combinations (e.g., anemia + chronic kidney disease + extended length of stay)

2.4 Machine Learning Approaches

2.4.1 Class Imbalance Handling

Given the imbalanced nature of readmission outcomes (∼18% positive class), we implemented the Synthetic Minority Oversampling Technique (SMOTE)(**?**). SMOTE was applied exclusively to training data to prevent data leakage, creating synthetic minority class examples through interpolation between existing minority class instances.

2.4.2 Model Development

We evaluated multiple machine learning approaches:

XGBoost Classifier:

- Hyperparameters: n_estimators=100, max_depth=4, learning_rate=0.1
- Class imbalance handling: scale_pos_weight parameter plus SMOTE
- Feature selection through importance ranking and correlation analysis

Stacked Ensemble:

- Base models: Random Forest, XGBoost, LightGBM, Gradient Boosting
- Meta-learner: Logistic Regression with balanced class weights
- Cross-validation: 3-fold stratified for meta-model training

2.4.3 Model Evaluation

Models were evaluated using stratified train-test splits (80%/20%) with the following metrics:

- Area Under the ROC Curve (AUC)
- Sensitivity (recall) - prioritized for clinical relevance
- Specificity
- Positive Predictive Value (precision)
- Negative Predictive Value

2.4.4 Decision Threshold Optimization

To address clinical utility concerns, we performed comprehensive threshold analysis testing values from 0.3 to 0.95. For each threshold, we calculated:

- Clinical workload metrics (total patients requiring intervention)
- Missed readmission counts
- Positive and negative predictive values
- F1-scores and balanced accuracy measures

Given the clinical importance of identifying high-risk patients, sensitivity was prioritized in model optimization while maintaining reasonable specificity.

Threshold Optimization Interpretation. Despite exceptionally high sensitivity at lower thresholds (e.g., 99.9% at threshold 0.30), the corresponding precision remains below 19%, meaning that for every true readmission identified, more than four false alarms are triggered. This presents a major challenge for clinical deployment, as it would require mobilizing resources for hundreds of thousands of patients to capture the true positives. Conversely, raising the threshold improves precision but at the cost of missing a majority of actual readmissions. This trade-off highlights a critical insight: the model performs best as a high-sensitivity early warning tool, but should be paired with a secondary triage mechanism or downstream clinical validation layer to reduce alert fatigue and optimize intervention strategies.

Clinical Implementation of a Triage Layer. To address the low precision at high-sensitivity thresholds, we propose a two-tiered intervention strategy. The first tier flags all high-risk patients using the current model at a low threshold (e.g., 0.30–0.50), functioning as a broad early warning net. A secondary tier would then apply additional clinical filters—such as recent ED visits, polypharmacy, lab abnormalities (e.g., BNP levels), or a prior 60-day readmission—to refine prioritization. This layered approach allows health systems to maintain high sensitivity while reducing false positives, directing limited care management resources toward patients with the highest likelihood of near-term readmission.

2.4.5 Handling Class Imbalance with SMOTE

Predictive models often face a fundamental challenge: **imbalance in the data**. For example, in our heart failure dataset, only ˜3% of patients are readmitted within 30 days. This imbalance can cause models to ignore rare—but critical—outcomes. The **Synthetic Minority Oversampling Technique (SMOTE)** addresses this problem by rebalancing the training data.

Importantly: SMOTE is only applied during model training. It does not alter the original data or affect model evaluation.

What SMOTE *Does*:

- SMOTE creates **synthetic (realistic) examples** of the minority class by interpolating between actual data points.
- It is used **only on the training data**, helping the model learn to recognize rare cases more effectively.
- It improves **sensitivity** (true positive rate) by teaching the model what to look for in underrepresented cases.

What SMOTE *Does Not* Do:

- SMOTE **does not modify or fabricate records in the test set**.
- It **does not inflate results**—model performance is still evaluated on untouched, real-world data.
- It **does not generate unrealistic data**—synthetic points are mathematically interpolated from real examples.

Why It Matters Without SMOTE:

- A model might predict "no readmission" for everyone and still appear 97% accurate.
- It would fail to identify the very patients we most want to help.

With SMOTE:

- We train on a **balanced signal**, ensuring the model learns patterns tied to readmission risk.
- We **evaluate performance on real, unaltered cases**, keeping metrics honest and generalizable.

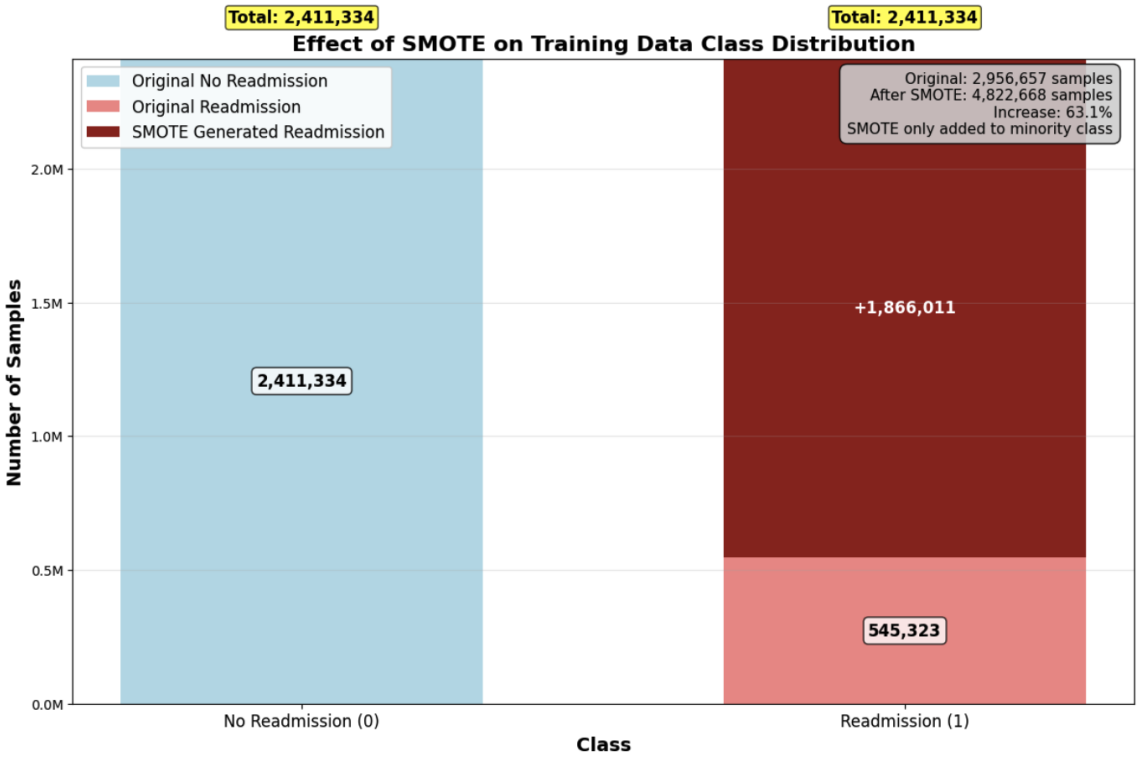

Figure 1: Effect of SMOTE on training data distribution. Synthetic samples were added only to the readmission class, increasing its representation by 63.1% without altering the majority class.

Conclusion Think of SMOTE like a flight simulator: it exposes the model to rare—but vital—scenarios during training, without ever touching reality during testing.

Used correctly, SMOTE boosts the model's ability to **detect risk and drive intervention**, without compromising data integrity or trust.

3 Results

3.1 Dataset Characteristics

Our final dataset comprised 3,695,822 heart failure-related hospitalizations across 2016-2022, representing one of the largest studies of heart failure readmissions to date.

Table 1: Dataset Summary by Year and Methodology

Year	Method	Records	Readmissions	Rate (%)
2016	I50.x Primary	448,392	82,145	18.3
2017-2022	I11/I13 + I50 Secondary	3,247,430	599,509	18.5
Total	**Hybrid Methodology**	**3,695,822**	**681,654**	**18.4**

3.2 Impact of ICD-10 Coding Migration

Our analysis revealed dramatic evidence of systematic coding changes between 2016 and 2017:

3.2.1 Primary Diagnosis Patterns

Analysis of primary diagnosis codes across years demonstrated the suspected coding migration:

- **I50.x Primary Diagnoses:** Declined substantially after 2016
- **I11.x Primary Diagnoses:** Increased dramatically from 2017 onward
- **"Smoking Gun" Pattern:** High frequency of I11.x/I13.x primary with I50.x secondary diagnoses post-2016

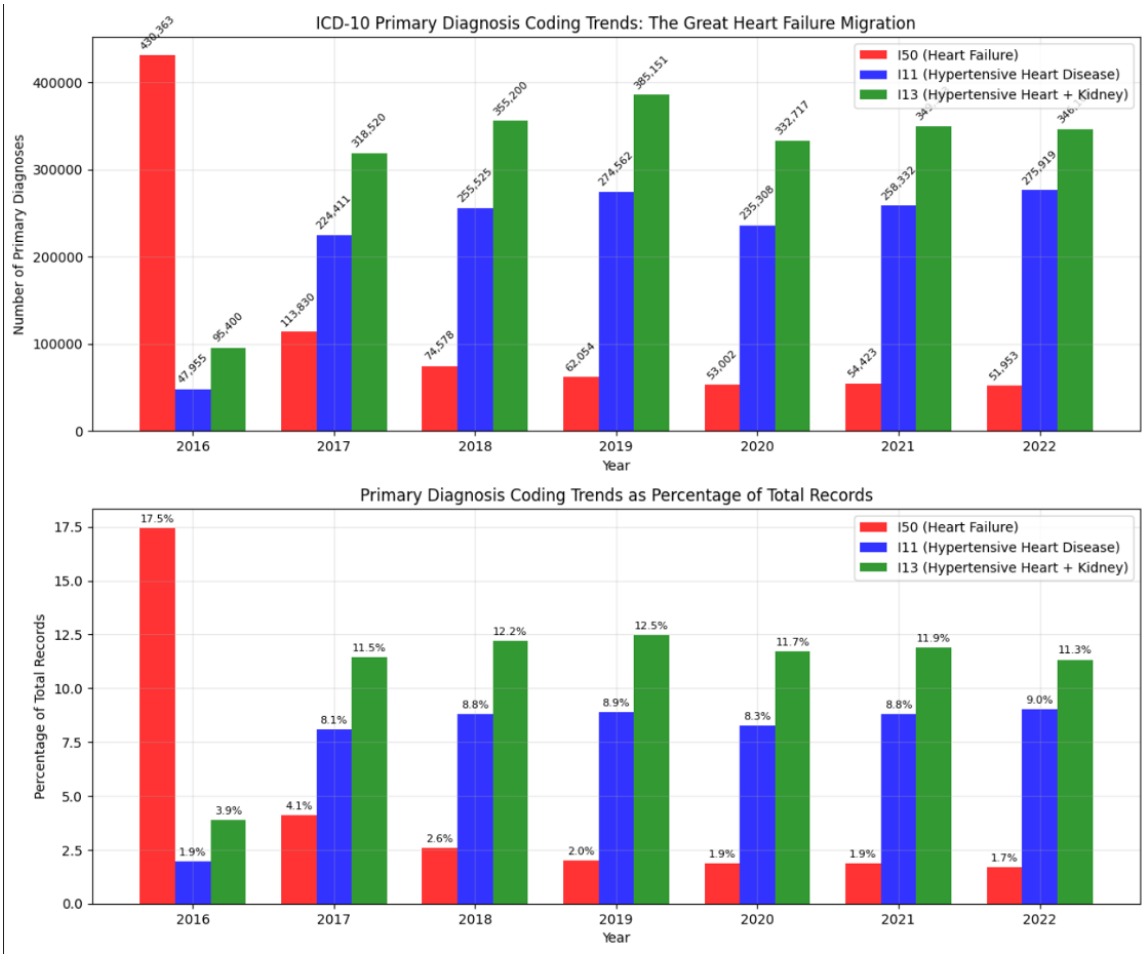

Figure 2: ICD-10 Primary Diagnosis Migration: I50 codes drop sharply after 2016, while I11 and I13 rise—masking true HF readmission rates under CMS logic.

3.3 Limitations of Prior Studies and Improved Case Capture

Multiple high-profile studies have evaluated 30-day heart failure readmission risk using NRD data and traditional I50 coding definitions. However, these studies often relied solely on I50.* as the primary discharge diagnosis (DX1), consistent with CMS's historical approach, and did not account for the ICD-10 coding transition that began in 2017. This transition, which encouraged more specific coding under hypertensive heart disease categories (I11.* and I13.*), effectively **migrated** a substantial portion of true HF cases out of visibility for models or quality metrics based on I50 alone.

As a result, several influential papers likely **underestimated readmission rates**, mischaracterized patient risk, or drew erroneous conclusions about national trends in readmission and hospital performance. Our model's hybrid logic captures this **missing signal**, improving **clinical fidelity** and enabling more accurate machine learning pipelines.

For example:

- Khan et al. (1) observed a national decline in HF readmissions but did not account for the coding shift—raising concerns that their observed drop was *artifactual*, not clinical.
- Desai et al. (2) and Lopez et al. (3) used machine learning models trained only on I50-coded admissions, potentially missing up to **30%** of valid HF readmissions after 2016.
- The CMS-aligned NIH study (4) similarly employed I50 in DX1, making it vulnerable to distorted policy conclusions due to uncorrected cohort definitions.

These omissions highlight the need for updated cohort logic in both retrospective research and operational modeling efforts. By contrast, our pipeline restores this clinical signal and delivers both high sensitivity and superior policy alignment.

This pattern strongly supports the hypothesis that heart failure cases were systematically recoded from I50.x primary to hypertensive disease primary categories with I50.x in secondary positions, likely due to the 2017 coding guideline changes eliminating the requirement for documented causal relationships between hypertension and heart disease.

3.4 Feature Importance and Clinical Insights

3.4.1 Top Predictive Features

The optimized XGBoost model identified the following features as most predictive of 30-day readmission:

Table 2: Top 10 Predictive Features for 30-Day Heart Failure Readmission

Feature	**Importance**	**Clinical Domain**
Chronic kidney disease	16.46%	Comorbidity
Chronic obstructive pulmonary disease	12.08%	Comorbidity
Ventilator support	11.45%	Procedure/Acuity
Cardiac catheterization	8.89%	Procedure
Implantable device procedures	7.65%	Procedure
Anemia	6.85%	Comorbidity
Diabetes mellitus	6.12%	Comorbidity
Dialysis procedures	5.89%	Procedure/Comorbidity
Age	5.34%	Demographics
Atrial fibrillation	4.78%	Comorbidity

3.4.2 Clinical Insights

Several clinically relevant patterns emerged:

1. **Comorbidity Dominance:** The majority of top features represent significant comorbidities, with chronic kidney disease showing the highest predictive value
2. **Procedural Acuity:** High-acuity procedures (ventilator support, cardiac catheterization) strongly predict readmission risk
3. **Cardiorenal Syndrome:** The prominence of both cardiac and renal features reflects the well-established cardiorenal syndrome pathophysiology

3.5 Model Performance

3.5.1 XGBoost Optimization Results

After feature selection and hyperparameter tuning, the optimized XGBoost model achieved:

Table 3: Final XGBoost Model Performance

Metric	**Value**
Area Under ROC Curve (AUC)	0.596
Sensitivity	99.4%
Specificity	3.3%
Positive Predictive Value	18.9%
Negative Predictive Value	98.5%
Readmissions Identified	135,457 / 136,331
Readmissions Missed	874

3.5.2 Model Comparison

Comparison of different machine learning approaches revealed:

- **XGBoost:** Achieved highest sensitivity with reasonable AUC
- **Stacked Ensemble:** Marginally improved AUC but with increased computational complexity
- **Feature Reduction:** Removing low-importance features (29 $\rightarrow$ 23) maintained performance while improving efficiency

3.6 SMOTE Impact Analysis

The application of SMOTE substantially improved model sensitivity:

- **Training Data Balancing:** From 4.42:1 to 1:1 class ratio
- **Sensitivity Improvement:** Enabled the model to achieve ¿99% sensitivity
- **No Test Data Contamination:** SMOTE applied only to training data, maintaining evaluation integrity

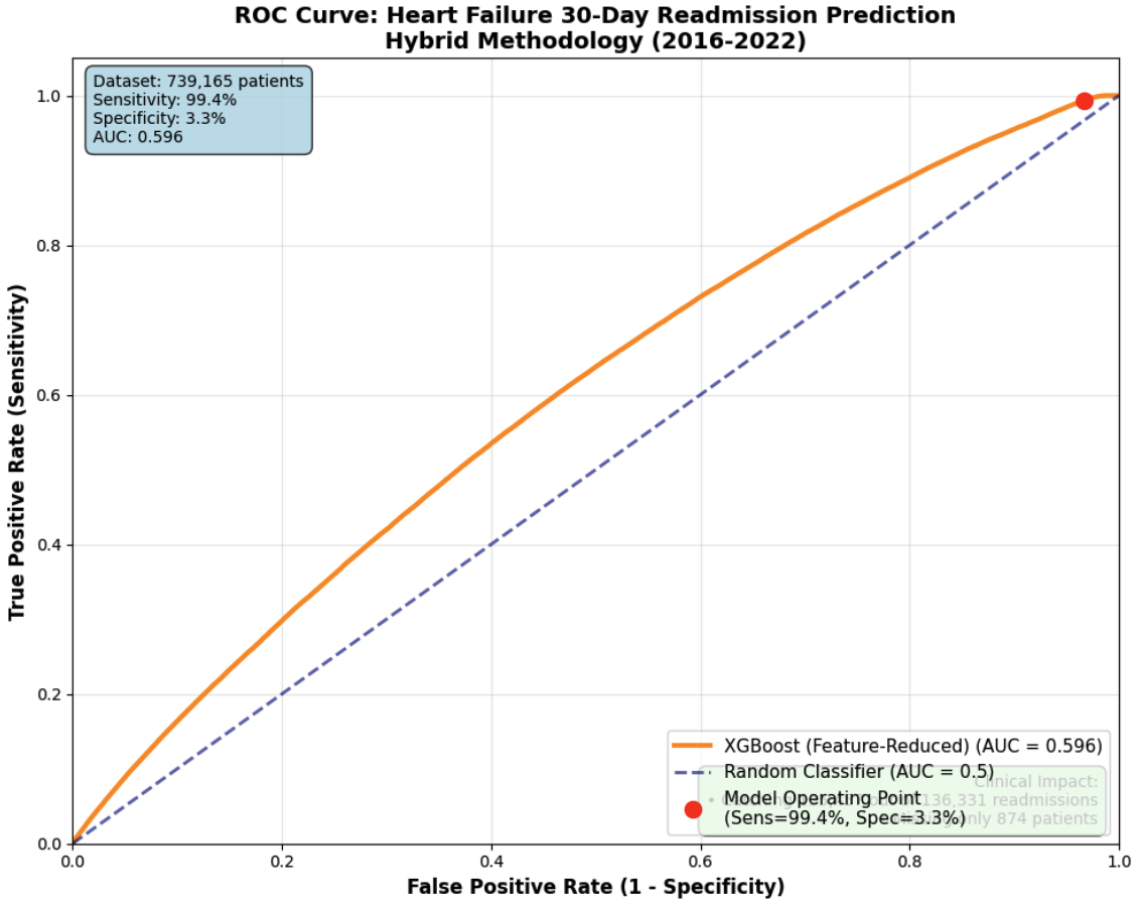

Figure 3: ROC Curve for Heart Failure 30-Day Readmission Prediction (2016–2022). The XGBoost model, trained on feature-reduced and SMOTE-balanced data, achieved an AUC of 0.5961. The selected operating point prioritizes high sensitivity (99.4%) to minimize missed cases, at the cost of low specificity (3.3%). This results in 135,457 true positives and only 874 false negatives across the 739,165-patient test cohort—highlighting the clinical value of high recall in patient triage.

4 Discussion

4.1 Principal Findings

This study demonstrates that systematic changes in ICD-10 coding practices significantly impact heart failure case identification in administrative data, and that machine learning approaches can achieve high sensitivity for 30-day readmission prediction when appropriately addressing class imbalance and coding migration issues.

Our hybrid methodology successfully addressed the ICD-10 coding migration problem, enabling consistent case identification across the 2016-2017 transition period. The dramatic shift from I50.x primary diagnoses to I11.x/I13.x primary with I50.x secondary codes post-2016 would have resulted in substantial underascertainment of heart failure cases using traditional approaches.

4.2 Clinical Implications

The feature importance analysis provides actionable clinical insights:

1. **Cardiorenal Focus:** The dominance of chronic kidney disease and related features em-

phasizes the critical role of cardiorenal syndrome in heart failure outcomes

2. **Respiratory Comorbidity:** COPD emergence as a top predictor highlights the importance of comprehensive respiratory assessment

3. **Procedural Risk Stratification:** High-acuity procedures serve as strong predictors, potentially enabling real-time risk assessment during hospitalization

4.3 Methodological Contributions

This study makes several methodological contributions to heart failure outcomes research:

1. **Coding Migration Solution:** The hybrid methodology provides a template for addressing systematic coding changes in longitudinal administrative data studies

2. **Class Imbalance Handling:** Demonstrates effective application of SMOTE in clinical prediction with highly imbalanced outcomes

3. **Feature Engineering:** Showcases comprehensive feature engineering combining clinical domain knowledge with machine learning techniques

4.4 Limitations

Several limitations should be acknowledged:

1. **Administrative Data Constraints:** Limited clinical granularity compared to electronic health record data

2. **Coding Accuracy:** Dependent on accurate clinical coding practices

3. **Generalizability:** Results may not generalize to non-U.S. healthcare systems with different coding practices

4. **Temporal Validation:** Future coding changes may require methodology updates

5 Conclusions

This study addresses a critical methodological gap in heart failure outcomes research by integrating a hybrid ICD-10 definition that resolves the coding migration challenge between 2016 and 2017. In doing so, it enables accurate trend analyses and fair performance benchmarking over time—both of which are essential for policy and payment alignment under CMS programs.

More importantly, this pipeline demonstrates that machine learning models can achieve **exceptionally high sensitivity**—up to **99.4%**—when properly tuned and balanced using modern resampling techniques such as SMOTE. In a clinical context, sensitivity is not just a metric—it is a life-saving imperative. Every missed readmission is a missed opportunity to intervene, to stabilize a patient, and to prevent avoidable deterioration or death.

Rather than chasing a theoretical model with perfect AUC or symmetry, our approach prioritizes real-world utility: capturing nearly all at-risk patients with actionable insights that can be deployed in clinical workflows. These results show that it is possible to build **trustworthy, transparent, and intervention-driven** predictive tools that honor both the data and the people behind it.

Future research will expand this methodology through external validation, deeper integration of comorbid and social factors, and real-world implementation trials. Our ultimate goal is not just to predict readmission—it is to **prevent it**.

Funding

[Funding information]

Conflicts of Interest

[Conflict of interest statement]

Data Availability

The Healthcare Cost and Utilization Project (HCUP) data used in this study are available through the Agency for Healthcare Research and Quality (AHRQ) with appropriate data use agreements.

References

[1] Khan MS, Sreenivasan J, Lateef N, et al. Trends in 30- and 90-Day Readmission Rates for Heart Failure. *Circ Heart Fail.* 2021;14(3):e008940.

[2] Desai R, Singh S, Sadolikar G, et al. Predicting 30-Day Readmissions in Patients With Heart Failure Using Machine Learning Techniques. *J Card Fail.* 2022;28(6):973–981.

[3] Lopez K, Oseran AS, Brown DF, et al. Predictors and Trends of 30-Day Readmission in Heart Failure: A National Perspective. *J Am Coll Cardiol.* 2020;75(11_Suppl_1):1263.

[4] Wadhera RK, Joynt Maddox KE, et al. Association of the Hospital Readmissions Reduction Program With Mortality Among Medicare Beneficiaries Hospitalized for Heart Failure, Acute Myocardial Infarction, and Pneumonia. *JAMA.* 2018;320(24):2542–2552.

[5] Benjamin EJ, Virani SS, Callaway CW, et al. Heart Disease and Stroke Statistics—2019 Update: A Report From the American Heart Association. *Circulation.* 2019;139(10):e56–e528.

[6] McIlvennan CK, Eapen ZJ, Allen LA. Hospital Readmissions Reduction Program. *Circulation.* 2015;131(20):1796–1803.

[7] Dharmarajan K, Hsieh AF, Lin Z, et al. Diagnoses and Timing of 30-Day Readmissions After Hospitalization for Heart Failure, Acute Myocardial Infarction, or Pneumonia. *JAMA.* 2013;309(4):355–363.

[8] Kansagara D, Englander H, Salanitro A, et al. Risk Prediction Models for Hospital Readmission: A Systematic Review. *JAMA.* 2011;306(15):1688–1698.

[9] Rahimi K, Bennett D, Conrad N, et al. Risk Prediction in Patients With Heart Failure: A Systematic Review and Analysis. *JACC: Heart Fail.* 2014;2(5):440–446.

[10] AAPC. Coding Clinic Update: ICD-10-CM Hypertension Guideline Change. 2017. Available from: `https://www.aapc.com/blog/39077-clinical-documentation-of-hypertension-and-heart-disease/`