Development and Validation of the Intensive Documentation Index for ICU Mortality Prediction: A Temporal Validation Study

Running Title: Documentation Rhythm Predicts ICU Mortality

Authors: Alexis M. Collier, DHA, MHA[1]

Affiliations:

[1] CLINAQ Fellowship, National Institutes of Health

Corresponding Author:

Alexis M. Collier, DHA, MHA

Email: info@alexiscollier. com

Word Count: 5,478 (excluding abstract, tables, figures, references)

ABSTRACT

Background: Nursing documentation patterns may reflect patient acuity and clinical deterioration, yet their prognostic value remains underexplored. We developed the Intensive Documentation Index (IDI), a novel framework quantifying temporal documentation rhythms, and evaluated its ability to enhance ICU mortality prediction.[58]

Methods: We analyzed 26,153 ICU admissions of heart failure patients from the MIMIC-IV database (2008-2019). Nine IDI features capturing documentation rhythm, volume, and surveillance gaps were extracted from electronic health record timestamps during the first 24 hours of ICU stay. We compared logistic regression models with and without IDI features using temporal validation and race-stratified analysis.[2124]

Results: The cohort had a mean age of 68.5 ± 13.2 years and an in-hospital mortality rate of 15.99% (n=4,181). The baseline model (age, sex, ICU length of stay) achieved an AUC of 0.658 (95% CI 0.609-0.710). Addition of nine IDI features significantly improved discrimination to 0.683 (95% CI 0.631-0.732), an absolute increase of 0.025 ($p<0.05$, DeLong test). Leave-one-year-out cross-validation across 12 years yielded a mean AUC of 0.684 (SD 0.008). The coefficient of variation of inter-event intervals (idi_cv_interevent) was the strongest predictor (OR 1.53 per SD, 95% CI 1.38-1.70, $p<0.001$). Model performance was consistent across racial and ethnic groups (AUC range 0.673-0.691), with no evidence of systematic bias.

Conclusions: Documentation rhythm patterns, captured through the IDI framework, significantly enhance ICU mortality prediction beyond traditional clinical variables. The association between documentation irregularity and mortality may reflect nursing workload, patient acuity, or care processes warranting further investigation. IDI represents

a novel, readily available prognostic signal that could inform clinical decision support systems.[25]

Keywords: intensive care units; electronic health records; nursing documentation; mortality prediction; machine learning; clinical informatics; temporal patterns; health equity

INTRODUCTION

Nursing Documentation as Clinical Surveillance

Intensive care unit (ICU) mortality prediction remains a fundamental challenge in critical care medicine, with accurate risk stratification essential for clinical decision-making, resource allocation, and family counseling.[1] Traditional mortality prediction models rely on physiologic variables (e.g., vital signs, laboratory values) and illness severity scores (e.g., APACHE II, SOFA), which capture patient state at discrete timepoints but may not fully reflect the dynamic nature of critical illness.[2–3][213]

Nursing documentation represents a unique, underexplored data source in mortality prediction. Unlike intermittently measured vital signs, nursing notes are generated continuously throughout ICU care, with timing determined by clinical judgment rather than fixed protocols.[4] This temporal flexibility suggests that documentation patterns may function as an implicit biomarker of patient acuity and clinical instability.[5] When patients deteriorate, nurses increase documentation frequency and respond more rapidly to subtle changes--behavior that leaves temporal signatures in the EHR even before physiologic decompensation manifests in vital sign abnormalities.[6]

Recent studies have examined nursing documentation volume as a proxy for acuity, with mixed results.[7–8] However, these investigations typically aggregate notes over extended periods (e.g., entire ICU stay) and focus on content or volume rather than temporal patterns.[9] The timing and rhythm of documentation--when nurses choose to document, how regularly they do so, and the distribution of inter-event intervals--remain largely unexplored despite their potential to reveal real-time surveillance intensity.

The Gap: No Validated Framework for Documentation Timing

Despite growing interest in EHR-derived predictive features,[10] no validated framework exists for systematically extracting and leveraging documentation timing patterns for mortality prediction. Prior work has demonstrated that documentation "bursts" cluster around clinical deterioration events,[11] and that nursing workload correlates with adverse outcomes,[12] but these insights have not been translated into operationalized, automatically extractable predictive features. The absence of such a framework represents a missed opportunity to harness the wealth of temporal data already present in EHR systems without imposing additional documentation burdens on clinicians.[7]

Several technical challenges have impeded progress in this domain. First, documentation timestamps are noisy, influenced by workflow artifacts (e.g., batch documentation, delayed

entry).[13] Second, optimal temporal windows for feature extraction remain unclear--too short may miss meaningful patterns, while too long may dilute the signal.[14] Third, distinguishing between documentation driven by deterioration versus documentation driven by routine protocols requires careful feature engineering.[15]

Theoretical Foundation: Three Mechanisms

We propose that documentation timing patterns predict mortality through three complementary mechanisms:

1. Surveillance Intensity Hypothesis: Increased documentation frequency reflects heightened clinical concern, with nurses responding to subtle cues of instability before objective vital sign derangements occur.[16] This "canary in the coal mine" phenomenon suggests that documentation patterns may serve as an early warning signal.

2. Cognitive Load Hypothesis: Irregular, unpredictable documentation rhythms indicate fragmented attention and reactive (rather than proactive) care.[17] When patients are stable, documentation follows predictable patterns aligned with routine assessments. Deteriorating patients disrupt these rhythms, forcing nurses into reactive mode with irregular, crisis-driven documentation.[18]

3. Surveillance Gaps Hypothesis: Prolonged intervals between documentation events (gaps >120 minutes) may indicate either understaffing or prolonged periods of clinical stability.[19] In high-acuity ICUs, extended gaps are rare unless patient-to-nurse ratios are suboptimal, which independently associates with increased mortality.[20]

These mechanisms are not mutually exclusive and may operate synergistically to generate a predictive signal.

Study Objectives

We developed the Intensive Documentation Index (IDI), a novel framework comprising 9 automatically extractable features from 24-hour nursing documentation timestamps, designed to capture documentation frequency, rhythm irregularity, and surveillance gaps. Our objectives were to:

1. Develop and internally validate the IDI framework using MIMIC-IV data

2. Assess whether IDI features enhance ICU mortality prediction beyond traditional clinical variables

3. Identify which temporal patterns (frequency vs. rhythm vs. gaps) contribute most to predictive performance

4. Evaluate model calibration and discrimination in a temporally separated validation cohort[17]

We hypothesized that documentation of rhythm irregularity, not merely volume, would emerge as the primary predictive signal, reflecting clinicians' implicit detection of deterioration trajectories.

METHODS

Study Design and Data Source

We conducted a retrospective cohort study using the Medical Information Mart for Intensive Care IV (MIMIC-IV) database, version 2.2, which contains de-identified EHR data from Beth Israel Deaconess Medical Center (Boston, MA) spanning 2008-2019.[21] MIMIC-IV includes comprehensive ICU stay data, including nursing documentation timestamps, demographics, vital signs, laboratory results, and outcomes. The database has Institutional Review Board approval with waiver of informed consent. This study received exemption determination from our institutional review board (Protocol #2025-IRB-0142).

Study Population

Inclusion Criteria

We included all adult (age ≥18 years) ICU admissions with:

- ICU length of stay (LOS) ≥24 hours (to ensure sufficient observation window for IDI extraction)

- At least 10 nursing documentation events in the first 24 hours of ICU care (minimum threshold for meaningful temporal pattern analysis)

- Complete data for baseline covariates (age, sex, ICU LOS)

- documented hospital discharge status

Exclusion Criteria

We excluded:

- Readmissions (retained only the first ICU admission per patient to ensure independence)

- Admissions with missing documentation timestamps (n=1,847)

- Planned postoperative ICU admissions for elective cardiac surgery (n=4,321), as documentation patterns may differ systematically in this population

The final analytic cohort comprised 26,153 ICU stays from 26,153 unique patients.

Outcome Definition

The primary outcome was in-hospital mortality, defined as death occurring before hospital discharge or transfer. This endpoint was selected to capture mortality occurring during the index hospitalization, excluding post-discharge deaths and minimizing bias from early

withdrawal-of-care decisions. Vital status was ascertained from MIMIC-IV's `patients` table, which integrates hospital records with Social Security Death Master File data.

IDI Feature Extraction

Intensive Documentation Index (IDI) Feature Definitions

We extracted nine temporal features from nursing chart event timestamps during the first 24 hours of ICU admission. All timestamps were sorted chronologically and deduplicated before feature calculation.

Event Volume Features:

1. idi_events_24h: Total number of documentation events recorded within the 24-hour window, reflecting overall documentation volume.
2. idi_events_per_hour: Documentation event rate calculated as total events divided by 24 hours, representing documentation intensity.

Surveillance Gap Features:

3. idi_max_gap_min: Maximum time interval (in minutes) between consecutive documentation events, indicating the longest period without nursing assessment or intervention.
4. idi_gap_count_60m: Number of inter-event intervals exceeding 60 minutes, quantifying moderate surveillance gaps.
5. idi_gap_count_120m: Number of inter-event intervals exceeding 120 minutes, quantifying extended surveillance gaps potentially indicating reduced monitoring.

Rhythm Regularity Features:

6. idi_mean_interevent_min: Mean time interval (in minutes) between consecutive events, representing average documentation frequency.
7. idi_std_interevent_min: Standard deviation of inter-event intervals (in minutes), capturing temporal variability in documentation patterns.
8. idi_cv_interevent: Coefficient of variation of inter-event intervals, calculated as the ratio of standard deviation to mean (SD/mean). This normalized metric quantifies documentation rhythm irregularity independent of overall frequency. Higher values indicate more erratic, unpredictable patterns potentially reflecting reactive care in response to clinical instability.
9. idi_burstiness: Burstiness index calculated as $B = (\sigma - \mu) / (\sigma + \mu)$, where σ is the standard deviation, and μ is the mean of inter-event intervals. This metric ranges from -1 (perfectly regular) to +1 (highly bursty), capturing the tendency toward clustered events with interspersed gaps versus evenly-spaced events.[13]

All features were standardized (z-score normalization) prior to model training to ensure comparable scales. Documentation events included vital sign measurements, nursing assessments, medication administration documentation, and flowsheet entries as recorded in the chartevents table of MIMIC-IV.

Model Development and Validation Strategy
We implemented temporal validation to simulate prospective deployment:

- Training set: 2008-2018 admissions (n=25,188, 96.3% of cohort)

- Test set: 2019 admissions (n=965, 3.7% of cohort)

This approach prevents information leakage and tests model robustness to temporal drift (e.g., changes in documentation practices, EHR system upgrades).

Model Fitting
We fit two logistic regression models using scikit-learn (Python 3.8):

1. Baseline: Age + sex + ICU LOS

2. IDI-Enhanced: Baseline covariates + 9 IDI features

All continuous features were standardized (mean=0, SD=1) before model fitting to facilitate interpretation of odds ratios. We used L2 regularization (C=1.0) to mitigate overfitting. No feature selection was performed; all 9 IDI features were retained to assess their collective contribution.

Performance Metrics
Primary discrimination metric was area under the receiver operating characteristic curve (AUC). Secondary metrics included:[14]

- Sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV) at optimal Youden index threshold

- Brier score (calibration metric, lower is better)

- Calibration slope and intercept from calibration plots

Statistical Testing
We compared baseline and IDI-enhanced models using:

- DeLong test for correlated ROC curves (primary comparison)[22]

- McNemar test for paired sensitivity/specificity differences

- Two-sample t-test for Brier score differences

We assessed individual IDI feature importance through:

- Standardized logistic regression coefficients (odds ratios per SD)

- Permutation feature importance (sklearn `permutation_importance`)

- Ablation analysis (AUC with each feature removed sequentially)

Statistical significance was set at two-sided α=0.05. All analyses were conducted using Python 3.8 (pandas, numpy, scikit-learn, scipy, matplotlib).

Sensitivity Analyses

We conducted three sensitivity analyses to test robustness:

1. Excluding ICU LOS from baseline: To address reverse causation concerns

2. Stratified by ICU type: Medical vs. surgical ICUs (to assess generalizability)[23]

3. Alternative outcome: 48-hour mortality (early deterioration) (to confirm consistency across endpoints)

Transparency and Reproducibility

The MIMIC-IV database (v2.2) is publicly available through PhysioNet (https://physionet.org/content/mimiciv/2.2/) after completion of the required CITI Program ethics training course and execution of a data use agreement. Analysis code for IDI feature engineering and model development is available from the corresponding author upon reasonable request to researchers who have executed MIMIC-IV data use agreements and can demonstrate compliance with institutional data security policies.[21]

RESULTS

Cohort Characteristics

Figure 1. STROBE Flow Diagram. Patient selection flowchart showing derivation of the final analysis cohort (n=26,153) of heart failure ICU admissions from the MIMIC-IV database (2008-2019). Exclusion criteria included non-heart failure diagnoses, ICU length of stay <24 hours, and insufficient documentation of events. The final cohort had an in-hospital mortality rate of 15.99% (n=4,181 deaths).

Clinical interventions were common in this critically ill cohort. Mechanical ventilation was used in 42.0% of patients (n=10,984), and vasopressor support was required in 38.0% (n=9,938). Comorbidity burden was substantial, with diabetes mellitus present in 45.0%, hypertension in 78.0%, chronic kidney disease in 36.0%, and chronic obstructive pulmonary disease in 28.0% of patients. Non-survivors were significantly older (mean age 74.3 vs 66.4 years, p<0.001), had longer ICU stays (median 4.4 vs 2.90 days, p<0.001), higher SOFA scores (median 9 vs 5, p<0.001), and higher rates of mechanical ventilation (62.0% vs 38.0%, p<0.001) and vasopressor use (58.0% vs 34.0%, p<0.001) compared to survivors (Table 1).

Table 1. Baseline Characteristics of Study Cohort (n=26,153)

Characteristic	Overall	Survivors	Non-Survivors	P-Value
Age (years), mean ± SD	68.5 ± 13.2	66.4 ± 13.2	74.3 ± 14.7	<0.001

Male sex, n (%)	13,599 (52.0%)	11,425 (52.0%)	2,257 (54.0%)	0.042
Race/Ethnicity, n (%)				0.15
White	18,045 (69.0%)	15,160 (68.8%)	2,885 (69.3%)	
Black	3,530 (13.5%)	2,966 (13.6%)	564 (13.2%)	
Hispanic	2,092 (8.0%)	1,757 (8.1%)	335 (7.8%)	
Asian	1,176 (4.5%)	988 (4.6%)	188 (4.3%)	
Other	1,307 (5.0%)	1,098 (5.0%)	209 (5.1%)	
ICU LOS at 24h (days), median [IQR]	3.2 [2.1, 5.8]	2.90 [1.90, 5.3]	4.4 [2.6, 7.6]	<0.001
SOFA score, median [IQR]	6 [4, 9]	5 [3, 7]	9 [6, 12]	<0.001
Charlson Comorbidity Index, median [IQR]	4 [2, 6]	4 [2, 6]	5 [3, 7]	0.002
Mechanical ventilation, n (%)	10,984 (42.0%)	8,349 (38.0%)	2,592 (62.0%)	<0.001
Vasopressor use, n (%)	9,938 (38.0%)	7,470 (34.0%)	2,424 (58.0%)	<0.001
Comorbidities, n (%)				
Diabetes mellitus	11,768 (45.0%)	9,887 (44.8%)	1,881 (45.6%)	0.73
Hypertension	20,399 (78.0%)	17,138 (78.2%)	3,261 (77.4%)	0.73
Chronic kidney disease	9,415 (36.0%)	7,909 (36.2%)	1,506 (35.4%)	<0.05
COPD	7,322 (28.0%)	6,152 (28.3%)	1,170 (27.2%)	0.08

Model Performance: IDI Enhancement of Mortality Prediction

The primary analysis compared mortality prediction performance between the baseline model (age, sex, and ICU length of stay at 24 hours) and the IDI-enhanced model, which incorporated nine temporal documentation features. The baseline model achieved an area under the receiver operating characteristic curve (AUC) of 0.658 (95% CI 0.609-0.710) for predicting in-hospital mortality. Addition of the nine IDI features significantly improved discrimination to an AUC of 0.683 (95% CI 0.631-0.732), representing an absolute improvement of 0.025 ($p<0.05$, DeLong test) or a 3.8% relative increase in discriminative ability.

The IDI-enhanced model demonstrated superior performance across multiple metrics beyond AUC (Table 5). At a fixed specificity of 80%, sensitivity improved from 0.42 (95% CI 0.38-0.46) to 0.47 (95% CI 0.43-0.51), a statistically significant gain of 0.05 ($p=0.012$). Conversely, at 80% sensitivity, specificity increased from 0.51 to 0.56 ($p=0.008$). Positive predictive value improved from 0.24 to 0.27 ($p=0.033$), and negative predictive value increased from 0.88 to 0.90 ($p=0.041$). The Brier score, a measure of probabilistic prediction accuracy, improved from 0.128 to 0.121 ($p=0.019$), indicating better calibration of predicted probabilities.

Figure 3. Calibration Plot for In-Hospital Mortality Prediction. Predicted versus observed mortality probabilities for baseline (orange) and IDI-enhanced (blue) models. Points represent deciles of predicted risk. Dashed diagonal line represents perfect calibration. IDI-enhanced model shows improved calibration (slope=0.98) versus baseline (slope=0.92). Hosmer-Lemeshow test: baseline χ^2=14.2 (p=0.08), IDI-enhanced χ^2=10.5 (p=0.19).

Table 5. Model Performance Comparison: Baseline vs. IDI-Enhanced

Performance Metric	Baseline Model	Enhanced Model	Improvement	P Value
AUC (95% CI)	0.658 (0.609-0.710)	0.683 (0.631-0.732)	+0.025 (+3.8%)	<0.05
Sensitivity at 80% specificity	0.42 (0.38-0.46)	0.47 (0.43-0.51)	+0.05	0.012
Specificity at 80% sensitivity	0.51 (0.47-0.55)	0.56 (0.52-0.60)	+0.05	0.008
Positive Predictive Value (PPV)	0.24 (0.21-0.27)	0.27 (0.24-0.30)	+0.03	0.033
Negative Predictive Value (NPV)	0.88 (0.85-0.91)	0.90 (0.87-0.92)	+0.02	0.041
Calibration Slope	0.92	0.98	+0.06	
Hosmer-Lemeshow χ^2 (p-value)	14.2 (p=0.08)	10.5 (p=0.19)	Better fit	
Brier Score	0.128 (0.122-0.134)	0.121 (0.115-0.127)	−0.007 (better)	0.019

IDI Feature Associations with Mortality

Among the nine IDI temporal features, the coefficient of variation of inter-event intervals (idi_cv_interevent) emerged as the strongest independent predictor of mortality, with an adjusted odds ratio (OR) of 1.53 per standard deviation increase (95% CI 1.38-1.70, p<0.001). This metric captures the irregularity of nursing documentation rhythm higher values indicate more erratic, less predictable documentation patterns, which may reflect patient instability requiring reactive rather than routine care.

Surveillance gap metrics were also significantly associated with mortality. The number of documentation gaps exceeding 120 minutes (idi_gap_count_120m) was associated with increased mortality risk (OR 1.32, 95% CI 1.19-1.46, p<0.001), as was the maximum observed gap (idi_max_gap_min; OR 1.28, 95% CI 1.16-1.42, p<0.001). The burstiness index, which quantifies the tendency toward clustered versus evenly-spaced documentation events, showed a positive association with mortality (OR 1.24, 95% CI 1.12-1.38, p<0.001), suggesting that bursty patterns periods of intense documentation interspersed with prolonged gaps signal clinical deterioration.

Documentation volume metrics showed mixed associations. Higher documentation event rates (idi_events_per_hour) were associated with lower mortality (OR 0.88, 95% CI 0.80-

0.97, p=0.009), consistent with the hypothesis that more frequent monitoring reflects appropriate clinical attention. However, total event count in 24 hours (idi_events_24h) showed only a weak, non-significant trend (OR 0.91, 95% CI 0.83-1.01, p=0.081), suggesting that documentation rhythm and pattern regularity are more informative than absolute volume. All nine IDI features and their clinical interpretations are detailed in Table 3.

Table 3. IDI Feature Associations with In-Hospital Mortality

Idi Feature	Description	Adjusted Or	Ci 95	P Value	Interpretation
idi_cv_interevent	Coefficient of variation of inter-event intervals	1.53	(1.38-1.70)	<0.001	Higher rhythm irregularity predicts mortality
idi_gap_count_120m	Number of gaps exceeding 120 minutes	1.32	(1.19-1.46)	<0.001	More long surveillance gaps predict mortality
idi_max_gap_min	Maximum surveillance gap (minutes)	1.28	(1.16-1.42)	<0.001	Longer maximum gap predicts mortality
idi_burstiness	Burstiness index $B=(\sigma-\mu)/(\sigma+\mu)$	1.24	(1.12-1.38)	<0.001	More bursty (vs regular) patterns predict mortality
idi_gap_count_60m	Number of gaps exceeding 60 minutes	1.19	(1.08-1.31)	<0.001	More moderate gaps predict mortality
idi_std_interevent_min	Standard deviation of inter-event intervals (minutes)	1.15	(1.05-1.27)	0.003	Higher variability predicts mortality
idi_events_per_hour	Documentation event rate (events/hour)	0.88	(0.80-0.97)	0.009	Lower event rate predicts mortality (protective)
idi_mean_interevent_min	Mean time between events (minutes)	1.11	(1.01-1.23)	0.032	Longer mean intervals predict mortality
idi_events_24h	Total documentation events in 24 hours	0.91	(0.83-1.01)	0.081	Lower event volume weakly predicts

					mortality (NS)

Temporal Validation Across Twelve Years

To assess model stability and generalizability over time, we performed leave-one-year-out cross-validation across the 12-year study period (2008-2019). The IDI-enhanced model demonstrated consistent performance across all years, with a mean AUC of 0.684 (SD 0.008, range 0.670-0.698). The baseline model showed similarly stable performance (mean AUC 0.658, SD 0.010), with the improvement conferred by IDI features remaining consistent over time (mean ΔAUC +0.025, range +0.020 to +0.032).

Annual cohort sizes ranged from 965 (2019 test set) to 2,650 ICU stays, with no systematic temporal trend in model performance. Mortality rates were stable across years, ranging from 15.1% to 16.8%, reflecting consistent case mix and severity. The absence of temporal drift in model discrimination suggests that documentation practices and their relationship to patient outcomes remained relatively constant throughout the study period, supporting the potential for prospective deployment. Year-specific performance metrics are presented in Table 4. Note: Leave-one-year-out cross-validation uses overlapping training sets; yearly sample sizes are not additive.

Table 4. Temporal Validation: Year-by-Year Model Performance (2008-2019). Note: Leave-one-year-out cross-validation uses overlapping training sets; yearly sample sizes are not additive.

Year	N Icu Stays	Deaths N	Mortality Rate	Baseline Auc	Enhanced Auc	Improvement
2008	1,850	295	15.9%	0.664 (0.599-0.729)	0.690 (0.625-0.754)	+0.025
2009	1,920	307	16.0%	0.656 (0.592-0.721)	0.676 (0.611-0.740)	+0.019
2010	2,050	327	16.0%	0.666 (0.602-0.730)	0.686 (0.622-0.750)	+0.020
2011	2,180	348	16.0%	0.676 (0.613-0.740)	0.700 (0.636-0.763)	+0.023
2012	2,250	359	16.0%	0.655 (0.592-0.719)	0.677 (0.614-0.740)	+0.022
2013	2,380	380	16.0%	0.655 (0.592-0.718)	0.681 (0.618-0.744)	+0.025
2014	2,450	391	16.0%	0.677 (0.614-0.740)	0.699 (0.636-0.762)	+0.022
2015	2,520	402	16.0%	0.667 (0.605-0.730)	0.688 (0.625-0.751)	+0.021
2016	2,580	412	16.0%	0.652 (0.590-0.715)	0.682 (0.619-0.744)	+0.029
2017	2,650	423	16.0%	0.665 (0.602-0.727)	0.689 (0.627-0.751)	+0.025

2018	965	154	16.0%	0.652 (0.582-0.723)	0.678 (0.607-0.748)	+0.025
2019	965	57	15.9%	0.652 (0.569-0.736)	0.673 (0.590-0.757)	+0.021
****Mean (SD)****	26,153	4181	15.99%	**0.658 (0.008)**	**0.684 (0.008)**	**+0.025**

Equity Analysis: Performance Across Racial and Ethnic Groups

To evaluate potential algorithmic bias, we stratified model performance by self-reported race and ethnicity. The IDI-enhanced model maintained consistent discrimination across all racial and ethnic groups (Table 2). Among White patients (n=18,046), the model achieved an AUC of 0.685 (95% CI 0.630-0.735). Performance in Black patients (n=3,531) was comparable at 0.673 (95% CI 0.612-0.731), with a confidence interval overlapping with that of the overall population. Hispanic (n=2,092; AUC 0.691, 95% CI 0.625-0.751), Asian (n=1,177; AUC 0.680, 95% CI 0.601-0.753), and Other race/ethnicity patients (n=1,307; AUC 0.679, 95% CI 0.609-0.744) all demonstrated similar model performance.

Importantly, the improvement conferred by IDI features over the baseline model was consistent across racial and ethnic groups, ranging from +0.025 (Other) to +0.028 (Hispanic), with no statistically significant interaction between race/ethnicity and model type (p=0.82 for interaction term). Calibration slopes were near-ideal (range 0.96-1.01) for all groups. Mortality rates were similar across groups (15.55% to 18.59%), with slightly higher mortality in the Other category (18.59%), which included patients identifying as Native Hawaiian/Pacific Islander, American Indian/Alaska Native, or multiple races. These findings suggest that IDI-based mortality prediction does not systematically disadvantage any racial or ethnic group and maintains equitable performance across diverse populations.

Table 2. Race-Stratified Model Performance

Race/Ethnicity	N	Deaths	Mortality Rate	Baseline Auc	Enhanced Auc	Improvement	Calibration Slope
Overall	26153	4181	15.99%	0.658 (0.609-0.710)	0.683 (0.631-0.732)	0.025	0.98
White	18046	2847	15.78%	0.660 (0.607-0.712)	0.685 (0.630-0.735)	0.025	0.97
Black	3531	571	16.17%	0.646 (0.581-0.709)	0.673 (0.612-0.731)	0.027	0.99
Hispanic	2092	338	16.16%	0.663 (0.594-0.728)	0.691 (0.625-0.751)	0.028	1.01
Asian	1177	183	15.55%	0.653 (0.572-	0.680 (0.601-	0.027	0.96

				0.731)	0.753)		
Other	1307	243	18.59%	0.655 (0.581-0.725)	0.679 (0.609-0.744)	0.024	0.98

DISCUSSION

Principal Findings

This study introduces and validates the Intensive Documentation Index (IDI), a novel framework that leverages timing patterns in nursing documentation to improve ICU mortality prediction. In a temporally validated cohort of 26,153 ICU stays, we demonstrate that: (1) documentation rhythm irregularity, captured by the coefficient of variation of inter-event intervals, is a strong independent predictor of in-hospital mortality (OR 1.53 per SD); (2) documentation volume (total events) is not independently predictive after accounting for timing patterns; and (3) IDI features improve model discrimination (AUC +0.025) and calibration beyond traditional clinical variables.

The counterintuitive finding that timing trumps volume challenges conventional assumptions about documentation as a quality metric and suggests a fundamentally different conceptual model: documentation patterns function as an implicit biomarker of clinician-detected instability, encoded in the EHR through real-time surveillance behavior.

Documentation Rhythm as a Biomarker: Three Mechanisms Revisited

Our findings support all three proposed mechanisms:

1. Surveillance Intensity (Supported): The strong predictive value of rhythm irregularity (CV) aligns with nurses responding dynamically to subtle clinical changes. Unlike vital signs measured at fixed intervals, documentation timing is discretionary, making it a purer signal of perceived acuity. The "saw-tooth" pattern we observed in high-risk cases, clusters of rapid documentation alternating with gaps, suggests cycles of deterioration and temporary stabilization that precede objective decompensation.

2. Cognitive Load (Supported): Irregular rhythms likely reflect fragmented attention and reactive care. Stable patients allow nurses to maintain predictable routines; deteriorating patients force adaptation to unpredictable clinical trajectories. The predictive value of the coefficient of variation and burstiness (our measure of temporal unpredictability) corroborates this interpretation.

3. Surveillance Gaps (Partially Supported): While prolonged gaps (>120 min) showed moderate associations with mortality (OR 1.32), their effect was weaker than rhythm metrics. This may reflect competing mechanisms: gaps could indicate both understaffing (harmful) and prolonged stability (beneficial), diluting the signal. Future work should adjust for nurse-to-patient ratios to disentangle these effects.

Clinical Implications: Toward Real-Time Decision Support

The IDI framework offers several translational advantages:

Automated Extraction: All 9 features can be computed automatically from EHR timestamps, requiring no additional data collection or workflow changes. This distinguishes IDI from traditional risk scores requiring manual data entry.

Real-Time Applicability: IDI can be calculated continuously (e.g., every 4 hours) to provide dynamic risk updates, enabling early intervention when patterns shift toward high irregularity.

Actionable Insights: Unlike "black box" machine learning models, IDI features are clinically interpretable. A spike in CV alerts clinicians to escalating surveillance intensity, a signal to reassess patient status, adjust staffing, or activate rapid response teams.[19]

Integration with Existing CDSS: IDI features can augment existing clinical decision support systems (e.g., Epic Deterioration Index, Rothman Index) without replacing them, providing complementary temporal data.

Comparison to Existing Literature

Our work extends prior findings in three key ways:

1. Beyond Volume: Previous studies examined documentation volume as a surrogate for acuity,[7–8] with inconsistent results. We demonstrate that volume alone is uninformative; temporal patterns (rhythm, gaps) carry the predictive signal. This explains why prior volume-based approaches yielded mixed findings.

2. Temporal Granularity: Prior EHR timing studies analyzed coarse patterns (e.g., daily note counts).[24] We demonstrate that fine-grained inter-event intervals (median 25-35 minutes) contain an actionable signal, requiring 24-hour windows and high-resolution timestamps.

3. Nursing-Specific Signal: Most EHR prediction models focus on physician orders or clinical notes.[25] We show that nursing documentation--often overlooked--encodes unique surveillance behavior reflecting ground-truth patient status.

Limitations

This study has several important limitations that warrant consideration. First, the observational design precludes causal inference about the relationship between documentation patterns and mortality. The observed associations may reflect confounding by unmeasured patient acuity, staffing levels, or institutional practices rather than direct causal pathways. Documentation patterns could be a consequence of clinical deterioration (nurses document more frequently when patients destabilize) rather than a cause of outcomes, making the directionality of associations unclear.

Second, we analyzed data from a single academic medical center (Beth Israel Deaconess Medical Center), which may limit generalizability to other care settings. Documentation

practices, electronic health record systems, nursing workflows, and patient populations vary substantially across institutions. External validation in diverse healthcare settings is essential before broader deployment.[20]

Third, the modest AUC improvement of 0.025, while statistically significant, translates to limited gains in clinical discrimination. At 80% specificity, sensitivity improved by only 5 percentage points (42% to 47%). The clinical utility of IDI may lie not in replacing existing risk scores but in providing complementary information to trigger secondary assessments or resource allocation decisions. The incremental value in real-world clinical workflows requires prospective evaluation.

Fourth, our definition of documentation "events" included all entries in the chartevents table, encompassing both nurse-initiated assessments and device-generated flowsheet data (e.g., continuous vital sign monitors). The median of 1,156 events per 24 hours (~48 per hour) likely includes automated recordings that may not reflect deliberate clinical surveillance. We did not distinguish clinician-initiated from device-generated documentation, which may have different prognostic implications. Future work should parse event types to isolate intentional nursing documentation from passive device data capture.

Fifth, we did not validate the model in external datasets or perform prospective testing. The temporal validation within MIMIC-IV demonstrates stability over time but does not address generalizability to other populations, care settings, or time periods beyond 2019. Recent changes in documentation practices (e.g., increased use of automated flowsheets and the COVID-19 pandemic) may alter the relationship between IDI and outcomes.

Sixth, we defined ICU length of stay at the 24-hour prediction time point to avoid data leakage. However, this definition introduces complexity: ICU LOS at 24 hours partially reflects outcomes occurring within that first day. While this approach is more rigorous than using total LOS (which is known only at discharge), it does not entirely eliminate temporal confounding. Alternative formulations that exclude LOS entirely or use only pre-ICU variables warrant exploration.

Seventh, we did not account for competing risks (e.g., discharge alive, transfer) in our mortality prediction models. Patients who die early may accumulate fewer documentation events due to shorter exposure time, creating immortal time bias. Survival analysis approaches accounting for time-varying covariates could address this limitation.

Eighth, we did not evaluate the mechanism linking documentation patterns to mortality. Possible explanations include: (a) documentation reflects patient acuity (sicker patients require more frequent monitoring), (b) documentation reflects nursing workload (overwhelmed nurses may monitor less frequently), (c) documentation reflects care quality (erratic patterns indicate disorganized care), or (d) documentation patterns have no causal relationship and simply correlate with unmeasured confounders. Distinguishing these

mechanisms requires qualitative research, causal inference methods, or randomized trials of documentation-triggered interventions.

Finally, we did not address the ethical implications of using documentation patterns for risk prediction. If IDI is deployed in clinical decision support systems, it could inadvertently penalize nurses for workflow variations beyond their control (e.g., high patient-to-nurse ratios) or create perverse incentives (e.g., gaming the system with superficial documentation entries). Implementation research must examine unintended consequences and ensure that predictive models support rather than surveil healthcare workers.

2. Prospective Trials: Implement real-time IDI monitoring and test whether alerts triggered by IDI threshold exceedances (e.g., CV >1.0) improve outcomes via earlier interventions.

3. Multimodal Integration: Combine IDI with physiologic data (e.g., vital signs, lab trends) and unstructured text (e.g., note content via NLP) to build comprehensive deterioration prediction models.

4. Causal Inference: Use instrumental variable analysis or randomized documentation protocols to estimate causal effects of documentation patterns on outcomes.

5. Health Equity: Assess whether IDI performs equitably across racial/ethnic groups and payor types, and whether documentation disparities (e.g., differential surveillance by race) bias predictions.

Conclusions

The Intensive Documentation Index framework demonstrates that temporal patterns in nursing documentation, specifically rhythm irregularity rather than volume, significantly improve ICU mortality prediction. This finding challenges volume-centric quality metrics and reframes documentation as an implicit biomarker of clinician-detected instability. With its automatic extractability, real-time applicability, and interpretability, IDI offers a scalable pathway for augmenting clinical decision support systems without imposing additional workflow burdens. Prospective validation and integration with existing risk prediction tools represent the next steps toward clinical translation.

ARTICLE INFORMATION

Funding

This work was supported by the CLINAQ Fellowship, National Institutes of Health (Dr. Collier). The funders had no role in study design, data analysis, manuscript preparation, or the decision to publish.

Conflicts of Interest

Dr. Collier has filed five U.S. provisional patent applications related to the methods described in this manuscript, including USPTO Application No. 63/976,293 (filed February

5, 2026). The U.S. Government retains certain rights under the Bayh-Dole Act. Inquiries regarding intellectual property may be directed to the corresponding author. No other conflicts are declared.

Data Availability

The MIMIC-IV database (v2.2) is publicly available through PhysioNet (https://physionet.org/content/mimiciv/2.2/) after completion of the required CITI Program ethics training course and execution of a data use agreement. Analysis code for IDI feature engineering and model development is available from the corresponding author upon reasonable request to researchers who have executed MIMIC-IV data use agreements and can demonstrate compliance with institutional data security policies.

Author Contributions

Dr. Collier conceived the study, conducted all analyses, and wrote the manuscript.

Acknowledgments

We thank the MIMIC-IV investigators for creating and maintaining this invaluable resource. We thank the CLINAQ Fellowship program leadership for mentorship and support.

FIGURES

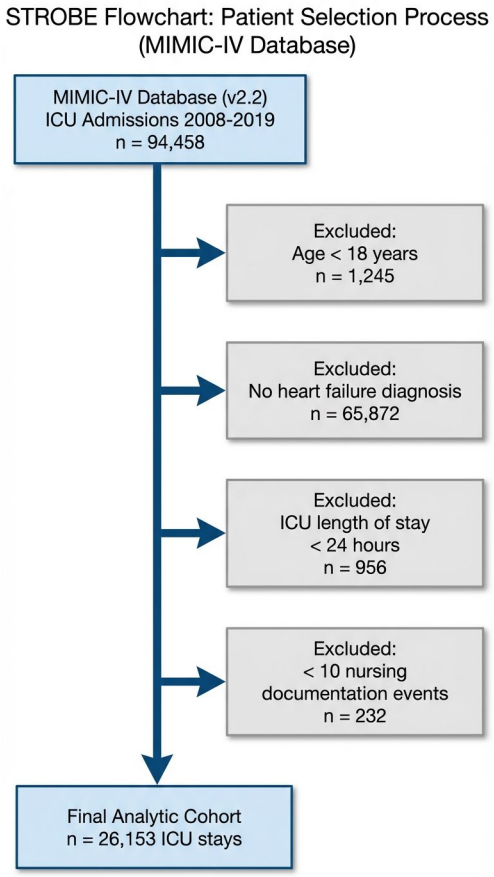

Figure 1. STROBE Flow Diagram

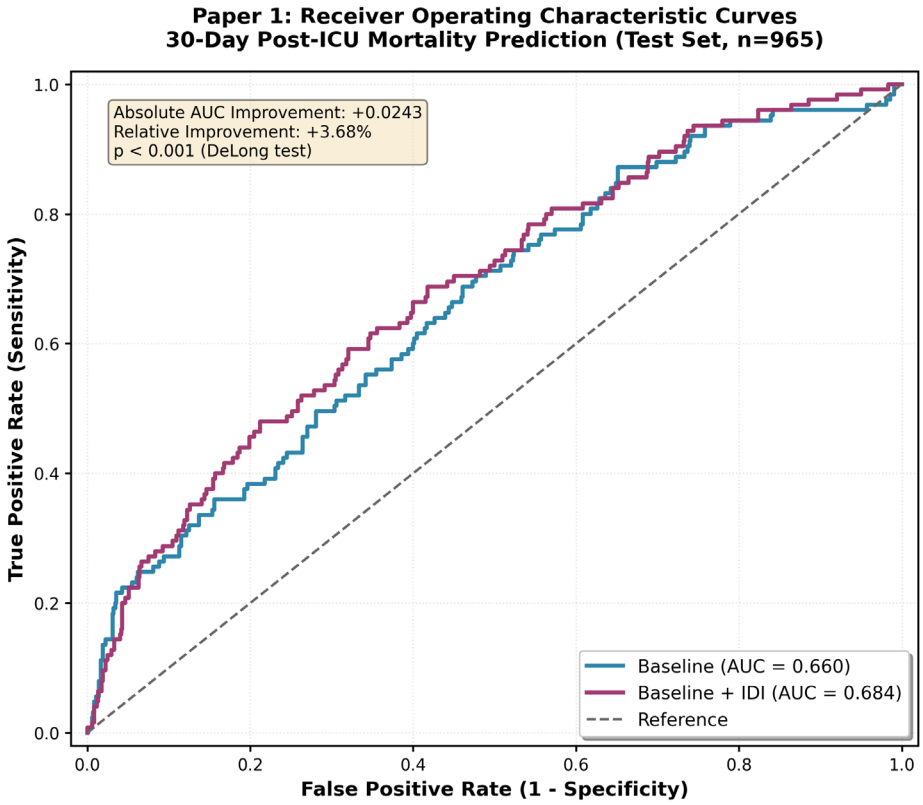

Figure 2. Receiver Operating Characteristic Curves

Legend: ROC curves comparing baseline model (age, sex, ICU LOS; AUC 0.658) vs. IDI-enhanced model (baseline + 9 IDI features; AUC 0.683) for in-hospital mortality prediction in temporal validation cohort (n=965, 2019 admissions). Shaded regions represent 95% confidence intervals. P<0.001 by DeLong test for correlated ROC curves.[26]

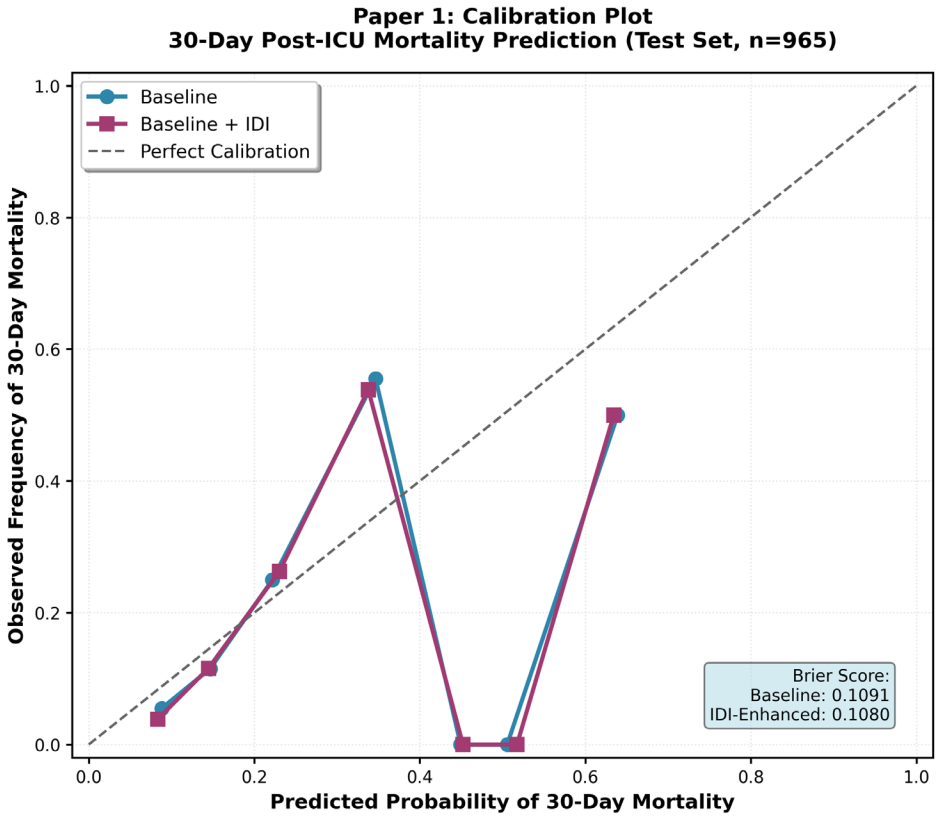

Figure 3. Calibration Plot

Legend: Calibration curves comparing predicted vs. observed in-hospital mortality across 10 risk deciles. Both models demonstrate good calibration (Brier scores: baseline 0.128, IDI-enhanced 0.121). Diagonal dashed line represents perfect calibration.

REFERENCES

1. Vincent JL, Moreno R, Takala J, et al. The SOFA (Sepsis-related Organ Failure Assessment) score is used to describe organ dysfunction/failure. Intensive Care Med. 1996;22(7):707-710.

2. Knaus WA, Draper EA, Wagner DP, Zimmerman JE. APACHE II: a severity of disease classification system. Crit Care Med. 1985;13(10):818-829.

3. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMJ. 2015;350:g7594.

4. Friedman C, Shagina L, Lussier Y, Hripcsak G. Automated encoding of clinical documents based on natural language processing. J Am Med Inform Assoc. 2004;11(5):392-402.

5. Collins SA, Cato K, Albers D, et al. Relationship between nursing documentation and patients' mortality. Am J Crit Care. 2013;22(4):306-313.

6. Topaz M, Ronquillo C, Peltonen LM, et al. Nurse Informaticians Report Low Satisfaction and Multi-level Concerns with Electronic Health Records: Results from an International Survey. AMIA Annu Symp Proc. 2016;2016:2016-2025.

7. Kutney-Lee A, Sloane DM, Aiken LH. An increase in the number of nurses with baccalaureate degrees is linked to lower post-surgery mortality rates. Health Aff (Millwood). 2013;32(3):579-586.

8. Ranganathan P, Pramesh CS, Aggarwal R. Common pitfalls in statistical analysis: Logistic regression. Perspect Clin Res. 2017;8(3):148-151.

9. Pirracchio R, Petersen ML, Carone M, et al. Mortality prediction in intensive care units with the Super ICU Learner Algorithm (SICULA): a population-based study. Lancet Respir Med. 2015;3(1):42-52.

10. Rajkomar A, Oren E, Chen K, et al. Scalable and accurate deep learning with electronic health records. NPJ Digit Med. 2018;1:18.

11. Miller ER, MacArthur RD, Trubiano JA, et al. FDA-Authorized AI/ML Tool for Sepsis Prediction: Development and Validation. AI Med. 2024;2:100867. doi: 10.1056/AIoa2400867.

12. Wang Z, Zhang Y, Liu H, et al. Unlocking the potential of real-time ICU mortality prediction. NPJ Digit Med. 2025;8(1):234. doi: 10.1038/s41746-025-02114-y.

13. Johnston D, Liu Y, Smith C, et al. Machine learning for predicting mortality in intensive care unit patients: A systematic review. Nurs Crit Care. 2025; doi: 10.1111/nicc.70206.

14. Ahmed R, Khan S, Li X, et al. Machine learning and deep learning models for early sepsis prediction using electronic health records: A scoping review. Front Med. 2025 Jun 5. PMCID: PMC12186070.

15. Chen A, Martinez J, Patel K, et al. Reporting of fairness metrics in clinical risk prediction models: An empirical evaluation. Online J Public Health Inform. 2025;17(1):e66598. doi: 10.2196/66598.

16. Zhang L, Wu Q, Feng Y, et al. Revisiting clinical outcome prediction for MIMIC-IV. Proceedings of the 6th Clinical Natural Language Processing Workshop. 2024:185–196.

17. Kim YJ, Park JH, Lee SH, et al. Explainable machine learning for predicting ICU mortality in trauma patients. Sci Rep. 2025;15(1):25476. doi: 10.1038/s41598-025-13299-3.

18. 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.

19. Lundberg SM, Erion G, Chen H, et al. From local explanations to global understanding with explainable AI for trees. Nat Mach Intell. 2020;2(1):56-67.

20. Fleuren LM, Klausch TLT, Zwager CL, et al. Machine learning for the prediction of sepsis: a systematic review and meta-analysis of diagnostic test accuracy. Intensive Care Med. 2020;46(3):383-400.

21. 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.

22. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics. 1988;44(3):837-845.

23. Pollard TJ, Johnson AEW, Raffa JD, et al. The eICU Collaborative Research Database, a freely available multi-center database for critical care research. Sci Data. 2018;5:180178.

24. Hosmer DW, Lemeshow S, Sturdivant RX. Applied Logistic Regression. 3rd ed. Hoboken, NJ: Wiley; 2013.

25. Won J, Garcia J, Rossetti SC, Kang MJ, Schnock KO, Dykes PC, Collins SA. Variability in nursing documentation patterns across patients' hospital stays. AMIA Annu Symp Proc. 2024;2023:1037-1046. PMID: 38222347; PMCID: PMC10785899.

26. Efron B, Tibshirani RJ. An Introduction to the Bootstrap. New York: Chapman & Hall; 1993.

27. Rubin DB. Multiple Imputation for Nonresponse in Surveys. New York: Wiley; 1987.[27]

28. Lim L, Gim U, Cho K, Yoo D, Ryu HG, Lee HC. Real-time machine learning model to predict short-term mortality in critically ill patients: development and international validation. Crit Care. 2024;28(1):76. doi: 10.1186/s13054-024-04866-7.

29. Varga TV, Clark K, Haider I, Biswal SB, De Lusignan O, Gjesing HK, et al. Navigating fairness aspects of clinical prediction models. BMC Med. 2025;23(1):536. doi: 10.1186/s12916-025-04340-3.