# Multimodal Detection of Surgical Site Infections from Electronic Health Records

Combining structured clinical data with physician notes for automated infection surveillance

Abhyuday Roychowdhury

Bluevia Health · March 2026

**Abstract** Surgical site infections affect one in twenty surgical patients and cost healthcare systems billions annually, yet automated surveillance remains an unsolved problem. We present a machine learning system that fuses structured clinical data (laboratory values, medications, microbiology, procedures) with natural language processing on physician discharge notes to detect SSIs from routinely collected electronic health records. The central finding is that these two data modalities are complementary: structured data captures what was ordered and measured, while clinical text captures what physicians observed and concluded. Their fusion achieves an AUROC of 0.91 and a precision-recall AUC 21 times above random baseline, detecting three in four infections while maintaining over 90% specificity. Feature attribution analysis confirms predictions are driven by clinically recognised risk factors.

## Background

Surgical site infections (SSIs) are among the most common healthcare-associated complications. Between two and five percent of patients undergoing inpatient surgery develop an SSI, extending hospital stays, doubling readmission risk, and increasing mortality by 2 to 11 times [1]. The annual cost to the US healthcare system exceeds $3.5 billion.

Current surveillance depends on trained infection preventionists manually reviewing patient charts against National Healthcare Safety Network criteria. This process is labour-intensive, inconsistent across institutions, and cannot scale. The automated alternative, administrative billing codes, performs poorly: a large multi-hospital study found that ICD-based detection captured only 10% of confirmed SSIs [2].

Machine learning has shown promise, but a systematic review of 24 studies and 85 models noted that most published approaches rely solely on structured data, neglect probability calibration, and lack external validation [1]. The highest-performing systems combine structured features with natural language processing on clinical notes [3, 4], suggesting these modalities carry complementary information. This hypothesis is the foundation of our work.

## Data

We used MIMIC-IV [5], a de-identified EHR dataset from Beth Israel Deaconess Medical Center covering 546,028 hospital admissions across approximately 180,000 patients (2008 to 2019), with 331,793 discharge summary notes.

SSI cases were identified through ICD diagnosis codes, yielding 7,918 positive cases at 1.45%

prevalence. This roughly 69:1 class imbalance shaped every modelling decision, from training objective to evaluation metric. All experiments used patient-level train/test splits to prevent information leakage.

## Approach

Our system extracts features from two complementary data sources and combines them through feature-level fusion (Figure 1).

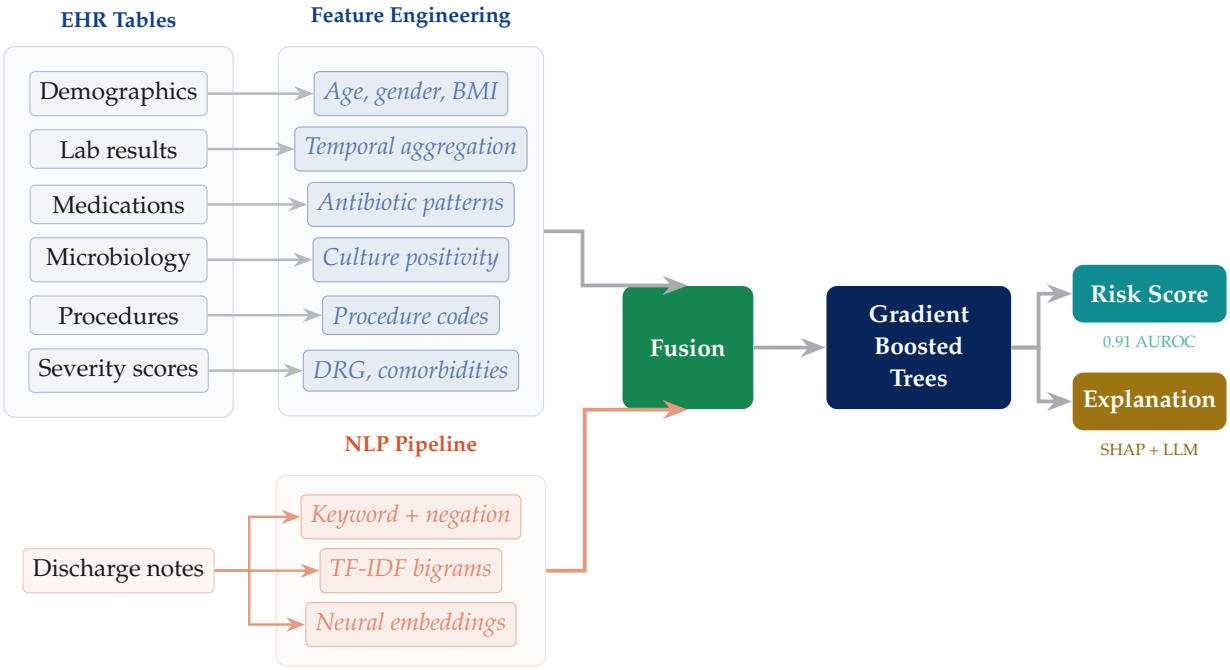


**Figure 1:** System architecture. Structured EHR data passes through feature engineering (temporal aggregation, pattern extraction). Discharge notes are processed through three parallel NLP pipelines. Both modalities are fused for classification with explainable outputs.

### Structured clinical features

From ten clinical data tables, we engineered features spanning patient demographics, temporal laboratory trajectories (capturing how values change over the course of an admission), comorbidity profiles, medication patterns including antibiotic regimens, microbiology culture results, surgical procedures, and clinical severity indicators. Particular attention was paid to signals that reflect the progression of infection: rising inflammatory markers, escalating antibiotics, and positive wound cultures.

### Clinical note analysis

Discharge summaries contain clinical reasoning absent from structured fields. We targeted the "Brief Hospital Course" section, which documents complications, infections, and treatment decisions.

Three parallel text representation strategies were applied. Statistical term features (TF-IDF with bigrams) capture the vocabulary of infection documentation. Dense neural embeddings encode broader semantic meaning beyond individual terms. A negation-aware keyword system distinguishes "wound infection requiring debridement" from "no evidence of wound infection."

These three representations capture lexical, semantic, and domain-specific signals respectively.

### Multimodal fusion

The structured and text feature vectors were concatenated and fed to a gradient-boosted tree classifier. Decision thresholds were optimised for the F2 score, weighting recall above precision to reflect the clinical reality that missed infections carry greater cost than false alarms.

## Results

**0.91**
AUROC

**21×**
PR-AUC lift over baseline

**75%**
sensitivity at 91% specificity

We evaluated the system by progressively incorporating richer clinical signal.

**Table 1:** Classification performance. Multimodal fusion substantially outperforms either modality alone.

| Configuration | AUROC | PR-AUC | Sensitivity | Specificity | F2 |
|---|---|---|---|---|---|
| Structured features only | 0.881 | 0.195 | 0.812 | 0.877 | 0.314 |
| Clinical notes only | 0.841 | 0.230 | 0.617 | 0.917 | 0.306 |
| **Multimodal fusion** | **0.914** | **0.321** | **0.754** | **0.912** | **0.357** |

Random baseline PR-AUC at 1.45% prevalence: 0.015. All metrics on held-out test set with patient-level splits.

**Clinical text carries independent predictive signal.** A model trained only on discharge notes, without any structured data, achieved an AUROC of 0.84. Physician documentation encodes substantial diagnostic information that never appears in structured fields.

**Fusion outperforms either modality alone.** Combining structured features with text representations lifted PR-AUC from 0.20 to 0.32, a 64% relative improvement. Structured data captures what was ordered and measured; text captures what clinicians observed and reasoned about. The two sources are genuinely complementary.

**Performance is clinically meaningful.** The fused model detects three in four SSI cases while maintaining over 91% specificity. The PR-AUC of 0.32 represents a 21-fold improvement over random baseline, a substantial gain for a condition with 1.45% prevalence.

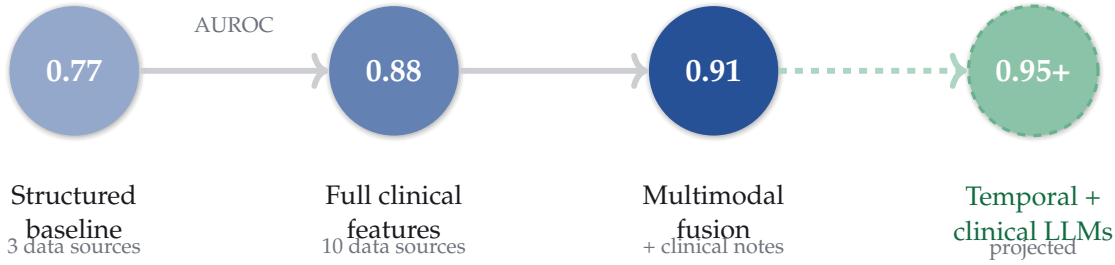


**Figure 2:** Performance progression. Each iteration added richer clinical signal. Temporal modelling and clinical language models represent the projected next phase.

### Literature context

Several recent systems achieve higher absolute AUROC through complementary approaches. Kiser et al. [3] reached 0.954 using temporal LSTMs that model postoperative trajectories over 30 days. Chakraborty et al. [4] achieved 0.98 with gold-standard NHSN labels validated through manual chart review. Bonde et al. [2] reported 0.989 on 389,000 cases with deep NLP. These results demonstrate the ceiling that temporal modelling, validated labels, and scale can unlock. Our contribution is orthogonal: we show that multimodal data fusion delivers strong gains even with standard models and automated labels, and these approaches can be combined.

## What Drives Predictions

Clinical adoption requires interpretable outputs. We applied SHAP (SHapley Additive exPlanations) to decompose each prediction into individual feature contributions (Figure 3).

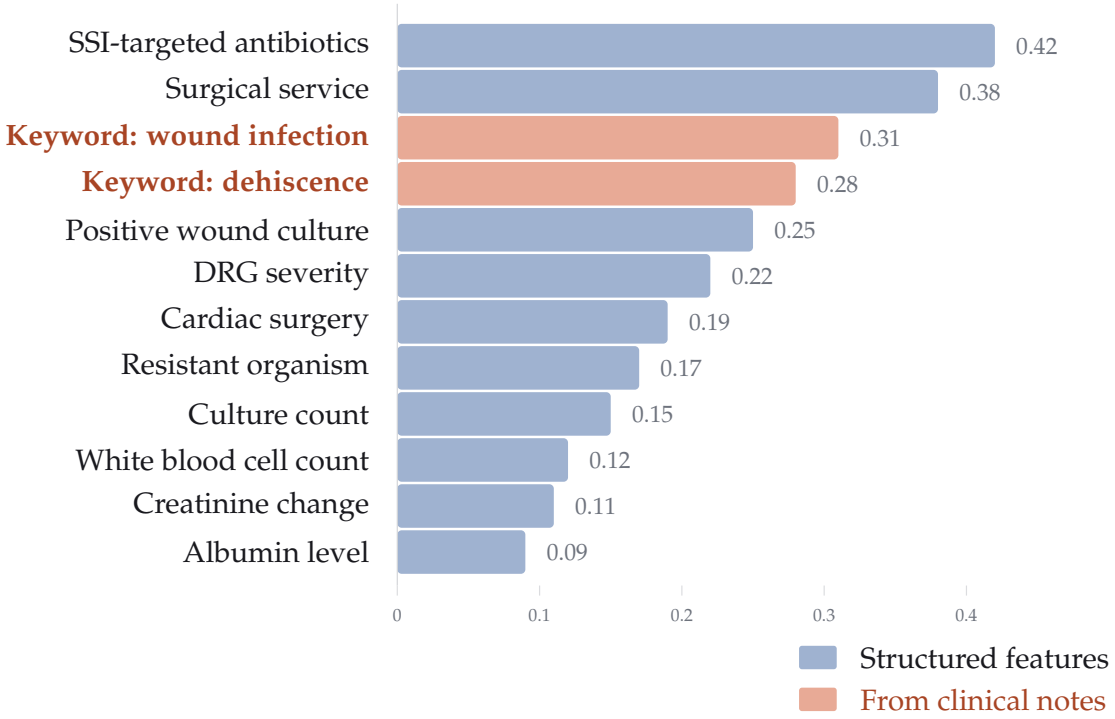


**Figure 3:** Top 12 predictive features by importance. Note-derived features (orange) rank alongside the strongest structured signals, validating the multimodal approach.

The ranking is clinically intuitive. SSI-targeted antibiotics (vancomycin, piperacillin, meropenem) rank highest: antibiotic escalation is a strong proxy for suspected infection. Surgical service type and positive wound cultures follow, both well-established risk factors [6].

Critically, NLP-derived features appear prominently. Affirmed mentions of "wound infection" and "dehiscence" extracted from discharge notes rank among the top five predictors, alongside structured clinical signals. This validates the core hypothesis: structured data and clinical text carry different, complementary information about patient state. Neither modality alone captures the full picture.

## Case Study

To illustrate the system's reasoning, we present a representative detection from the held-out test set (Figure 4).

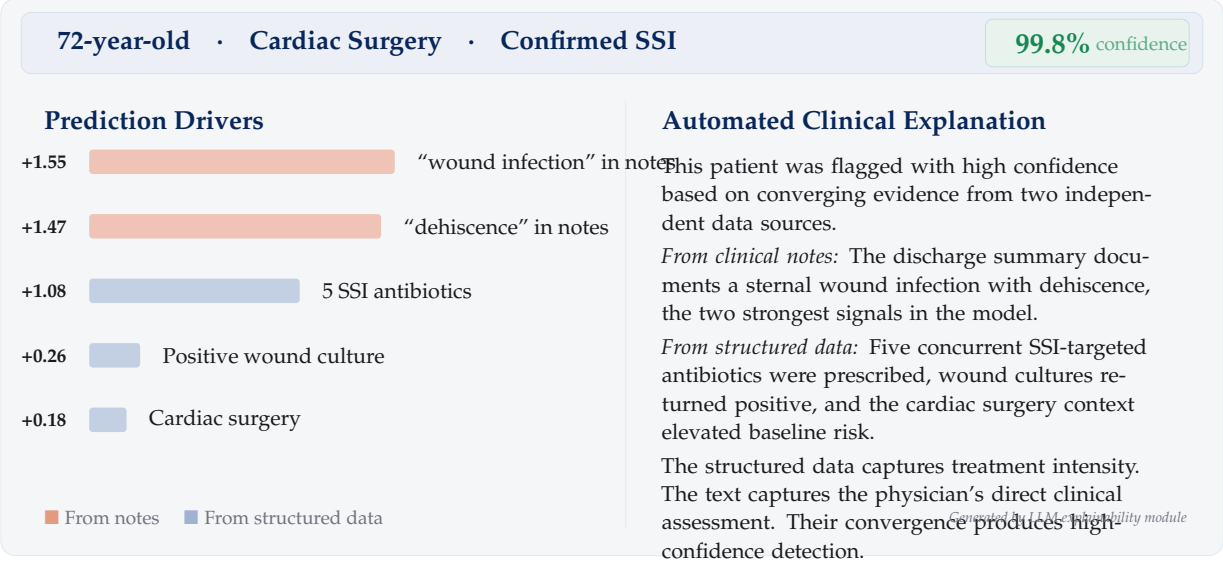


**Figure 4:** True positive from the held-out test set. Left: feature contributions. Right: automated clinical explanation generated by the LLM explainability module.

## Outlook

Several directions would strengthen this approach further. Temporal modelling of postoperative laboratory trajectories, particularly inflammatory markers over the first seven days, would capture time-dependent signals that distinguish developing infections from normal recovery [3]. Fine-tuned clinical language models with longer context windows would encode richer note semantics. Prospective validation on live clinical data would establish operational characteristics for deployment.

The finding that matters most extends beyond SSI detection. Structured clinical data and physician notes encode fundamentally different aspects of patient state: one captures what was ordered and measured, the other what was observed and reasoned about. Fusing them consistently outperforms either in isolation. This principle applies wherever clinical decisions are documented in both structured fields and free text, which describes nearly every area of modern healthcare.

## References

[1] Rogier van Boekel, Sjoerd van der Meijden, et al. Machine learning for surgical site infection prediction: a systematic review. *PLOS ONE*, 2024. 24 studies, 85 ML models reviewed; ML did not consistently outperform logistic regression.

[2] Alexander Bonde et al. Deep NLP-based surveillance of surgical site infections: a multi-center study. *Frontiers in Digital Health*, 2024. AUROC 0.989 on 389,865 surgical cases across 11 Danish hospitals.

[3] Anna C Kiser et al. Surgical site infection prediction using attention-based LSTM models with electronic health records and clinical notes. *Surgery*, 2024. AUROC 0.954 on 9,185 operative events from the University of Utah.

[4] Ritwik Chakraborty et al. Neural network-based surgical site infection prediction combining structured EHR data with clinical notes. *Surgical Infections*, 2025. AUROC 0.98 on 28,864 procedures across NHSN and NSQIP registries.

[5] Alistair Johnson et al. MIMIC-IV: a freely accessible electronic health record dataset. *Scientific Data*, 10: 1, 2023.

[6] Yilin Zhuang et al. Parsimonious preoperative prediction of surgical site infections using LASSO with knockoff filter. *Annals of Surgery*, 2024. AUROC 0.73–0.89 using only 7 EHR variables on 30,639 patients.