CAUTI Feature Exclusion Rationale (Leakage Control)

Overview

In this project, **Catheter-Associated Urinary Tract Infection (CAUTI)** is identified using **ICD diagnosis codes assigned at discharge for billing purposes**.
As a result, **there is no reliable clinical timestamp for CAUTI diagnosis during the admission**.

Because of this, feature selection **cannot rely on temporal ordering** (before vs after diagnosis).
Instead, features must be evaluated using **causal and logical reasoning** to prevent **label leakage**.

This document explains **why specific feature categories must be excluded** from the modeling dataset.

Core Principle

> **A feature must be dropped if it could only exist because CAUTI occurred, or because clinicians knew CAUTI occurred.**

Even if such a feature appears in the dataset during the admission, it represents **post-event knowledge** and would not be available at prediction time.

1. Identifier Columns

Dropped Features

- `subject_id`
- `hadm_id`

Rationale

These are **technical identifiers**, not clinical signals.
They provide no predictive value and can introduce: - memorization - data leakage across splits - spurious correlations

2. Outcome & Administrative Features

Dropped Features

- `length_of_stay`
- all `discharge_location_*` columns

Rationale

These are **consequences of the admission**, often influenced by CAUTI itself. Using them would allow the model to indirectly infer the outcome.

Example: - CAUTI increases length of stay - Model learns "long stay CAUTI"

This violates causal modeling principles.

3. Diagnosis / Infection Flags (Hard Leakage)

Dropped Features

- `other_uti_present`
- `has_cauti_history` *(when derived from current admission)*

Rationale

These variables explicitly encode: - presence of UTI - knowledge of infection

They are **direct proxies for the label** and result in **near-perfect leakage**.

4. Treatment & Clinical Reaction Features

Dropped Features

- `catheter_removal`
- `catheter_removal_replacement`
- `antibiotics_per_admission`
- `recent_antibiotic_use`
- `pain_documented`

Rationale

These features represent **actions taken by clinicians in response to infection or suspicion of infection**.

They cannot occur in a counterfactual world where CAUTI never happened.

Using them allows the model to learn: > "Treatment was given → therefore infection exists"

This is invalid for prediction.

5. Laboratory & Microbiology Features (Soft Leakage)

Dropped Features

- `urinalysis_wbc`
- `urinalysis_rbc`
- `blood_wbc`

- `creatinine`
- `procalcitonin_measured`
- `urine_culture_performed`
- `blood_culture_performed`
- `gram_negative_organisms_present`
- `gram_positive_organisms_present`
- `fungi_present`
- `blood_crp_measured`
- `cfu_count_measured`

Rationale

These features reflect **physiological response to infection** or **diagnostic confirmation**.

Because CAUTI diagnosis time is unknown: - Labs may have been drawn **after infection onset** - Cultures are often ordered **because infection is suspected**

Including them would introduce **silent leakage**, even if timestamps exist.

6. Physiological Measurements & Monitoring

Dropped Features

- `oliguria`
- `urine_output_measured`

Rationale

These represent **clinical deterioration or monitoring triggered by illness**. They are downstream effects rather than baseline risk factors.

7. Vitals (Post-Infection Effects)

Dropped Features

- `temperature`
- `heart_rate`
- `resp_rate`
- `o2sat`
- `bp_systolic`
- `bp_diastolic`

Rationale

Vital sign abnormalities often occur **after infection onset**. Without a diagnosis timestamp, they cannot be safely constrained to a pre-infection window.

Using full-admission vitals introduces **temporal leakage**.

8. Diagnostic Test Indicators

Dropped Features

- `nitrite_tested`
- `nitrite_positive`

Rationale

Urinalysis nitrite testing is typically ordered **because UTI is suspected**. These variables encode clinician suspicion and should not be used as predictors.

Summary Table

Category	Reason for Dropping
Identifiers	Non-predictive
Outcomes	Consequences of CAUTI
Diagnosis flags	Encode label directly
Treatments	Reaction to infection
Labs & cultures	Reflect infection effects
Vitals	Post-onset physiology
Tests ordered	Encode clinical suspicion

Final Modeling Philosophy

Because CAUTI labeling is retrospective:

- **All features must be valid at admission or during routine care**
- **No feature may depend on knowing CAUTI occurred**
- **Causality takes precedence over availability**

This ensures the resulting model: - is clinically defensible - avoids label leakage - reflects real-world deployment constraints - withstands peer review and audit scrutiny