CogniCrest

# Anonymization Pipeline — Technical Benchmark Report

Healthtech · Clinical Data for Research Partnerships · synthetic data, generated for demonstration purposes

**Important note:** This report uses a fully synthetic dataset generated for demonstration purposes only — no real customers, patients, policyholders, or employees are involved, and no real company's data was used. This is a technical benchmark, not a client case study.

## Clinical Data for Research Partnerships

**PROBLEM** A diagnostics provider wants to share imaging metadata and outcomes with a research partner without exposing patient identity.

**APPROACH** De-identify direct identifiers, generalize quasi-identifiers like age and location, validate against k-anonymity thresholds before data leaves the boundary.

**GOAL** Research moves forward. Patient identity stays protected.

## Methodology

A synthetic healthtech dataset of 420 patient records was generated, following the end-to-end process common to clinical data anonymization: ingestion, schema normalization, field classification, policy selection, privacy transformation, risk evaluation, utility evaluation, and a governance gate before delivery. Direct identifiers were tokenized. Quasi-identifiers (age, gender, region, diagnosis category) were generalized using an adaptive suppression ladder until every record achieved k-anonymity of at least 5. Free-text clinician notes were screened for embedded names and addresses and redacted separately from the structured fields.

## Field Classification

Fields are classified into four categories: direct identifiers, quasi-identifiers, sensitive attributes, and operational fields.

| Field | Classification |
|---|---|
| patient_id | Direct identifier |
| full_name | Direct identifier |
| email | Direct identifier |
| phone | Direct identifier |
| street_address | Direct identifier |
| age | Quasi-identifier |

| | |
|---|---|
| gender | Quasi-identifier |
| zip_code | Quasi-identifier |
| diagnosis_category | Quasi-identifier |
| treatment_outcome | Sensitive attribute |
| treatment_cost | Sensitive attribute |
| clinician_note | Sensitive attribute (free text) |
| prior_visits_12mo | Operational field |
| high_risk_outcome | Target / operational field |

## Text Redaction — Clinician Notes

Free-text clinical notes are a well-known privacy gap in healthcare data: patient names, home addresses, and case details are routinely typed directly into notes fields, bypassing whatever protection was applied to the structured columns. The pipeline screens narrative fields separately and redacts embedded direct identifiers before the note is included in any anonymized release.

### *BEFORE:*

```
Patient Emma Wilson, residing at 118 Birch Lane, Denver, presented for cardiovascular evaluation.
Outcome: Stable. Prior visits (12mo): 4.
```

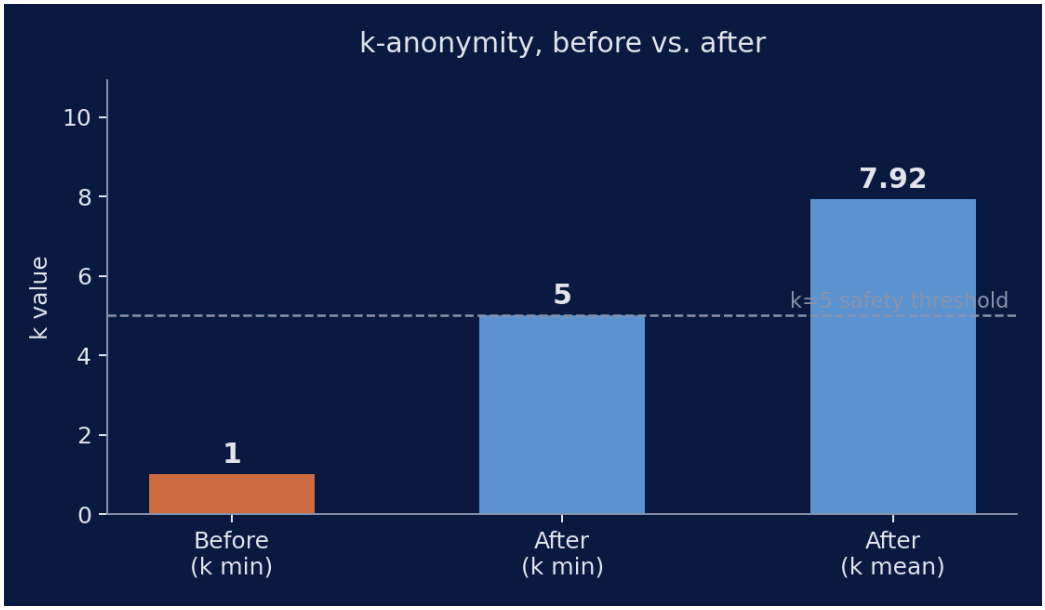
### *AFTER:*

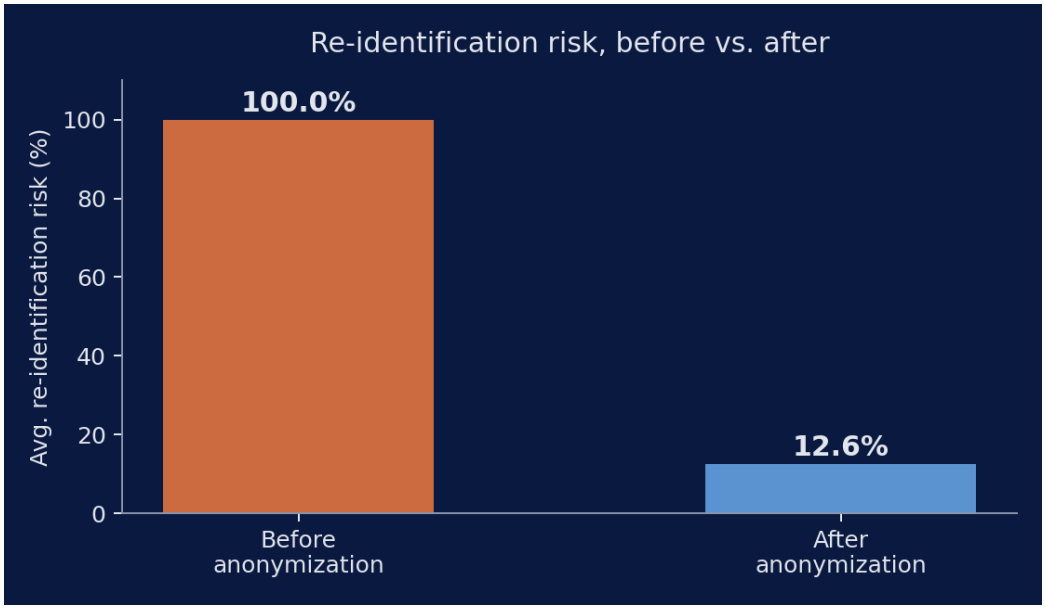
```
Patient [NAME REDACTED], residing at [LOCATION REDACTED], Denver, presented for cardiovascular
evaluation. Outcome: Stable. Prior visits (12mo): 4.
```

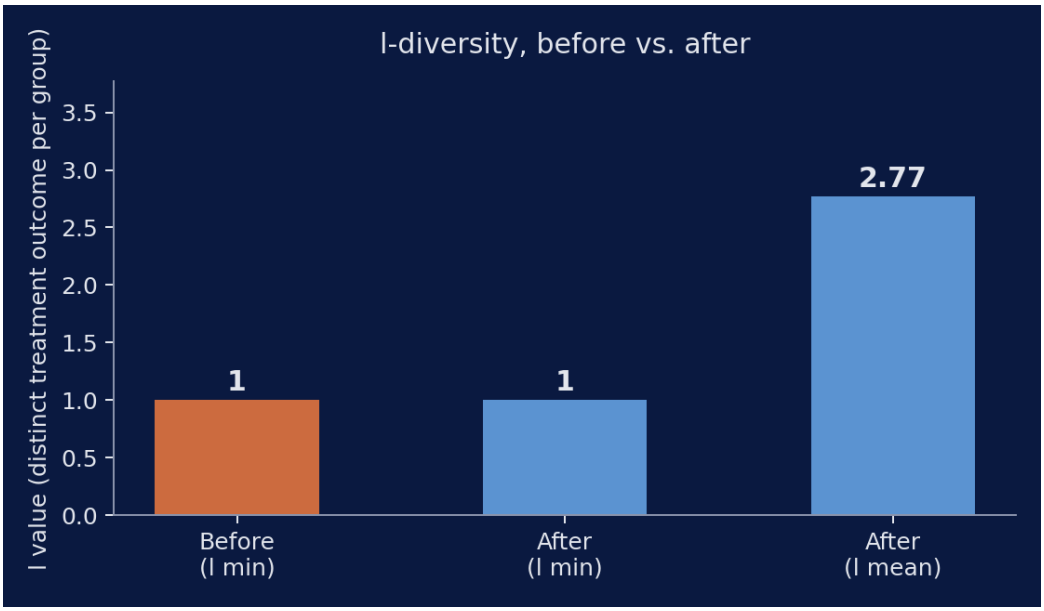
Note that city-level location ("Denver") was deliberately retained — it's already generalized to region elsewhere in the pipeline and provides useful signal for clinical research, while the specific street address (a direct identifier on its own) is removed.

## Anonymity Metrics

Two metrics are reported together deliberately. k-anonymity alone can be satisfied while still leaking sensitive information: if every record in a group shares the same sensitive value, an attacker who narrows a target to that group learns the answer with certainty even without picking out the exact record. l-diversity checks for that specific gap by measuring how many distinct treatment outcome values exist inside each k-anonymous group.






| Metric | Before | After |
|---|---|---|
| Minimum k (k-min) | 1 | 5 |
| Mean k (k-mean) | 1.0 | 7.92 |
| Average re-identification risk | 100.0% | 12.6% |
| Minimum l-diversity (l-min) | 1 | 1 |
| Mean l-diversity (l-mean) | 1.0 | 2.77 |
| Groups with l=1 (attribute disclosure risk remains) | 100.0% | 3.8% |

**Honest finding:** 3.8% of equivalence classes still have l=1 after anonymization — meaning k-anonymity was satisfied for those groups but attribute disclosure risk on treatment outcome was not fully eliminated. This is a known, real limitation of k-anonymity alone, not a bug: it's exactly why both metrics are measured and reported together rather than stopping at k-anonymity and calling the dataset safe. A production release gate would flag these specific residual groups for additional generalization or exclusion before delivery to a recipient sensitive to attribute disclosure.

## Utility Metrics

### *Coverage Retention*

100% of input records are present in the anonymized output. Where a record's quasi-identifier combination couldn't reach k-anonymity through normal generalization, the pipeline generalized its fields further rather than deleting the row — no record is ever silently dropped.

### *KPI Fidelity*

Treatment costs are heavily right-skewed (true mean $5,530 driven up by a small number of high-cost cases; median is far lower). Two reconstruction methods were tested against this specific challenge:

| Method | Reconstructed Mean | Error % |
|---|---|---|
| Naive band midpoint (no aggregate published) | $7,471.63 | 35.1% |
| Published bucket mean (aggregation release) | $5,530.28 | 0.0% |

**Honest finding:** guessing from the band midpoint alone overestimates the true average treatment cost by 35.1% — even quantile-based bins leave some distortion because the top bin still spans a long tail of rare, high-cost cases. This is exactly why generalized row-level data (used for clinical research datasets) and aggregate analytics releases (used for cost and utilization reporting) are treated as separate outputs with separate techniques, not one dataset trying to serve both. Publishing the true per-bucket average alongside the generalized label — standard practice for aggregate reporting releases — reconstructs the KPI exactly, with no row-level values exposed.

---