CogniCrest

Anonymization Pipeline — Technical Benchmark Report

Insurtech · Claims Data for Actuarial Modeling · 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.

Claims Data for Actuarial Modeling

PROBLEM An insurer needs to share claims history with an external actuarial consultant, but policyholder PII can't leave their systems unprotected.

APPROACH Anonymize policyholder identifiers and sensitive claim details while preserving the statistical patterns actuarial models depend on.

GOAL Modeling proceeds on schedule. Policyholder data stays in bounds.

Methodology

A synthetic insurtech dataset of 430 policyholder records was generated, following the end-to-end process common to insurance anonymization platforms: 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, occupation, household size) were generalized using an adaptive suppression ladder until every record achieved k-anonymity of at least 5. Free-text adjuster 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
policyholder_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
occupation	Quasi-identifier
household_size	Quasi-identifier
injury_category	Sensitive attribute
fraud_flag	Sensitive attribute
claim_amount	Sensitive attribute
adjuster_note	Sensitive attribute (free text)
policy_type	Operational field
vehicle_type	Operational field
premium	Operational field
prior_claims	Operational field
policy_tenure_years	Operational field
high_severity_claim	Target / operational field

Text Redaction — Adjuster Notes

Free-text claim narratives are a well-known privacy gap in insurance data: names, addresses, and incident details are often 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:

```
Contacted Sarah Johnson at 4821 Oak Street, Chicago regarding auto claim. Vehicle involved: Sedan. Injury noted: Whiplash. Prior claims on file: 2.
```

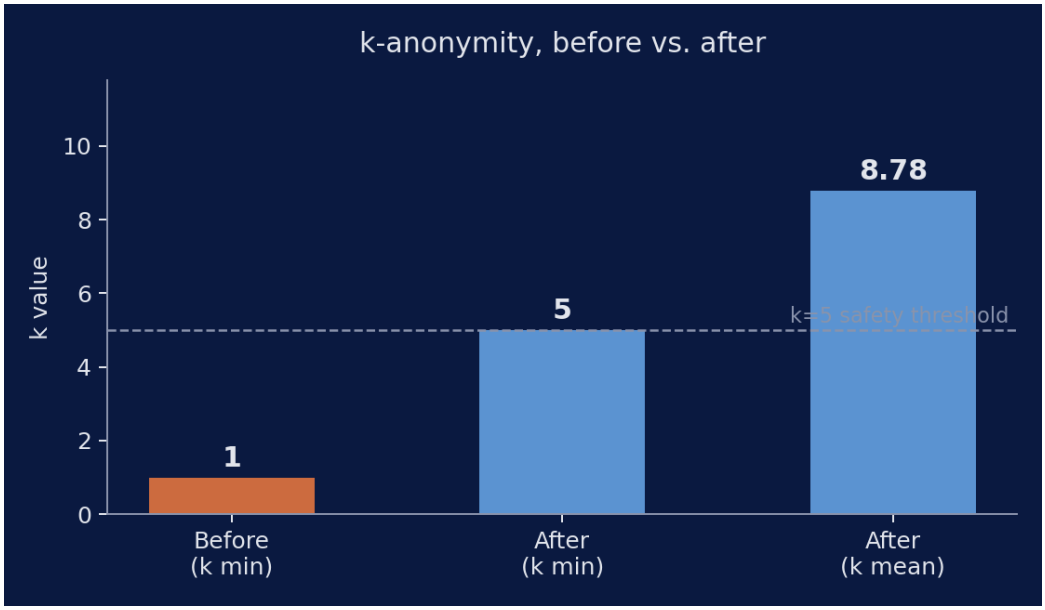
AFTER:

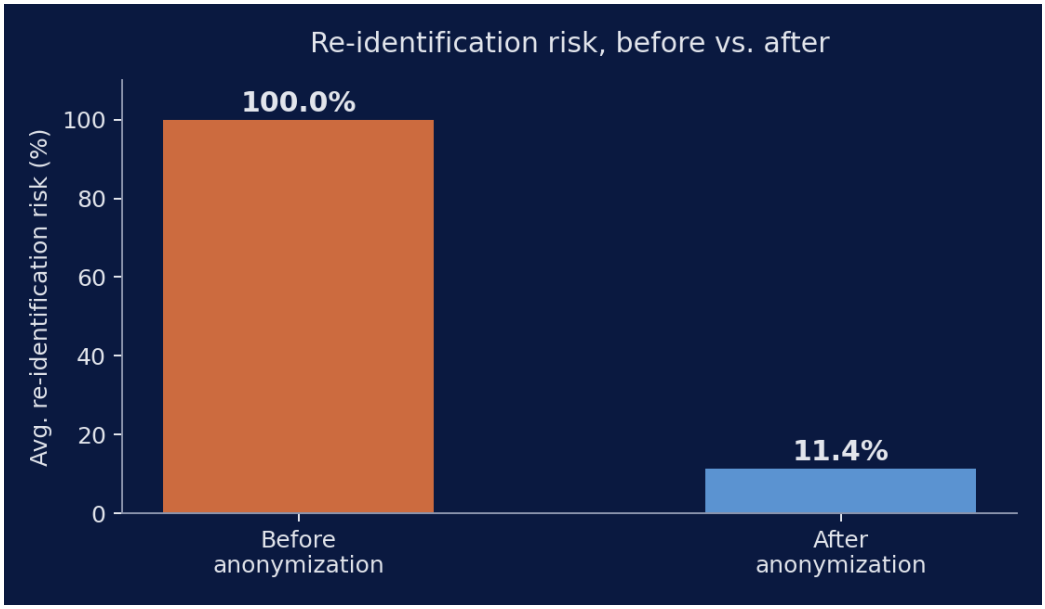
```
Contacted [NAME REDACTED] at [LOCATION REDACTED], Chicago regarding auto claim. Vehicle involved: Sedan. Injury noted: Whiplash. Prior claims on file: 2.
```

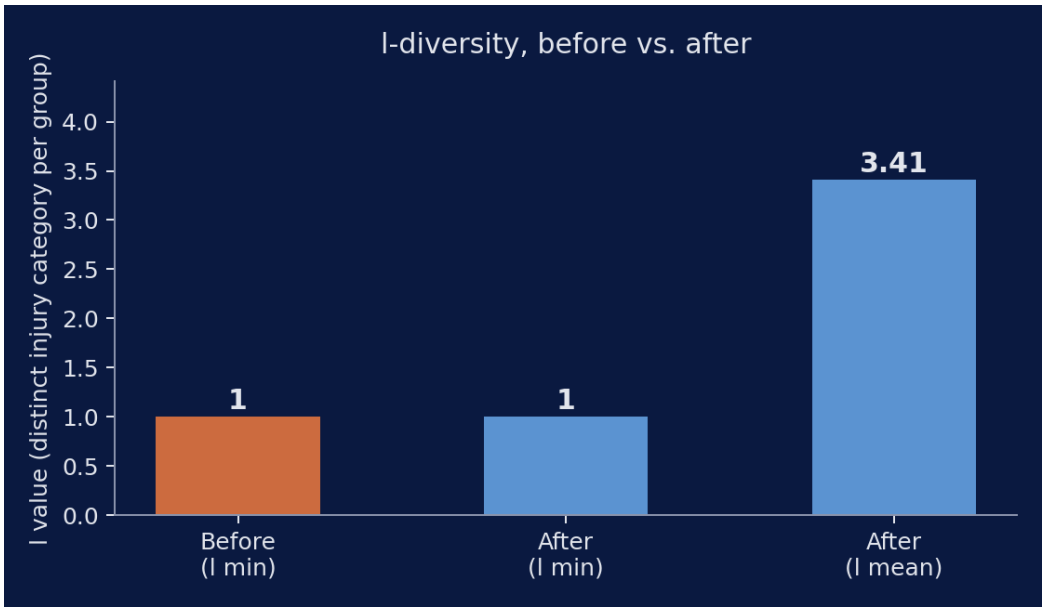
Note that city-level location ("Chicago") was deliberately retained — it's already generalized to region elsewhere in the pipeline and provides useful signal for claims analytics, while the specific street address (a direct identifier on its own) is removed.

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 injury category values exist inside each k-anonymous group.







Metric	Before	After
Minimum k (k-min)	1	5
Mean k (k-mean)	1.0	8.78
Average re-identification risk	100.0%	11.4%
Minimum l-diversity (l-min)	1	1
Mean l-diversity (l-mean)	1.0	3.41
Groups with l=1 (attribute disclosure risk remains)	100.0%	2.0%

Honest finding: 2.0% of equivalence classes still have l=1 after anonymization — meaning k-anonymity was satisfied for those groups but attribute disclosure risk on injury category 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

Claim amounts are heavily right-skewed (true mean $3,983 driven up by a small number of large claims; median is far lower). Two reconstruction methods were tested against this specific challenge:

Method	Reconstructed Mean	Error %
Naive band midpoint (no aggregate published)	$7,799.29	95.79%
Published bucket mean (aggregation release)	$3,983.49	0.0%

Honest finding: guessing from the band midpoint alone overestimates the true average claim amount by 95.79% — quantile-based bins don't fully fix this, because the top bin still spans a long tail of rare, large claims. This is exactly why generalized row-level data (used for fraud-detection model training) and aggregate analytics releases (used for portfolio and loss-trend 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.
