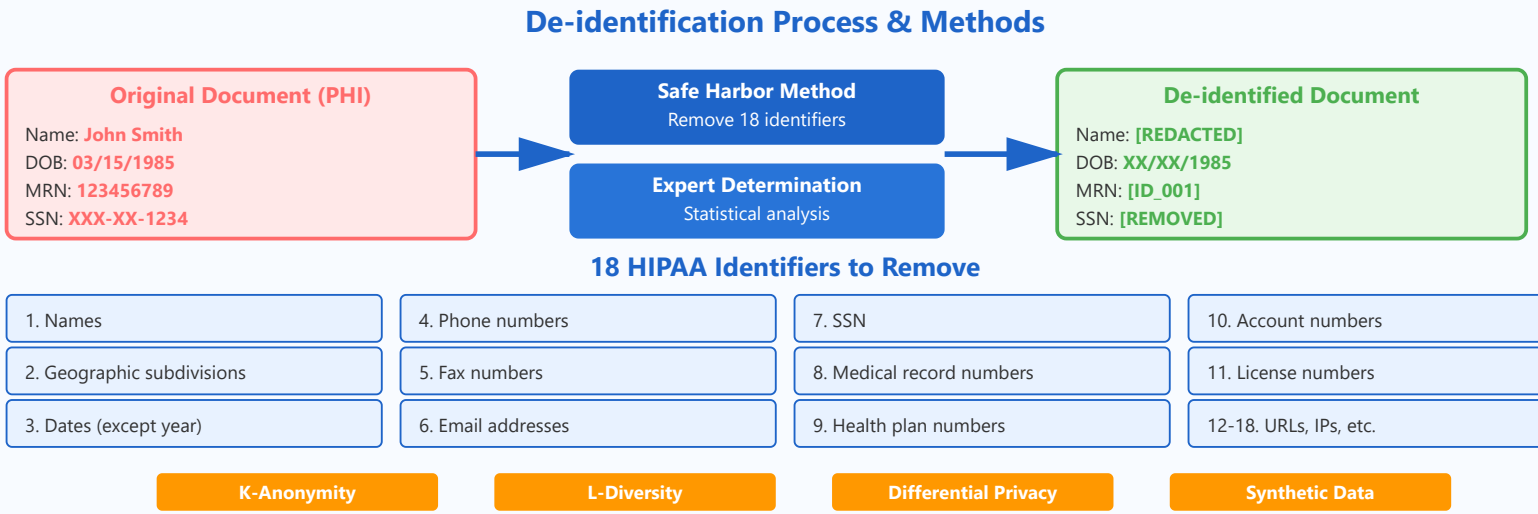
De-identification



Safe Harbor Method

- Remove all 18 identifiers
- Dates → year only
- Ages >89 → grouped as 90+
- Geographic: first 3 digits ZIP
- No statistical expertise needed

Expert Determination

- Statistical risk assessment
- Very small re-ID risk
- Retains more data utility
- Requires certified expert
- Document methodology

Automated Tools

- NER for PHI detection
- Philter, Scrubber, BoB
- Date shifting algorithms
- Validation required
- Hybrid human-AI review

Privacy Techniques

- K-anonymity (grouping)
- L-diversity (variety)
- Differential privacy (noise)
- Synthetic data (GANs)
- Utility vs privacy tradeoff

Safe Harbor Method - Detailed Overview

What is the Safe Harbor Method?

The Safe Harbor method is a HIPAA-compliant de-identification approach that requires the removal or modification of 18 specific types of identifiers from health information. It provides a clear, rule-based framework that doesn't require statistical analysis or expert determination. Once all 18 identifiers are properly removed, the data is considered de-identified under HIPAA regulations.

Practical Example: Patient Record Transformation

BEFORE (Contains PHI)

Name: Sarah Johnson

AFTER (De-identified)

Name: [REMOVED]

DOB: June 15, 1978

Address: 456 Oak Street, Boston, MA 02101

Phone: (617) 555-0123

Email: sarah.j@email.com

MRN: 987654321

SSN: 123-45-6789

Diagnosis: Type 2 Diabetes

Visit Date: March 3, 2024

DOB: Year 1978 only

Address: Boston, MA 021XX

Phone: [REMOVED]

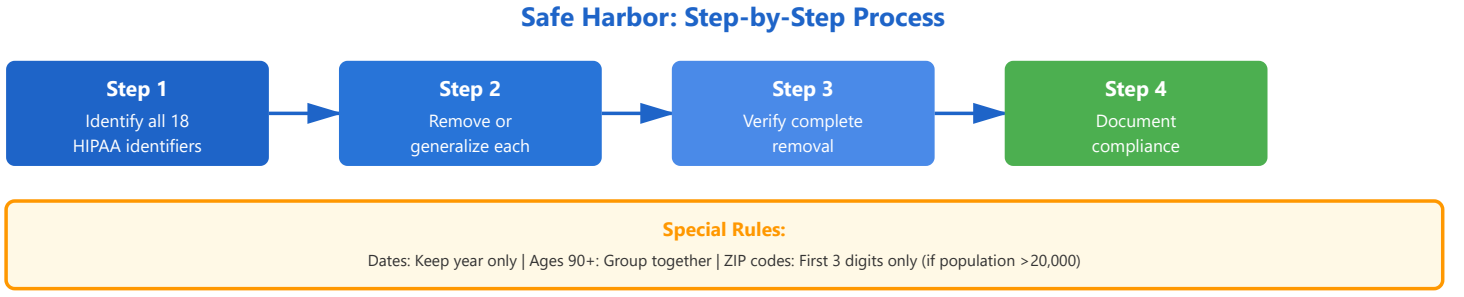
Email: [REMOVED]

MRN: [ID_12345]

SSN: [REMOVED]

Diagnosis: Type 2 Diabetes

Visit Date: Year 2024 only



📌 Key Considerations

- No statistical expertise or software required - follows clear rules
- Most conservative approach - may remove more data than necessary
- Dates can retain year for temporal analysis
- Small geographic areas (ZIP codes with <20,000 people) must be completely removed
- Ages over 89 must be aggregated to protect elderly individuals

✅ Advantages

✓ Clear, objective criteria

⚠️ Limitations

⚠ Reduces data utility

- ✓ No expert needed
- ✓ Legally defensible
- ✓ Easy to implement
- ✓ Widely accepted

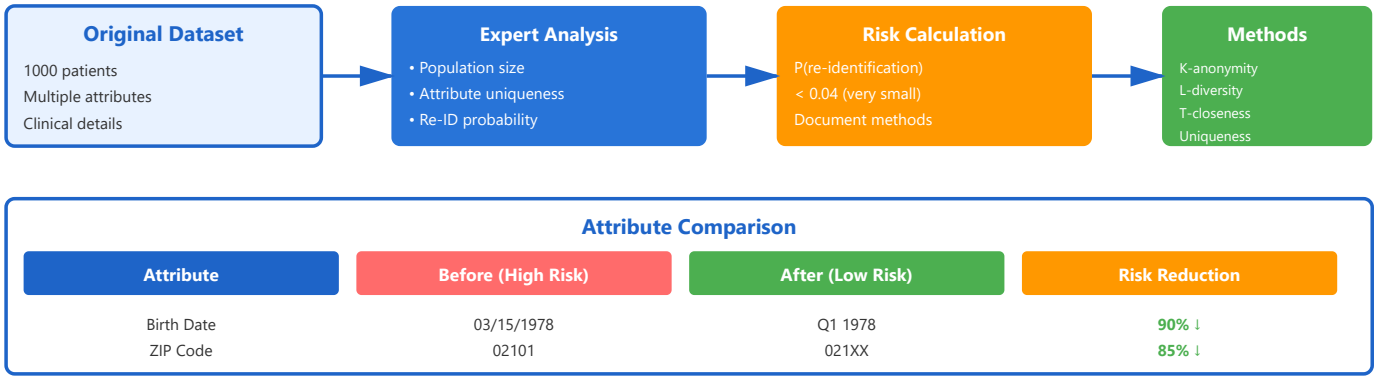
- ⚠ May be overly conservative
- ⚠ Limited temporal precision
- ⚠ Geographic restrictions
- ⚠ Not suitable for small datasets

📊 Expert Determination - Detailed Overview

What is Expert Determination?

Expert Determination is a HIPAA de-identification method that relies on statistical analysis by a qualified expert to assess re-identification risk. Unlike Safe Harbor's rigid rules, this approach allows for more flexible data retention while ensuring that the risk of re-identifying individuals is very small. The expert must have appropriate knowledge and experience in statistical and scientific methods for rendering information not individually identifiable.

🔍 Practical Example: Risk Assessment Process



Expert Qualification Requirements

Required Expertise

- Advanced degree in statistics or related field
- Experience in privacy risk analysis
- Knowledge of HIPAA regulations
- Understanding of re-identification methods
- Ability to document methodology

Deliverables

- Formal risk assessment report
- Statistical methodology documentation
- Probability calculations
- Justification for retained data
- Certification statement

Key Considerations

- Allows for more granular data retention than Safe Harbor
- Risk threshold: probability of re-identification must be "very small"
- Requires comprehensive documentation of methods and assumptions
- Must consider both internal and external data sources for linkage
- More expensive and time-consuming than Safe Harbor

✅ **Advantages**

- ✓ Retains more data utility
- ✓ Flexible approach
- ✓ Tailored to specific datasets
- ✓ Better for research needs
- ✓ Can adapt to context

⚠️ **Limitations**

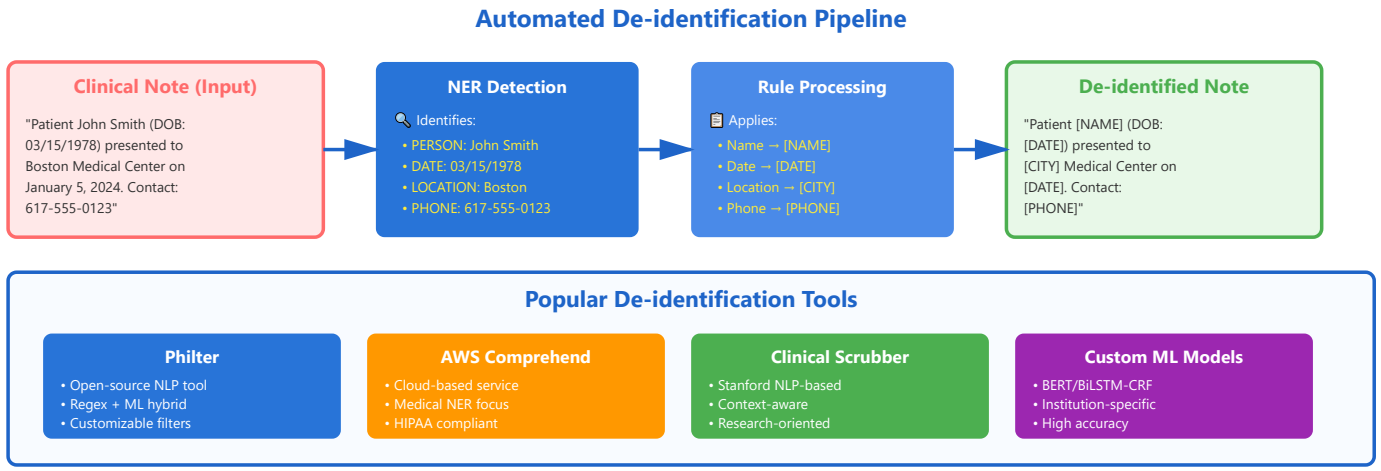
- ⚠ Requires qualified expert
- ⚠ Higher cost
- ⚠ Time-intensive process
- ⚠ Complex documentation
- ⚠ Subjective elements

Automated De-identification Tools - Detailed Overview

What are Automated De-identification Tools?

Automated de-identification tools use Natural Language Processing (NLP), machine learning, and rule-based algorithms to automatically detect and remove or mask PHI from unstructured clinical text. These tools can process large volumes of medical documents efficiently, identifying entities like names, dates, locations, and medical identifiers. They typically combine multiple techniques including Named Entity Recognition (NER), regular expressions, and dictionary lookups.

🔍 Practical Example: Automated Tool Pipeline



Common De-identification Techniques

🎯 Technique Types

Redaction: Complete removal → [REDACTED]

Substitution: Replace with fake → "John" → "Patient_A"

Date Shifting: Consistent offset → preserves intervals

Generalization: 02101 → 021XX

Pseudonymization: Reversible hashing

⚙️ Implementation

NER Models: BERT, BiLSTM-CRF, spaCy

Regex Patterns: Structured identifiers

Dictionary Lookup: Name databases

Context Analysis: Surrounding words

Validation: Human review + metrics

📌 Key Considerations

- No tool is 100% accurate - always requires validation and human oversight
- Performance varies by document type (clinical notes vs. discharge summaries)
- Context matters: "May" (month) vs. "may" (verb), "Paris" (name) vs. "Paris" (city)
- Trade-off between precision (avoiding false positives) and recall (catching all PHI)

- Regular updates needed to handle new patterns and edge cases

✅ Advantages

- ✓ Fast processing
- ✓ Scalable to large datasets
- ✓ Consistent application
- ✓ Reduces human error
- ✓ Cost-effective long-term

⚠️ Limitations

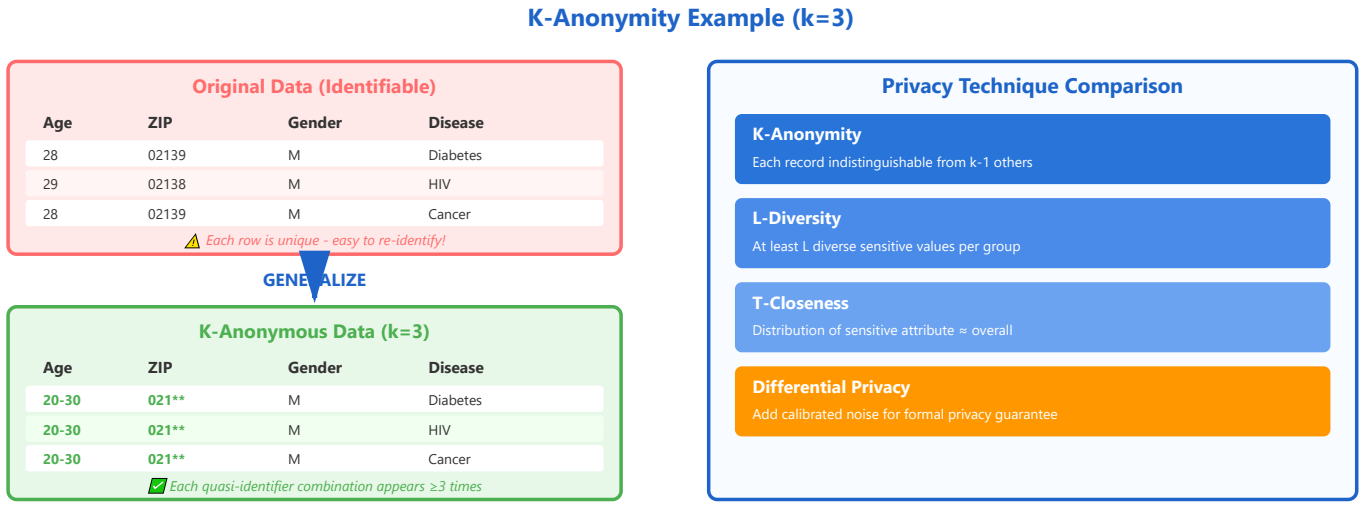
- ⚠ Not 100% accurate
- ⚠ Context challenges
- ⚠ Requires training data
- ⚠ Initial setup costs
- ⚠ Validation needed

⚖️ Advanced Privacy Techniques - Detailed Overview

What are Advanced Privacy Techniques?

Advanced privacy techniques go beyond simple identifier removal to provide mathematical guarantees about privacy protection. These methods address the challenge that even de-identified data can potentially be re-identified through linkage attacks or inference. Techniques like k-anonymity, l-diversity, differential privacy, and synthetic data generation provide formal privacy guarantees while attempting to maintain data utility for analysis and research.

🔍 Practical Example: K-Anonymity in Action



Detailed Technique Breakdown

1 K-Anonymity

Ensures each record is indistinguishable from at least k-1 other records based on quasi-identifiers (attributes that could be used for re-identification). Achieved through generalization and suppression.

How It Works

- Groups similar records together
- Generalizes attributes (age 28 → 20-30)
- Suppresses specific values when needed
- k=3 means groups of ≥3 records

Limitations

- Vulnerable to homogeneity attack
- Background knowledge attack possible
- Doesn't protect sensitive values
- Can reduce data utility significantly

2 L-Diversity

Extends k-anonymity by ensuring each equivalence class has at least L "well-represented" values for sensitive attributes. Protects against homogeneity and background knowledge attacks.

Example

Problem: Age 20-30, ZIP 021** all have HIV

Solution (L=3): Same group must have ≥3 different diseases

HIV, Diabetes, Cancer in each group

Prevents inference from group membership

Variants

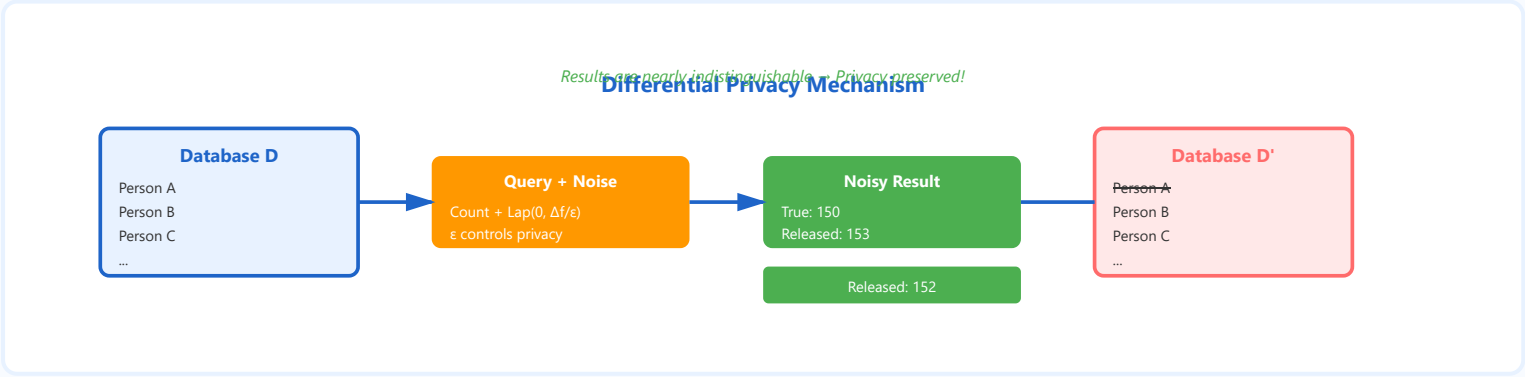
Distinct L-diversity: L distinct values

Entropy L-diversity: Shannon entropy ≥ log(L)

Recursive (c,L)-diversity: Most frequent value bounded

3 Differential Privacy

Provides the strongest mathematical privacy guarantee. Ensures that the presence or absence of any single individual's data doesn't significantly affect the output of an analysis. Achieved by adding carefully calibrated random noise.



Key Parameters

ε (epsilon): Privacy budget (smaller = more private)

Applications

- Census data (US Census 2020)

δ (delta): Probability of failure

Sensitivity: Max change from one record

Typical: ε = 0.1 to 1.0

- Medical statistics release
- Machine learning model training
- Location data aggregation

4 Synthetic Data

Creates artificial datasets that preserve statistical properties of the original data without containing any real individuals' information. Modern approaches use Generative Adversarial Networks (GANs) or Variational Autoencoders (VAEs).

Generation Methods

DP-GAN: GAN with differential privacy

PATE-GAN: Teacher ensemble approach

Bayesian Networks: Capture correlations

CTGAN: Conditional tabular GAN

Benefits

- No real patient data released
- Preserves statistical relationships
- Unlimited sharing possible
- Good for ML training

Key Considerations

- Trade-off between privacy and utility is fundamental - stronger privacy often means less useful data
- Multiple privacy attacks exist: linkage, inference, homogeneity, background knowledge
- No single technique is perfect - often combined for stronger protection
- Privacy guarantees depend on attacker's knowledge and capabilities
- Evaluation metrics: information loss, privacy risk, computational cost

Advantages

✓ Mathematical privacy guarantees

Limitations

⚠ Complex implementation

✓ Protection against linkage attacks

✓ Flexible trade-offs

✓ Research-proven methods

✓ Enables data sharing

⚠ Reduces data utility

⚠ Requires expertise

⚠ Computational overhead

⚠ Parameter tuning challenging