De-identification Protocol for Healthcare Data

Overview

This protocol outlines the procedures for de-identifying Protected Health Information (PHI) to comply with HIPAA Privacy Rule Safe Harbor method (45 CFR §164.514(b)(2)).

Purpose

To ensure that healthcare data used for AI training and research is properly de-identified, protecting patient privacy while maintaining data utility.

Scope

This protocol applies to all patient health and dental records that will be used for:

- AI model training
- Research and analysis
- Educational purposes
- Synthetic data generation

De-identification Methods

Safe Harbor Method

Remove all 18 HIPAA identifiers and ensure no actual knowledge that remaining information could identify an individual.

Expert Determination Method

Have a qualified expert determine that the risk of re-identification is very small.

For this project, we will use the Safe Harbor Method.

---

18 HIPAA Identifiers to Remove

1. Names

- □ Patient full name
- □ Patient nicknames or aliases
- □ Relative names
- □ Employer names
- □ Healthcare provider names (when serving as patient identifier)

Action: Replace with anonymous identifiers (e.g., P00001234)

2. Geographic Subdivisions Smaller than State

- ☐ Street address
- ☐ City
- ☐ County
- ☐ Precinct
- ☐ ZIP code (if population < 20,000)

Action:

- Remove or generalize to state level
- For ZIP codes: Keep only first 3 digits if combined population ≥ 20,000
- Use "00000" for all ZIP codes with population < 20,000

3. Dates (Except Year)

- ☐ Dates of birth
- ☐ Admission dates
- ☐ Discharge dates
- ☐ Date of death
- ☐ Service dates
- ☐ Any other dates directly related to an individual

Action:

- Convert to age in years (for individuals ≥ 89, aggregate to "90+")
- Use only year for dates when precision not required
- Calculate time intervals instead of specific dates when possible

4. Telephone Numbers

- ☐ Home phone
- ☐ Mobile phone
- ☐ Work phone
- ☐ Fax numbers

Action: Remove entirely or replace with dummy values

5. Fax Numbers

- ☐ All fax numbers

Action: Remove entirely

6. Email Addresses

- ☐ Personal email
- ☐ Work email
- ☐ Any other electronic addresses

Action: Remove entirely

7. Social Security Numbers

- □ Full SSN
- □ Partial SSN

Action: Remove entirely, replace with anonymous ID if needed

8. Medical Record Numbers

- □ MRN from any healthcare facility
- □ Patient account numbers

Action: Replace with de-identified study IDs

9. Health Plan Beneficiary Numbers

- □ Insurance member IDs
- □ Policy numbers

Action: Remove or replace with anonymous codes

10. Account Numbers

- □ Financial account numbers
- □ Billing account numbers

Action: Remove entirely

11. Certificate/License Numbers

- □ Driver's license numbers
- □ Professional license numbers

Action: Remove entirely

12. Vehicle Identifiers and Serial Numbers

- □ VIN numbers
- □ License plate numbers

Action: Remove entirely (rarely applicable to health data)

13. Device Identifiers and Serial Numbers

- □ Medical device serial numbers
- □ Implant identifiers

Action: Remove or generalize to device type only

14. Web URLs

- □ Personal websites
- □ Social media profiles

Action: Remove entirely

15. IP Addresses

- □ IPv4 addresses
- □ IPv6 addresses

Action: Remove from any logs or audit trails

16. Biometric Identifiers

- □ Fingerprints
- □ Retinal scans
- □ Voice prints
- □ Facial photographs
- □ Other biometric data

Action: Remove images or convert to de-identified features only

17. Full-Face Photographs

- □ Any full-face photo
- □ Comparable images

Action: Remove entirely or apply facial de-identification algorithms

18. Other Unique Identifying Numbers or Codes

- □ Any other characteristic that could uniquely identify
- □ Rare diagnoses
- □ Unique combinations of characteristics

Action: Assess and remove or generalize

De-identification Workflow

Step 1: Data Inventory

1. Identify all data fields in source dataset
2. Map each field to HIPAA identifier categories
3. Document fields requiring de-identification

Step 2: Automated De-identification

```python
# Pseudocode for automated de-identification
def deidentify_record(record):
    # Remove direct identifiers
    record.remove_fields(['name', 'ssn', 'mrn', 'phone', 'email'])

    # Generate anonymous ID
    record['patient_id'] = generate_anonymous_id()
    # Generalize geographic data
    record['zip_code'] = generalize_zip(record['zip_code'])
    record.remove_fields(['address', 'city'])

    # Convert dates to ages/years
    record['age'] = calculate_age(record['date_of_birth'])
    if record['age'] >= 89:
        record['age'] = '90+'
    record.remove_fields(['date_of_birth', 'admission_date',
'discharge_date'])

    # Remove other identifiers
    record.remove_fields(['device_serial', 'ip_address'])

    return record
```

Step 3: Manual Review

1. Review automated de-identification results
2. Check for narrative text fields that may contain identifiers
3. Assess rare characteristics or combinations
4. Verify compliance with all 18 identifiers

Step 4: Expert Review

1. Privacy officer reviews de-identified dataset
2. Statistical analysis to ensure no residual identification risk
3. Assess uniqueness of combinations (k-anonymity, l-diversity)

Step 5: Documentation

1. Document all de-identification steps taken
2. Create mapping between original and de-identified IDs (store securely)
3. Record date of de-identification
4. Sign off by privacy officer

Quality Assurance

Validation Checks

- □ Verify all 18 identifiers removed
- □ Confirm no dates earlier than year
- □ Check all ages ≥ 89 aggregated to 90+
- □ Validate ZIP codes (first 3 digits only, or 00000)
- □ Scan free-text fields for names, addresses, etc.
- □ Verify anonymous IDs cannot be reversed

Testing

- □ Run automated identifier detection tools
- □ Attempt re-identification with public data sources
- □ Statistical disclosure control assessment
- □ Document testing results

Special Considerations for AI Training Data

Minimum Necessary Standard

- Only de-identify data actually needed for AI training
- Remove fields not required for model development
- Use aggregated data where individual records not needed

Synthetic Data Alternative

- Consider generating synthetic data based on de-identified data
- Provides additional privacy protection layer
- See Module 4 for synthetic data generation techniques

Model Output Privacy

- Ensure AI models don't memorize and reproduce PHI
- Implement differential privacy techniques
- Test for data leakage in model predictions

Re-identification Risk Management

Prohibited Actions

- Do not attempt to re-identify de-identified data
- Do not combine with other datasets that could enable re-identification
- Do not share de-identification keys or mapping tables

Access Controls

- Limit access to de-identified data to authorized personnel only
- Maintain audit logs of all access
- Require data use agreements for external researchers

Incident Response

- If potential re-identification occurs, immediately report to Privacy Officer
- Assess breach risk and follow HIPAA breach notification requirements
- Document incident and remediation steps

Compliance Certification

I certify that the de-identification procedures outlined in this protocol comply with the HIPAA Privacy Rule Safe Harbor method and that all reasonable steps have been taken to prevent re-identification.

Privacy Officer: ___________________________ Date: __________

Data Steward: ___________________________ Date: __________

Project Lead: ___________________________ Date: __________

References

- 45 CFR §164.514(b) - HIPAA Privacy Rule De-identification Standard
- HHS Guidance on De-identification of Protected Health Information
- NIST Privacy Framework
- ISO/IEC 20889:2018 Privacy Enhancing Data De-identification Terminology

---

This protocol is for educational purposes. Organizations should work with legal counsel and privacy professionals to develop customized de-identification procedures.