IHE Work Item Proposal (Short)

# Proposed Work Item: de-identification handbook update – two/multi stages process of de-identification

Proposal Editor: **Alan Zhang, Essien Ge, Lisson Zhang, Martin Rosner, Christopher Melo**

Work item Editor: Alan Zhang

Date: 2024-12-12

Version: 0.1

Domain: ITI

# The Problem

The needs on secondary use of clinical data with certain personal information is increasing, especially in AI-enabled medical device development and product registration. For example, FDA and NMPA require high representative data including demography information used during algorithm training/testing. Data users care the usability of data, and from their view, De-Identification should be done per needs of secondary data use.

A typical workflow of de-identification in the AI-enabled medical device algorithm development is below.



Figure 1 An example workflow of de-identification in the AI-enabled medical device algorithm development

The example workflow above consists of typical activities of de-identification collaborated between data holder (hospital in this case) and data receiver (medical device manufacture). The workflow usually starts from exporting DICOM data from the medical modality deployed in hospital. To ensure that patient privacy is properly protected, de-identification function on the medical modality is enabled when exporting the DICOM data. The de-identification function on the modality usually follows the specification of the basic profile” of DICOM PS3.15 E.1. Next, hospital transfer the de-identified DICOM data to a data receiver (a medical device manufacture) based on a pre-defined data sharing protocol/agreement. Afterwards, security & privacy experts from a data receiver perform a residual re-identification risk check. Any issues of risk and data utility will be further negotiated between the data holder and the data receiver. Finally, de-identified DICOM (passed the check) will be released to the end uses for the purpose of secondary use.

**Problem description**

As the de-identification becoming a commonly accepted regulatory requirement regarding data privacy protection, de-identification standards from different parties are becoming available. However, these standards are not well aligned with each other. Applying these standards is therefore a challenging and risky task, as a comprehensive de-identification effort requires the knowledge and skills of multiple functional teams, such as data recipients/analytics, privacy law, privacy engineering, etc. A major concern with applying these standards is that a de-identification workflow that starts with the most stringent de-identification policy may result in de-identified data that is less useful, or in some cases, unusable at all.

For example, in the medical imaging domian, the current utility of features on the medical devices following certain standards (e.g, basic profile” of DICOM PS3.15 E.1) is limited to certain cases and cannot be applied to scenarios where patient demographic information is required. Typically, hospitals use the de-identification function in modalities which follow DICOM basic de-identification profiles in PS3.15, according to which patient demography information like age, weight, are all directly removed. However, they are needed according to regulatory requirements of FDA, NMPA etc. to prove data representativeness or do stratified performance.

**Cause of the problem**

Applying the most stringent de-identification policy at the beginning of a workflow/process is a natural consequence of a single-stage de-identification process. A single-stage de-identification process usually consists of activities fully functional for any given use case. For example, given a well identified de-identification requirement, a step of performing de-identification function would include all the transformations of direct identifiers and quasi-identifiers. As a result, implementors of single-stage de-identification prefer a stringent or conservative style of de-identification function (data transformations regarding direct and quasi-identifiers).

In the medical imaging domian, most modality/system manufactures provide the strictest De-ID option as default (or even the only) option, for example, imaging device De-Identification feature based on the “basic profile” of DICOM PS3.15 E.1. Such “One click completion” removes most of the demography information, which is required by AI medical devices developer to meet the regulatory requirement of AI medical device safety check.

**Impacts of the problem**

The de-identification process ends up generating unusable de-identified datasets. Unusable de-identified datasets lead to rework of de-identification, resulting in wasted work and project delays (a few months for some cases). Addressing this issue will reduce the effort and delivery time for sharing data in healthcare innovation and promote patient safety.

# Key Use Case

A typical problematic DICOM data sharing use case scenario is illustrated as Figure 2:



Figure 2 An example of problematic single-stage de-identification process

1. A physician performs de-identification function on the modality/PACS system when exporting/querying health data for secondary use (e.g, RAD:[TCE](https://wiki.ihe.net/index.php/Teaching_File_and_Clinical_Trial_Export)). The de-identification function is usually implemented by following the basic de-identification profile specified in DICOM PS3.15/Attribute Confidentiality Profiles[[1]](#footnote-2). The basic de-identification profile is the most conservative one, which leads to a de-identified DICOM data without containing some of the mandatory data attributes required by AI algorithm development team per regulatory requirements, like, patient age, patient weight etc.
2. The hospital transfers the de-identified data to the medical device manufacture.
3. A De-Identification expert from the medical device manufacture checks the de-identified DICOM data and identifies the missing mandatory DICOM data attributes.
4. The De-identification expert communicates the missing data attributes issue and alternative solution with the physician and AI algorithm developers. Usually, rework on implementing, deploying, and performing de-identification is required.

Figure 3 below shows a process of de-identification starting with a slight de-identification and ending up with a fully de-identified usable DICOM data, namely, two/multi stages of de-identification.



Figure 3 An example of multiple stage de-identification process

1. Stage 1: A physician performs de-identification function on the modality/PACS system when exporting/querying health data for secondary use (e.g., RAD:[TCE](https://wiki.ihe.net/index.php/Teaching_File_and_Clinical_Trial_Export)) which follows a minimal de-identification profile (first stage of de-identification service)
2. The physician submits the minimal de-identified DICOM data for further de-identification.
3. Stage 2: Experts from the hospital perform additional/advanced de-identification function (transforming data) according to the well-defined de-identification requirement.
4. A De-Identification expert from the medical device manufacture checks the de-identified DICOM data and performs further de-identification actions according to the identified gaps between de-identification requirements and the checking results.
5. De-Identification expert shares the fully de-identified DICOM data with the AI algorithm development team.

# Standards & Systems

**Systems:**

1. Raw data repositories/stores. For example, image modality, PACS, EHR.
2. De-identification tools. For example, de-identification integrated within modality system, standalone on-premises de-identification tools, de-identification services on the cloud, etc.
3. De-identified data repositories. For example, Health/Clinical Data Lakes.

**Standards**:

1. ISO 25237:2017 Health informatics — Pseudonymization
2. ISO IEC 27559 Information security, cybersecurity and Privacy enhancing data de-identification framework
3. DICOM/PS3.15/Attribute Confidentiality Profiles
4. HL7 CDA, FHIR Bulk Data

# Discussion

**Philosophy of solution design**

* Separation of concerns (SoC): dividing a complex de-identification process into multiple smaller, more manageable modules
* Prioritize the current deployment environment. A full de-identification process involves the integration of medical devices, independent de-identification tools (including on-premis and cloud), types of users from both hospitals and medical device manufactures.

**Proposal of solution**

This proposal (second use case scenario) introduces a method to resolve the problem of generating unusable de-identified data during the de-identification process of healthcare data. The proposed method is based on a concept of two/multi stages of de-identification process.

* Stage 1: Preliminary De-Identification: Apply common de-identification policies/practices which are applicable to most (if not all) the cases, for example, transforming direct identifiers.
* Other Stages: Advanced De-Identification: Apply de-identification policies/practices which are required by a specific case, for example, shifting patient hospital visit date, generalizing age, applying K-Anonymity with K=5, etc.

**Ideas of handbook update**

* Chapter 1. Intended audience:

**Issue:** Assuming the responsibilities of *IHE Profile editors* and healthcare *information technology implementers* are different, it’s unclear how *IHE Profile editors* and healthcare *information technology implementers* benefit from the handbook.

**Proposed update: i**dentify the benefits that can be gained from the handbook for *IHE Profile editors* and *healthcare information technology implementers*, respectively.

* Chapter 2. De-Identification, Pseudonymization, and Relinking (including Chapter 2.2 Definitions)

**issue:** " *Anonymization and pseudonymization are the two types of de-identification*" may not be accurate now.

**proposed update:** the concept/definition of anonymization, de-identification and pseudonymization need to be updated according to other sources of information, including ISO IEC 20889-2018, NIST SP 800-188, GDPR, PIPL etc.

* Chapter 2.1 General Approach

**issue:** The approach of starting by allowing no data does not support enabling de-identification services for multiple projects/cases. " *This approach starts by allowing no data, which requires that the project team justify that each attribute is required to fulfill the use case objectives."*

**proposed update:** two/multi stages of de-identification. The first stage (preliminary de-identification) is the starting point of the process and can be leveraged by multiple data collection projects/cases.

* Chapter 2.3 De-identification Background

**issue:** the examples in Chapter 2.3.1 is not relevant to IHE profiles.

**proposed update:** include IHE profiles that requires de-identification services as examples, like, RAD:[TCE](https://wiki.ihe.net/index.php/Teaching_File_and_Clinical_Trial_Export), ITI:[MPQ](https://profiles.ihe.net/ITI/TF/Volume1/ch-25.html), [XDR](https://profiles.ihe.net/ITI/TF/Volume1/ch-15.html), etc.

* Chapter 5. Process

**Issue**: The process consists of six steps without specifying the actors which makes the adoption of the process a bit challenging (responsibility assignment of de-identification is unclear).

**proposed update:** Identify essential actors and assign the steps to actors considering two/multiple stages of de-identification service.

* Chapter 6 De-Identification and Pseudonymization for IHE Profile Editors

**Issue:** single-layered of de-identification profile can lead to inconsistencies in applying common practices of de-identification due to the duplication between different data collection cases.

**Proposed update:** Multi-layered de-identification profiles can separate de-identification concerns and standardize how common de-identification practices are applied.

* Chapter B.3 DICOM De-identification

**Issue:** the referenced file is out of dated.

**Proposed update**: update the referenced files.

1. https://dicom.nema.org/medical/dicom/current/output/chtml/part15/chapter\_E.html [↑](#footnote-ref-2)