PHILIPS

China specific Methodology

Document history			
Version	Date	Author	Description
1	March, 2024	Oleksandr Tomashchuk, Renate van Kempen, Alan Zhang	Initial version of the methodology

PHILIPS

Table of Contents

1 Purpose

The purpose of this document is to guide de-identification practitioners to perform de-identification tasks appropriately in Philips China and ensure the de-identification practices are comply with the relevant regulatory requirement in China.

2 Scope and applicability

This document is part of an overall de-identification approach of Philips and focuses on compliance with China's regulatory requirements. Therefore, the scope of application of this document is limited to de-identification activities conducted within Philips China for datasets that originate from China.

Cross-border data transfers are not in scope for this document. Please refer to the "Risk Assessment Methodology" for the methodology used when cross-border data transfers are applicable.

3 Normative references

The following documents constitute essential provisions of this document through normative references in the text. Among them, for dated reference documents, only the version corresponding to the date applies to this document; for undated reference documents, the latest version (including all amendments) applies to this document:

- PIPL Standards:

GB/T 35273 - 2020, Information security technology - Personal information security specification
GB/T 37964 - 2019, Information security technology - Guide for de-identifying personal information.

  - GB/T 42460 - 2023, Information security technology - Guide for evaluating the effectiveness of personal information de-identification.

- Re-identification Risk Assessment Methodology v1.0
- Overarching document, being Data De-identification Methodology v0.2

4 China specific terminology

4.1 completely public sharing

Once the data is released, it is difficult to recall and is generally released directly to the public through the Internet.
[SOURCE: GB/T 37964-2019, 3.12]
Note 1: Same as the English term The Release and Forget Model. [SOURCE: GB/T 37964-2019, 3.12]

4.2 controlled public sharing

Restrict the use of data through a data usage agreement.
[SOURCE: GB/T 37964-2019, 3.13]
Note 1: For example, the agreement prohibits the information recipient from launching re-identification attacks on individuals in the data pool, prohibits the information recipient from associating with external

data sets or information, and prohibits the information recipient from sharing data sets without permission. [SOURCE: GB/T 37964-2019, 3.13]
Note 2: Same as the English term The Data Use Agreement Model. [SOURCE: GB/T 37964-2019, 3.13]

4.3 enclave public sharing

Shared within the physical or virtual territory, data cannot flow outside the territory.
[SOURCE: GB/T 37964-2019, 3.14]
Note 1: Same as the English term The Enclave Model. [SOURCE: GB/T 37964-2019, 3.14]

4.4 Sensitive attribute

Once leaked, illegally provided or misused, personal information may endanger personal and property safety, and can easily lead to damage to personal reputation, physical and mental health, or discriminatory treatment.
[SOURCE: GB/T 35273 - 2020, 3.2, modified, replaced the term “personal sensitive information” with “sensitive attribute]
Note 1: Sensitive attributes include ID number, personal biometric information, bank account, communication records and content, property information, credit information, credit investigation, accommodation information, health and physiological information, transaction information, and individuals of children under 14 years old (inclusive) Information etc.
Note 2: Please refer to the appendix B of GB/T 35273 – 2020 for the determination methods and types of personal sensitive information.
Note 3: Designating an attribute as sensitive depends on the application context, and such a designation is an input to the design of the de-identification process in a specific use case [ISO/IEC 20889-2018, 3.34, note 1].
Note 4: prevent sensitive attribute from being associated with any one data subject during a potential re-identification attack [SOURCE: GB/T 37964-2019, 3.10, note].

5 Abbreviations

Abbreviations	Description
PIA	Privacy Impact Assessment
PII	Personally Identifiable Information
REID-RAR	Re-identification Risk Assessment Report

6 Overview

As a guide of performing de-identification in Philips China, this document is structured in a way with respect to the regulatory requirements of the Chinese market and globally applied Philips de-identification method and practices. From the regulatory requirement point of view, this document serves as a whole picture of those requirements for the readers, and as a guide to fulfilling those requirements in Philips. On the other hand, as a special track of Philips de-identification, the globally

applied Philips de-identification method and practices won't be repeated in this document. Whenever applicable (or required), it will be referenced in this document.

This document touches the following topics:

- Regulatory requirement. Captured the regulatory requirements of de-identification at three levels, namely, legislative, administrative, and standards.
- De-Identification process. Describes the activities that need to be performed for a given de-identification service request.
- Analyze the context. Identify the risks by analyzing the purpose of collecting data, data flow, and data recipients.
- Data assessment. Understand the features of the data, evaluate the potential attack types of exploiting those features, determine reasonable goals of de-identification.
- Assess re-identification risk. Calculating the probability of re-identification given an attack.
- Risk mitigation. Measures need to be performed to mitigate re-identification risk including transform identifiers and security defense.
- Governance. Establishment of Principles/Policies, People, Process, and technology. Periodically review data processing activities and re-evaluate risks.

7 Regulatory requirement

Regulatory requirement is a broader term in this guide, which covers all the requirements at three different levels, legislative requirement, government administrative regulations, and standard specifications.

7.1 Legislative requirements

China's Personal Information Protection Law (PIPL)[1], adopted on Aug. 20, 2021, at the 30th Session of the Standing Committee of the 13th National People's Congress, is the first national-level law comprehensively regulating issues in relation to personal information protection. PIPL, Article 51, Item 3 stipulates that organizations should take corresponding encryption, de-identification, and other security technical measures for personal information. At the same time, PIPL Article 75 Article clarifies the definition of de-identification. Beyond article 51 and 75, the rest of the articles of PIPL impact the de-identification process in many ways. 13Appendix A summarizes the relevant requirements, and the column "guide" indicates the measures Philips is taking to fulfil the regulatory requirement.

7.2 government administrative regulations

To support the implementation of laws, government administrative regulations usually further emphasize compliance with legal requirements from different perspectives. However, most of them won't touch upon too much detail in many areas, such as technical requirement, which leaves room for standardization. Appendix B lists examples of those regulations, note that this list is not exhaustive.

[1] https://www.china-briefing.com/news/the-prc-personal-information-protection-law-final-a-full-translation/

Standard specifications

There are many standards relevant to de-identification in China. Depending on the content, these standards can be classified into two groups, A) high-level requirement; B) guide for de-identification. High-level requirement of de-identification takes de-identification as one of the measures of security and privacy when dealing with personal information in various scenarios or at different levels. For example, GB/T 35273-2020 (Personal information security specification chapter 6.2,), and GB/T 37973-2019 (Big data security management guide, Chapter 8.4.2, Chapter 8.5.2).
In this guide, high-level requirement type of standards will be out-of-scope and we will focus on the standards with detailed level guide (or requirement) of de-identification. The provincial level standards of de-identification considering their complexity and limited applicable geographical area are also out-of-scope since they follow the national standards. Leaving the following national standards to be the regulatory requirement:

1. GB/T 37964-2019 Guide for de-identifying personal information.
2. GB/T 42460 - 2023, Guide for evaluating the effectiveness of personal information de-identification.
3. GB/T 39725-2020 Guide for health data security.

Details can be found in Appendix C, and the column “guide” indicates the measures Philips is taking to fulfil the regulatory requirement.

De-Identification process for China

De-identification process is a response to de-identification service, containing the activities need to be performed. Depending on the type of de-identification services, the de-identification process can be different. There are three types of de-identification service, A) Asset analysis; B) Risk assessment without dataset; C) Risk assessment and transform identifiers. Since the Chinese standards of de-identification only consider the de-identification type C (risk assessment and transform identifiers, therefore this document won’t touch upon de-identification services of type A and B. See additional de-identification process from chapter xx of overarching document.

For the de-identification process of dealing with de-identification service type C, the following activities are included in general:

1. Analyze the context
2. Data assessment
3. Assess re-identification risk
4. Risk mitigation
5. Governance

The subsequent chapters will explain each of the activities above.

8 Analyze the context

The dataset's context refers to the environment in which the data is stored and transferred. To understand the complete situation, there are four core elements that need to be analyzed, why (purpose of collecting data), who (data recipients), what (data content), and how (data flow).

8.1 Purpose of collecting data

See chapter XX of overarching document.
[guide: Why we need to know the purpose? Where can we get the original description purpose? What if the original source of the purpose is not clear (Reviewer, for example, from PA don't think the original description of the purpose is not clear)?]

8.2 Data recipients

Data recipients are individuals, groups and/or organizations who will use the data. PIPL Article 17 (1) requires data custodian to inform the individual the name and contact information about the data recipients. GB/T 42460 – 2023 6.1 b considers data recipients as one of the elements of data sharing context.
The characteristic of data recipients impacts risk identifying. De-identification could be applied at any point in the data lifecycle: from designing the data collection, internal reuse of data, making data available to the external partners. Therefore, data recipients can be internal or external to the data custodian (e.g., Philips). Depending on the relationship, the relevant risks can go from low to high. In addition, knowing the background/profile of the data recipients is also important for determining the background knowledge which is a core factor for identifying indirect identifiers and context risks.

It is recommended that the following information need to be analyzed and documented in the risk assessment report:

1) Name of the data recipients at organization level. It's better to use the formal precise name so that the reviewers/auditor will clearly know who the data recipients is. It could be the department/business within Philips or the partners of Philips (depending on the use cases).
2) Basic profile of data recipients at individual level. The basic profile contains name, email, role in the project.

8.3 Data flow

Data flow depicts how the data is collected and shared with the data recipients. Most data situations are dynamic, that is they involve a set of processes by which data are moved from one data environment to another. Risks come from the data environments and movements as well. That's why regulatory requirements of processing personal data emphasize the legal liability of data custodian (personal information processor), like, PIPL Article 51. The relationship between the data custodian and data recipients is also a critical factor for determining how much security protection measures should be taken.

The legal responsibility may be different regarding the various relationships between data custodian and data recipients. For example, if the data recipients and the data custodian are in a same organization, they may follow the same enterprise level data security and privacy framework. In other cases, like,

involving external party to process personal data and sharing the data with party/parties outside the country from where the data is collected, PIPL requires additional responsibilities in Article 21 and Section 3 (Article [38-43]). Thus, the main purpose of analyzing data flow is to ensure your processing is compliant with legal requirement.

The key components of data flow include:

1. Data source. It's the organization from where the data is collected. The following aspects need to be considered:
    a. Basic profile of the data source. Basic profile includes information like, formal name of the data source. Who are the people that may support the data collection process?
    b. Data types. For the single data source cases, data types are the same one as specified in the data content section. However, for the cases that data is collected from multiple data sources, different data types could be collected from different data sources.
    c. Commitment of performing de-identification. In most of cases, if the data is collected from an organization, an agreement (e.g. a contract or data use agreement) is required. The commitment of performing de-identification may be described in the agreement.
2. Data recipients. Along the flow, the data recipients are the final destinations of the data. For the cases with multiple data recipients, how each of them receives the de-identified data need to be illustrated. For each of the data recipients, their data security and privacy measures should be analyzed.
3. Data environment. Data environment is the place that data stored and processed. Data may be processed in multiple environments of a single organization to enable the data flow. Data risks need to be documented and highlighted. For example, in a data ingestion enabling AI model training case, the environments within Philips may include A) Data collection environment; B) Data landing zone; C) data de-identification zone; D) AI model training zone. Each data environment needs to be analyzed and documented. Each environment probably will have a different configuration of the same core features:
    a. people,
    b. other data including the released de-identified data previously,
    c. infrastructure,
    d. and governance processes.
4. Domain specific regulatory requirement in China. The general regulatory requirement has been identified and specified in **Error! Reference source not found.**. For a particular case, domain specific regulatory requirement needs to be further analyzed, which are the relevant policies, laws, regulations and standards of region or industry.

9 Data assessment

9.1 Data content

Data content is the full (personal) scope of the data and information collected to achieve the defined business purpose. PIPL Article 6 requires collection of personal information shall be limited to the minimum scope for the purpose of processing and shall not be excessively collected.

One of the challenges of minimizing the number of attributes in the data collected is that it's not easy to make decision which data attributes can be removed., .Data recipients tend to ask for whatever is possible, which can lead to a significant number of mitigations to be implemented. There is no ideal

solution for resolving these conflicts since data recipients are considered to be the experts to justify why certain data attributes are mandatory for the defined business purpose. However, it's important to highlight those data attributes that may require a huge de-identification effort, also timewise, leading to . the recipients to consider removing some of the data/data attributes to speed up the process. Types of the data that are challenging and time consuming to de-identify properly:

1) Longitudinal data.
2) Free text data.
3) Binary data in a proprietary format.
4) Imaging data.
5) Unstructured/Semi-structured data

Except for the scope of collecting data, data subject, data type, data attribute type and properties of dataset are also important for risk assessment, therefore need to be analyzed as well.

9.1.1 Data subject

Who are the data subjects and why are they in the dataset? In some cases, there might be different types of data subjects, for example, in a CRF (Case Report Form), both the patient as well as the physician could be included.

Describing characteristics of data subject precisely is necessary for risk assessment. These characteristics impact the estimation of applicable population and the population bins which is a critical consideration when applying theoretical maximum risk calculating approach (see Risk Assessment Methodology chapter 2, Approach 1). The characteristics are usually determined during the initiation stage of the project, and should be documented already, for example, the inclusion and exclusion criteria specified within the study protocol. Here are some of the examples of the attributes being used to describe the characteristics of the data subjects:

1. Age range.
2. Geographical areas.
3. Gender
4. Medical conditions

9.1.2 Data type[2]

There could be different types of data that need to be collected; structured, for example excel or csv files or unstructured, for example images or free text files.

The challenge is the completeness of capturing data types. People used to describe the main element of the data needs to be collected and overlooked/documented the auxiliary information. Re-Identification risks might exist in that auxiliary information. Some of the examples are:

1) File name including path. Name of the physician (or other medical staff), Diagnostic disease, DICOM Instance UID could be part of the file name or path.
2) Patients' demographic data linked with Images.
3) CRF (Case report form). a paper or electronic questionnaire specifically used in clinical trial research.

[2] It's the type of a collection of data attributes as a complete concept, not the data value type for a single simple data attribute, like patient name, birthdate.

9.1.3 Data attribute type

For each data type, describe all the data attributes by using a predefined data dictionary structure. The data dictionary structure includes data attribute name, description, data value type, sample data value etc. (updated formal guide?).

Documenting the data attribute types may take a lot of time. This information usually is collected from the data recipients. The assumption is that data recipients define the business purpose and data collection requirements, and they have the domain knowledge to interpret each data attribute. In some other cases, like, public dataset, the data attribute descriptions may exist together with the public dataset and relevant papers. De-identification service team may need to collect and document that information and confirm it with data recipients. The reason is the data recipients are not allowed to access the dataset included the metadata during the risk assessment stage. Metadata and a small portion of the sample data could be shared with the data recipient to help with documenting this information accurately, and it can be done only with the permission from privacy legal and the help from de-identification service team.

One of the key tasks of analyzing data attribute type is to identify identifiers including direct identifiers (DI) and quasi/indirect identifiers (QI). For the sensitive attributes, it should be marked as SA (sensitive attributes). Refer to **Error! Reference source not found.** for the details of the guide.

9.1.4 Dataset properties

The properties of a dataset can potentially increase or decrease the risk of re-identification. When describing properties of dataset, consider including:

1. age of data (time span data is collected). The older the data, the harder it is to identify people correctly from them.
2. number of data subjects. The number of data subjects is a major concern from regulators' point of view. For example, in China, PIPL Article 58 requires additional legal liability if the number of data subjects is big enough. In addition, in some cases to calculate the probability of acquaintance (T2), the number of data subjects is essential.
3. volume size of dataset. This is important because it impacts the methods and time of processing and transmission of the data.
4. data quality. All data contains some level of error which offers some degree of protection. Although the custodian is likely to want to minimize this error for the sake of providing the most useful data possible. (See ISO/IEC 27559 7.2.5).

9.2 Attack modeling

9.2.1 Select data sharing model

According to the analysis of data sharing situation, decide how the de-identified data will be shared with the data recipients. There are three types of data sharing model can be considered A) completely public sharing (The Release and Forget Model); B) controlled public sharing (Data Use Agreement Model); C) enclave public sharing (The Enclave Model). In Philips China, most cases fall into the category of B) controlled public sharing (Data Use Agreement Model).

9.2.2 Determine attack type

Based on the purposes (or what can be achieved) of re-identification attacking, there are three types of re-identification attacks; identity attack, membership attack, and attribute attack. Depending on the background knowledge of the adversary, Identity attacks usually can be further categorized into three risk models, as specified in the chapter 7.3.1 of ISO/IEC 20889:2018,

1) Prosecutor risk: attack where the adversary knows that a target individual entity is in the data.
2) Journalist risk: attack where the adversary does not, or cannot, know if a target individual entity is in the data.
3) Marketer risk: attack on all entities (rather than a target individual entity) that can be in the data.

GB/T 37964 – 2019 (Chapter 4.3.2) specifies five types of attack,

a) Re-identify a record as belonging to a specific personal information subject.
b) Re-identify the personal information subject of a specific record.
c) Associate as many records as possible with their corresponding personal information subjects.
d) Determine whether a specific personal information subject exists in the data set.
e) Infer a sensitive attribute associated with a set of other attributes.

Among those attacks a), b), and c) are identity attacks, and equivalent to the attacks 1), 2), 3) specified in the ISO/IEC 20889:2018 respectively. Attack d) is membership attack and attack e) is attribute attack.

Although GB/T 37964 – 2019 covers all the three types of attacks (identity, membership, and attribute) this guidance focuses mainly on identity attacks. More detailed technical requirements are not specified in GB/T 37964-2019. In addition, GB/T 42460 – 2023 only specifies the guidance of evaluating effectiveness of personal information de-identification regarding identity attack.

9.2.3 Identify data privacy model

Risk calculation is a core part of re-identification risk assessment. The calculation relies on privacy model, and two formal privacy measurement models are K-Anonymity and differential privacy (See **Error! Reference source not found.**). K-Anonymity is widely used privacy model, especially in healthcare industry. The current regulatory requirement in China and Re-identification Risk Assessment Methodology of Philips are mainly based on K-Anonymity. Thus, applying K-Anonymity privacy, by default, is recommended unless it's not technically possible.

9.3 Determine de-identification goals

9.3.1 General goals

General goals of de-identification should reflect the idea of balancing the protection of privacy risks and data usefulness in a realistic way. Pursuing anonymous data is out of the scope of de-identification. The general goals specified in GB/T 37964 – 2019 (chapter 4.1) confirms the same idea through three independent goals specified as the following list:

a. Delete or transform direct identifiers and quasi-identifiers to prevent attackers from directly identifying or combining with other information to identify the original personal information subject based on these attributes.
b. Control the risk of re-identification, select appropriate models and technologies based on the available data and application scenarios, control the risk of re-identification within an acceptable range, ensure that the risk of re-identification will not increase with the release of new data, and

ensure that data potential collusion among recipients does not increase the risk of re-identification.

c. On the premise of controlling the risk of re-identification, combined with business objectives and data characteristics, select appropriate de-identification models and technologies to ensure that the de-identified data set meets its intended purpose as much as possible (useful).

9.3.2 Determine specific goals.

Compared with the general goals of de-identification (See **Error! Reference source not found.**), specific goals are the implementation of it based on the concrete requirement of a particular case. In other words, specific goals are about how to achieve general goals under given context specified in 8. To determine specific goals, the following steps are recommended:

1. Determine minimum meaningful data requirement. Based on the business purpose (See 8.1) and the data content (See 8.3), identify the minimum meaningful data requirement for each of the data attribute. For those data attributes that won't contribute to the business purpose, should be marked. For the indirect identifiers, an appropriate level of generalizations needs to be specified. For example, instead of using an accurate age, rather use a range of age. The determinization of minimum meaningful data requirement should be confirmed with data recipients.
2. Determine risk level and risk thresholds.
    a. Level of identifiability. There are four levels of identifiability specified in GB/T 42460 – 2023 chapter 4. According to the criteria of the definition, level 3 is the applicable for most of the cases, in which, direct identifiers are eliminated but quasi-identifiers are included, and the risk of re-identification is below the acceptable risk threshold. Level 4 requires exclude all the identifiers including indirect identifiers, which is probably not possible for most of cases considering minimum meaningful data requirement. Therefore, level 3 is the recommended level of identifiability.
    b. Data risk thresholds. Threshold values of re-identification risk of the equivalence class (τ) is recommended in the GB/T 42460 – 2023 D.1.5. They are specified for each of data sharing model. Specifically, Release and Forget Mode: 1/20, Data Use Agreement Model: 1/5, Enclave Model: 1/3. Based on the τ, the risk threshold of the equivalence class ($DR_a = \frac{1}{|J|}\sum_{j\in J} I(\theta_j > \tau)$) is recommended as zero, meaning any equivalent class has a risk greater than or equal to the threshold of re-identification τ is not acceptable.
    c. Overall threshold. It is advisable to set the overall acceptable risk threshold to 0.05 (See GB/T 42460 – 2023 D.1.5)

10 Assess re-identification risk

10.1 Introduction

Depending on the ranges of the quantified de-identification risk value (a probability See **Error! Reference source not found.**) , the levels of re-identification can be classified into four (See GB/T 42460 – 2023 chapter 4):

- Level 1: The direct identifier is included, and the data subject can be directly identified in a specific operational context.
- Level 2: The direct identifier is eliminated, but the quasi-identifier is included. Moreover, the re-identification risk is higher than or equal to the acceptable risk threshold.
- Level 3: The direct identifier is eliminated, but the quasi-identifier is included. Moreover, the re-identification risk is lower than the acceptable risk threshold.
- Level 4: No identifier is included.

According to GB/T 42460 – 2023 chapter 4, there are two approaches to assess the risk, qualitative evaluation, and quantitative evaluation. Qualitative evaluation result can only be either level 1 (no direct identifier) or level 4 (no identifier including direct and indirect), which is consistent with the Re-identification Risk Assessment Methodology v1.0 (chapter 4 Re-identification risk assessment based on analysis of available information). To determine whether the risk is level 2 or level 3, quantitative evaluation is required. The whole idea of the quantitative evaluation is similar to the statistical risk assessment method mentioned in the Risk Assessment Methodology v1.0. However, there are some differences at the detail level.

10.2 qualitative evaluation

Following the steps below to perform qualitative evaluation:

1. Create an identifier list including direct identifiers and indirect identifiers. This should be done based on the 9.1.3 Data attribute type and 9.3.2 Determine specific goals. Although, the data attribute type classified each data attribute into one of the four types (DI, QI, NI, SA), the scope can be reduced based on de-identification technologies applied to each attributes specified in Determine specific goals. For example, patient birthdate can be classified as QI in 9.1.3, but applying date shifting technology to it could also be determined in 9.3.2. After applying date shifting, patient birthdate is no longer a QI.
2. If the identifier list does not contain any identifiers, rate it as level 4 (Low).
3. If the identifier list contains any direct identifier, rate it as Level 1 (High).

10.3 quantitative evaluation

Re-Identification risk can be quantified by investigating the probability of identification given a threat and the probability of a threat being realized. The probability of identification given a threat is usually called data risk, and the probability of a threat being realized is considered as context risk. Given the threat (deliberate attempt, inadvertent attempt, and data breach), how the adversary will attack (re-identify) the data subject will also influence the method of measuring the probability of the attack (re-identify). The method of measuring the probability of the attack is called data risk metrics.

10.3.1 Calculating the overall re-identification risk

The overall re-identification risk is a function of the probability of an attack (context assessment) and the probability of successfully identifying a data principal given that there is an attack (data assessment). It can usually be quantified following a standard risk model where the probability of identification (the likelihood of a risk) is given by the probability of identification given a threat times the probability of a threat being realized.

Figure 1 components of overall risk

As depicted on Figure 1, the overall risk R is determined by the calculation based on two components – data risk (DR) and context risk (CR). This calculation is performed with the usage of the following formula (See Re-identification Risk Assessment Methodology v1.0):

$$DR \times CR = R \qquad \textit{(Equation 10.1)}$$

However, GB/T 42460 – 2023 puts additional restrictions on data risk, except for the overall risk threshold, a set of data risk thresholds are also recommended. In other words, if the data risk (DR_a) > 0, then regardless of the context risk, the overall risk should be considered as 1 (See GB/T 42460 – 2023 D1.5).

Therefore, it is recommended to follow the following steps to calculate the overall risk:

$$R = \begin{cases} 1, & DR_a > 0 \\ DR \times CR, & DR_a = 0 \ or \ Not \ Available \end{cases} \qquad \textit{(Equation 10.2)}$$

10.3.2 Data Risk

The data risk represents the component of the overall risk which is imposed by the data which is present in the dataset itself. To calculate the data risk, it is necessary to classify all variables within the dataset into direct identifiers (DIs), indirect identifiers (QIs, because they are also called quasi-identifiers), and non-identifiers (NI).

10.3.2.1 Direct identifiers (DIs)

Direct identifiers are variables that could be used by themselves to identify a data subject. Examples are a data subject's full name or any part of it, phone number, home address, email address, etc. They also include variables that are unique codes assigned to data subjects that can be used to link with other datasets, such as a Patient ID. If direct identifiers are present in the dataset, it is possible to re-identify a data subject directly, thus the data risk is 100%. In case direct identifiers are present in the dataset, then it is recommended to delete them or apply measures that are considered as equivalent to deletion (e.g. hashing [3] with random salt) under applicable regulations. In addition, de-identification methods (aggregation, subsampling, top and bottom coding, etc.) can be applied to direct and indirect identifiers for mitigating the risk of directly identifying a data subject.

10.3.2.2 Indirect identifiers (QIs)

Indirect identifiers are variables that cannot be used to identify data subjects directly, but can be used in combination with other indirect identifiers or additional information for re-identification. Indirect identifiers satisfy the conditions of being knowable, distinguishable, and replicable. Knowable means

[3] At Philips SHA-512 hash function is considered as secure for the usage as of April 2023. Please, make sure to consult with the de-identification experts regarding secure hashing functions when in doubt.

that the information contained within the variable can be knowable for an adversary from public sources or personal interactions. Distinguishable means that information contained in the variable has sufficient variation among data subjects present in the dataset. Replicable means that information in the variable is stable over time for each data subject. Examples of indirect identifiers are gender, age, race, height, etc. Calculation of data risk is based on indirect identifiers.

10.3.2.3 Non-identifiers (NIs)

Non-identifiers are variables that cannot be used for re-identification of a data subject. Examples are heart rate value, blood oxygen saturation measurements, technical parameters of a device, etc. Considering that their influence on re-identification risk is negligible, they can remain in the dataset as long as the data minimization principle is respected.
Besides direct, indirect and non-identifiers, dataset can also contain fields with free text or other information (e.g., images, video, audio) that might contain combinations of DI, QI, and NI, but cannot be effectively classified in most of the cases, due to pending data collection, or lack of access to values in the original dataset. However, it can be possible to analyze this kind of information on presence of identifiers in it if the data is already collected and it is possible to have access to it, but in these cases an additional investigation on feasibility of such an activity considering available tools and resources should be in place.

10.3.2.4 Data risk metrics

There are different types of metrics of measuring data risks associated with K-Anonymity privacy model. Data risk metrics are associated with the privacy model applied and the type of attack (or risk type). Two formal privacy measurement models are K-Anonymity and differential privacy. This guide only applicable to K-Anonymity privacy model. In other words, calculating data risk for a differential privacy model is out of the scope. Data risk metrics are only valid for a given type of attack. Table below lists the data risk metrics specified in GB/T 42460 and Re-identification Risk Assessment Methodology v1.0.

Table 1 Data risk metrics

Risk Metric	Interpretation	Type of attack
$\Theta_j = \frac{1}{f_j}$	re-identification risk of a single record	All
$DR_a = \frac{1}{\lvert J\rvert}\sum_{j\in J} I(\theta_j > \tau)$	Risk threshold of the equivalence class (See GB/T 42460 – 2023 D.5).	Prosecutor
$DR_b = \max_{j\in J}\theta_j$	The maximum probability of re-identification in the data set among all records (See GB/T 42460 – 2023 D.2, Re-identification Risk Assessment Methodology v1.0 Approach 3).	Prosecutor
$DR_c = \frac{1}{\lvert J\rvert}\sum_{j\in J}\theta_j$	Average value of the re-identification risk of the equivalence class (See GB/T 42460 – 2023 D.3).	Prosecutor
$DR_t = \frac{1}{AP \times \prod_{i=1}^{n} MPB_i}$	Theoretical maximum risk based on finding within publicly available information the smallest population distribution bin for every indirect	Journalist

	identifier (See Re-identification Risk Assessment Methodology v1.0 Approach 1).	
$DR_i = \max(SR_1, SR_{2,\ldots,} SR_m)$	Dataset-wise maximum risk per data subject calculated by identifying a population bin size for every data subject for every indirect identifier based on their corresponding values (See Re-identification Risk Assessment Methodology v1.0 Approach 2).	Journalist

Here is the interpretation of the metrics in the table above.

re-identification risk of a single record

The re-identification risk of all records in an equivalence class is the same, can be calculated by using the equation below:

$$\Theta_j = \frac{1}{f_j} \qquad \textit{Equation (10.3)}$$

Where:
Θ_j- Re-identification risk of the equivalence class.
f_j - Size of the equivalence class.

Risk threshold of the equivalence class

$$DR_a = \frac{1}{|J|}\sum_{j\in J} I(\theta_j > \tau) \qquad \textit{Equation (10.4)}$$

Where:
DR_a- Risk threshold of the equivalence class
J- Collection of equivalence classes
$|J|$- Number of equivalence classes
Θ_j- Re-identification risk of the equivalence class.
τ- Threshold value: 1/20 for data release through completely public sharing; 1/5 for data release through controlled public sharing; 1/3 for data release through enclave public sharing
$I(\theta_j > \tau)$- Judge whether θ_j is larger than τ. If yes, the value is 1, otherwise the value is 0.

The maximum probability of re-identification

$$DR_b = \max_{j\in J} \theta_j \qquad \textit{Equation (10.5)}$$

Where:
DR_b- Maximum value of the re-identification risk of the equivalence class
Θ_j- Re-identification risk of the equivalence class
J- Collection of equivalence classes

Average value of the re-identification risk

$$DR_c = \frac{1}{|J|}\sum_{j\in J} \theta_j \qquad \textit{Equation (10.6)}$$

Where:

DR_c- Average value of the re-identification risk of the equivalence class
Θ_j- Re-identification risk of the equivalence class
J- Collection of equivalence classes
$|J|$- Number of equivalence classes

Theoretical maximum risk

$$DR_t = \frac{1}{AP \times \prod_{i=1}^{n} MPB_i} \quad \textit{Equation (10.7)}$$

Where:
DR_t- Theoretical maximum risk based on finding within publicly available information the smallest population distribution bin for every indirect identifier.
AP- is the applicable population.
n- is the number of indirect identifiers.
MPB_i- the minimum population bin of the ith indirect identifiers, which can be calculated via the
$MPB = \min_{m \in M}(PB_m)$ Equation (10.8)

The minimum population bin

$$MPB = \min_{m \in M}(PB_m) \quad \textit{Equation (10.8)}$$

Where:
M- is a set of all population bins
PB_m- is the mth population bin of a given indirect identifier.

Dataset-wise maximum risk per data subject

$$DR_i = max(SR_1, SR_{2,\ldots,} SR_m) \quad \textit{Equation (10.9)}$$

Where:
DR_i- Dataset-wise maximum risk per data subject calculated by identifying a population bin size for every data subject for every indirect identifier based on their corresponding values.
SR- is the risk for a particular data subject.

10.3.2.5 Calculating data risk

Calculating data risk relies on picking the right re-identification risk metrics. As specified in the 10.3.2.4, risk metrics are associated with privacy model and type of attack, which has to be determined in setting the specific goals (See 9.3.2). Data sharing model can also be considered when selecting risk metrics, for example, an average risk metric (DR_c) is recommended for Data Use Agreement Model and Enclave Model in GB/T 42460 – 2023. The following step is recommended for selecting appropriate risk metrics for calculating:

1. Confirm key factors, privacy model, data sharing model and type of attacks. These key factors are determined during setting specific goals stage. Before calculating the data risk, these factors can be further confirmed. By default, K-Anonymity privacy model is applied, because this guide will be only applicable to K-Anonymity.
2. When the type of attacks is Prosecutor

   a. If the data sharing model is Data Use Agreement Model or Enclave Model, pick DR_c as data risk metric, $DR = DR_c$
   b. If the data sharing model is Release and Forget Model, pick DR_b as data risk metric, $DR = DR_b$.
3. When the type of attack is Journalist
   a. If the dataset needs to be assessed is not available and the smallest population distribution bin for every indirect identifier is available, use DR_t as data risk metric, $DR = DR_t$
   b. If the dataset needs to be assessed is available and the smallest population distribution bins for most of the indirect identifier are available, use DR_i as data risk metric, $DR = DR_i$. There is a limitation of using this method, refer to Re-identification Risk Assessment Methodology v1.0 Approach 2 for the details.
4. Calculate DR_a when the data set needs to be assessed is available and the any one of the conditions below is satisfied:
   a. The type of attack is Prosecutor.
   b. The number of data subjects included in the data set is large enough. Although in some cases, Journalist attack is the major concern, stakeholders like data owner, privacy legal, may have their concerns on the data set, especially when the number of data subjects is large. Unfortunately, there is no standardized quantity to define the concept of "large enough". According to the past practices in Philips, for the cases that having the number of data subjects greater than a few thousands (although, compared with the prevalent population is still small), applying the risk metrics of Prosecutor is recommended.

10.3.3 Calculating context risk

The context risk represents the component of the overall risk which is imposed by the environment in which data is stored and transferred. Furthermore, it covers scenarios of possible attacks in a way that also reflects motives and capabilities of adversaries as well as contractual controls and taken security measures. GB/T 42460 – 2023 D.1.4 specifies a method of calculating context risk, which is identical to the method described in Re-identification Risk Assessment Methodology v1.0 Context risk. Therefore, to calculate the context risk, just follow Re-identification Risk Assessment Methodology v1.0 Context risk.
The context risk consists of three elements and is set equal to the largest of these three elements. The elements are probability of deliberate attempt ($T1$), probability of inadvertent attempt ($T2$), and probability of data breach ($T3$). In case the dataset is made public, then the calculation of the context risk does not take place as it is equal to 100%. If the dataset is not made public, then the context risk is calculated with the usage of the following formula:

$$CR = \max(T1, T2, T3)$$

For the detail information, refer to Re-identification Risk Assessment Methodology v1.0 Context risk.

11 Risk mitigation

11.1 Transform identifiers

See chapter XXX of overarching document.

11.2 Defense in Depth

In general, it should follow Security Management Framework (SMF)[4] when performing de-identification. There are some specific topics worth emphasize, like, data transfer, encryption, secrets management, etc.

11.2.1 Secrets management

In dealing with de-identification there are some secrets need to be managed, the secrets scope should include the items in the list below but not limited to it.

1. passwords for encryption
2. salted keys for hashing.
3. random seed date for date shifting, etc.
4. patient code index/mapping

Secrets should be managed in a secure manner. GB/T 39725-2020 suggests an internal designated management needs to be considered. It should follow Password Security and Access guide[5] to manage those sorts of secrets. Specifically:

1. Use strong passwords. Strong passwords are essential to computer security because they are the first line of defense of user accounts.
2. Use different passwords for different purposes.
3. Using a Password Manager to store your passwords. KeePass is a Password Manager which offers complex and secure passwords and memorizes them for you.

11.2.2 Data Transfer/Sharing in China

There are two methods that can be used to transfer data from our partners to Philips de-id service team or between de-id service team and data recipient in a secure way:

A) Philips Enterprise Cloud[6]

B) Hardware Encrypted USB Storage Device[7].

11.2.2.1 Transfer data via Philips Enterprise Cloud

Follow the steps below to transfer data via Philips Enterprise Cloud:

1. Data recipient requests de-id service team to send a data collection notification to the email address shared by the data source organization, e.g., hospital.

[4] https://share.philips.com/sites/STS020170831152504/gs/Library/Forms/Domain%20view.aspx
[5] https://share.philips.com/sites/STS020170831152504/gs/SitePages/Passwords.aspx
[6] https://philips.service-now.com/itportal?id=kb_article&sysparm_article=KB0015166&sys_id=9330ba121b90d5d443deffb5464bcbc3&spa=1
[7] https://philips.service-now.com/itportal?id=kb_article&sysparm_article=KB1507578&sys_id=370acbd31bfd9d10728a6428bd4bcb3a&spa=1

2. De-ID service team sends data collection notification via SDT[8] to the email address provided by the data source organization. The notification includes the web address and password of uploading dataset, and suggested password of data encryption.
3. Data source organization upload the dataset by following tasks:
    a. encrypts the dataset (see **Error! Reference source not found.** for details) using the password provided by de-id service team.
    b. uploads the encrypted dataset via the web address and password specified in the data collection notification.
    c. Reply the data collection notification email to de-id service team specifying the status of data uploading and the password of data encryption.
4. De-id service team validates the collected dataset and report the validation results to the data recipient.

11.2.2.2 Transfer data via Hardware Encrypted USB Storage

Follow the steps below to transfer data via hardware encrypted USB storage:

1. Data recipient applies for a management-approved business reason for using removable storage media9 (or reuse an existing one)
2. Data recipient sends the hardware encrypted USB storage to the data source organization, e.g., hospitals.
3. Data recipient requests de-id service team to send a password for encrypting the dataset to the data source organization.
4. De-id service team sends a password to the data source organization via the SDT for encrypting dataset.
5. Data source organization encrypts the dataset using the password provided by the de-id service team (see **Error! Reference source not found.** for details).
6. Data source organization copies the encrypted dataset to the hardware encrypted USB storage.
7. Data source organization sends the hardware encrypted USB storage to the de-id service team of Philips.
8. De-id service team validates the collected dataset and report the validation results to the data recipient.

11.2.3 Data encryption

The source data organization encrypts the data using the AES-256 method offered by 7zip. To install 7zip, refer to the following list and select the correct one according to the type of OS.

- Windows: https://www.7-zip.org
- Linux (ubuntu): https://www.digitalocean.com/community/tutorials/install-7zip-ubuntu

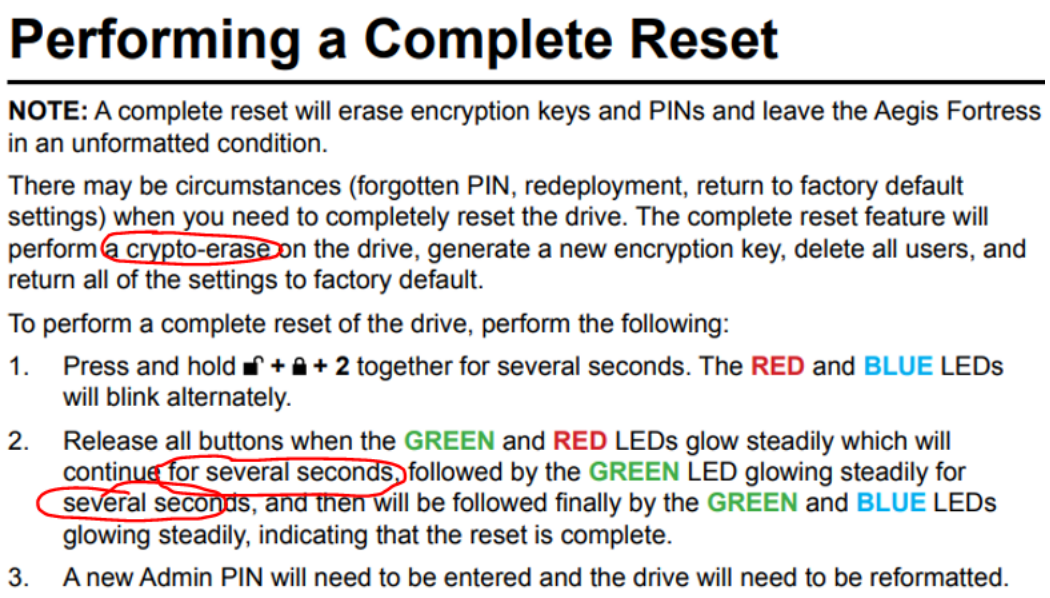
[8] https://philips.service-now.com/itportal?id=kb_article&sysparm_article=KB0016871&sys_id=d522eb208798d114cbaf53d83cbb350d&spa=1

[9] https://philips.service-now.com/itportal?sys_kb_id=6807270bdb810950617e9605f39619998id=kb_article_view&sysparm_rank=1&sysparm_tsqueryId=cc05c7731b654550a427ca286e4bcb8a&sys_id=6807270bdb810950617e9605f3961999

11.2.4 Data Disposal

Unsecure data disposal may lead to data loss and data compromise. Therefore, it is suggested to follow the Data Disposal guidelines[10] to perform data disposal. Depending on the cases, different tools can be used:

1. External USB HDD or a hard disk. Use DiskWipe to perform a DoD-level secure data disposal.
2. Encrypted USB hard disk from Philips. Follow the instructions of performing a complete reset in user manual (see figure below).



Performing a Complete Reset

NOTE: A complete reset will erase encryption keys and PINs and leave the Aegis Fortress in an unformatted condition.

There may be circumstances (forgotten PIN, redeployment, return to factory default settings) when you need to completely reset the drive. The complete reset feature will perform a crypto-erase on the drive, generate a new encryption key, delete all users, and return all of the settings to factory default.

To perform a complete reset of the drive, perform the following:

1. Press and hold + + **2** together for several seconds. The **RED** and **BLUE** LEDs will blink alternately.
2. Release all buttons when the **GREEN** and **RED** LEDs glow steadily which will continue for several seconds, followed by the **GREEN** LED glowing steadily for several seconds, and then will be followed finally by the **GREEN** and **BLUE** LEDs glowing steadily, indicating that the reset is complete.
3. A new Admin PIN will need to be entered and the drive will need to be reformatted.

12 Governance

12.1 General principles

See chapter XX of overarching document.

12.2 Role responsibilities and people management

Individual roles and responsibilities within a custodian should be determined to avoid tasks and duties assignments resulting in a conflict of interests in an organization. In general, it is recommended to follow Human Resource Security Policy of Philips.

12.2.1 Role responsibilities

GB/T 37964-2019 suggests three roles policy manager, executor, and supervisor.

1. Policy manager. A joint team consisting of experts from privacy legal, enterprise IT, and external consulting agency (Privacy Analytics) is taking the role as policy manager of de-identification. The policy regarding the regulatory requirement in China is also maintained as part of the whole policy via this joint team.

[10] https://share.philips.com/sites/STS020170831152504/gs/Library/Infographic%20Data%20Disposal.pdf?CT=1705993987060&OR=OWA-NT&CID=6720b507-a890-531a-4dd5-7479ccfd675e

2. Executor. A dedicated (independent from data recipients) de-identification team is established (Refer to the website for the details), which includes dedicated resources located in China to deal with de-identification requests in China.
3. Supervisor. Privacy counsel/Privacy officer is responsible for supervising compliance with privacy rules (See Privacy Rules for Customer, Supplier and Business Partner Data Article 13). De-identification policy is part of the privacy rules.

12.2.2 people management

It is also important to ensure that all relevant staff are suitably trained and understand their responsibilities for data handling, management, sharing and releasing. GB/T 37964-2019 chapter 6.2 recommends the following actions:

1. Refining job requirements for personal information de-identification, including technical ability requirements and security and confidentiality requirements.
2. When recruiting, applicants should be inspected in accordance with relevant laws, regulations, ethics, and corresponding job requirements.
3. The work contracts or supplementary documents for personnel working in positions of de-identification should clearly indicate their understanding of job responsibilities and the security and confidentiality requirements they must bear.
4. Organizations should regularly conduct business and security training to ensure that personnel working in de-identification positions receive adequate and up-to-date training, ensure that personnel in positions meet training requirements, maintain appropriate skills, and can perform relevant tasks related to personal information de-identification as required.
5. When employees in positions where personal information is de-identified resign, appropriate confidentiality requirements should be added to the resignation confidentiality agreement based on the importance of the data involved.

It is suggested that actions above should be considered in the people management of de-identification.

12.3 Validate and approve

See chapter XX of overarching document.

12.4 Monitor and audit

See chapter XX of overarching document.

12.5 documents/records

Documents and records are a formal way of communicating during the whole process of de-identification. What need to be documented can be very broad. For example, GB/T 42460 – 2023 6.6 a suggested a scope covering the following topics:

1. Evaluation solution. Data sharing situation, evaluation method, and timeline.
2. Identifier report. The process of identifying identifiers and the results.
3. Calculating method of re-identification risk. The process of determining risk threshold and the calculating result.
4. Re-identification risk assessment report. The process of assessment and the result.
5. Assessment records. Various records including communication and negotiation.

The documents and records could also be requested by external government department when they perform their duties on personal information protection. PIPL Article 63 states that the State Cyberspace Administration (SCA) may take the following measures when performing the duties of personal information protection: inquiring, investigation, copying contracts, records etc.

Table 2 types of documents/records

stage	Document	Type	owner	Comments
Preparation	De-identification service request	records	Privacy Counsel	PDA/PCA is ready before submission
	Questionnaire	records	Data recipients	As an attachment of service request
Evaluation	Re-identification Risk Assessment report (draft)	document	De-id Service team	Cover the topics 1-5
	Extra information (communication & negotiation)	records	De-id Service team	As an attachment of REID-RAR
	Software/scripts including configuration parameters	records	De-id Service team	---
	Secrets generated/used	records	De-id Service team	See **Error! Reference source not found.**
	Datasets list including the dataset need to be assessed, other datasets used, de-identified dataset.	records	De-id Service team	The secrets of accessing those datasets should also be managed as secrets.
Release	Re-identification Risk Assessment Report (final)	document	De-id Service team	---

The documents should be stored securely in a place managed by de-identification service team. Re-identification risk assessment report should be shared within a privacy legal and trusted representative of data recipients. Any further sharing the report needs to be approved by de-identification experts and privacy counsel.

Figure 2 Performing a Complete Reset

12.6Security Incident management

Follow the Security Incident Management Policy of Philips.

13 Appendix

Appendix A Legislative requirements and implementation guide

Index	requirement	guide
Article 6	Processing personal information shall be for a definite and reasonable purpose, shall be directly related to the purpose of processing, and shall be processed in a manner that has the least impact on individual rights and interests.	See 8.1
Article 17 (I)	personal information processor shall truthfully, accurately, and completely inform the individual of the following matters: (I) the name and contact information of the personal information processor;	See 8.2
Article 13	A personal information processor may not process personal information unless: (1) the individual's consent has been obtained; (2)-(7) circumstances which do not require consent.	See 8.1
Section 2 Articles [28-32]	Rules for Processing Sensitive Personal Information	See **Error! Reference source not found.**
Article 36	The personal information processed by a State organ shall be stored within the territory of the People's Republic of China …	See 12.2.1
Article 51 (I)	formulating internal management system and operational procedures;	See 12.2.1
Article 51 (II)	managing personal information by classification;	See 8.3
Article 51 (III)	taking corresponding technical security measures such as encryption and de-identification	See **Error! Reference source not found.**
Article 51 (IV)	reasonably determining the authority to process personal information and conduct security education and training for employees on a regular basis;	See 12.1
Article 51 (V)	formulating and organizing the implementation of emergency plans for personal information security incidents	See **Error! Reference source not found.**
Article 54	A personal information processor shall regularly audit whether its processing of personal information is in compliance with provisions of laws and administrative regulations.	See 12.4
Article 56	The personal information protection impact assessment shall include the following: … legitimate, justifiable, and necessary…	See 10
Article 63	Departments (The state Cyberspace Administration) performing duties of personal information protection may take the following measures when performing the duties of personal information	See **Error! Reference source not found.**

	protection: inquiring, investigation, copying contracts, records etc.	
Article 73 (I)	A personal information processor refers to any organization or individual that independently determines the purpose and method of processing in personal information processing activities.	Similar/equivalent to the term **Error! Reference source not found.** in this guide.
Article 73 (III)	De-identification refers to the process in which personal information is processed so that it is impossible to identify certain natural persons without the use of additional information.	Similar/equivalent to the term **Error! Reference source not found.** in this guide.
Article 73 (IV)	Anonymization refers to the process in which the personal information is processed so that it is impossible to identify a certain natural person and unable to be recovered.	Similar/equivalent to the term **Error! Reference source not found.** in this guide.

Appendix B government administrative regulations

Administrative regulation	index	requirements
Opinions of the CPC Central Committee and the State Council on Building Basic Governing Systems for Data to Give Better Play to Data	Article 16	Promote various departments and industries to improve standard systems such as metadata management, data desensitization, data quality, and value assessment.
Administrative Measures for the Compliance Audits of Personal Information Protection	Article 22 (2)	The evaluation includes but is not limited to whether security technical measures such as encryption and de-identification are adopted.
Administration of Cybersecurity in Medical and Health Institutions	Article 22 (1)	Take prevention and control measures such as data desensitization, data encryption, and link encryption to prevent data from being leaked during the data collection process.
Guidelines for Technical Review of Medical Device Cybersecurity (version 2022)	Chapter 2.4 (9)	Medical device cybersecurity capabilities include: … (9) Data De-Identification and Anonymization (DIDT): The ability of products to directly remove and anonymize personal information contained in data.

Appendix C Standard Specification and implementation guide

Appendix .C.1 GB/T 37964-2019

index	Requirement	guide
4.1 General goal	1. Transform direct identifiers; 2. Risk level control; 3. Ensure data usefulness	See **Error! Reference source not found.**

4.2 principles	1. compliance; 2. Security control and usefulness; 3. Technology + management; 4. Automation (tool); 5. Continuous improvement	See **Error! Reference source not found.**
4.3.2 re-id attacks	1. identity attacks (prosecutor, journalist, marketer); 2. Membership attack; 3. Attribute attack	See **Error! Reference source not found.**
4.5 sharing mode	1. Release and Forget (Risk: High); 2. Data Use Agreement (Risk: Medium); 3. Enclave (Risk: Low)	See 9.3.2
5.2.2 determine situation	1. domain specific regulatory requirement; 2. Policy; 3. Data source; 4. Business background; 5. Data usage; 6. External data.	See 8
5.2.3 specific goals	Balance the acceptable risk and data usefulness. Select appropriate re-id risk assessment model and risk level according to characteristic of data and business.	See 9.3.2
5.2.4 work plan	Develop a de-identification implementation plan and get approval.	To be discussed
5.3 identify identifiers	Methods for identifying identifiers include table lookup identification, rule determination, and manual analysis.	See **Error! Reference source not found.**
5.4 transform identifiers	1. data pre-processing; 2. Select privacy model and technologies according to characteristic of data and business; 3. Execute transformation of identifiers	See 10
5.5 validate and approve	Validate risk level and data usefulness, approve de-identified data.	See 12
5.6 monitor and audit	1. Document each step of execution; 2. review documents/records; 3. document review activity; 4. Continuously monitor changes of data situation	See 12.4

Appendix .C.2 GB/T 42460 – 2023

index	Requirement	guide
4. Classification of the Effectiveness	Identifiability of personal data is divided into four levels. • Level 1: The direct identifier(s) is included. • Level 2: Only quasi-identifier(s) is included, and risk is higher (or equal) acceptable threshold. • Level 3: Only quasi-identifier(s) is included, and risk is lower than the acceptable threshold. • Level 4: No identifier is included	See 10.1
6.1 a	Determine the dataset to be evaluated	See 8.3
6.1 b	Determine the context in which the dataset is used, including business scenarios, organizations, personnel, systems, other existing data, etc.	See 9.2
6.1 c	Set up an evaluation team, consisting of personal information protection compliance experts, de-identification technology experts, relevant bus	See 12.1
6.1 d	Conduct preliminary research, including detailed research on the context in which data is used.	See 9.2

6.1 e	Determine the evaluation basis, including relevant laws, regulations, and standards, etc.	See 9.2
6.1 f.1	The dataset and the context in which it is used shall be considered when determining the plan for calculating the re-identification risk, which can be based on the K-anonymity model or the differential privacy model;	See 9.3.2
6.1 f.2	The acceptable risk threshold shall meet the corresponding security requirements and the needs of application	See 9.3.2
6.1 g	Develop an evaluation plan	To be discussed (do we have a plan?)
6.2 a	Identify identifiers in accordance with 5.3 in GB/T 37964-2019, and form a list of identifiers (including direct identifiers and quasi-identifiers);	See **Error! Reference source not found.**
6.2 b	If no identifier is included, it is rated as Level 4, and the evaluation is ended.	See 10.2
6.2 c	If any direct identifier in the list is included, it is rated as Level 1, and the evaluation is ended.	See 10.2
6.3 a	Quantitatively calculate the re-identification risk according to the plan for calculating the re-identification risk as determined in 6.1 f); See Annex D for the plan for calculating the re-identification risk based on K-anonymity model and the example of evaluation.	See 10.3
6.3 b	Compare the calculated re-identification risk result with the acceptable risk threshold. If the re-identification risk result is lower than the acceptable risk threshold, it is rated as Level 3. Otherwise, it is rated as Level 2, and the evaluation is ended	See 10.3
6.4 a	Draw a conclusion on classification of de-identification effectiveness in light of the results of quantitative evaluation and qualitative evaluation	TBD
6.4 b	Obtain an approval from the top management regarding the conclusion	TBD
6.5 a	Understanding and confirmation of the data sharing purpose and the data sharing context;	See 8
6.5 b	Establishment of a mechanism for notifying major changes in the data context	See 12.4
6.5 c	Mutual exchange of information and opinions on re-identification risk metrics	See 12
6.5 d	Opinions expressed by related parties on re-identification risks	See 12
6.5 e	Plan for regular/irregular reevaluation	See 12.4
6.6 a	The evaluation process documents include process documents and result documents followed, referred to and generated in the evaluation process, including but not limited to the following: 1) Evaluation plan	See **Error! Reference source not found.**

	2) Identifier identification report 3) Plan and results for calculating the re-identification risk. 4) Evaluation report: 5) Evaluation records: various records in the evaluation process, including communication & consultation records, etc.	
D.1.2	Calculate the re-identification risk of each record: $\theta_j = \frac{1}{f_j}$ Formula (D.1)	See 10.3
D.1.3	Calculate the re-identification risk of the dataset: $R_b = \min_{j \in J} \theta_j$ Formula (D.2) $R_c = \frac{1}{\lvert J \rvert} \sum_{j \in J} \theta_j$ Formula (D.3)	See 10.3
D.1.4	Calculate the probability of context re-identification attack. a) Release and Forget data sharing model: pr(context) = 1 b) Data Use Agreement and Enclave data sharing models: pr(context) = max (p(Deliberate attempt), p(Inadvertent attempt), p(Data breach))	See 10.3
D.1.5	Calculate the overall re-identification risk. a) Data risk of equivalence class threshold (τ) control: Release and Forget model (Threshold =1/20), Data Use Agreement model (Threshold=1/5), Enclave model (Threshold = 1/3) b) Data risk of equivalence class control: $R_a = \frac{1}{\lvert J \rvert} \sum_{j \in J} I(\theta_j > \tau)$, c) if $R_a > 0$, then the overall risk $R = 1$ d) if $R_a = 0$, for Release and Forget model, the overall risk $R = R_b \times pr(context)$ Formula D.6 e) if $R_a = 0$, for Data Use Agreement and Enclave data sharing models, the overall risk $R = R_c \times pr(context)$	See 10.3

Appendix .C.3 GB/T 39725-2020

Index	Requirement	guide
10.2 a	It should remove attribute information of personal data that can uniquely identify individuals or information that will have a significant impact on individuals after disclosure, such as: Name; Id card/driver's license number; Phone number, fax, email; Medical insurance number, medical record file number, account number; Biometrics (fingerprints, voice, and other information unrelated to the purpose of the application); Photos; Hobbies, beliefs, etc.	See **Error! Reference source not found.**

10.2 b	Information that can be indirectly related to individuals in personal attribute data, such as date of birth, clinic time, inspection time, treatment/cure time, hospitalization and discharge time, work unit, etc., should be generalized.	See **Error! Reference source not found.**
10.2 c	Names and other identifying information of medical staff should be deleted	See **Error! Reference source not found.**
10.2 d	The minimum number of people with the same value of all attributes in the dataset should be more than 5;	See 9.3.2
10.2 e	For cases that need to be traced back to patients, it should establish a patient code index within the controller;	See **Error! Reference source not found.**
10.2 f	Various parameter configurations used in the de-identification process, such as time shift range, patient code index, various personal code generation rules should be strictly confidential, limited to the internal designated management of the controller	See **Error! Reference source not found.** and 12.2.1
10.2 g	In the case of re-identification to the subject, it should be handled by an internal person of the controller, and the processing process shall be strictly confidential;	See 12.2.1
10.2 h	Data recipients should be prohibited from participating in de-identification related work;	See 12.2.1
10.2 i	signing a data use agreement is required to restrict the purpose and retention of data use and data protection measures;	See 9.3.2
10.2 j	In the controlled public sharing mode, users should record data usage and be audited by the controller	See 12.4

Appendix D Classify data attributes

Data attributes can be classified as DI (Direct Identifier), QI (Quasi/Indirect Identifier), and NI (Quasi/Indirect Identifier), and SA (Sensitive Attribute). SA can be considered as a special NI because it's not widely available, but SAs need additional protection.

Classifying data attributes is critical for the whole risk assessment. There are multiple reasons for that. First, any DI results in a 100% re-identification risk. Thus, the classification of DIs must be complete, meaning no DI can be overlooked. However, it's not easy even for the apparently simple case, because the DIs can easily go to some of the not obvious places, like, file path, semi-structured/unstructured attributes. Second, the number of QIs will impact risk assessment and the efforts of de-identifying QIs a lot when applying K-Anonymity privacy model.

Appendix .D.1 Identify DIs (Direct Identifiers)

Direct Identifiers are attributes in microdata can individually identify data subjects under specific circumstance (See **Error! Reference source not found.**). Follow the steps specified in GB/T 37964-2019 chapter 5.3 to identify DIs.

1. Look up the whitelist. Many of the data attributes have already been recognized as direct identifiers. The first step is to match the data attribute with the existing whitelist. There are two whitelists need to be checked:

a. The 18 HIPAA Identifiers. HIPAA Identifiers that are considered personally identifiable information[11].
    1. (A) Names
    2. (B) All geographic subdivisions smaller than a state, including street address, city, county, precinct, ZIP code, and their equivalent geocodes.
    3. (C) All elements of dates (except year) for dates that are directly related to an individual, including birth date, admission date, discharge date, death date, and all ages over 89 and all elements of dates (including year) indicative of such age, except that such ages and elements may be aggregated into a single category of age 90 or older
    4. (D) Telephone numbers
    5. (L) Vehicle identifiers and serial numbers, including license plate numbers
    6. (E) Fax numbers
    7. (M) Device identifiers and serial numbers
    8. (F) Email addresses
    9. (N) Web Universal Resource Locators (URLs)
    10. (G) Social security numbers
    11. (O) Internet Protocol (IP) addresses
    12. (H) Medical record numbers
    13. (P) Biometric identifiers, including finger and voice prints
    14. (I) Health plan beneficiary numbers
    15. (Q) Full-face photographs and any comparable images
    16. (J) Account numbers
    17. (R) Any other unique identifying number, characteristic, or code, except as permitted by paragraph (c) of this section [Paragraph (c) is presented below in the section “Re-identification”]; and
    18. (K) Certificate/license numbers

b. GB/T 42460 – 2023 Appendix A (Sample Direct Identifiers)
    1. Name
    2. ID number
    3. passport number
    4. driver's license number
    5. address
    6. email address
    7. phone number
    8. fax number
    9. bank card number
    10. license plate number
    11. vehicle identification number
    12. social insurance number
    13. health card number
    14. medical record number
    15. device identifier
    16. biometric identification number
    17. Full-face photographs and any comparable images

[11] https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html

18. Account number, certificate/license numbers
19. Internet Protocol (IP) address number and network universal resource locator (URL), etc.

2. Identify direct identifiers automatically based on rules. Some of the DIs are not easy to be captured by only analyzing the data dictionaries. They may exist in some data attributes which are unlikely to be considered as DIs according to the interpretation of the business meaning, for example, patient name could show up in the diagnostic conclusion. If those rules can be defined, leveraging automation to further capture those DIs is highly recommended.
3. Expert analysis method. Expert analysis relies on the deep understanding the business situation and related data structures. The key is to uncover those "hidden variables" (not easily to capture) of the data structure. For example, data records relationships, abnormal data points.

Appendix .D.2 Identify QIs (Quasi/Indirect Identifier)

Quasi identifiers, by definition (See **Error! Reference source not found.**), they are attributes in microdata, combined with other attributes, can uniquely identify a data subject. A large amount of data loss may occur if the information classified as QI is de-identified, and the data that can be used for actual analysis will be greatly reduced. Accordingly, selecting QIs appropriately and providing information that is needed for data analysis are important issues associated with de- identification.

QI selection is inconsistent. Because no exact criteria have been defined for de-identification until now, QIs are selected by subjective judgments stemming from the experience of the person in charge. The difficult part of selection is judgments about the possibility of combining with other attributes (including background knowledge and plausible additional data sources). Some argue that, essentially, all fields in a database are quasi-identifiers because all fields in a database can be used to re-identify individuals. Other suggest that only information in public databases should be considered quasi-identifiers.

The following steps are recommended for identifying quasi identifiers:

1. Lookup the whitelist/blacklist. There are many data attributes have been widely used as quasi-identifiers. Checking whether or not the data attributes are within the whitelist is the easiest thing to do. Here are the quasi identifiers listed In GB/T 42460—2023 Annex B:
    a. Gender
    b. Date of birth or age
    c. Date of an event (e.g., hospital admission, operation, discharge from hospital, visit-related date)
    d. Geographic scope (e.g., postal code, building name, region)
    e. Ethnic origin
    f. Nationality, place of origin
    g. Language
    h. Aboriginal identity
    i. Visible minority group status
    j. Occupational information such as position, employer, department, etc.
    k. Marital status
    l. Level of education.
    m. Number of years of schooling.
    n. Total income
    o. Religious belief
2. identify quasi-identifiers through attribute correlation. For example, in the birth registration information base, the date of birth and the date of discharge of an infant are highly correlated, and the date of birth is recognized as a common quasi-identifier.

3. Expert analysis method: Check basic properties of data attributes. The concept of quasi identifier can be extended into three basic properties, namely, replicable, distinguishable, and knowable[12]. It means a data attribute shouldn't be considered as a quasi identifier unless it satisfies the condition of replicable, distinguishable and knowable at the same time.
    a. Replicable: A field is replicable if it is sufficiently stable over time, so the same values for the data subject. For example, Glucose Level is not replicable.
    b. Distinguishable: A field is distinguishable if it has sufficient variation to distinguish among individuals. For example, breast cancer in breast cancer database is not distinguishable (all the records are breast cancer).
    c. Knowable: what would you know of a close friend or close relative? Demographic & socio-economic information are knowable to the public. Event dates, number of children, diagnosis are knowable to a (close) acquaintance.
4. Identify quasi identifiers automatically based on rules. Similar to the situation in direct identifies, some of the quasi identifiers cannot be captured only by looking at the data dictionary. A rule-based data value check automation is a practical way to identify the quasi identifiers in those not obvious places.

Appendix .D.3 Identify SAs (Sensitive Attributes)

By definition (See 4.4)., sensitive attribute is the information, once leaked, illegally provided or misused, personal information may endanger personal and property safety, and can easily lead to damage to personal reputation, physical and mental health, or discriminatory treatment. A similar concept in GDPR is special categories of personal data (see Art. 9 GDPR). However, the scope defined in GB/T 35273 – 2020 is much broader. Therefore, in Philips China, we should follow the definition and method of determination of sensitive personal information specified GB/T 35273 – 2020.

Below are the data attributes considered as sensitive attributes (See GB/T 35273 – 2020 Annex B)

- **Personal information property**

Bank account, authentication information (password), bank deposit information (including amount of funds, payment and collection records), real estate information, credit records, credit information, transaction and consumption records, bank statement, etc., and virtual property information such as virtual currency, virtual transaction and game CD Keys.

- **Physiological and health information**

The records generated in connection with medical treatment, including pathological information, hospitalization records, physician's instructions, test reports, surgical and anesthesia records, nursing records, medicine administration records, drug and food allergy, fertility information, medical history, diagnosis and treatment, family illness history, history of present illness, history of infection.

- **Personal biometric information**

Personal gene, fingerprint, voice print, palm print, auricle, iris, and facial recognition features, etc.

- **Personal identity information**

ID card, military officer certificate, passport, driver's license, employee ID, social security card, resident certificate, etc.

- **Other information**

Sexual orientation, marriage history, religious preference, undisclosed criminal records, communications records and content, contacts, friends list, list of chat groups, records of whereabouts, web browsing history, precise location information, accommodation information, etc.

[12] A method recommended by PRIVACY ANALYTICS.

14 References

[1] Henriksen-Bulmer J, Jeary S. Re-identification attacks—A systematic literature review. International Journal of Information Management. 2016;36(6):1184-1192. doi:10.1016/j.ijinfomgt.2016.08.002
[2] ISO/IEC 27559:2022 Information security, cybersecurity and privacy protection – Privacy enhancing data de-identification framework.
[3] ISO/IEC 20889:2018 Privacy enhancing data de-identification terminology and classification of techniques.
[4] NIST SP 800-188 De-Identifying Government Data Sets
[5] Emam, K. E. (2013, May 6). Guide to the De-Identification of Personal Health Information. CRC Press. http://books.google.ie/books?id=UuZH_aa4py0C&printsec=frontcover&dq=Guide+to+the+De-Identification+of+Personal+Health+Information&hl=&cd=1&source=gbs_api