NOTES FOR ETHICS COMMITTEES

MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA
UNIVERSITY OF NEW ZEALAND

Participant Data – The life cycle of Data

At the beginning of your project, your collected data may be 'identifiable' yet at the end of your project, you may produce an aggregated dataset. Alternatively, perhaps you have created an 'anonymous' survey where you do not intend to collect any identifiers. Please review the following document to understand the different types of data that can be retained in a Research Project.

The following documents are important to consider ensuring that data collection is in line with both legal and ethical requirements in New Zealand:

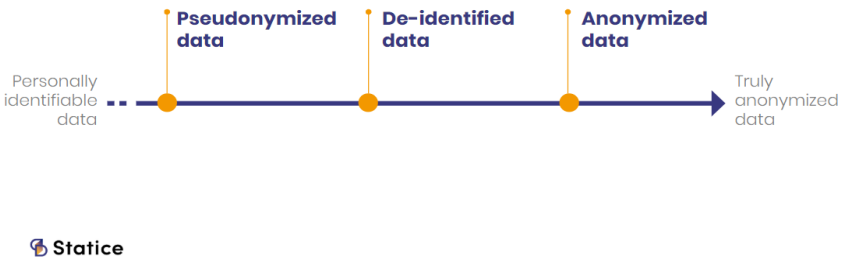
- Privacy Act 2020
- Health Information Privacy Code 2020
- National Ethical Standards for Health and Disability Research and Quality Improvement (NEAC Standards)

The first type of data is "Identifiable information". This is where information contains direct or indirect identifiers. These are usually explicitly personal and can be easily traced back to a particular individual in research outputs. Direct identifiers can include:

- Individual's name
- Phone Number
- Street Address
- NHI (medical) number
- Online identification for example Twitter Handle, Facebook name
- Identification numbers for example Community Service Card number, Driver's license number
- Car license plate
- PO BOX number
- GPS location

What if you want to ensure that your collected data cannot be directly linked back to an individual?

There are three scenarios of removing identifiable information you can consider. Consider the spectrum below when understanding the lifecycle of information:



Last updated: October 2021

Figure 1 – A spectrum of Data Privacy[1]

The following describes these three scenarios further:

SCENARIO 1: Re-identifiable (or Pseudonymised data)

Where direct identifiers are eliminated but indirect identifiers remain intact. For example:

- Employer's name
- Employee's name
- Clinical (medical) notes
- Date of Birth
- IP address can be linked back to the individual (for example, using Qualtrics will collect IP address unless it is switched off)

Data becomes re-identifiable (or Pseudonymous) when information cannot be directly attributed to an individual without the provision of further information. For example:

1. Data is **Key-coded**, and the curator retains the key to the code:
   - For example, John Doe – Diabetes may become key JD-D-1 (specific coding)
   - For example, Jane Doe – Depression may become JaD-Dep-12
2. Data is **Pseudonymous or protected pseudonymous** - Unique, artificial pseudonyms replace direct identifiers, and the unique code is not used elsewhere. These can be protected by technical, organizational, or legal controls that limit the access to 'crack the code' by the researchers or third parties
   - John Doe, Diabetes may become key HGAP67B (unique, unlinked coding)

SCENARIO 2: De-identified information

Where direct or indirect identifiers have been removed or manipulated to break the linkage to real world identities such as:

- Encrypted NHI or study codes
- Year of birth OR Age in years at a given date
- Specific event dates
- Gender
- Ethnicity (Level 2 as defined by Statistics New Zealand)
- Mesh block or suburb
- Deprivation index

De-identification is the process of preventing an individual's identity from being compromised by removing all personally identifiable information that is collected.

[1] https://www.statice.ai/post/how-anonymous-is-anonymized-data

1. Data is suppressed, generalised, perturbed, swapped etc. to remove identifiers (both direct and indirect).
    - For example, where a participant may have a GPA of 3.2, the data indicates this is between 3.0 – 3.5.
    - For example, John Doe aged 37 may become Male – age bracket 30 - 40

SCENARIO 3: Anonymous Data

Where direct and indirect identifiers have been removed and manipulated together into mathematical and technical guarantees.

Anonymisation is not as simple as we might think. Whether an individual data item can be considered anonymous or not requires case-by-case evaluation. The collected material can contain detailed information on individuals (e.g., rare diseases) or enough different types of data which makes them indirectly identifiable. It may be that all information REVIEWED AS A COLLECTIVE can indeed make the individual re-identifiable.

There are two forms of 'anonymous data':

1. **Anonymous** – For example, 'noise' is calibrated to a data set to hide whether an individual is present or not.
    - No immediate identifiers are collected
    - Disclosure of the bare minimum data set for purpose
    - Use of 5–10-year bands rather than dates
    - Blurring of geographic data (by area unit or city)
    - Exclusion of low-frequency characteristics useful for re-identification (e.g., rare medical conditions)
2. **Aggregated anonymous**
    - Aggregation of ethnicity data (level 1 as defined by Statistics New Zealand). For example, 50.85% of New Zealanders are women
    - Strong consideration of more technical assessments or approaches such as k-anonymity ≥5, federated learning, differential privacy.