TEN Recommendations For Regulating De-identification

Khaled El Emam
Main Drivers

• There are many efforts going on globally in privacy law reform, or there are new privacy laws taking effect, and these need to address the concept of identifiability
• Multiple guidelines, standards, and opinions are being developed or updated on how to generate non-identifiable data
• There has been extensive experience with generating and using non-identifiable data over the last decade within existing regimes and following existing standards and guidelines
• We wanted to capture key learnings and summarize them as input into all of these processes
Context for generating non-identifiable data

- The assumption is that data will be used for a secondary purposes
- This secondary purpose is different than the purposes that the data subjects had originally consented to (primary purpose)
- Secondary purposes can involve training AI or machine learning models
- The non-identifiable data may be disclosed to third parties
- No assumptions are made about the technology that is used to generate non-identifiable data
- Pseudonymous data is not non-identifiable data
Definitions

We will use the definitions from CANON (Canadian Anonymization Network)

<table>
<thead>
<tr>
<th>DEFINITIONS – Spectrum of Identifiability</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Identified information:</strong> Information which, by itself, directly identifies an individual.</td>
</tr>
<tr>
<td><strong>Identifiable information:</strong> Information for which there is a serious possibility in the circumstances that it could be associated with an identifiable individual.</td>
</tr>
<tr>
<td><strong>Non-identifiable information:</strong> Information for which there is no serious possibility in the circumstances that it could be associated with an identifiable individual.</td>
</tr>
</tbody>
</table>
Identifiability spectrum and risk thresholds

Identifiable Data (Probability=1)  
Identifiability Threshold  
Not Identifiable Data (Probability=Ø)  

Personal Information  
Not Personal Information
Recommendations

• The recommendations consist of:

• Principles (3):
  – General considerations that apply across the board

• Practices (7):
  – Specific practices that have not worked well, or that should be followed
Reduce uncertainty

• Uncertainty = no decisions being made / “paralysis”
• Focus on the big issues (like the ones we cover here)
• Flexibility elsewhere is important
Create incentives

• Organizations respond to incentives
• Removing incentives (or creating disincentives) for implementing good practices means very few organizations will implement them
• For example, if there is no clear benefit to creating and processing non-identifiable data then organizations will find other ways which may be less privacy protective
• Imposing unattainable standards is also a disincentive
Ensuring sufficient data utility is an important incentive
Recognize the broad benefits of non-identifiable data

- The starting point should be that the processing of non-identifiable data can be beneficial for society and beneficial economically, including by commercial actors and for commercial purposes.
- The emphasis on, for example, the marketing and advertising uses of data or on a few actors distorts the trade-offs when developing guidance and standards.
Obtaining consent for generating non-identifiable data

- In some cases, generating non-identifiable data is treated explicitly as a permitted use (e.g., Ontario’s PHIPA)
- When this issue is left ambiguous it creates uncertainty
- If consent is required then an organization might as well obtain consent for the secondary purposes
- There is strong evidence of consent bias
Anticipated adversary or all possible adversaries

• Many contemporary risk assessment methods need to make assumptions about the background knowledge of an adversary
• An “anticipated” adversary can be defined
• “Any adversary” makes it necessary to treat non-identifiable data as if it is being publicly released
• This is a very high standard when data will be used or disclosed in a non-public manner
Destroying identifiable data

- Some regulatory guidelines have stated that if the original (identifiable) data exists then a dataset cannot be generated that is not non-identifiable
- The definition of “exists” is not clear
- This is a very high standard that would limit many beneficial uses of data (e.g., the ability to conduct health research)
- Reasonable steps can be defined to separate access to identifiable data from non-identifiable data, and these should be encouraged
Threshold definitions need to be precise

- Terms such as “impossible” and “irreversible” imply zero risk, which is an unrealistic standard
- The qualitative terms that are often used to describe when data becomes non-identifiable are difficult to interpret in practice, especially as datasets are becoming more complex
- Example terms such as “reasonable”, reasonably likely”, “very low”, very small”, “serious possibility”, and “acceptably small”
- There are precise precedents that can be suggested
Regulation of uses of non-identifiable data

• Prescribing or proscribing uses of non-identifiable data will be problematic because not all uses can be anticipated and “acceptable” uses will change over time

• A better option is to require ethics reviews on uses of data, models, and decision making from models; this is an approach that has worked well in other domains
Consideration of context

• Many contemporary models for ensuring that the risk of re-identification is below some threshold use controls to manage residual risk (after applying data transformations)

• This approach can work well but it does need guard rails to ensure that it is implemented in a credible manner (e.g., there needs to be some oversight)
A common approach that has worked well in practice is risk-based anonymization.

- **Data Transformations**
  - Generalization
  - Suppression
  - Addition of noise
  - Microaggregation

- **Controls**
  - Security controls
  - Privacy controls
  - Contractual controls
Consequences of re-identification attacks

• Making re-identification an offence under certain conditions is generally a good idea as it adds another layer of protection.

• Allowances for legitimate re-identification is necessary.

• Whit hat attacks need to go through ethics reviews, and there is a need for standards to ensure that they are reported accurately.
Thank you

• Replica Analytics develops the Replica Synthesis software – generator of privacy protective synthetic health data and simulator exchange
  – For more information on our synthetic data solutions:
    • Visit our website www.replica-analytics.com
    • Message us via the website contact page
QUESTIONS