FINAL

Replica Analytics

10 Recommendations for Regulating Non-identifiable Data

September 2021
Khaled El Emam and Mike Hintze

Contents

10 Recommendations for Regulating Non-identifiable Data

Khaled El Emam[1,2,3] and Mike Hintze[4]

[1]Replica Analytics, [2]University of Ottawa, [3]CHEO Research Institute, [4]Hintze Law[*]

1. Introduction

There is considerable activity today in regulatory development and updates around the privacy rights of individuals with respect to their personal data in Europe, Canada, and US, among other jurisdictions. This includes guidelines, standards, and regulator orders and opinions. Some key ones are summarized in the sidebar "Some of the key laws and guidance in progress". Many of these efforts need to (re)define what non-identifiable data is and how its development, use, and disclosure should be regulated.

Over the last twelve months we have had many discussions with individuals and organizations leading these initiatives to provide them one perspective on how non-identifiable data should be regulated. This is informed by almost two decades developing data analytics and privacy enhancing technologies and deploying them in practice globally.

The purpose of this article is to distill the key points that we have made during these discussions into ten specific recommendations, in terms of principles and practices. These are not intended to be comprehensive of all of the issues that need to be considered to regulate non-identifiable data. However, they are a "top ten" list of themes and issues that we believe ought to be considered. Ultimately, this will be up to legislators and regulators to assess.

We will not cover obvious things, like the need for standardized terminology. It is clear that terminology remains a challenge and additional harmonization is needed there. For our purposes, we will use the terminology defined by the Canadian Anonymization Network, which uses the terms "identified," "identifiable," and "non-identifiable" to described the different states of data identifiability [1].

Some of the key laws and guidance in progress

The following are some examples of on-going legislative and regulatory efforts that would require the definition and regulation of non-identifiable data.

Canada: Bill C-11 will replace PIPEDA with an updated federal privacy Law. Quebec Bill 64 will update the provincial privacy law.

Europe: European Data Protection Board updates to the 2014 guidance on anonymization techniques

United States: Regulations and guidance under newly enacted state privacy laws (California, Virginia, Colorado) with many other new state laws under consideration.

Our starting point is that data needs to be rendered non-identifiable to enable secondary analyses, such as building AI models. Also, we are treating pseudonymous data as a form of identifiable data (i.e., the term "identifiable data" will also include pseudonymous data). The method for rendering the data non-identifiable

[*] Replica Analytics is a client of Hintze Law PLLC, and support for Mike Hintze's contribution to this Report has been provided by Replica Analytics.

will not be particularly important for understanding the issues raised, and implementing the suggested principles and practices below.

It is not assumed that following these recommendations is easy or that a universally ideal solution that optimizes successfully on all of them is possible. But if the considerations below can provide a frame of reference to discuss options and explicitly examine trade-offs, then the objectives of this exercise would have been met.

2. Principles

There are some relevant principles to consider when developing or updating regulations for non-identifiable data. These principles should guide the practices and what to prescribe and proscribe to ensure that undesired side-effects do not emerge.

2.1 Reduce Uncertainty

It is important that regulations provide answers to the difficult questions. Many of the points below represent difficult questions. The more precise the answers the better. Precision gives organizations certainty regarding the rules to follow. Leaving aspects of the process of generating, using, and disclosing non-identifiable data ambiguous creates uncertainty, and organizations do not like uncertainty. In practice, uncertainty will result in no action being taken – essentially paralysis. This means that data will not be rendered non-identifiable, and data will not be used for secondary purposes. This also means that potentially economically or socially beneficial activities will not be undertaken because the uncertainty represents heightened risk, and the safest thing to do in the face of high and unknown risks is to avoid them.

Therefore, a reduction in uncertainty to the extent possible is important. Addressing the difficult questions is important. The argument here is not for the elimination of flexibility, but rather that the key parameters should be defined.

2.2 Create Incentives

Regulatory regimes should not put in place disincentives for good behavior and implementing beneficial practices. For example, if meeting regulatory requirements would result in non-identifiable data that are of very low utility, then that would act as a disincentive for generating non-identifiable data. If friction is created such that there are obvious alternatives to achieving the same business objectives that are cheaper or faster, then that also places disincentives for generating non-identifiable data.

In fact, there should be incentives for organization to create non-identifiable data. There should be incentives to process non-identifiable data in a responsible way. More concretely, this means that the benefits of doing a good job generating non-identifiable data should be higher than not doing so. If there are no obvious benefits, then a task will likely not be performed. We will provide some examples below.

2.3 Recognize and Calibrate the Broad Benefits of Non-identifiable Data

There are many industries that generate, use, and disclose non-identifiable data. While the public discourse around the uses of personal data have been focused on marketing and advertising use cases, there are many other use cases, such as health research. For citizens there will be differential benefits for each of these use cases and therefore they should not be treated as one. For example, is the benefit of singling out an individual for the purpose of delivering a new truck advertisement the same as the benefit of singling out a cancer patient

to recommend a potentially life-saving clinical trial? The benefits should be considered as part of defining what is deemed an acceptable application of non-identifiable information.

Acceptable benefits can span those that are beneficial for society (e.g., public good), and those that are commercially beneficial for organizations. The processing of non-identifiable data for commercial gain should not be treated as inherently negative.

3. Practices

The following are specific practices that have proven controversial because, in some cases, they have been defined or interpreted in a manner that goes against the principles above.

3.1 Enable the Creation of Non-identifiable Data without Consent

A key issue is whether individual consent is required for the creation of non-identifiable information. Some statutes, such as Ontario's PHIPA [2], are explicit that the creation of non-identifiable information is a permitted use and therefore does not require additional consent. Some statutes, including the European GDPR, strongly suggest that legal bases other than consent can be used for the creation of non-identifiable data [3]. Other statutes are silent on this issue. Some privacy organizations and advocates have made the case for consent under these circumstances.

It is important to first disentangle the uses of non-identifiable data from the discussion of creating non-identifiable information. The former is addressed separately below.

If an organization needs to obtain consent for creating non-identifiable data, then they might as well obtain consent for processing the identifiable data. For example, if an academic medical center needs to go back and obtain patient consent to reuse their data to study the impact of air pollutants on infant weight, then instead of asking for consent to create non-identifiable data for that study they can just ask for consent to use the original data or a pseudonymized version of the dataset for that study. The creation of non-identifiable data takes time and money and entails some risk because the risk of re-identification is not zero. Why would someone do that if they can just get consent to use the original data? Of course, using the original or pseudonymized data for the purpose of the study means that the data is still personal information, which represents a higher privacy risk to the patients.

Furthermore, it is known that consent bias exists in that consenters and non-consenters differ in systematic ways [4]. This means that unbiased population-level analyses would be more difficult to conduct under the consent requirement.

In this case we would have created a disincentive to implement a privacy protective measure, which is less favorable to the data controller and the data subject. Producing non-identifiable data is in the interests of both the data controller (because it reduces the risk of handling identified data) and the data subject (because it's a means of protecting the data subject's fundamental rights and freedoms).

3.2 Clarify Whether Destroying Original (Identifiable) Data is Necessary

When data is rendered non-identifiable, this is produced through some transformations applied to the original identifiable data. Some guidelines and opinions that have appeared over the last decade have argued that if there is a copy of the identifiable data then it is not possible to claim that a dataset is non-identifiable. It Is not always clear whether this means that the identifiable data exists within the same organization or anywhere even outside the organization. The latter is of course a much more conservative definition of identifiability. If

the original data exists in either of these, does it then need to be destroyed for the transformed data to be legitimately claimed to be non-identifiable ?

The Article 29 Working Party [5] and the Irish DPA [6] have both issued guidance indicating that as long as the raw data exists in its original (identifiable) format, a version of that data that has been transformed in order to make it non-identifiable would still be considered personal data because of the possibility of re-identification using the raw data.

Specifically, the Article 29 Working Party Opinion on Anonymization Techniques [5] indicates:

> *"the means likely reasonably to be used to determine whether a person is identifiable" are those to be used "by the controller or by any other person". Thus, it is critical to understand that when a data controller does not delete the original (identifiable) data at event-level, and the data controller hands over part of this dataset (for example after removal or masking of identifiable data), the resulting dataset is still personal data. Only if the data controller would aggregate the data to a level where the individual events are no longer identifiable, the resulting dataset can be qualified as anonymous. For example: if an organisation collects data on individual travel movements, the individual travel patterns at event level would still qualify as personal data for any party, as long as the data controller (or any other party) still has access to the original raw data, even if direct identifiers have been removed from the set provided to third parties. But if the data controller would delete the raw data, and only provide aggregate statistics to third parties on a high level, such as 'on Mondays on trajectory X there are 160% more passengers than on Tuesdays', that would qualify as anonymous data.*

The Irish DPA's 2019 guidance on anonymization states, "If the source data is not deleted at the same time that the 'anonymised' data is prepared, where the source data could be used to identify an individual from the 'anonymised' data, the data may be considered only 'pseudonymised' and thus still 'personal data', subject to the relevant data protection legislation." [6]

Interpreted literally this means that academic medical centers, as an example, cannot practically operate because they need the original data to provide care, but also use the non-identifiable information for research purposes. Unless of course they cease to use non-identifiable data for health research and only use identifiable information, which brings us back to the situation where we are elevating operating risks for the organization and doing a disservice to the patients themselves because we are encouraging the broad processing of their identifiable information. Plus, the use of identifiable information for research would entail obtaining consent for every use, which would increase the costs and add to the barriers in running secondary health research projects.

This stipulation is therefore quite challenging to operationalize in practice without undesirable side effects. Lesser requirements would be to require that the individuals processing identifiable and non-identifiable information need to be different (i.e., they cannot be the same individuals). One can stipulate that the reporting lines for these individuals have to be different up to a C-level executive, for example. Or that the datasets reside in different databases and there are other internal policies and procedures designed to prevent the re-linking or re-identification of the non-identifiable information. One can even stipulate that the entities processing identifiable and non-identifiable datasets are physically separated (e.g., different floors or buildings). This way there are some concrete separations in place between the entity processing identifiable and non-identifiable data, and would allow the management of risk in a reasonable manner, but still allowing the reasonable and beneficial processing of non-identifiable data.

3.3 Risks Should be Assessed for An Anticipated Adversary

An adversary is an individual or entity that can potentially launch a re-identification attack on a non-identifiable dataset. When performing a re-identification risk assessment, the background knowledge of the adversary is an important consideration.

Some guidelines specify that the re-identification risk should be low against an "anticipated" adversary. The US HIPAA Privacy Rule [7], § 164.514 indicates that the risk of re-identification needs to be very small, such that the information could not be used, alone or in combination with other reasonably available information, by an *anticipated recipient* to identify an individual who is a subject of the information. Others specify that the risk should be low against *any* adversary. The risk levels are very different between these two.

In the former case an analyst can anticipate the background knowledge of an anticipated adversary reasonably well, and account for these in a risk analysis when generating non-identifiable data. This is particularly true when non-identifiable data are being disclosed for non-public purposes. This approach is supported by a number of leading privacy regimes, including the GDPR. As recognized by the European Court of Justice in the *Breyer* case, the risk of re-identification must take into account the "means likely reasonably to be used" by an adversary [8]. Such an assessment necessarily requires one to anticipate which potential adversaries may have opportunity and means to successfully re-identify the information.

By contrast, requiring that the risk be assessed against any adversary essentially equates to a data disclosure being public all the time because "any" adversary means even those who will be very unlikely to ever get access to the non-identifiable data, but they should still be accounted for. Treating the context of disclosure as a public release by default is quite a conservative approach in that it will result in extensive data perturbations and transformations. If the data will be disclosed publicly then that is a reasonable assumption. But if the data will not be disclosed publicly then it is a very conservative approach.

Public data that is rendered non-identifiable will generally have low levels of utility. When data utility deteriorates to a high degree, then it is not useful for the secondary purpose anymore. Thus, an organization would either not be able to use non-identifiable data under such conditions and abandon an analysis project or resort to obtaining consent and using identifiable data. Imposing very conservative requirements will have practical consequences on the ability to make use of non-identifiable data. If the data will have low utility then organizations will not go through the process and either not use and disclose the data for secondary purposes, or work with identifiable data instead.

3.4 Define Acceptable Thresholds

An important component of the process of creating non-identifiable data is deciding on an acceptable risk threshold. That is when the measured risk in the data is below the threshold then the data is considered to be non-identifiable.

In guidance or regulatory documents the use of terms such as "impossible" to re-identify and "irreversible" implies that this threshold is zero. Having a zero-risk threshold is not a practical standard because the risk will never be zero – there will always be some risk. By setting standards that are not practical means that organizations will either not attempt to create non-identifiable data or will use identifiable data instead.

If zero is not a standard, then what should the standard be? Many different terms have been used to characterize that threshold, such as "reasonable", "reasonably likely", "serious possibility", "very low", "very small", or "acceptably small" risk. Subjective translation of these terms to quantitative values proves to be difficult, with a large variation in how they are interpreted, according to recent surveys [9]. The imprecision leads to uncertainty, which acts as a barrier to adoption.

It turns out that there are many precedents for what is deemed an acceptable threshold [10]. Some of these precedents have been repeatedly used and have established de facto standards in some domains. For example, specific thresholds have been defined by the European Medicines Agency and Health Canada for creating non-identifiable clinical trial data [11], [12].

Having precise thresholds increases certainty and makes it easier to implement methods for creating non-identifiable data. A threshold is key in determining when information becomes non-identifiable, and therefore it is a key consideration in the overall process and for having confidence that supervisory authorities will deem the information to be non-identifiable.

3.5 Require Ethics Review Rather than Regulate Specific Uses of Non-identifiable Data

Information, whether identifiable or non-identifiable, can be used for good purposes or inappropriate purposes. For example, a machine learning model can be constructed from a dataset and that model can be used to make decisions that are beneficial or discriminatory to individuals. The appropriateness of the purposes of data processing are orthogonal to the identifiability status of the data.

Because there are examples of inappropriate purposes does not mean that all purposes are inappropriate. On the other hand, defining all possible appropriate purposes in advance is not possible because these purposes will evolve and change over time. And what is deemed appropriate or inappropriate at one point in time can also change in the future, which means these designations are not static. Norms evolve and change over time.

The way to manage the risk of inappropriate uses of non-identifiable data is to set up an ethics review process. This review process would be staffed appropriately and would account for contemporary cultural and societal norms to judge whether a particular data use is appropriate or not.

One can stipulate what this ethics review process should look like, its terms of reference, how it should be staffed, requirements for establishing independence from the business, transparency in its decision-making, and any reporting requirements. Organizations would then be required to establish this oversight mechanism for their data uses, whether these are for identifiable or non-identifiable data.

The ethics review process that is often used as a reference is the academic research one. However, in practice this has been seen as too slow, especially in commercial settings where the drive for innovation and rapid iteration can be intense. Efficiencies have been achieved with commercial ethics review organizations, and these may provide another template.

3.6 The Data Processing Context and Controls Should be Considered

Common methods for assessing the risk of re-identification involve accounting for the context. The anticipated adversaries discussed in 3.3 above is one contextual factor. But others include the security, privacy, and contractual controls that are in place to process the non-identifiable data [13]. If these controls are high, then the overall risk of re-identification is reduced. Considerations of context in a regulatory setting have to account for multiple competing requirements and also to learn from experiences implementing this model in the real world. Some of these are outlined below.

We have seen some organizations argue that pseudonymous data with high controls constitutes non-identifiable data. The residual risk from pseudonymous data is going to be high. The argument has been made that controls can manage that residual risk. A legitimate case can be made that this approach takes the concept of managing residual re-identification risk through controls a bit too far because pseudonymous data remains

highly identifiable. In fact, most known re-identification attacks were performed on pseudonymous data [14]. This model does not pass the smell test and has faced resistance.

On the other side of ledger, when data goes through minimal transformations and there is a strong reliance on controls to manage residual risk, this requires many controls to be put in place. At some point the economics of implementing so many controls to manage residual risk become difficult to justify – we may end up implementing as many controls to manage non-identifiable data as we are managing identifiable data. One of the main benefits for organizations to process non-identifiable data was to be able to operate at a lower control environment which is less restrictive and less expensive to operate.

That controls are important for managing the residual risk is clear. Without controls then the amount of data perturbations and transformations will be so extensive that the data utility will diminish. This will be more so for complex datasets. Therefore, there needs to be allowance for the use of controls to manage residual risks but at the same time this needs to be throttled so that the overall risk management model continues to have credibility.

3.7 Define the Consequences of Re-identification Attacks

We have seen re-identification attacks on non-identifiable datasets occur on a regular basis. Many of these re-identification attacks have been performed by academics and the media [14]. This does not mean that others are not performing re-identification attacks, it is just that we do not necessarily hear about them.

It is reasonable that re-identification attacks without the approval of the data controller / custodian be considered an offence. For example, under Section 171 of the UK Data Protection Act of 2018, re-identification of de-identified data is listed as an offence when done “without the consent of the controller responsible for de-identifying the personal data” [15]. Monetary penalties can be issued to those who contravene the Act.

Under the California Privacy Rights Act (CPRA) the definition of “deidentified” data requires that a business that possesses de-identified data “publicly commits to maintain and use the information in deidentified form and not to attempt to reidentify the information.” The Virginia Consumer Data Protection Act has similar requirements for de-identified data. Such public commitments not to re-identify data are enforceable under state and federal consumer protection law that prohibit deceptive trade practices.

An exception for testing and research purposes (e.g., white hat attacks) is reasonable. For example, the CPRA rule cited above has an exception that allows re-identification attempts “solely for the purpose of determining whether its deidentification processes satisfy the [law’s] requirements.” However, it is important that re-identification research is performed with the approval of a research ethics board. It is not clear that all of the published re-identification attacks have gone through an ethics review. As for all research with data that is or would become identifiable, ethics oversight is important.

The quality of the published re-identification attacks is another important issue. In a number of cases the claims made by the adversaries in their reports were not consistent with their data. The use of imprecise language may be at fault here, but it is clear that notoriety is gained by exaggeration and repetition. For example, demonstrating uniqueness in a sample dataset is not a re-identification - it is an insufficient condition for sure. Presenting that as a successful re-identification attack would not be accurate, but it is a common claim. We need better standards for reporting and interpreting re-identification attacks.

4. Conclusions

This brief report provided ten recommendations for regulating the generation and processing of non-identifiable data. It is based on our experiences developing and implementing privacy enhancing technologies,

as well as working with organizations developing such technologies. Given the number of current legislative and regulatory efforts on developing new privacy laws, guidelines, opinions, and standards, or to update current ones, this is a good point in time to capture experiences and use them to inform these efforts. This report is our contribution.

5. References

[1] Khaled El Emam, Paige Moura, Vance Locton, Elizabeth Jonker, Adam Kardash, and CANON Steering Committee, "Practices for Generating Non-identifiable Data," Canadian Anonymization Network, 2021. [Online]. Available: https://deidentify.ca/wp-content/uploads/2021/08/CANON-OPC-Project-Final-Report-v9.pdf

[2] Government of Ontario, *Personal Health Information Protection Act*. 2004.

[3] K. El Emam, M. Hintze, and R. Boardman, "Does de-identification require consent under the GDPR and English common law?," *Journal of Data Protection & Privacy*, vol. 3, no. 3, Jul. 2020, [Online]. Available: https://www.ingentaconnect.com/content/hsp/jdpp

[4] K. El Emam, E. Jonker, E. Moher, and L. Arbuckle, "A Review of Evidence on Consent Bias in Research," *American Journal of Bioethics*, vol. 13, no. 4, pp. 42–44, 2013.

[5] Article 29 Data Protection Working Party, "Opinion 05/2014 on Anonymization Techniques," Apr. 2014.

[6] "Guidance Note: Guidance on Anonymisation and Pseudonymisation." Data Protection Commission (Ireland), 2019. Accessed: Aug. 30, 2021. [Online]. Available: https://www.dataprotection.ie/dpc-guidance/anonymisation-and-pseudonymisation

[7] U.S. Department of Health and Human Services, *45 CFR Parts 160, 162, and 164; Health Information Portability and Accountability Act, Privacy Rule*. [Online]. Available: https://www.hhs.gov/sites/default/files/ocr/privacy/hipaa/administrative/combined/hipaa-simplification-201303.pdf

[8] *Patrick Breyer v Bundesrepublik Deutschland*. 2016. Accessed: Sep. 17, 2021. [Online]. Available: https://curia.europa.eu/juris/document/document.jsf?docid=184668&doclang=EN

[9] A. Mauboussin and M. J. Mauboussin, "If You Say Something Is 'Likely,' How Likely Do People Think It Is?," *Harvard Business Review*, Jul. 03, 2018. Accessed: Apr. 15, 2019. [Online]. Available: https://hbr.org/2018/07/if-you-say-something-is-likely-how-likely-do-people-think-it-is

[10] K. El Emam, L. Mosquera, and J. Bass, "Evaluating Identity Disclosure Risk in Fully Synthetic Health Data: Model Development and Validation," *JMIR*, vol. 22, no. 11, Nov. 2020, Accessed: Oct. 13, 2020. [Online]. Available: https://www.jmir.org/2020/11/e23139

[11] European Medicines Agency, "External guidance on the implementation of the European Medicines Agency policy on the publication of clinical data for medicinal products for human use," Sep. 2017.

[12] Health Canada, "Guidance document on Public Release of Clinical Information," Apr. 01, 2019. https://www.canada.ca/en/health-canada/services/drug-health-product-review-approval/profile-public-release-clinical-information-guidance.html

[13] K. El Emam, *Guide to the De-Identification of Personal Health Information*. CRC Press (Auerbach), 2013.

[14] K. El Emam, E. Jonker, L. Arbuckle, and B. Malin, "A Systematic Review of Re-identification Attacks on Health Data," *PLoS ONE*, vol. 6, no. 12, 2011, [Online]. Available: http://bit.ly/2hYogS0

[15] *Data Protection Act 2018 (UK)*. Accessed: Aug. 31, 2021. [Online]. Available: https://www.legislation.gov.uk/ukpga/2018/12/section/171?view=plain