WHITEPAPER

Differential Privacy and Risk Metrics for Creating Safe Data

PRIVACY ANALYTICS
an IQVIA company

Advancing the Safe Sharing of Data with Truthful Statistics

Devyani Biswal, Methodology Architect, and Luk Arbuckle, Chief Methodologist

Sensitive data can be reused in many ways to improve healthcare services, uncover new insights and opportunities that can influence healthcare strategies, and develop data products that address societal health needs. Health data can be particularly sensitive as it can reveal a lot about an individual's medical history and lifestyle. Disclosure risk metrics allow such data to be transformed and reused in novel ways while protecting individual contributions and preserving the integrity and truthfulness of data and analytical outputs. Emerging technologies can leverage established benchmarks to enable the safe and responsible use of health-related data.

Introduction

Individual health data has immense value if it can be utilized for beneficial or innovative purposes, including evidence-based process improvement and policy development. Personal health data can include any data created or used in delivering healthcare services or medical reporting, such as demographics, medical history, laboratory results, and any health-related activities. The reuse of any personal data can be done safely and responsibly for the benefit of people and society. Some examples where the reuse of accurate health data is critical include:

- Disease risk analysis and services to provide appropriate medical treatments and preventative measures.
- Health insurance risk modeling to ensure proper coverage and pricing.
- Enabling analytics internally to innovate or inform on policy and program design.

There are many dimensions to the safe and responsible reuse of data,(1) which can also be thought of in terms of defense in depth, ie, protecting data from unauthorized access and misuse through layers of administrative and technical controls. Such considerations have included privacy-enhancing technologies, which are increasingly included as core components in privacy and data protection frameworks.(2) For more information on emerging technologies in this space, see *Advancing Privacy-Enhancing Technologies*.(3)

Technical privacy models are one such control as they are used to assess the risk of disclosure and determine appropriate data transformations that will eliminate those risks.(4) For example, data can be transformed so that the identifiable features of information about people look *similar*, and are therefore clustered (via a *similarity metric*).

Differential privacy is a technical privacy model that protects individuals by requiring that the information contributed by any individual does not significantly affect the output. More specifically, differential privacy is a mathematical property that defines an adjustable information limit. It combines the concepts of outliers and ambiguity into a single mathematical definition.

- Outliers are individuals who stand out in the data and could be singled out, therefore representing a vulnerability.

- Ambiguity is the uncertainty in what can be learned from data sharing, and introduces the concept of adjusting how much can be learned.

- Differential privacy is ambiguity between the same result with and without an outlier (and any other data subject).

By augmenting differential privacy with a framework of risk metrics and other associated benchmarks, we can enable safe and responsible data sharing. Risk metrics are essential tools as they allow organizations to measure and manage the potential risks associated with various data sharing strategies.

Differential Privacy and Risk Metrics

A consistent measure of identifiability (or re-identification risk) is needed to ensure data are safe for sharing. While many technical privacy models exist, and can guide the selection of data transformations for a given data sharing scenario, these models are complex and provide different measures of protection against a variety of possible disclosures.(5)

Mechanisms that are differentially private protect outputs (queries or datasets) by incorporating a level of uncertainty through randomness (eg, noise injection, permutation, shuffling). The randomness produces indistinguishable outputs up to a defined information limit. The *privacy budget* is a form of information limit on how much can be inferred or learned from a dataset and is governed by the sensitivity of a dataset, or the individuals in the dataset. Therefore, the same privacy budget can result in different levels of protection for different datasets due to the differences in their level of sensitivity.

A coherent risk metric can be used for potential disclosures of information and to direct the amount and type of randomization that is needed. To leverage well-established benchmarks, involving strong precedents from the data sharing and use of reputable public organizations such as national statistical organisations (as shown in Figure 1), minimum group sizes can inform the level of randomness needed to produce safe data**.**

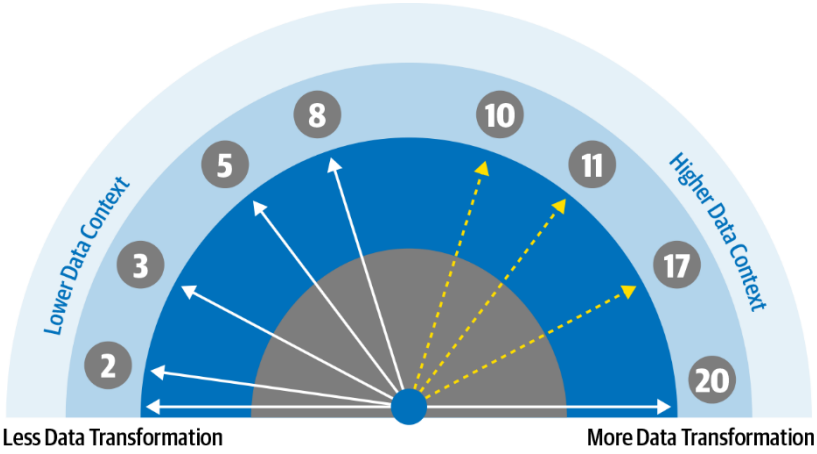

Figure 1: Group sizes are benchmarks based on past precedents for different sharing scenarios (from low to high risk contexts) used by reputable public organizations.(6)

Benchmarks, representing statistical thresholds, ensure there is an objective way to assess safe data sharing.(7) For example, a minimum group size of 10 requires that all information for a group of data subjects with the same set of identifying values be represented by at least 10 individual contributors, such as 10 women aged 45.

By leveraging minimum group sizes to inform the level of randomness required, the privacy budget of differential privacy can be determined in such a way to meet existing benchmarks and eliminate the probability of singling out an individual's contribution of data. Incorporating the notion of group sizes as a risk threshold will determine the privacy budget for that dataset.

By construction, meeting the definition of differential privacy will limit the information that can be inferred or learned from the dataset. Risk metrics alleviate concerns that exists with differential privacy of variable protection across datasets, for the same privacy budget, while ensuring that individuals are hidden in the data and analytics while providing a form of plausible deniability.

Because risk measurement invariably requires the use of statistical methods, any risk measurement technique will be based on a model of plausible attacks, and models make assumptions about the real world. Therefore, risk measurement will always imply a series of assumptions that need to be made explicit. Furthermore, because of the statistical nature of risk measurement, there will also be uncertainty in these measurements and this uncertainty needs to be taken into account.

Three kinds of risks need to be managed in sharing safe data, of which detailed metrics can be derived:

- Prosecutor risk (attack of entity in population): The adversary has background information about a specific person that is known to them, and uses this background information to search for a matching record in the shared data.

- Journalist risk (attack of entity in sample): The adversary doesn't know the particular individual in the shared data, which is a subset of a larger public dataset, but does know that all the people in the data exist in a larger public dataset.

- Marketer risk (dataset attack): The adversary is less concerned if some of the records are misidentified. Here the risk pertains to everyone in the data. Marketer risk is always less than prosecutor or journalist risk, and is therefore often ignored.

If a population registry has information about individuals who are known to be in the shared data, an adversary may target the highest risk data subjects. In this case, the maximum of the risk metric is taken across all data subjects when there are no controls in place to prevent such an attack (e.g., public data sharing). On the other hand, if an adversary will not target the highest risk data subjects because there are controls in place to prevent such an attack, but is trying to find information about a specific individual, the risk metric is averaged across all data subjects (e.g., private data sharing).

Figure 2, for individual versus platelet count, illustrates how group sizes are considered across individual contributions, with ranges that indicate the precision around each randomized value (using a 95% coverage probability). The vertical dotted lines are the ranges representing the similarity that achieve an equivalent group size, based on what an adversary can infer from the randomness, and the vertical solid lines are the ranges of randomization applied to the data. The red dots are an example of randomized values.

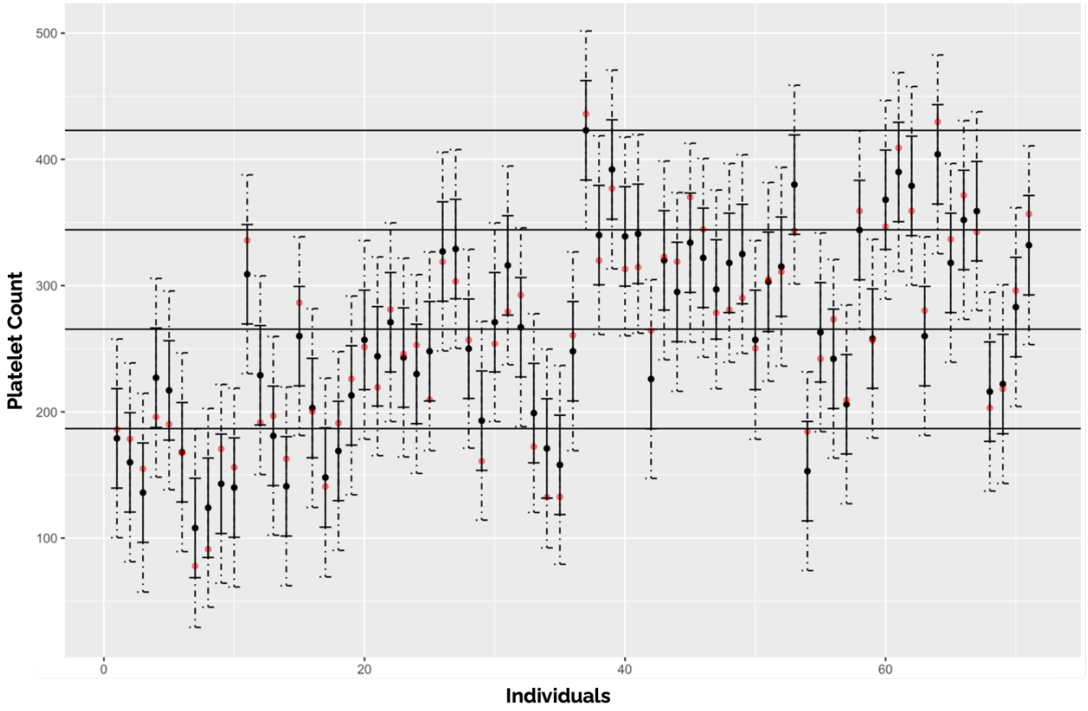

Figure 2: Equivalent group sizes and differential privacy through randomization. Dotted lines represent similarity; solid lines represent randomization (for 95% coverage probabilities).

Creating Differentially Private Health Data

Table 1 provides a fictitious example of a small sample of personal health data of five individuals from a larger dataset. This sample is for illustration purposes only and is greatly simplified to demonstrate the concepts of interest. A privacy budget will set an information limit on how much can be inferred or learned from the dataset. The privacy budget could be set very high, however, and the risk of singling out an individual could also be high as a result. A challenge faced is how much randomization, such as noise addition, is sufficient to protect the data and ensure the dataset is effectively anonymized.

Table 1: Example of personal health data

Individual	Age	Weight (kg)	Height (cm)	White blood cell count (k/uL)	Platelet count (k/uL)
1	24.3	65.0	145	4.5	365
2	25.5	63.0	160	4.8	350
3	27.6	75.0	180	6.0	390
4	29.8	85.0	192	7.4	420
5	30.1	90.0	198	7.1	420

To maintain the integrity of data, randomization can be done in a way that preserves the underlying assumptions of the data and the basis of statistical inference, thereby producing truthful statistics.

Table 2 extends the fictitious example to include ranges that indicate the precision around each value once noise is added to the entries in the table. Overlap between the entries of individual health data can be seen due to the inclusion of confidence intervals. For example, individuals 1 and 2 share similar profiles, as do individuals 4 and 5. Individual 3 has data that is deemed sensitive, and therefore requires special treatment; two rows are included to explain this further.

Table 2: Example with coverage intervals for randomization

Individual	Age	Weight (kg)	Height (cm)	White blood cell count (k/uL)	Platelet count (k/uL)
1	(23.2-25.3)	(60-70)	(140-150)	(4.0-5.0)	(355-375)
2	(24.5-26.5)	(58-68)	(155-165)	(4.3-5.3)	(340-360)
3 (outlier)	(26.6-28.6)	(70-80)	(175-185)	(5.5-6.5)	(380-400)
3 (tuned)	(25.6-29.6)	(65-85)	(170-190)	(5.0-7.0)	(370-410)
4	(28.8:30.8)	(80-90)	(187-197)	(6.9-7.9)	(410-430)
5	(29.1:31.1)	(85-95)	(193-203)	(6.6-7.6)	(410-430)

The third individual's health data is repeated twice: 3 (outlier) and 3 (tuned). In 3 (outlier), the same level of uncertainty is proposed but the individual can still be singled out since there is no overlap with the data of other individuals (thereby deeming their data sensitive); in 3 (tuned), the uncertainty is adjusted to ensure the sensitive data of the individual cannot be singled out and that the resulting dataset is differentially private. In practice, few records of data require an adjusted level of randomization using this localized approach. Alternatively, the level of uncertainty of all records can be globally adjusted equally to achieve the same result.

The overlap between the data of individuals, introduced to tune the degree of randomness and protect against singling out, is no longer evident once the data is randomized and shared as a differentially private dataset, as shown in Table 3. This example is only a sample from a larger dataset that would demonstrate greater variation than shown here. The example has been simplified to maintain structure, otherwise such a small dataset would require significant randomization to be differently private.

Table 3: Example with differentially private data

Individual	Age	Weight (kg)	Height (cm)	White blood cell count (k/uL)	Platelet count (k/uL)
1	24.1	66.3	145	4.4	371
2	26.4	64.7	161	5.3	348
3	27.7	72.3	172	5.4	399
4	29.9	85.5	193	7.7	419
5	30.1	91.2	202	7.1	425

Controlled Environments

The use of a secure and controlled data environment reduces potential re-identification risks so that randomization is minimized and the most useful data can be shared. Mitigating controls and recipient trust can limit the possibility of a deliberate attempt at, or accidentally, identifying data subjects. This is determined by assessing potential threats, or all the means reasonably likely to be used to identify data subjects, including re-identification opportunities. Re-identification opportunities over time can also be contemplated, with data retention, disclosure, and periodic assessments all being important considerations.(8)

Mitigating controls and recipient trust are only referring to the data environment, independent of the data (eg, there could still be a risk of singling out). These factors reduce the threat landscape by limiting access to data, limiting access to external sources of information, and limiting what users can do with the data they are authorised to use. They determine the data sharing scenario and benchmarks used, as previously shown in Figure 1. For more information, see the *Solutions for Complex Data Environments*.(9,10)

The combination of strong controls, recipient trust, and randomization driven by established risk metrics provide defence in depth and can result in a remote risk of re-identification so that data can be deemed effectively anonymized. The technical and organisational measures protect against singling out, linkability, and inference with other available data. Periodic assessments will determine if updates are needed to the technical or organisational measures. Norms or expectations may also change with regards to benefits, risks, and tolerance, resulting in a need to adapt the strategy with the shifting baseline of data protection and data enablement.

Conclusions

The safe and responsible sharing of patient health data is enabled through the use of established benchmarks and emerging technologies. A variety of data sharing scenarios have produced strong precedents for reusing sensitive data. The use of differential privacy as a property of data randomness and protection of individual contributions is emerging as a common theme for data protection, one that can be made consistent with existing risk metrics in producing safe data. By enabling the reuse of sensitive consumer data, organizations can drive novel evidence-based insights that transform policy, products, and society.

About Privacy Analytics

Established in 2007, Privacy Analytics enables organizational leaders to deploy transformative privacy and data protection solutions. With proven technology and unmatched expertise, we empower organizations to gain new insights from their most sensitive data, while alleviating data protection and privacy concerns throughout their data life cycle. To learn more about our design and engineering services, or schedule a time to speak with an expert, contact info@privacy-analytics.com.

References

1. National Institute of Standards and Technology. NIST Privacy Framework: A Tool for Improving Privacy Through Enterprise Risk Management [Internet]. Gaithersburg, MD; 2020 p. 39. Available from: https://www.nist.gov/privacy-framework

2. Tim Sparapani, Justin Sherman. Privacy Tech Buyer Framework [Internet]. Washington, DC: Future of Privacy Forum; 2022. Available from: https://fpf.org/wp-content/uploads/2022/04/FPF-Privacy-Tech-Buyer-Framework-R5-singles-1.pdf

3. Arbuckle L, Collins J. Advancing Privacy-Enhancing Technologies. Privacy Analytics (an IQVIA company) [Internet]. 2022; Available from: https://privacy-analytics.com/resources/articles/advancing-privacy-enhancing-technologies/

4. Wagner I, Eckhoff D. Technical Privacy Metrics: A Systematic Survey. ACM Comput Surv. 2018 Jun 12;51(3):57:1-57:38.

5. International Organization for Standardization. Privacy Enhancing Data De-identification Terminology and Classification of Techniques (Standard No. ISO/IEC 20889:2018) [Internet]. Vernier, Geneva; 2018 p. 46. Available from: https://www.iso.org/standard/69373.html

6. Arbuckle L, El Emam K. Building an Anonymization Pipeline: Creating Safe Data [Internet]. Sebastopol, CA: O'Reilly Media; 2020. 148 p. Available from: https://www.oreilly.com/library/view/building-an-anonymization/9781492053422/

7. International Organization for Standardization. Information Security, Cybersecurity and Privacy Protection – Privacy Enhancing Data De-identification Framework (ISO/IEC Standard No. 27559: 2022) [Internet]. Vernier, Geneva; 2022 p. 22. Available from: https://www.iso.org/standard/71677.html

8. Borel S, Arbuckle L. Good Governance for Anonymized Data [Internet]. Privacy Analytics. 2023. Available from: https://privacy-analytics.com/resources/articles/good-governance-for-anonymized-data/

9. Arbuckle L, Muhammad Oneeb Rehman Mian. Engineering Risk-Based Anonymisation Solutions for Complex Data Environments. Journal of Data Protection & Privacy. 2020;3(3):334–43.

10. Arbuckle L, Ritchie F. The Five Safes of Risk-Based Anonymization. IEEE Security & Privacy. 2019 Oct;17(5):84–9.