PRIVASA

# New privacy-preserving methods to facilitate data-driven research, development, and decision-making

## 1. Summary

Data is the raw material that fuels evidence-based decision-making as well as research, development and innovation activities. For example, data concerning human health, consumption and mobility contain societally relevant insights. Researchers, product developers and decision-makers are increasingly interested in individual choices and challenges, making aggregate datasets unfit for their needs. However, the utilization of data must not jeopardize people's fundamental rights, such as the right to privacy. The more detailed individual-level information datasets contain, the greater the associated privacy risks. Therefore, using and sharing of individual-level data always requires special attention to protective measures.

Currently, the inefficient exploitation of data reserves is hindering business-oriented research, development and innovation activities. Improving the operating conditions for businesses is a socially and economically valuable target, which can be promoted by supporting data-driven product-development and co-creation. We believe that bridging the gap between aggregated statistics and individual-level data becomes possible with time, resources and expertise invested. Synthetic data can help narrow this gap and gradually build more flexible forms of collaboration to answer the data requirements in different stages of research and development. Scientific research and education would also benefit from having easy access to up-to-date, realistic datasets for testing and teaching.

To incorporate the use of synthetic data into RDI processes, we must increase our understanding of the potential use cases and the requirements those impose. For example, the quality and privacy criteria for synthetic data may vary greatly depending on whether the data is used for testing system functionalities or formulating preliminary research hypotheses in a secure environment. Additionally, many alternative methods exist for generating synthetic data, and not all of them are designed to protect privacy. Therefore, the privacy implications of synthetic datasets must always be evaluated on a case-by-case basis, taking into account the intended context of use.

It is also important to emphasize that synthetic data should not be used in situations where achieving real-world accuracy is of highest priority, such as in clinical decision-

making. In most cases, synthetic data represents an intermediate step to enable the first stages of research or development work and helps identify potential challenges early on.

Evidence-based RDI activities and decision-making call for clear rules on how data assets can be exploited. As with any emerging technology, the wider adoption of synthetic data also requires research, conceptualization, testing, and reassessment of existing practices. Technological breakthroughs, such as generative AI, should not be seen merely as a risk to privacy but also recognized for their value in enhancing privacy. Enabling legislation has been identified as one of the key areas for development in, for example, recently published growth and competitiveness vision for the Finnish health sector[1] and the Sotedigi toolkit produced in collaboration with the business sector organizations[2].

All of this is possible through collaboration between research institutions, businesses, and public data controllers. Against this backdrop, the PRIVASA project was launched in 2021, and the lessons and observations from the three-year project have shaped these policy recommendations. We wish to thank all PRIVASA partners for the shared journey.

**Key words:** co-innovation, health data, machine learning, privacy, secondary use, synthetic data generation

# 2. Policy recommendations

**Policy recommendations from the PRIVASA project:**

1. Risk-based privacy benchmarks to support responsible data sharing.
2. Enabling legislation to support business-driven innovation.
3. Flexible operating models to support collaboration between companies and research institutions, also internationally.
4. Incorporating synthetic data into the strategic development of national R&D activities.

[1] The Finnish health sector growth and competitiveness vision 2030, Sitra working paper (2023)
[2] Sotedigin työkalupakki - Kohti vaikuttavampaa sosiaali- ja terveydenhuoltoa (2023) [In Finnish]

## Recommendation 1 | Risk-based privacy benchmarks to support responsible data sharing.

**Need:** Researchers and developers need practical guidelines for demonstrating the required level of data protection to data controllers and data protection authorities. The data protection laws are based on a dichotomic view of anonymous and non-anonymous data, whereas it would be more realistic to view data protection as a continuum with low-risk data at one end and high-risk data at the other. Contrary to traditional legal interpretation, the strength of data protection achieved cannot be determined based on high-level methodologies like pseudonymization or anonymization. This results in uncertainty in the interpretations and application of law, which slows down the adoption of new, potentially superior solutions.

**Solution:** Quantitative metrics that describe the level of data protection, such as differential privacy, would promote the development of sustainable data strategies. A risk-based approach can encourage organizations to develop new and innovative ways to address data protection challenges, while also fostering technological and business development. The public sector could be the forerunner in exploring and adopting new practices, paving the way for other organizations. Illustrative metrics would also facilitate transparent communication with data subjects, increasing awareness about the risks related to personal data processing and providing better opportunities for individuals to influence on how their personal data is used.

## Recommendation 2 | Enabling legislation to support business-driven innovation.

**Need:** In the era of personalized products and services, Finnish companies require individual-level data for their development activities. The current data protection legislation creates a divide between different types of actors. One group is formed by risk tolerant companies whose data strategy is based on meeting the minimum legal requirements. Specifically, global corporations have been issued warnings for illegal and unethical data protection practices. At the other end are risk-averse companies and public administrations that prioritize responsibility and place a strong emphasis on preventive measures. Finland cannot afford the stagnation and relocation of R&D activities abroad; responsible actors must be ensured the opportunity to utilize individual-level health data in a manner that respects the privacy of individuals.

**Solution:** Synthetic data with privacy guarantees could provide companies with individual-level information instead of the current aggregated statistical data. Legislation should recognize the opportunities created by emerging technologies. In some cases, carefully formed individual-level datasets, such as differentially private synthetic data, may offer even stronger data protection than unfiltered aggregate variables derived from the original data.

## Recommendation 3 | Flexible operating models to support collaboration between companies and research institutions, also internationally.

**Need:** Alongside the regular process of obtaining data permits, there is a need for streamlined alternatives to support low-threshold experimentation before deciding on further investments (in time, money or other resources). More flexible testing and piloting of data-driven solutions would significantly enhance the ability of small and medium-sized enterprises to collaborate with research institutions. The importance of international cooperation is also increasing with the accelerating rate of technological development. Supporting these activities requires not only rapid and cost-effective access to relevant datasets but also secure protocols for data sharing and collaborative work.

**Solution:** Synthetic test datasets accessible to companies could accelerate the development of products and services that promote public health. The recently published ready-made datasets by Findata serve as an example of steps taken to this direction, and we propose that similar datasets could also be produced and shared (with permission) as synthesized versions, possibly facilitating wider use. In recent years, similar types of initiatives have been launched by the Clinical Practice Research Datalink (United Kingdom and Northern Ireland) and the U.S. Census Bureau (United States). The exploitation of various test datasets could be facilitated by developing seamless service pathways around them and, where possible, by offering industry-relevant testbed activities. The benefits of synthetic data for both data controllers and processors can be multiplicative, for example, if using of synthetic data translates into simpler data management processes.

## Recommendation 4 | Incorporating synthetic data into the strategic development of national R&D activities.

**Need:** Finland could explore the applicability of synthetic data as a tool for national R&D activities. The emerging opportunities should be specifically considered with respect to national registers and other data resources. Different types of solutions have already been sought after internationally. For example, Research Data Scotland (United Kingdom and Northern Ireland) leads active strategic work with synthetic data. South Australian Health (SA Health, Australia) has recently initiated a collaboration with a company called Gretel AI to explore the possibilities of synthetic data. In January 2023, the United Nations Economic Commission for Europe (UNECE) published official guidelines on the use of synthetic data as a tool for providing statistical information. In March 2023, the Financial Conduct Authority (FCA, United Kingdom and Northern Ireland) established a specialist group focused on synthetic data, providing references from the financial sector.

**Solution:** Questions related to synthetic data should primarily be integrated into the activities of existing expert groups, but specifically in later stages, nominating a dedicated multidisciplinary expert group could be justified. The proposed actions would support Finland in maintaining its position as an internationally recognized leading country that supports data-driven decision-making, research, and development. This issue is particularly timely as the European data spaces are emerging. In addition to decentralized analytics, there is likely market potential in synthetic data, promising significant opportunities for Finnish companies, research institutions, and public administration organizations.

### Synthetic data is generated data that can be used for improving data protection.

In this policy recommendation, we refer to synthetic data as modeled datasets that mimic real observations. Synthetic datasets follow the structure and correlations of real observations. Often, synthetic datasets are modeled sets of observations based on the original data.

Improving privacy is just one of the many reasons why synthetic data is generated. Therefore, not all synthetic datasets are suitable for sharing, as they may contain sensitive information present in the original data. Generating synthetic datasets in a way that ensures the privacy of individuals in the original data requires appropriate methods.

From a privacy perspective, different types of datasets form a continuum, with original datasets containing personally identifiable information at one end. Pseudonymized data lacks the most obvious identifiers, such as names and social security numbers. Anonymization aims to prevent indirect identification of individuals by combining information, such as birth month and postal code. However, the privacy provided by traditional anonymization methods has been proven insufficient in many contexts. At the other end of the continuum are fabricated datasets that are not based on real observations in any way.

### What is differential privacy?

Differential privacy is a mathematical framework that enhances data privacy when data is analyzed or shared. The privacy protection offered by differential privacy is based on statistical noise, which ensures that individual data points do not stand out when comparing analysis results. The level of privacy applied to analysis or synthetic data generation can be adjusted with a privacy parameter. More noise means better privacy protection but also reduces data quality, limiting potential use cases. In other words, the same inaccuracy that protects the privacy of individuals in the dataset also increases the margin of error in statistical testing.

A significant advantage of differential privacy is the mathematical guarantees that allow the strength of privacy protection to be quantified. Although technical details do not eliminate the need for case-specific judgment in privacy-related issues, defining indicative thresholds could support the responsible use of health data.

# Contact information

Coordinator contacts from the University of Turku:
**Antti Airola**, Associate Professor, Department of Computing
**Tapio Pahikkala**, Professor, Department of Computing

Project leads for the Turku University of Applied Sciences
**Elina Kontio**, Principal Lecturer, Research Group Leader, Engineering and Business, ICT
**Mojtaba Jafaritadi**, Principal Lecturer and research lead, Engineering and Business, ICT

Project leads for VTT Technical Research Centre of Finland:
**Mika Hilvo**, Research Team Leader, Health Data Analytics
(Mark Van Gils and Miguel Bordallo at the beginning of the project.)