PUBLIC CONSULTATION

Advancing Privacy-Enhancing Technologies

PRIVACY ANALYTICS
an IQVIA company

Safe Data Enablement for Health Services & Research through Privacy-Enhancing Data Sharing and Analytics: A Spectrum of Perspectives

Luk Arbuckle, Chief Methodologist and Privacy Officer
Jordan Collins, Data Privacy Solutions Business Leader

We would like to thank the Office of Science and Technology Policy for this opportunity to provide an industry perspective and help inform the development of a national strategy on privacy-enhancing data sharing and analytics, along with associated policy initiatives (as published in the Federal Register on 06/09/2022 and available online at federalregister.gov/d/2022-12432, and on govinfo.gov). We believe that public consultation is an important step in producing an informed strategy that is more likely to produce solutions that are scalable and proportionate to the intended uses cases and needs.

Abstract

Since 2007, Privacy Analytics has been providing services and software in privacy-enhancing data sharing and analytics for organisations in the consumer and healthcare industries. We are particularly interested in safe data enablement for health services & research. From our 15 years of experience working in this space, we agree that there is a tremendous opportunity to use data safely and responsibility to the benefit of people and society. We are pleased that the request for information acknowledges the need to consider operational challenges and incentives to adoption.

In this response, we provide an overview of the landscape of tools considered for privacy-enhancing data sharing and analytics with a view towards the needs and perspectives of different stakeholders involved in health services improvement and health research for the full lifecycle of data. While there are many such tools, we focus our attention on the challenges of integration and interoperability, especially for complex health data and analytical pipelines, and the impact on end users driving to improve health outcomes. We believe that, in order to drive the adoption of safe, useful, and timely data and analytics at scale, there is an increasing need for the integration and deployment of suites of tools that are interoperable and complimentary.

Introduction

Health data is often described as some of the most sensitive since it deals with the intimate details of a person's body and mind. The information itself exists before it is measured, inferred, or assessed through various machines, tests, or questions to patients, eventually recorded in a form that may be used in data sharing or analytics. Once captured and collected, health data is continuously updated, transformed, harmonized and restructured to meet the myriad needs of extracting the greatest insights from advanced statistical methods that are themselves continuously updated and improved on throughout the data lifecycle.

The importance of data sharing and analytics to drive evidence-based decision making has been highlighted by data science consortia and researchers alike. In the highly competitive and innovative fields of health research and treatment development, solutions for data sharing and analytics must contend with data that include a large number of variables, have spatial and

temporal dependence, and inconsistent or missing data (eg, due to non-response bias in surveys or complex data collection practices and linking challenges). Even static data is refactored and reharmonized to suit the various acrobatics of statistical analysis.

Take the pharmaceutical industry as an example of the complexities and challenges of working with health data. Besides requirements to maintain copies of data that support drug approvals, the data submission guidelines include thousands of variables needed to understand chemical and health interactions in the drug development process.(1) Pharmaceutical companies collect data in many different formats (eg, approximately 40% is still based on electronic capture from forms), with many steps before delivering to different functional groups in the drug development process, including thousands of mapping rules for billions of data points. Each company will have their own internal standards and constantly evolving schema that result in continuous data integration and harmonization.

It takes over a decade to develop a new medicine, at a cost in the billions of dollars with only a 0.01% chance of success for compounds in preclinical research.(2) Trial participants expect the effective use of data about them, including reuse, provided safeguards are in place. The safe reuse of trial data also reduces the burden on participants by making the best use of data already collected, accelerating research discoveries by improving access to data and analytics. Coupled with the regulatory recognition of research benefits to improve health outcomes, there is clearly a general consensus of the need and importance to make the best use of data and analytics in clinical research.(3)

In this response, we introduce a data lifecycle view and stakeholder perspectives to set up a baseline for discussion of expectations and needs. We then provide a categorization of tools and how they align with stakeholder perspectives before introducing a spectrum of tools that incorporate some of the main considerations that were established. Finally, we provide a brief view into those tools as we have seen them used in practice, culminating in a revised view of the data lifecycle with the various tools that may best fit the needs of users in that stage of the lifecycle. Our hope is that this landscaping exercise will drive a conversation so that more practical and applied methods get the attention they deserve.

Data Lifecycle

For the purposes of privacy enhancing data sharing and analytics, for sensitive data in particular, we consider a data lifecycle focused on those activities where there is a particular need for privacy considerations. Throughout this lifecycle, the goal of producing safe, useful, and timely data and analytics, often at scale, should be kept in mind so that proposed solutions are practical and more likely to be adopted in practice. The subject of legislative or regulatory incentives are out of scope for our review (eg, see the "Failure" of PETs in (4)).

We begin with a simple 5 step data lifecycle with general needs for thinking about privacy enhancing data sharing and analytics, which we will revisit after introducing a spectrum of perspectives and technologies. The intent here is to lay out where opportunities may exist for privacy technology that serve practical needs, without consideration for legislative or regulatory incentives. This list is only a summary and by no means exhaustive.

Get data

- Why and what to collect serves to define purpose and minimize the capture and collection to what is truly necessary at the very outset.

- Needs and wants are evaluated through strategic thinking to ensure the right breadth is included for future opportunities (including a full lifecycle view).
- Capture and collection is designed to ensure the right concentration of data will deliver the sharing and analytics that provide the desired strategic intent.

Link data

- Separation anxiety, in the form of stakeholder concerns with linking, is acknowledged while focusing on data capture or collection with strategic benefit (ie, demonstrating value).
- Identify use cases so that linking of data across capture and collection points have a defined purpose and are appropriately minimized (ie, demonstrating means).
- Match data subjects or insights, introducing opportunities to disassociate from the analytics (eg, prior distributions, transfer learning), based on actual needs.

(Re)Use data

- Isolation anxiety, in the form of stakeholder concerns with data enablement, is acknowledged while focusing on use and reuse with strategic benefit (ie, demonstrating value).
- Build trust from users, in the use and reuse of data (ie, primary and secondary purposes) through the principles of trustworthy data sharing and analytics.
- Drive adoption in the trustworthy use and reuse of data as standard practice and with suitable privacy defaults for data enablement.

Share data

- Win stakeholder trust in sharing useful and timely data and insights with demonstrable benefit to current and future data subjects (eg, reducing costs, collaboration).
- Safe design and delivery through trustworthy methods that safeguard what is shared and how, for the intended purposes.
- Share insights, wherever possible, as collaborators in understanding human health and improving health outcomes (ie, evidence based).

Open data

- Citizen data as a government commitment towards transparency, innovation, and accountability (eg, G8 Open Data Charter)
- Data transparency as a regulatory commitment to democratizing data, analytics, and insights for all stakeholders (eg, European Medicines Agency Policy 0070)
- Data trust as an asset management framework with independent stewards acting on behalf, and for the benefit, of a broader group stakeholders.

Stakeholder Perspectives

Having considered a data lifecycle for privacy enhancing data sharing and analytics, we will now turn our attention to different stakeholder types to understand needs and expectations. The 4 stakeholder types we consider represent an evolving set of perspectives, and the interplay between them will be important to explore the spectrum of tools we consider in the next section.

End user. We start by assuming the end user of data, responsible for making sense of data and performing analytics, believes information assets and technologies should be available to support

their functional role, and that they should be trusted. The end user is, after all, a professional with a job to do, a job that often requires specialized training and is respected in our modern world of evidence-based decision making. Consider the biostatistician, epidemiologist, or a health scientist as standard examples of end users that are pursing improvements in efficiencies and health outcomes. The interests of the end user include:

- Access to data that is timely and useful to meet their needs and objectives
- Data clean-up and preparation so that they can perform their analyses
- Access to suitable, and likely familiar and preferred, analysis tools
- Flexibility to develop their own custom or tailored analysis methods

Risk-based IT Security Officer. Information assets and technologies should be adequately protected, and in any modern environment there will be a role to ensure this takes place. This can be thought of as "trust, but verify" what the end user is able to do and is doing (ie, authorized uses only). To make this scalable, guardrails are set up in advance to ensure the end user can do their job but in a safe enough way. Risk-based implies that the tolerance around security controls will be commensurate with the sensitivity of the information and expectations of what the end users can and should be doing. The interests of the risk-based IT security officer include:

- Reducing information security risks through mitigating controls and audit logs
- Selectively restricting access to what data is needed, by whom, and only when they need it
- Preventing the loss of data, whether it be accidental or intentional

Privacy Officer. Whereas a security perspective will aim protect all organizational information assets, privacy has emerged as an additional consideration focused on societal rules of behavior, decorum, and civility.(5) Security plays an integral role, and the privacy perspective brings attention to how people represented in data feel regarding the uses of identifiable information about them. Ethics will often be a consideration, with the privacy officer concentrated on the applicable legal frameworks that attempt to codify these societal norms and values into an established set of principles for the responsible sharing and use of sensitive data. When privacy is also evaluated in terms of human rights, it is done in the context in which privacy issues or protections are being considered, and with regards to all other human rights, including economic, social, and cultural rights (ie, it is contextual). The interests of the privacy officer can therefore be summarized as:

- Reducing privacy risks from working with personal information and codified data
- Respecting legally established data subject rights based on jurisdiction
- Norms for trustworthy data processing, including governance and privacy technology

Zero-trust IT Security Officer. The concept of zero-trust, or trustless, security is a relatively recent IT architecture design principle in which a breach is always assumed and every request for data is verified as though it's coming from outside (ie, never trust, and always verify). This perspective has also been introduced as a technical solution to privacy, although it doesn't really address responsible uses of data. The interests of the zero-trust IT security officer include:

- Eliminating all information security risks with always on protection
- Risks from everywhere, and from everyone, assuming every interaction is a threat
- An end-to-end protection strategy with provable guarantees

As was mentioned, these represent an evolving set of perspectives that should, in theory at least, build on one another. However, we need to go back to the beginning and ask what end users think of all these added perspectives, and how they could limit their ability to achieve their goals of

delivering insights that improve efficiencies and health outcomes. While aiming to deliver trustworthy systems that address potential privacy concerns, we need to ensure that the ultimate goal of producing safe, useful, and timely data is still possible in the eyes of those professionals that derive value from working with sensitive data. Eg, do they believe their needs will be met? Will they be forced to change how they work? Is it even realistic to ask them to change how they work in practice?

Recall that the end users in health sciences include the preprocessing, or data engineering, required to prepare data for analytical uses across the data lifecycle, as well as providing the flexibility to develop novel and tailored analyses that achieve the greatest insights. This will inform the type of privacy enhancing solutions that are suitable for safe and responsible data sharing and analytics.

Spectrum of Tools

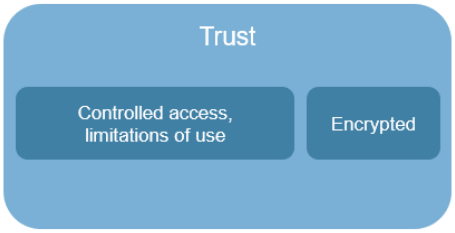
Before we consider a spectrum of privacy-enhancing technologies, we will first look at categories in which they may be considered. This will allow us to consider the different stakeholder perspectives that were introduced, and how they may perceive these tools that are intended to address their needs.

Trust Based

Traditionally a risk-based approach has been used, as described from the perspective of the risk-based IT security officer. For sensitive data, this has met the needs of the end users because it usually allows them to do any preprocessing that they require prior to doing their statistical analysis, and they can get access to the analytical software tools they need for analysis (including the most recent algorithms, and customization for more advanced modelling) From the end user's perspective, this is better termed a trust-based approach. We can therefore summarize this category of tools as:

- Trust (but verify) end users to do their job ethically and responsibly
- Provide access to authorized users/roles to minimize exposure
- Encrypt data in transit and at rest to avoid breaches

→ Confidentiality, in the form protecting information from unauthorized access, is the primary set of tools considered in this category.



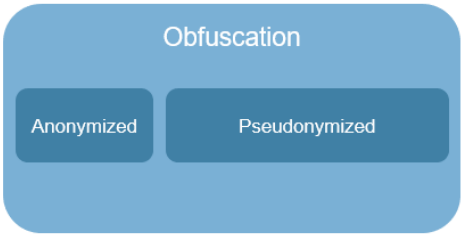
Obfuscation Based

While the trust-based tools serve an important role, primarily in protecting confidentiality, the sensitive data that is made available to end users may nonetheless raise privacy concerns. The well-established concept of data minimization can reduce some of these concerns, or more advanced methods may be used to remove the personal from data, especially for purposes other than what the data was originally collected for, known as secondary purposes. Legislation to

ensure non-identifiable information is used for secondary purposes is becoming increasingly common, although the HIPAA Privacy Rule has required this for large classes of health information since 2003.(6) We can summarize this category of tools as:

- Obfuscate to hide personal information, or confuse what the information actually contains
- Pseudonymize to hide pieces of information with a coded replacement
- Anonymize or de-identify to hide and confuse to create non-identifiable information

→ Disassociability, in the form of removing the personal from data, based on the use case and needs to ensure the appropriate degree of useful data is still available



The use of trust-based and obfuscation-based tools is well established in health services & research because they typically meet the broad set of needs for complex data and complex data pipelines that involve advanced statistical analysis by experts.(7) Privacy concerns nonetheless exist and have resulted in another category of emerging tools.

Limited to Zero Trust

Here we begin to combine the concepts of confidentiality and obfuscation into a single category of tools. We can start with limiting trust so that the input data to statistical analyses is protected.

From a security engineering perspective, and maintaining the confidentiality of inputs, we find tools that:

- Hide input data entirely from the end user or anyone else (eg, intruders)
- Hide the computations performed on data, for the purposes of statistical analysis, entirely from the end user or anyone else (eg, intruders)

While the input data and computations are entirely hidden through means of encryption (in a broad sense), in our experience there are simple ways to unpack and reveal the underlying input data unless additional measures are introduced to prevent this from happening. The most sophisticated approach would be a database reconstruction attack, although there are much easier ways to do this when a person has access to any computational function in the applicable software library that implements such approaches.

From a privacy engineering perspective, and disassociating the input data, we find tools that:

- Obfuscate input data to hide and confuse the information, rendering it non-identifiable in the circumstances of use (eg, anonymize or de-identify)
- Provide access to the outputs of statistical analyses only, which are less granular by definition (ie, statistics are summaries)

The key advantage of using these approaches is that the data feeding into an analytics pipeline is in this case non-identifiable. That can mean the difference between having access to the input data in the first place, such as for secondary purposes in which the use of personal information may not be permitted. To ensure the most statistically useful input data is available, however, reasonable assurance that the input data is non-identifiable will in part rely on access to statistical analyses only. Concerns around the ability to unpack and reveal the underlying input data will still exist, although this is less of a concern since the input data is at least partly minimized.

For these reasons, we can also extend to limiting trust so that the inputs and outputs to statistical analyses are protected.

From a security engineering perspective, and maintaining the confidentiality of input data and statistical outputs, we find tools that:

- Do all of the previous, ie, hide input data and computations entirely
- Limit operations to a set of protocols that are believe to avoid reconstructions of input data

From a privacy engineering perspective, and disassociating the input data and statistical outputs, we find tools that:

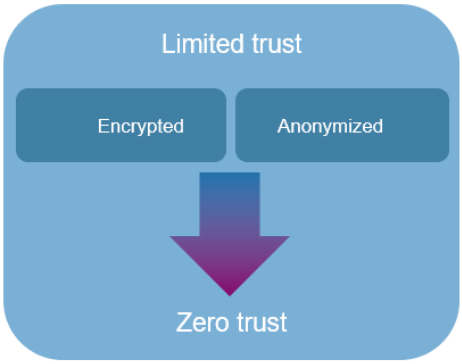
- Do all of the previous, ie, obfuscate input data to hide and confuse, and provide access to statistical outputs only
- Limit output disclosure by "checking" what the statistical outputs are and deciding whether or not they can be provided (using manual or automated means)

Both these security and privacy engineering perspectives, requiring limits imposed on operations that can be performed or checking statistical outputs, can be manual, daunting, and seemingly ad hoc no matter how well thought out they may be.(8) Which is why another class of tools are emerging.

Recall the zero-trust IT security officer assumes there is always a breach, and systems should be designed assuming they will be in a constant state of breach. This perspective has evolved to include provable security, and attempts at provable privacy (although much more complicated to define because privacy is a societal good without clear boundaries).

From a security and privacy engineering perspective, we find tools that are emerging in an attempt to provide both confidentiality and disassociability to inputs and outputs with provable guarantees in one form or another. In practice, and especially for health data, we consider tools that:

- Satisfy a narrow set of use cases due to their complexity, computational performance, possible need of specialized hardware, and significant restrictions on end users

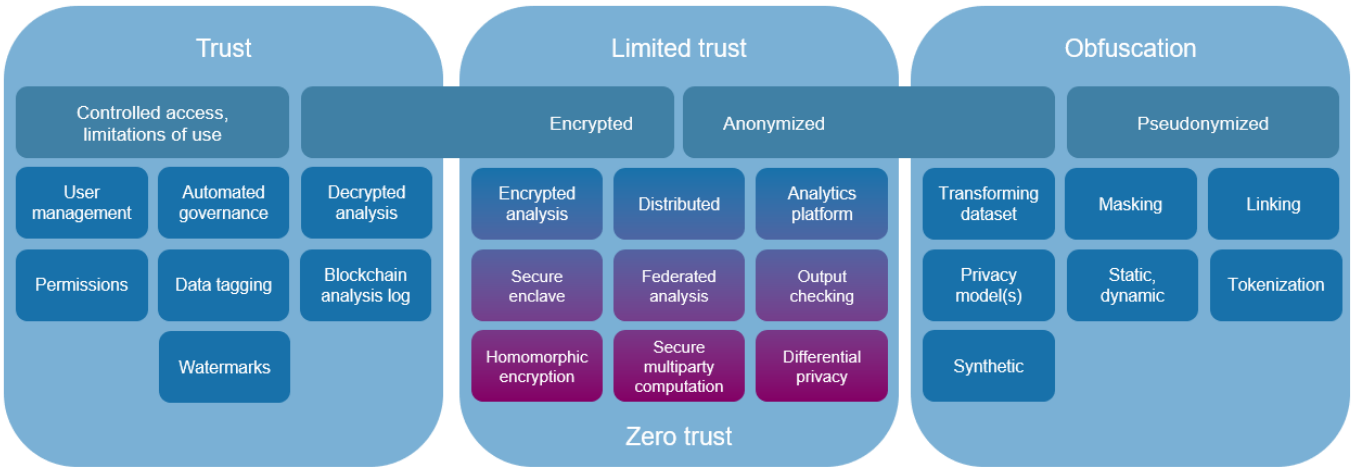


The health researcher, for example, would need to know in advance that they have clean, well formatted input data, with a limited set of predefined statistical analyses in order to use many of these tools. The reality of health data is that of complex structures, sparsity, and disparate formats that rarely line up across departments in the same organization let alone different sites. While study protocols are often defined in advance, for example to get ethics approval, they would rarely if ever define the exact statistical algorithms that will be used. In more sophisticated and established data pipelines, such as in drug development, even static data is refactored and reharmonized to suit the various acrobatics of statistical analysis.

The truth is that a great deal of analysis goes into deriving meaningful statistical results from health data: evaluating various forms of bias, understanding error distributions, inspecting outliers, testing assumptions, refining algorithms and methods, etc. While there are good examples of zero-trust tools being used, in our experience they are often academic or pilots of limited use more broadly, and ill-suited to the realities of health service improvement and research intended to meaningfully improve health outcomes.

Full Spectrum

With the above framing and description of categories, we now provide a more detailed view of the full spectrum of tools (inspired by (9)). We provide this diagram without delving deeper into each tool or subcategory, in the hopes that it will at least motivate discussion and exploration.

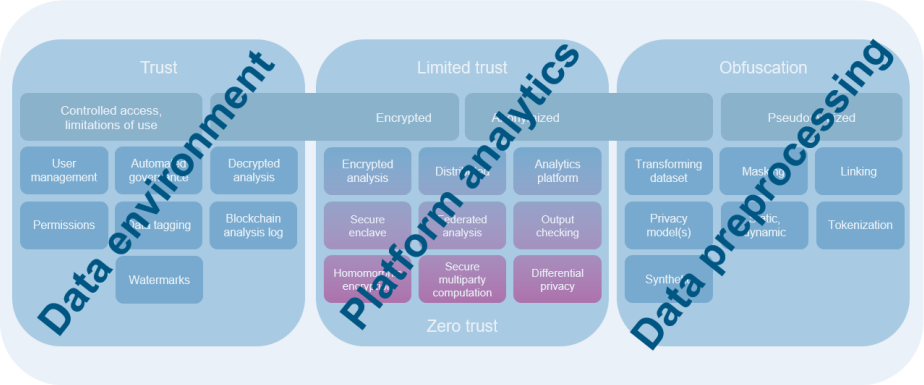


Tools in Practice

PETs and Data pipeline

Another way to think about these tools is in terms of their application areas in a data pipeline, ie, secure data environments (to preserve confidentiality), data preprocessing to reduce privacy risks (disassociation), and platform analytics to further protect through a combination of technical security and privacy controls. A view of privacy tech in the data pipeline, from sensitive information to insightful analytics, can be described as:

- *Data preprocessing:* privacy processing to disassociate or remove the personal from data, in preparation for statistical analyses, from pseudonymized to anonymized
- *Data environment:* protecting confidentiality of sensitive data in the environment in which the data will be used for statistical analyses, including the security posture, protecting against unauthorized access, and technical data governance (non-PETs related protections are out of scope for this discussion)
- *Platform analytics*: protecting confidentiality and possibly disassociating in an environment of limited trust, where access to the data by end users is through the platform only, with a minimal view into outliers, and minimal ability to clean or prep the data



PETs and Data Lifecycle

We can now revisit the data lifecycle we described in the introduction, and consider where different tools may be deployed from sensitive data to insightful analytics (a glossary is provided in the appendix that describes the different tools):

Get data

- Creation, collection by privacy model (anonymization)
- Pseudonymization
- Randomized response, local differential privacy
- Simulated data

Link data

- Federated statistics or AIML
- Secure linking and dataset anonymization
- Secure multiparty computation

(Re)Use data

- Analytics platform (output checking, global differential privacy)
- Anonymized, synthesized dataset
- Homomorphic encryption

Share data

- Anonymized, synthesized dataset
- Personal data store (ie, releasing control, allowing individual control)
- Secure enclave or environment

Open data

- Aggregation and statistical outputs
- Anonymized, synthesized dataset
- Redaction or summarization
- Simulated data

Conclusions

While we may have provided some sweeping generalizations to challenge certain ideas and perspectives, our goal is very much to drive a conversation so that more practical and applied methods get the attention they deserve. Our hope is that the adoption of tools the enable privacy-enhancing data sharing and analytics can be increased through an increased focus on real end users that need safe, useful, and timely data and analytics that can work with the complexity and scale of real health data and modern health challenges.

In our experience, it is the entire spectrum of privacy-enhancing tools that are needed for the safe enablement of data and analytics, depending on the needs of end users and the specific uses cases being deployed. The more practical approach is therefore, in our opinion, a combination of tools rather than any one tool. While this may seem obvious to some, it bears mentioning so that more effort is put towards the integration and deployment of suites of tools that are interoperable and complimentary.

We wish to thank you again for this opportunity to provide our views on the operational challenges with, and development needs for, privacy-enhancing data sharing and analytics. We hope that you have found our feedback helpful and insightful towards developing a national strategy on this topic. We look forward to participating in future consultations, such as exploring more detailed views on the interplay between different tools that support privacy-enhancing data sharing and analytics.

References

1. Clinical Data Interchange Standards Consortium. Study Data Tabulation Model (SDTM) v2.0 [Internet]. Clinical Data Interchange Standards Consortium. 2021. Available from: https://www.cdisc.org/standards/foundational/sdtm/sdtm-v2-0

2. PhRMA. Biopharmaceutical Research & Development: The process behind new medicines [Internet]. PhRMA. Available from: http://phrma-docs.phrma.org/sites/default/files/pdf/rd_brochure_022307.pdf

3. Stephen Bamford, Sarah Lyons, Luk Arbuckle, Pierre Chetelat. Sharing Anonymized and Functionally Effective (SAFE) Data Standard for Safely Sharing Rich Clinical Trial Data. Applied Clinical Trials [Internet]. 2022; Available from: https://www.appliedclinicaltrialsonline.com/view/sharing-anonymized-and-functionally-effective-safe-data-standard-for-safely-sharing-rich-clinical-trial-data

4. Office of the Privacy Commissioner of Canada. Privacy Enhancing Technologies: A Review of Tools and Techniques [Internet]. Gatineau, Canada; 2017. Available from: https://www.priv.gc.ca/en/opc-actions-and-decisions/research/explore-privacy-research/2017/pet_201711/

5. Daniel J Solove. "I've Got Nothing to Hide" and Other Misunderstandings of Privacy. San Diego Law Review. 2007;44:745.

6. Office for Civil Rights. Guidance regarding methods for de-identification of protected health information in accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule [Internet]. Washington, DC: Department of Health and Human Services; 2012 [cited 2021 Jul 7]. Available from: https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html#_edn9

7. 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.

8. O'Keefe C, Chipperfield J. A Summary of Attack Methods and Confidentiality Protection Measures for Fully Automated Remote Analysis Systems. International Statistical Review. 2013;81(3):426–55.

9. Mobey Forum. The Digital Banking Blindspot: Emerging Privacy Enhancing Technologies and their Role in Privacy Risk Mitigation and Business Innovation [Internet]. 2021 p. 17. Available from: https://mobeyforum.org/the-digital-banking-blindspot/

Glossary

Anonymization: transformation, including synthesis, of data with inclusion of privacy model and controls

Differential privacy: noise addition to produce indistinguishable outputs up to a defined information limit

Federated analysis: combining the insights from the analysis of data assets without sharing the data itself

Homomorphic encryption: encrypted data that can be analyzed without decryption of the underlying data

Output checking: verifying disclosure risk of analysis results conducted on confidential data

Privacy model: syntactic evaluation of data threats or formal proof of information limits

Secure enclave: isolated execution environment to ensure integrity of applications and confidentiality of assets (aka trusted execution environment, or confidential computing)

Secure multiparty computation: combining the encrypted insights from the analysis of data assets without decrypting the underlying insights themselves

Simulated data: artificially generated data by a theoretical, representative model

Synthetic data: artificially generated data by a statistical or learning model trained on real data

Tokenization: secure process of substituting sensitive data elements with non-sensitive and secure data elements