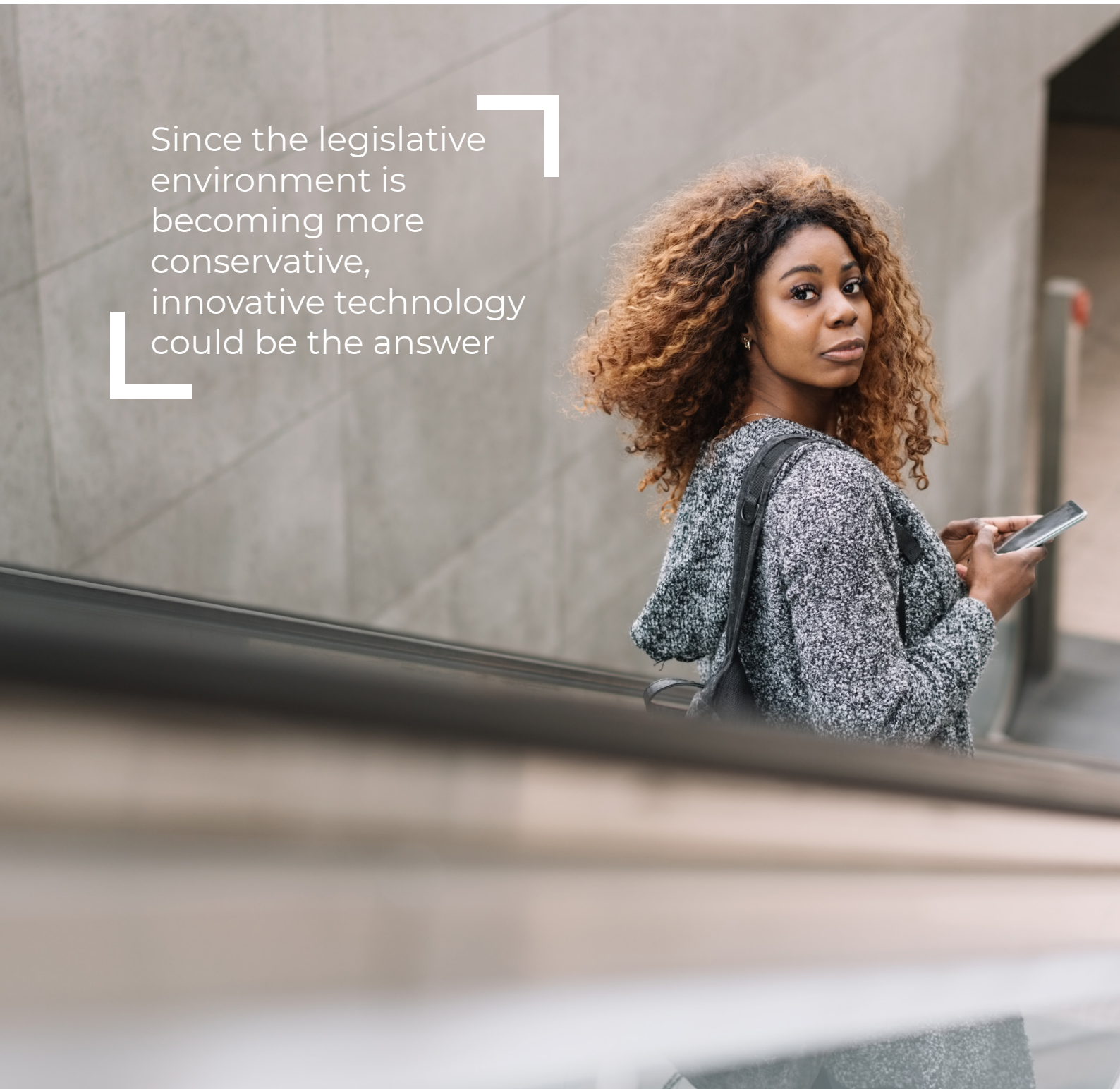
Cerner Enviza℠

Technology solutions for privacy challenges

Dr Jessica Santos and Tom Haskell

Privacy challenges have existed for some years now and intensified in the last decade. We have witnessed the enactment of privacy laws all over the world, including record-breaking fines for privacy breaches, tightening of regulators sanctions, wide spread of data localization[1] and restrictions on cross-border data transfer — all second to the increasing demand of consumers (data subjects) needing more privacy protection. So, what is a solution to these challenges? Since the legislative environment is becoming more conservative, innovative technology could be the answer. Here are some examples:



1. Homomorphic encryption

Homomorphic encryption is a form of encryption that permits users to perform computations on its encrypted data without first decrypting it. These resulting computations are left in an encrypted form which, when decrypted, result in an identical output to what would be produced if the operations had been performed on the unencrypted data. Homomorphic encryption can be used for privacy-preserving outsourced storage and computation. This allows data to be encrypted and outsourced to commercial cloud environments for processing, while encrypted[2]. Fully homomorphic encryption (FHE) is an encryption scheme that enables analytical functions to be run directly on encrypted data while yielding the same encrypted results as if the functions were run on plaintext[3].

For **example**, a medical researcher wants to compute descriptive statistics on a population of lung cancer patients at a hospital. The **complication** is the hospital is unable to share its private medical records with the researcher due to the HIPAA privacy rule. The **resolution** is that hospital encrypts its sensitive data using a fully homomorphic encryption scheme. As a result, the data can be computed on while still being protected.

How it works: The hospital homomorphically encrypts its medical records and sends them to the medical researcher's cloud computing environment or allows remote access. Because the data is encrypted, it is fully protected and private in the cloud. Next, the researcher runs its analytical functions on the homomorphically-encrypted data in the cloud, manipulating the data while it remains encrypted. Finally, the researcher downloads the encrypted output, and decrypts the result to reveal the plaintext answer. The sensitive medical record data is encrypted end-to-end and is only decrypted when revealing the final answer behind organizational firewalls.

The **advantage** of homomorphic encryption is the data remains **secure and encrypted at all times** which minimizes the likelihood that sensitive information gets compromised.

This also **eliminates trade-off** between data usability and data privacy as there is no need to mask or drop any features in order to preserve the privacy of the data, i.e., the traditional anonymization technique to remove all potentially personal identifiers. All features may be used in an analysis, without compromising privacy. Finally, it is also **quantum-safe**. Fully homomorphic encryption schemes are resilient against quantum attacks.

However, homomorphic encryption does have some performance limitations: between slow computation speed or accuracy problems, fully homomorphic encryption remains commercially challenging or infeasible for computationally heavy applications. Nevertheless, it is useful for cases that are not computationally intensive, like prediction using a pre-trained model or in conjunction with other privacy-enhancing technologies like secure multiparty computation.

2. Synthetic data

As the name suggests, synthetic data is data that is artificially created rather than being generated by actual events. It is often created with the help of algorithms and is used for a wide range of activities, including as test data for new products and tools, for model validation and in AI model training[4,5].

The biggest criticism for synthetic data is obvious, it is not real data. Few case studies indicate that regulatory bodies will accept synthetic data as part of evidence-based findings given the increasing scientific rigor towards the generation of regulatory grade real-world evidence (RWE) generation. Other limitations include missing outliers, heavy reliance on original data quality, and the fact that any synthetic models derived from data can only replicate specific properties of the data, i.e., they will only be able to simulate general trends.

However, synthetic data has several benefits over real data:

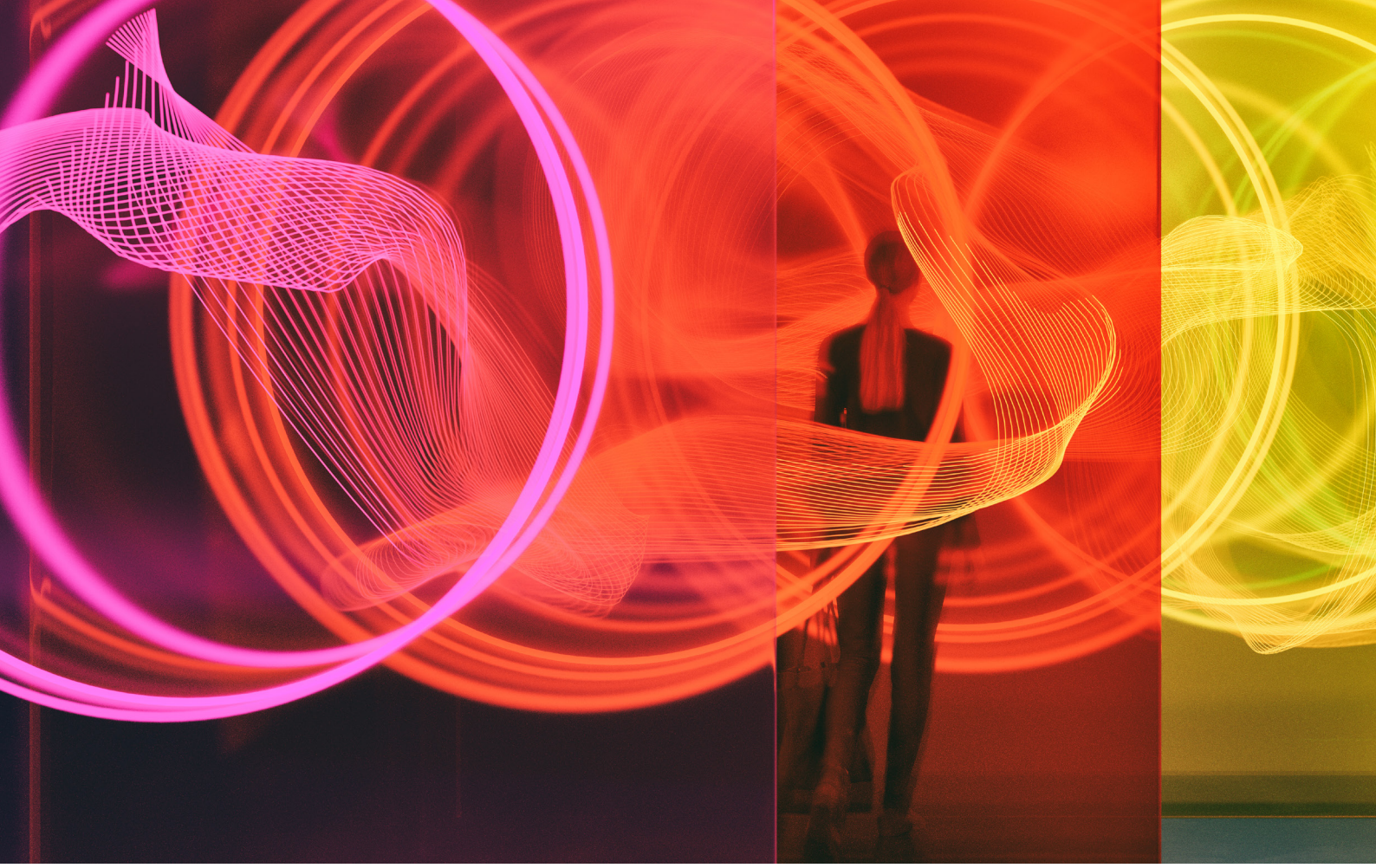
— **It can overcome real data usage restrictions:** Real data may have usage constraints due to privacy rules or other regulations. Synthetic data can replicate all the important statistical properties of real data without exposing it, thereby eliminating the issue.

— **It can simulate conditions not yet encountered:** Where real data does not exist, synthetic data is the only solution.

— **It provides immunity for some common statistical problems:** These can include item nonresponse, skip patterns and other logical constraints.

— **It focuses on relationships:** Synthetic data aims to preserve the multivariate relationships between variables instead of specific statistics alone.

— **It is able to work with small databases:** Traditional statistical machine learning techniques (e.g., deep learning) typically work for large datasets, but generally will not do well on small datasets. Synthetic data can utilize source data with a sample size of only a few hundred.



Synthetic data can be very useful as an option to meet specific needs or conditions that are not available in existing (real) data, especially in the following **use cases:**

— Privacy requirements limit data availability or how it can be used

— Data is needed for testing a product to be released, however, such data either does not exist or is not available to the testers (e.g., restriction on data rights)

— Training data is needed for machine learning algorithms, but such data can be scarce, expensive, or very hard to generate in real life (e.g., rare disease)

Researchers can also use different methods for creating synthetic data. **Fully synthetic data** does not contain any original data, re-identification of any single unit is almost impossible, and all variables are still fully available. **Partially synthetic data** that is sensitive is replaced with synthetic data, requiring heavy dependency on the imputation model where some disclosure could be possible. **Hybrid synthetic data** is derived from both real and synthetic data, and the relationship and integrity between other variables in the dataset are maintained while the underlying distribution of original data is investigated, and the nearest neighbour of each data point is formed.

Although user acceptance and validity are still a challenge for both researchers and regulatory bodies, synthetic data can be a useful option in the toolbox for algorithm generation and enhancements together with other instruments.

— **User acceptance is more challenging:** Synthetic data is an emerging concept, and it may not be accepted as valid by users who have not witnessed its benefits before.

— **Synthetic data generation requires time and effort:** Although it is easier to create than actual data, synthetic data is also not free.

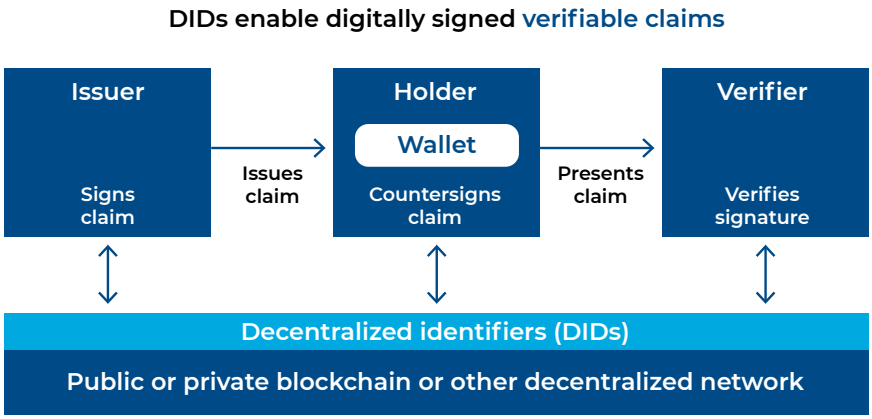
— **Output control is necessary:** Especially in complex datasets, the best way to verify that the output is accurate is by comparing synthetic data with authentic data or human-annotated data. This is because there could be inconsistencies in synthetic data when trying to replicate complexities within original datasets.

3. Self-sovereign identity

Self-sovereign identity (SSI) is an approach that gives individuals control of their digital identities. It is a term used to describe the digital movement that recognizes an individual should own and control their identity without the intervention of administrative authorities. It allows people to interact in the digital world with the same freedom and capacity for trust as they do in the offline world[6].

Everyone, even businesses and the internet of things (IoT), has unique sets of identifying information, personable identifiable information (PII), protected health information (PHI) or personal/confidential data. In the physical world, these are represented by cards and certificates that are held by the identity holder in their wallet or in a safe place like a safety deposit box, and they are presented when the person needs to prove their identity or something about their identity. SSI brings the same freedoms and personal autonomy to the internet in a safe and trustworthy system of identity management.

With SSI, the power to control personal data resides with the individual, rather than an administrative third party granting or tracking access to these credentials. SSI addresses the difficulty of establishing trust in an interaction. In order to be trusted, one party in an interaction will present credentials to the other parties, and those relying parties can verify that the credentials came from an issuer that they trust. In this way, the verifier's trust in the issuer is transferred to the credential holder. This basic structure of SSI with three participants is sometimes called "the trust triangle".[7]



It is generally recognized that for an identity system to be self-sovereign, users control the verifiable credentials they hold, and their consent is required to use those credentials[8]. This aims to reduce the unintended sharing of users' personal data and is contrasted with the centralized identity paradigm where identity is provided by some outside entity.[9]

Applied to the healthcare research environment, SSI can be used for patient identification, consent management, unique data verification, clinical monitoring, pharmacovigilance safety surveillance, and other purposes which traditionally all require the exchange of PII with the potential conflict of privacy requirement on data minimization.

With SSI, the biggest change is instead of building trust to endless external parties (e.g. a bank, a retailer, a healthcare provider to a potential business partner, or any data recipient), the trust will be established through a SSI system, which is decentralized. Similar to the rise of cryptocurrency, this paradigm shift might be seen as dramatic, but is not impossible.

4. Differential privacy

Differential privacy is a system for publicly sharing information by describing patterns of groups, not individuals, or a rigorous mathematical definition of privacy. In the simplest setting, consider an algorithm that analyses a dataset and computes statistics (such as the data's mean, variance, median, mode, etc.).[10] Such an algorithm is said to be differentially private if by looking at the output, one cannot tell whether any individual's data was included in the original dataset or not. In other words, the guarantee of a differentially private algorithm is that its behaviour hardly changes when a single individual joins or leaves the dataset. Anything the algorithm might output on a database containing some individual's information is almost as likely to have come from a database without that individual's information. Most notably, this guarantee holds for any individual and any dataset. Therefore, regardless of how eccentric any single individual's details are, and regardless of the details of anyone else in the database, the guarantee of differential privacy still holds. This gives a formal guarantee that individual-level information about participants in the database is not leaked.[11]

Consider an individual who is deciding whether to allow their data to be included in a database. For example, it may be a patient deciding whether their medical records can be used in a study, or someone deciding whether to answer a survey. A useful notion of privacy would be an assurance that allowing their data to be included should have a negligible impact on them in the future, which is the most important rationale behind privacy protection (that is to avoid potential harm and consequence). Although it is argued that absolute privacy is inherently impossible, what is being implied here is that that the chance of a privacy violation is small. This is precisely what differential privacy (DP) provides.[12]

Differential privacy formalizes how we define, measure and track the privacy protection afforded to an individual as functions of factors like randomization probabilities and the number of times their data is recorded. Differentially private algorithms are able to answer a large number of queries (e.g., percentage of patient with condition X on treatment Y who live in Z), so that researchers seeing these approximate answers can draw roughly the same conclusions as if they had the data themselves.

Differential privacy is not simply data aggregation. Aggregate-only restricts the analyst to run summary statistics like averages, counts and sums over a dataset. It may contain a minimum size constraint so that aggregate queries cannot be run on datasets containing only a few entries. Differential privacy augments aggregate-only policies by adding random noise into the analysis in order to obscure the impact of any single record. It is considered "fool proof" anonymization, providing a mathematical fail-safe to protecting privacy.[13]

It is worth remarking that differential privacy works better on larger databases, because as the number of individuals in a database grows, the effect of any single individual on a given aggregate statistic diminishes. It also has the potential limitations of reducing utility and losing accuracy[14], hence the trade-off between precise accuracy versus absolute effective privacy together with the expectations of the data subjects generates a level of trust from data providers.

5. Database tokenization (data linking without data sharing)

Database tokenization utilizes token-enabled algorithms to facilitate data linking without data sharing. The demand of generating insights from multiple data sources is extremely high due to the limited variables and scope of a single data source. For example, de-identified patient-level data linking has become extremely common in the medical research community mainly because more valuable evidence can be generated from multiple patient-level sources. Notably, if a researcher wants to understand the economic burden of illness for overweight patients from commonly available sources, they will need to link electronic health record (EHR) data (for BMI information) with insurance claims data (for cost information).

The biggest challenge of linking multiple healthcare datasets is usually these datasets come from multiple sources, and the data owners do not have the rights to provide PHI for these patients to third parties to facilitate the linkage due to privacy compliance. (e.g., HIPAA or GDPR). Traditional methods of linking databases are either through personal data (which has huge data protection challenges) or lookalike modelling (which is widely used by insurance companies, but its accuracy and projective power are questionable). Numerous companies have addressed these challenges by creating data tokenization software that is installed at the data owner's location. This software usually takes advantage of one-way hashing algorithms on various patient attributes (usually, some combination of name, sex, date of birth, geographic location, social security and other identifiable attributes) to create a de-identified token that can then be compared against similarly created tokens from other datasets to find matches without the PHI leaving the data owners' sites. If a patient is identified by the same token in different datasets, it is assumed that they are the same patients, and their data can be combined for analysis.

The biggest advantage of database tokenization is the original source data **will never need to leave the hosting location**, the data controller has full custody of their databases, and no accuracy will be lost during data linking as a result. Database tokenization is becoming more and more widely accepted in the industry as a valid method for linking patients, especially in healthcare research. Multiple companies have created such linking software and many of the available healthcare datasets have these tokens assigned, facilitating the linkage by healthcare researchers.

Like all methodologies, there are limitations to this approach. Some of the potential limitations of database tokenization are as follows:

— While linking datasets at the patient level is quite accurate, it is still a probabilistic approach, i.e., we have no way of validating that two patient records with the same token refer to the same person. There will always be a certain number of false positives and false negatives. While this relatively small amount of uncertainty can be acceptable for certain types of healthcare research (as long as it is documented), it is unclear whether it will be accepted by the FDA as part of submissions.

— Because linking multiple datasets will almost always require that the date of service be known, the HIPAA Expert Determination method for the de-identification of datasets must be used. Every combination of linked datasets needs to be examined by a statistical expert to ensure that there is minimal chance of re-identification. This process can be time-consuming and expensive.

The benefit of database tokenization, especially the conformance to data protection regulations, certainly outweighs its challenges for most researchers. It is generally accepted for US healthcare data at the moment, but has the potential to reach beyond its geographic and usage boundaries

Conclusion

All instruments noted above are gradually gaining awareness but are still in their infancy, and are a long way away from being accepted by regulatory bodies as a gold alternative standard to traditional methods. As each country/region's privacy regulation is getting stricter, sanctions on data breach are getting heavier and data sharing and transfers are becoming more limited, the demand of better, faster and more insightful data output has increased. Innovative technology could offer the necessary solutions for the future.

References

1. https://incountry.com/blog/data-residency-laws-by-country-overview/
2. https://eprint.iacr.org/2015/1192
3. https://inpher.io/technology/what-is-fully-homomorphic-encryption/
4. https://research.aimultiple.com/synthetic-data/
5. https://replica-analytics.com/
6. https://sovrin.org/faq/what-is-self-sovereign-identity/
7. https://en.wikipedia.org/wiki/Self-sovereign_identity
8. http://www.lifewithalacrity.com/2016/04/the-path-to-self-soverereign-identity.html
9. https://ec.europa.eu/futurium/en/system/files/ged/eidas_supported_ssi_may_2019_0.pdf
10. https://www.microsoft.com/en-us/research/project/database-privacy/?from=http%3A%2F%2Fresearch.microsoft.com%2Fen-us%2Fprojects%2Fdatabaseprivacy%2Fsensitivity.pdf
11. https://privacytools.seas.harvard.edu/differential-privacy
12. https://www.borealisai.com/en/blog/tutorial-12-differential-privacy-i-introduction/
13. https://www.immuta.com/articles/differential-privacy-a-sound-way-to-protect-private-data/#:~:text=Aggregate%2DOnly%20policies%20restrict%20the,over%20a%20dataset.&text=Differential%20Privacy%20(DP)%20augments%20aggregate,impact%20of%20any%20single%20record.
14. https://www.tau.ac.il/~saharon/BigData2018/privacybook.pdf

Authors:

Jessica Santos, PhD, CIPP

Global Head of Compliance and Quality, DPO

Jessica.santos@cernerenviza.com

Tom Haskell

Global Head of RWE Data Products

Tom.haskell@cernerenviza.com

About Cerner Enviza℠

Cerner Enviza aims to accelerate the discovery, development and delivery of extraordinary insights and therapies to improve everyday health for all people globally. By combining decades of innovation, life sciences knowledge and collaborative research, Cerner Enviza provides data-driven solutions and expertise that helps bring remarkable clarity to healthcare's most important decisions. For more information on Cerner Enviza, visit **www.cernerenviza.com**.

For more information, please contact **info@cernerenviza.com**