Questions for Hayo Schreijer and Joep Meindertsma (Dexes)

Elize Vlainic asks:
Could you explain who would perform the audits in your previous picture?

This question was answered by Hayo during the webinar in the first round of questions in the Q&A (after the presentations, before the panel discussion).

Chang Sun asks:
Who will host the personal data vaults? How do you search in the personal data vaults without going through their personal sensitive data?

This question was answered by Joep during the webinar in the second round of questions in the Q&A (after the presentations, before the panel discussion).

Yan Wang asks:
How to handle potential conflicts (in different business scenarios) among the separated agreements made with data onwers?

"We treat conditions for sharing data like legislation; higher legislation rules over lower legislation and rules. When a conflict arise that can't be solved by logic, the data space governance system kicks which a user has agreed to follow its ruling before entering the data space."

QUESTIONS FOR GUIDO VAN 'T NOORDENDE (WHITEBOX SYSTEMS)

Ronald Siebes asks:

Does this metaphor apply for this approach: with diamonds you need a certificate where you got it from in order to legally sell it or make jewelry with it. Would this then be possible with DNA data. To legally work with it, you need a certificate signed by the owner(s) that contains information what the certificate owner can do with it?

"Irrespective of the method for 'signing' statements or policies regarding what may be done with DNA data, it is imperitive to realize that there is no single party that can sign a statement that allows for (re-)usage of DNA data: the data is not only held by one person or patient, but by genetic/hereditary family members, which extends also to future generations.

The solution that I have sketched in my talk, is to keep track of data dynamically, and over the lifetime of the DNA (which is, by nature, several generations), by distributing unique references to the data and keeping track of all (distinct/unique) copies of those references instead of copying the data itself. Tracking of references should be done at a predictable location which can always be found by (future) generations, where legal constraints (e.g., on keeping audit logs) may be legally enforced and thus can be expected to be maintainable for a long period of time – that is, over multiple generations.

Obviously, DNA sharing should be purpose-bound and as limited as possible, where the recipient of the data (reference) accepts a strict and legally binding data management policy imposed by the source – to start with, "don't copy this data, request and pass a reference instead".

I believe that these requirements can only be held by medical professional organizations that (already) have to meet legal requirements that make them responsible for medical data over extended periods of time. Current legislation already imposes stringent restrictions on medical data sharing due to professional secrecy law; this legislation may require some strengthening (i.e., on how long logging and provenance data should be kept) but already forms a strong basis for the proposed solution.

Combined with general and medical-specific data protection rights, when patients keep track of where they were treated, untractable data sharing may become a thing of the past and may be prevented for DNA data."

Elize Vlainic asks:

This year during a Health-RI event, the Public Health Train was promoted. In what ways does 'White Box' work in a similar way/ collaborates with/ compliments/ competes with Personal Health Train services

"The Personal Genomic Locker (PGL) project is a project that originates from PHT. Whitebox Systems contributes to that project and I am appointed in a researcher role within this project. There are similarities in keeping data at the source and in application of FAIR principles (without loosing privacy), however, we find that processing DNA and the "personal" aspect of the PGL – see the discussion above – poses unique challenges and issues that are not addressed by the PHT directly. In particularly, how to track data after obtaining it, is an issue in the PHT project as it is an issue in many projects to date. The Whitebox "reference" scheme ("copy references, not data") may pose a solution here, for both PGL and PHT."

This question was also answered by Guido during the webinar in the first round of questions in the Q&A (after the presentations, before the panel discussion).

Freek Dijkstra asks:

Question for Guido: you emphasis keeping data at the source. However, most general practisionars or pharmacists are not (computer) security experts. How to solve that?

"Being a computer literate and keeping data at the source are orthogonal aspects and in effect quite unrelated. One can keep data at the source, irrespective of whether this is physically on-site at a doctor's practice or in a data center, as many current-day GP systems are.

The core issue that I am making is that data that originates from a medical professional intervention, i.e., that is created under medical professional secrecy constraints under the legal responsibility of a doctor or medical institute, should stay there. Or more precisely, it should remain under the legal responsibility of the medical professional (organization) where the data was/is initially stored. This is what I call the "source". The core point is that we should stop copying away data from that source, and instead should start referring to that source record. This way, the source can maintain responsibility over tracking and tracing who has access to (a reference to) this record. Importantly, such a health organization 'source' - or its legal successor! - will have a legal obligation to maintain audit trails of where data went for what purpose – in contrast to where a (research) party is allowed to make a copy of the data, after which it is easy to loose sight of where the data goes – particularly over a long period of time (multiple generations). The same applies to when patients make copies of data into some online "personal" data storage environment, which may have mixed incentives, and which may even cease to exist at some point in time, resulting in data that was shared from this environment becoming untractable. The point is that such a data management/copying model is not necessary, since it is straightforward to share references (pointers) to the original data instead of sharing copies of data, with the same or better results – ensuring data remains tractable over a long period of time.

In the proposed scheme, keeping data at the (medical professional) source is imperative because the medical professional domain can be held to constraints such as auditing, whereas the legal private domain (governed by the GDPR) holds much less stringent legal constraints. In particular, patients may share their data for research using a consent form that basically allows unrestricted sharing with other researchers. Such a consent may lead to this data escaping from sight not only to the original data 'owner' or subject (who gave consent), but also to his or her (future) family members – which in case of DNA data is a real problem, and a problem that we should worry about now instead of leaving it to future generations."

This question was also answered by Guido during the webinar in the second round of questions in the Q&A (after the presentations, before the panel discussion).

Questions for Sofie Hansen (Lygature)

Freek Dijkstra asks:

Podium does not process data itself, but helps in the access management. If you wanted to integrate with a data processing tool, would you consider a data modification (anonymization) tool, or prefer a tool that adds more control how the data is actually used after it is shipped/send to the researcher?

"Currently we leave this decision up to the organization that is providing the data. However, we are looking to connect Podium[1] to SRAM[2] which helps verify the identity of the requester."

Elize Vlainic asks:

Are there additional plans to add Patient Reported Outcomes for example, so all data from e.g. clinical research be included, to ensure not just general demographics can be added to the biological data, but checked/analyzed in combination?

This question was also answered by Sofie during the webinar in the second round of questions in the Q&A (after the presentations, before the panel discussion).

Linda Rieswijk asks:

In the worst case scenario; how do you avoid that copies are being used by other researchers? Is this also registered somewhere? Is there a "tracker" on the data?

This question was also answered by Sofie during the webinar in the first round of questions in the Q&A (after the presentations, before the panel discussion).

Jeroen Rombouts asks:

A question as a reaction to Sofie's answer. Would Podium and Whitebox work well with large non anonynimisable data (like video recordings in which facial expression and original sound is important)?

"Podium is a request management tool only. Data access and delivery happens outside of our tool e.g. via remote desktop."

[1] https://podium.bbmri.nl/

[2] https://wiki.surfnet.nl/display/SRAM/SURF+Research+Access+Management+-+SRAM+Home

QUESTIONS FOR WIM-KEES JANSSEN (SYNTHO)

Michiel Schok asks:
When training AI, do you need to know what algorithm is going to be used in data-analysis?

"No. To actually verify this, when using synthetic data for the first time, we make sure that we collaboratively define some metrics to evaluate the quality of the synthetic data in comparison to the original data for that specific use case or algorithm.

In addition to these custom metrics that we define together, we provide a detailed quality report (70+ pages document). This report consists of several descriptive statistics, such as distributions, correlations and multivariate distributions. Furthermore we show that even a machine learning model can barely distinguish original data from the synthetic data. This ensures that also the more deep 'hidden' relationships, not easily detectable by the human eye, are well captured in the synthetic data.

More on the evaluation of the data quality can be found in this blog: https://www.syntho.ai/synthetic-data-preserves-statistical-properties/"

This question was also answered by Wim-Kees during the webinar in the second round of questions in the Q&A (after the presentations, before the panel discussion).

Wim Schaasberg asks:
Syntho must have access to the original sensitive data. Is Syntho a black box program or are there operators who have access to the personal data?

"The Syntho Engine can be installed in our clients on-premise or private cloud environment. This safeguards that the data remains within their own secure IT environment."

Leonieke van den Bulk asks:
Does this approach also work for data with a big amount of variables? Is the AI still able to capture all the correlations correctly?

"Yes, if there is sufficient data. As rule of thumb, we advise an additional 500 rows per extra column (variable), although the exact ratio depends on data complexity."

This question was also answered by Wim-Kees during the webinar in the first round of questions in the Q&A (after the presentations, before the panel discussion).

Paul Leeraert asks:
Can AI data be reversed to real data when needed?

"Synthetic data does not have a 1-to-1 relationship with the original data, in contrast to data manipulated with classical anonymization techniques. As this is one of our unique selling points, it also implies that use cases that require re-identification of individuals are not supported by synthetic data."

Yan Wang asks:
Generating 'more' synthetic data? I'm not sure if there is any research ethics issue for certain disciplinary research, e.g. using data that didn't exist in reality may lead to quality/validity issue of the research. "The data quality is good, but they are not the right data" could be one argument there.

"That is correct. If the use case is statistical research, we always advice to use a synthetic dataset with a similar N. An example use cases where increasing N would be valuable is in (stress) testing with software, where you would like to evaluate how your application handles larger amounts of data."

Tiemen Folkers asks:
I agree with Yan Wang. Also, you say the data is comparable with real data. But can this synthetic data be tracked back to "subjects" if somebody doesn't know this is synthetic data. So real people are being "indentified" and perhaps even approached, while they never really where involved. So fake subjects could be linked to real persons. Did you take this in account?

"Yes. While we've never experienced a situation where we happened to generate a synthetic data string that matches a real individual, we've built in safety nets to control for both exact and close matches. Furthermore any direct identifiers are replaced by their 'dummy' counterparts."

Wim Schaasberg notes:
Zie ook Mu-argus een open source programma om micro data veilig te maken door gaten te maken in de data set.

"This an example of a classic anonymization technique. With classic anonymization, we imply all methodologies where one manipulates or distorts an original dataset to hinder tracing back individuals. Typical examples of classic anonymization that we see in practice are generalization, pseudonymization and row and column shuffling and in this example suppression or wiping.

Manipulating a dataset with classic anonymization techniques results in two key disadvantages:

1. *Distorting / manipulating a dataset results in decreased data quality (i.e. data utility) and may render a dataset unusable.*

2. *The privacy risk may be reduced, but will always be present because the manipulated data records preserve their 1-to-1 relationship with the original data. Today, individuals can easily be re-identified by having access to only a few attributes.*

This introduces the trade-off between data utility and privacy protection, where classic anonymization techniques always offer a suboptimal combination of both.

This blog illustrates the concept of classic anonymization techniques in comparison to synthetic data: https://www.syntho.ai/classic-anonymization/

And this blog describes the risk of a linkage attack as a result of having 1 to 1 relations in a dataset with applies classic anonymization techniques: https://www.syntho.ai/5-examples-why-removing-names-fails-as-anonymization/"

QUESTIONS FOR ROBERT GRIFFIOEN (SURF)

Barbara Vermaas asks:

When will the Surf pilot go to production. You mentioned next year, but is that at the start, or later on in the year....?

"At this moment, ZorgTTP is preparing their proposal including the planning, so until we agree on this proposal and the planning, nothing is certain. However, I can tell you that the beginning of next year is not realistic. The project is officially planned until the first of July and this is still a realistic date. So, to give you an indication, I think you should count on the beginning of Q4 2021."