# Data Protection Impact Assessment (DPIA) for vantage6

## The Personal Health Train Implementation

**Version history**

| *Version* | *Date* | *Authors* |
|---|---|---|
| 0.1 | August 27, 2018 | Gijs Geleijnse, Ferdinand Impens |
| 1.0 | February, 2020 | Gijs Geleijnse, Marilyn Maas, Ferdinand Impens, Daan Knoors, Jaap van Ekris |
| 1.1 | March 26, 2020 | Marilyn Maas, Ferdinand Impens, Gijs Geleijnse |
| 1.2 | April 9, 2020 | Marilyn Maas, Ferdinand Impens, Gijs Geleijnse |
| 1.3 | April 23, 2020 | Gijs Geleijnse |
| 1.4 | June 10, 2020 | Gijs Geleijnse, Daan Knoors |
| 2.0 | June 19, 2020 | Gijs Geleijnse, Daan Knoors |
| 2.1 | November 12, 2020 | Arturo Moncada-Torres |

**Participants**

- Ferdinand Impens – Security Officer (IKNL)
- Marilyn Maas – Legal advisor (IKNL)
- Daan Knoors – Clinical Data Scientist (IKNL)
- Arturo Moncada-Torres – Clinical Data Scientist (IKNL)
- Gijs Geleijnse – Sr Clinical Data Scientist (IKNL)
- Jaap van Ekris – Data Protection Officer (Palga)
- Jan Nygard – Innovation Manager (Cancer Registry of Norway)

**Client**

- Xander Verbeek – Head R&D (IKNL)

**Reviewer**

- Merle Hafkamp (Data Protection Officer)

# Contents

# A. Description of data processing features

## 1. Proposal

Analysis of separate databases and registries allow to create a better understanding of health and care. However, many questions and research projects require the combination of data from different databases and/or institutions. vantage6, the open source implementation of the Personal Health Train, is intended to facilitate such analyses while protecting the privacy of the patient.

For such analyses using data from multiple organizations and/or databases, we distinguish two scenarios (Fig. 1):

- **Horizontally-partitioned data,** where records from one organization are enriched with data from different patients, yet with similar features. An example is the combination of data from multiple cancer registries that cover different geographies and patients. Combining cancer registry data allows for inter-geographical comparisons and creates a large patient volume. The latter is particularly relevant for the research on rare cancers. As by construction, the databases contain data from different patients, matching identifiers between databases is not a concern for horizontally partitioned data.

- **Vertically-partitioned data**, where data on a selected group of patients is distributed across several databases. An example is the combination of data items on cancer patients from the Netherlands Cancer Registry and PALGA. For vertically partitioned data, the identifiers of the patient records from the databases involved should be matched.

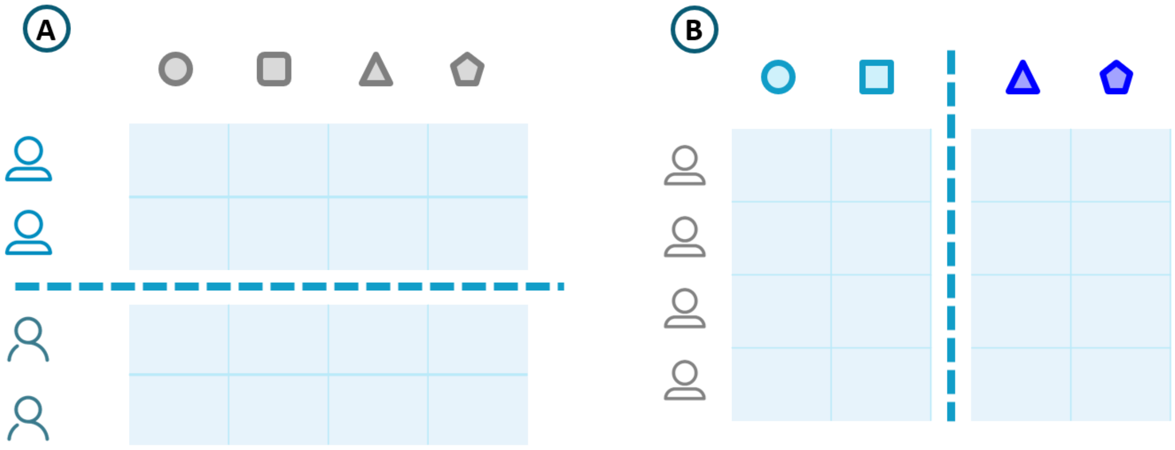

Figure 1. (A) Horizontally-partitioned data contains records from multiple organizations with the same features from different patients (e.g. cancer registry data from the Netherlands and Czech Republic).
(B) Vertically-partitioned data contains records with different features with the same patients (e.g. cancer registry data from the Netherlands and socio-economic data on these patients from CBS Statistics Netherlands). Image adapted from (1).

Traditionally, for both scenarios, datasets are prepared and shared with a researcher or analyst ("Client") after fulfilling the data request procedures at the data providing organizations involved. This means that patient-level data leave the organizations and is brought together on the machine of the data analyst.

In the recent years, concerns around ensuring patient privacy have increased, making data providing organizations more hesitant to share patient-level data with third parties. On the other hand, to progress our knowledge on healthcare in general and cancer in particular, there is an increasing need to combine both horizontally as well as vertically partitioned data.

### The Personal Health Train

The Personal Health Train (PHT) is the Dutch national initiative to provide a solution to patient-level data sharing concerns. The PHT is a paradigm to enable analyses of data from multiple organizations, without identifiable data leaving the organization. This privacy-by-design paradigm enables researchers to conduct their analyses, while not accessing or "seeing" individual patient records. By keeping data at the source, no copies of the datasets are generated that are shared with third parties.

vantage6 is the open source implementation of the PHT (1, 2). The development of vantage6 is currently coordinated by IKNL but allows contributions from all interested. Following the metaphor of the Personal Health Train, we identify:

- **Stations:** locations where data is hosted that is made available for analyses using the PHT. Data providing organizations can host a station themselves or they can work with an external party to host the data (e.g., a cloud provider).

- **Rails:** the technical infrastructure that connects the stations. vantage6 is the infrastructure that implements authentication and authorization, such that the right parties are connected in the right way.

- **Trains:** statistical analyses on the data stations. An analysis script is composed of multiple trains (e.g., containing descriptive statistics, collecting information for tables and figures as well as more advanced regression and machine learning analyses).

- **Journey:** A full study involving one or more stations with dedicated datasets connected via the rails with a researcher (Client), who can send a predefined selection of trains to these stations.

In this document, we assess the privacy impact for the vantage6. It is intended to be independent of specific collaborators, algorithms, and data sets. This document assumes a separate Privacy Impact Assessment for regular data requests and analyses and focuses solely on the privacy impact for the usage of vantage6 in data analysis projects.

## 2. Description of vantage6

For a journey on the PHT using vantage6, we distinguish the following computer architecture (Fig. 2):

- **Client:** a computer of a researcher, epidemiologist, or other professional requesting insights via a *journey*

- **Station:** a (virtual) machine where one or more datasets for a participating organization is stored and made available. With each journey, a dedicated dataset is associated.

- **Central server**: a machine where the journey is managed, and communication between stations is orchestrated and computations are performed on non-identifiable data and statistics received from the stations.

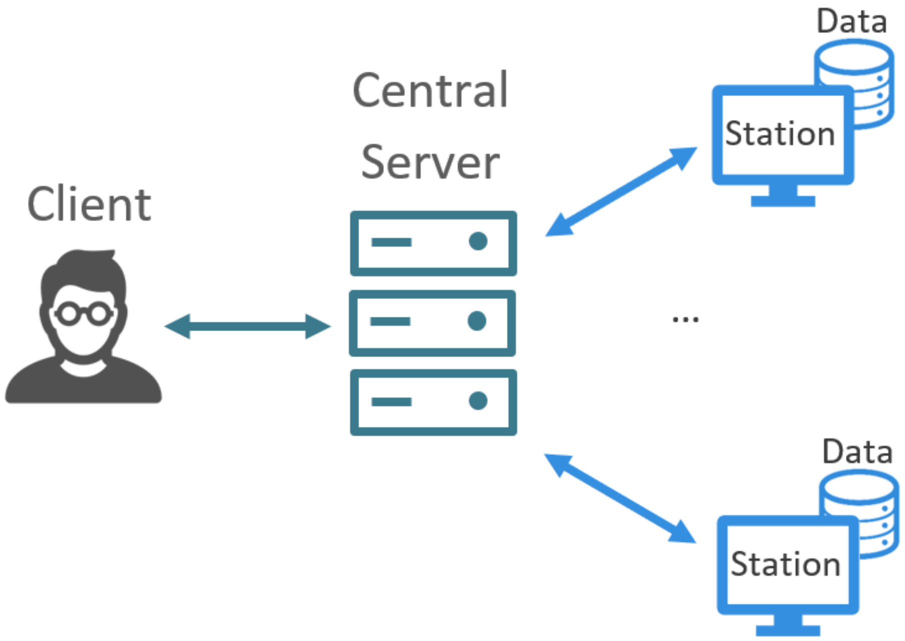

Figure 2. High level IT architecture of vantage6. The arrows denote communication between components over the internet. Image adapted from (1)

In practice, multiple organizations may be involved:

- The client's organization ("*Client or Data requesting Party*")
- The organization providing the PHT service and managing the central server ("*Central Server Manager*")
- The *Central Server Provider*, the party hosting the central server on behalf of the central server management (e.g., a cloud provider).
- The organizations hosting the data stations- whether or not - on behalf of the data providing organizations ("*Providers of PHT stations*")
- The data providing organizations ("*Data providing organizations*")

An example of such a set up for a journey can be found in Fig. 3.

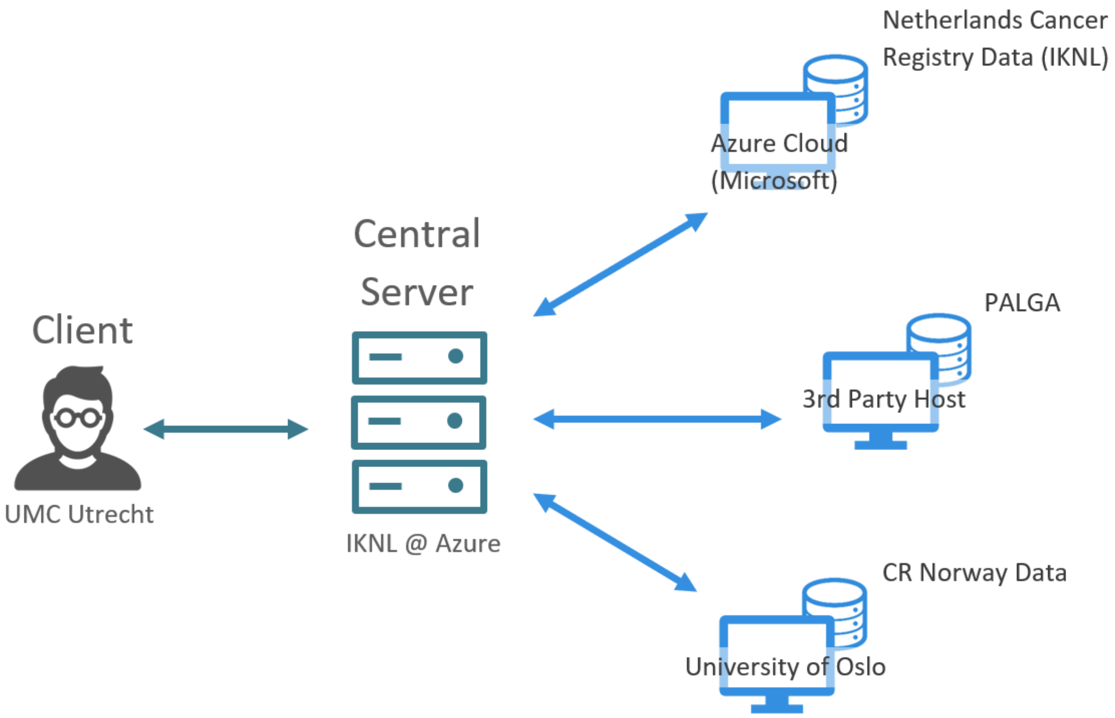

Figure 3. An example of a possible journey with collaborations between three data providers (IKNL, Palga, CR of Norway) and a client at a university

The roles in vantage6 correspond to the ones defined in the note by Bontje (3) (Fig. 4).

| | Role in (3) | Role in vantage6 |
|---|---|---|
| **Data requesting site** | Opdrachtgever van de PHT trein | Data requesting party / Client |
| **PHT Domain** | Aanbieder van de PHT trein (provider of the PHT train) | Central Server Manager |
| | Aanbieder van PHT station | Provider of PHT station |
| **Data provider site** | Aanbieder van PHT data | Data providing organization |

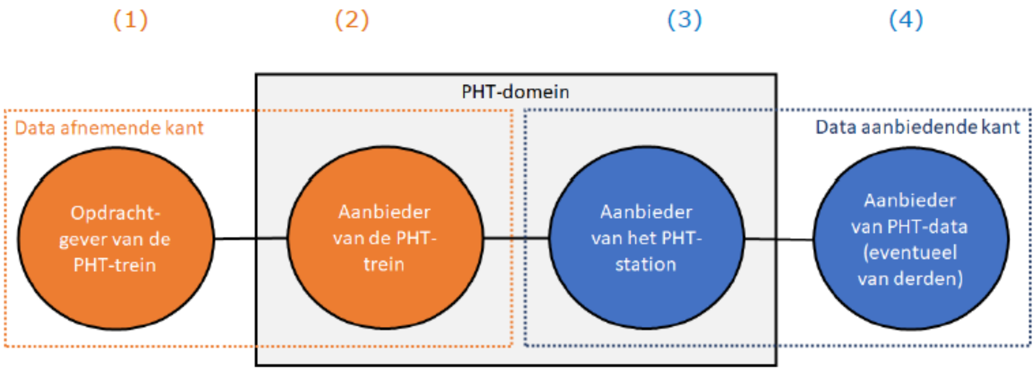

Figure 4. Roles as defined in "privacyaspecten van de personal health train" (3)

## 3. Trains in vantage6: Federated Learning and Multi-Party Computation

vantage6 trains can be categorized according to two different mathematical principles: Federated Learning and Secure Multi-Party Computation.

### Federated Learning

Federated learning is typically applied for horizontally partitioned data (i.e. organizations provide data from disjunct cohorts of patients, yet providing same characteristics/items). Federated learning is based on the mathematical principle of splitting a computation into (a) parts at the stations and (b) a central part. The stations share sub-computations with the central server.

For example, let's suppose one is interested in the average age of all patients in a cohort

$$Average = \frac{\sum_{i=1}^{n} age(i)}{n}$$

Here, $age(i)$ is the age of patient $i$ in a group of $n$ patients.

Now suppose that the group of $n$ patients is now divided over multiple organizations, that do not want to share the age of individual patients. We then implement a federated algorithm, where each station computes the following:

1. $\sum_{i=1}^{m} age(i)$ – The sum of the age of all patients in the cohort at the station
2. $m$ $-$ the number of patients in the cohort at the station

The central server now collects these statistics from all participating stations. To compute the average over the entire group, it takes the sum of sums of the ages (1) and divides this by the sum of the number of patients in each cohort (2).

This principle of splitting computations into a central- and a station-part is illustrated here by a simple example, but can be applied in complex computations as well (4–8).
Note that the technique of splitting computations as explained above only works for horizontally partitioned datasets.

For vertically partitioned data, federated learning can be used to approximate centralized calculations for some tasks. At the moment of writing, one algorithm (for logistic regression (1)) was included in the vantage6 library. However, such calculations cannot be generalized for all analyses required in epidemiological research.

### Multi-Party Computation

Similar to Federated Learning, Secure Multi-Party Computation (MPC) enables various organizations to perform a joint analysis without the need to share raw sensitive records. However, instead of mathematically decomposing an algorithm, MPC relies on a toolbox of cryptographic techniques that allows several different parties to jointly compute data, just as if they have a shared database. These techniques are used to protect the data, so it can be shared in a way that prevents the parties involved from ever being able to view the other party's data. However, the protection, if set-up correctly in the form of a protocol, allows one to still perform mathematical operations on this encrypted data. At the end, only the final result is revealed and the participating parties determine who is allowed to view the outcome of the computation.

Let's use the same example as above to calculate the average age of some participants through an MPC protocol. For this, we can use a secure sum algorithm.

Let us assume that there are 3 registries (A, B and C having respectively a, b and c patients) that want to know the total number of patients of all registries (N=a+b+c) but do not want to share their number of patients with the other registries. However it is accepted that the other registries learn the average of the other two registries. An algorithm that solves this problem is show in Fig. 5:

1. A generates a random number R
2. A add this number to its patient count x1=a+R
3. A shares x1 with registry B
4. B adds its patient count x2=x1+b
5. B shares x2 with registry C
6. C adds its patient count x3=x2+c
7. C shares x3 with registry A
8. A subtracts the random number N=x3–R
9. Optionally A shares the result N with B and C

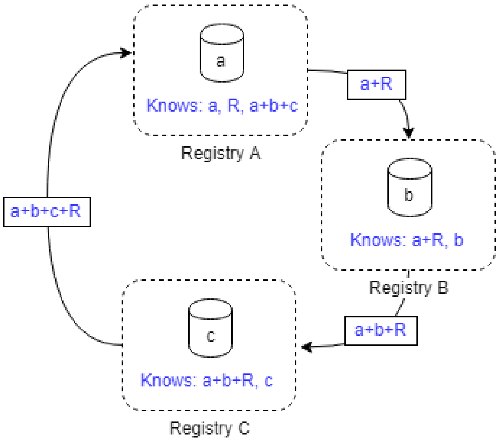

Figure 5. Example algorithm for MPC.

Having calculated the number of participants securely (i.e. the denominator of the average age formula), one can follow a similar process to obtain the sum of all ages in each registry.

Observe that, through this example we show how we can still share and do computations on encrypted numbers that do not reveal anything about their true values. With a bit more complexity this scenario extends itself to full databases, including varying column types and dimensions. Contrary to federated learning, MPC is a bit more flexible and can be used for both horizontally and vertically partitioned data. It has technical guarantees to perform all computations securely and thus keep all input and intermediate results encrypted. Only the final result is revealed and thus this prevents any inferences being made on intermediate output. Hence, the security properties provide superior privacy protection, but do come at the cost of added algorithm complexity, computation time and communication rounds, i.e. many MPC protocols require that trains need to pass by stations multiple times. Therefore, MPC solutions are more difficult to develop, interpret and often require a custom fit for specific use case and data dimensionality to achieve the required efficiency.

## Train Certification

In vantage6, a *train* is an analysis script implemented in a Docker container. Docker is a technology to execute a script on a machine without installation of additional software packages. If a programmer creates an analysis in a certain programming language (e.g., version 3.1.5 of the language R), Docker creates a virtual machine, i.e., it compiles and executes the analysis script as was generated on the machine of the programmer.

In order to transform a software script to a vantage6 train, a Docker container of the script is created. The Central Server Manager will send this Docker container to the data station in order to execute the analyses as was defined in the journey.

To certify that the Docker container corresponds to the corresponding script, vantage6 makes use of Docker Notary (https://docs.docker.com/notary/getting_started/). Docker Notary allows to verify the author of the Docker container. At the moment it is the *de facto* technology to implement this functionality.

## 3.1 vantage6 in Practice: Adoption in Workflow and Processes

To use vantage6, we identify two steps: (1) the deployment of vantage6 at the organization and (2) the process of using vantage6 for a single study/journey.

**Deployment of vantage6**

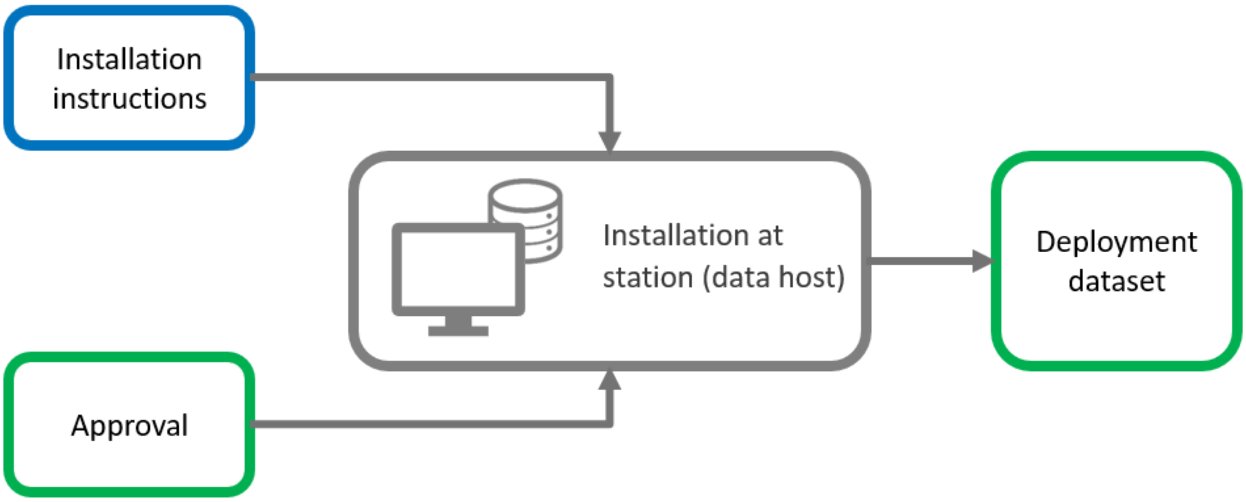
Installation instructions
Installation at station (data host)
Deployment dataset
Approval

Figure 6. A data providing organization will be working with the PHT service provider (in blue) for the technical installation of the vantage6 software. This will either be done within their own IT infrastructure or at a Data Hosting organization. Installation will take place after receiving the approval necessary. vantage6 is ready for usage once a dataset from the organization is (and may) be made available within the vantage6 infrastructure.

Approval may not only include approval for local usage and installation, but also a contract with other data providers providing a framework to facilitate studies using their respective data.

**Usage vantage6 for a single journey**

Once all participating organizations are prepared to partake in any journey using vantage6, researchers (“Clients” in grey) can utilize vantage6 to conduct their studies. We visualize this process in Fig. 7 - from study idea to execution.

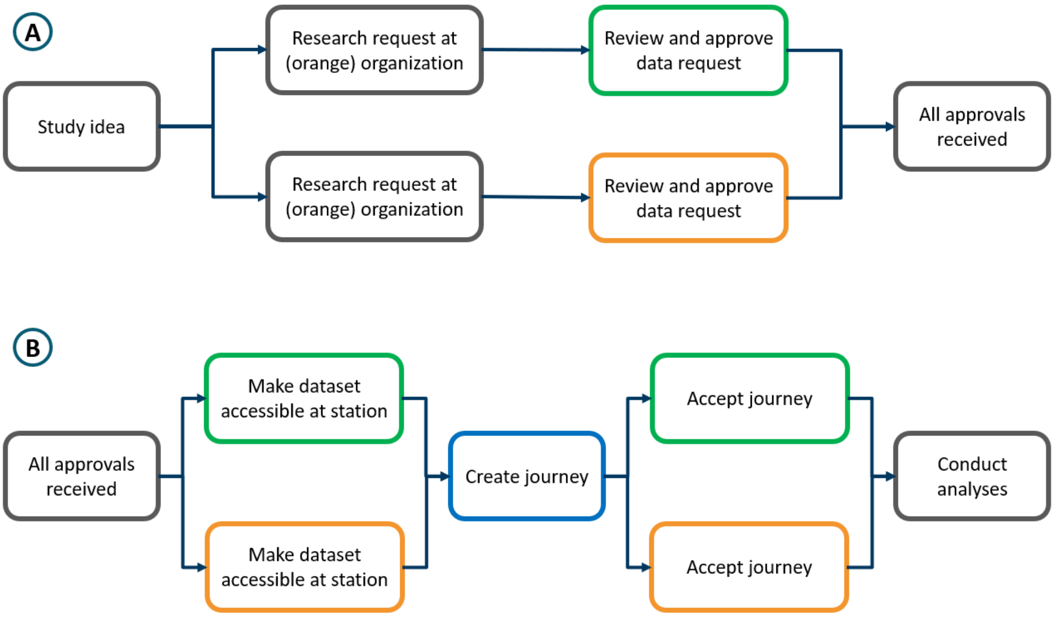
A
Study idea
Research request at (orange) organization
Research request at (orange) organization
Review and approve data request
Review and approve data request
All approvals received
B
All approvals received
Make dataset accessible at station
Make dataset accessible at station
Create journey
Accept journey
Accept journey
Conduct analyses

Figure 7. (A) Seeking approval according to data usage/research request processes at all participating organizations. In this example, we assume two (green and orange).
(B) From Approval to Analysis on vantage6. The organization managing vantage6 and the central server manager creates a “Journey” (defining trains, privacy aspects, stations, dataset, users) that is to be accepted by the data providing organizations.

Following the flow-chart, we envision the following steps:

1. A Client (researcher or user) has a study idea.
2. They file data usage requests to the designated bodies responsible for each of the data sets. The data usage requests defines the journey:
    a. Which data providing organizations are involved?
    b. Which datasets are requested (from all organizations)?
    c. Which trains (analyses) are to be conducted on these datasets?
    d. Who (besides the Client) should have control to execute the analyses?
3. These organizations all review the research/data usage requests
4. After receiving all requests, the Client requests the participating data providers to make the dataset accessible in their Stations. After doing so, these datasets cannot be accessed/analyzed by any outside party yet.
5. The PHT service provider (blue) defines the journey according to the specification in the data request. It associates the datasets made available in the stations with the journey. Moreover, the trains are selected that can be used (and can only be used) to analyze the data. Lastly, the user(s) of the system are identified and logins are created.
6. Before being able to conduct any analysis, all data providers are required to accept the journey. The specification of the journey is shared with these organizations and they are invited to review and compare with the original data request. The approvals are logged both locally at the station as well as at the central server.
7. After having all data providing parties have granted permission, the user(s) can execute their research by running the trains as defined in the journey.

For vertically partitioned data, we assume at this moment that patient IDs between the organizations have been matched. The datasets will be disconnected from the station after the time defined in the data request.

## 4. Data Processing

vantage6 is designed and being deployed to enable the privacy preserving processing of sensitive data. In case of IKNL, vantage6 is anticipated to be used to enable analyses of data from the Netherlands Cancer Registry. Partners in the Netherlands Personal Health Train coalition have requested an analysis of Personal Health Train Technology in respect to the GDPR (3,9). In particular, the legal role of the PHT service provider is assessed. However, Van Graafeiland and Bontje (9) do not consider Federated Learning as an analysis technique to enforce preservation of privacy. In this section, we will therefore argue that the technologies used in this PHT implementation differ from those discussed in these reports. The different technology used may therefore impact the legal status of the PHT service provider.

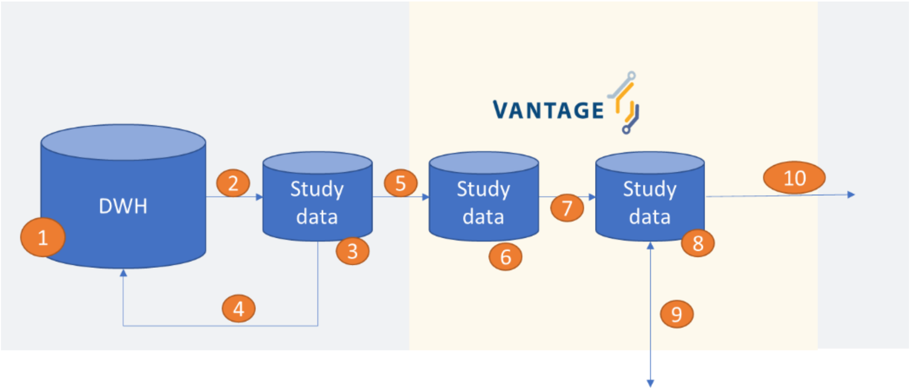

Figure 5. Flow chart of the data processing steps when analyzing a dataset with vantage6.

In Fig. 8, we present the data flow chart from the perspective of a single data providing organization. Here, we assume that

- Either a study/journey involving horizontally partitioned data (disjunct/different set of patients) is proposed, or
- The participating data providers with vertically partitioned data (same or overlapping set of patients, with different items) have a shared identifier that can be used by the trains to match patients in the different databases.

In this case, we identify the following flow of data:

1. We assume the data of data providing organization is stored in a Data Warehouse (DWH) or other central location, managed either by the organization itself or by a contracted data host. Storage of this data may be topic of a separate PIA and out of this scope here.
2. After having received permission for the user’s data request, the data providing organization extracts a dataset dedicated for the study, according to the specifications in the data request. This is a standard procedure for each data request and not specific to this PIA. At IKNL this is typically done by the NCR Analyst handling the data request.
3. The dataset is isolated from the DWH and ready to be shipped. At IKNL this is typically done by the NCR Analyst handling the data request.
4. To enable reproducibility of the study, the data extracted is stored in the DWH or other environment that is routinely used to store datasets associated with data usage requests. At IKNL this is typically done by the NCR Analyst handling the data request.
5. The data provider enables the data set to be accessible in vantage6. Currently at IKNL, this requires a file transfer via SFTP to the Microsoft Azure cloud hosting the vantage6 station. This is done by the colleague at Development responsible for the vantage6 development. Prior to the set, the dataset is received from the NCR Analyst via IKNL Transfer.
6. The data is now stored on the Azure cloud yet not associated with the journey.

7. The organization receives an invitation to approve the Journey. It makes the data now available and discoverable for the vantage6 trains associated with the study. This is currently implemented by the colleague at Development responsible for managing the Data Station.
8. The station will now send pull requests to the central server, requesting trains, i.e. tasks in the form of Docker containers.
9. If the central server has a train ready for the station, it will receive the Docker container. The Docker container will be executed. The results of the computation are sent back to the central server.
10. After the retention period defined in the data request, the dataset will be removed from the vantage6 data station. Via step 4, the dataset (still) can be recovered.

In the case shared patient IDs are required, but not available yet, a trusted third party is used to create pseudo identifiers. The following additional steps are foreseen (Fig. 9):

A. Next to the creating of the study data (1), the associated patient identifiers (e.g. local ID, name, date of birth, address location, data of diagnosis etc.) are extracted from the data warehouse. This is done according to the standard process used by the data providing for ID matching with other organizations.
B. The identifiers are isolated from the DWH and ready to be shipped to the Trusted Third Party (TTP)
C. The identifiers are sent to the TTP using the standard process
D. The identifiers are now stored at the TTP and ready to be processed
E. Pseudo IDs are generated using probabilistic matching with the identifiers provided by the other data providers.
F. A list of pseudo IDs is generated by the TTP, that is associated with the local IDs provided.
G. This list is sent back to the data provider

The process of ID matching using a TTP is therefore no different in conventional data analysis studies. For the remainder of the document, we consider these steps out of scope.

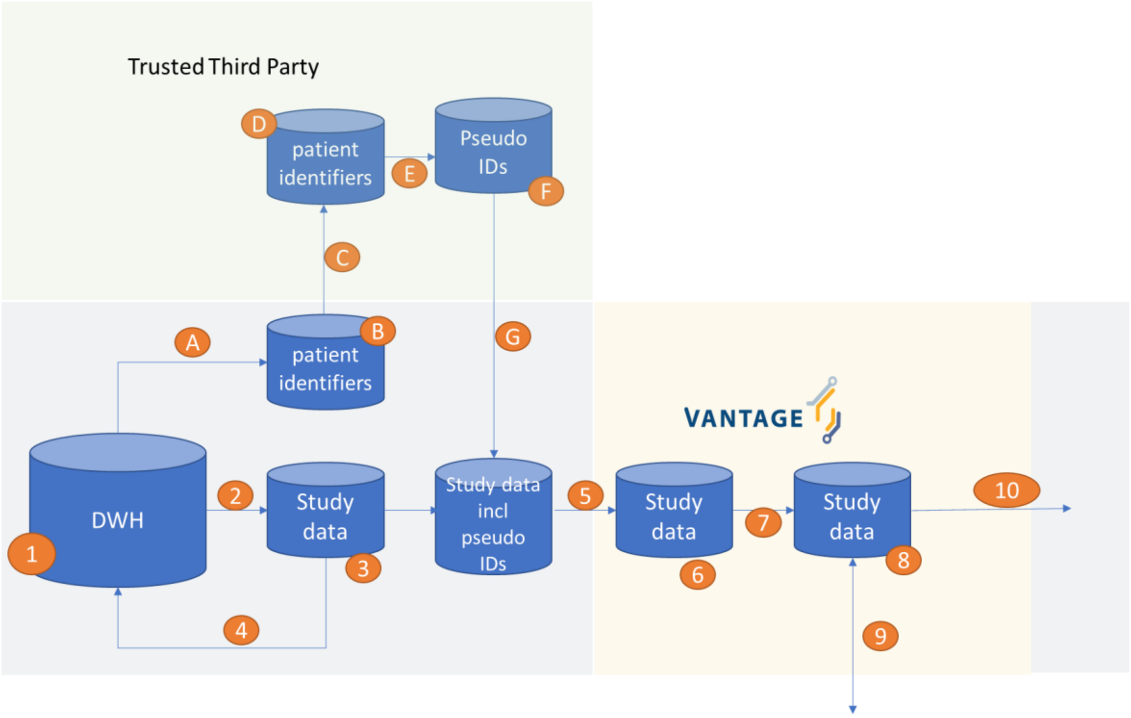

Figure 9. Flow chart of the data processing steps when analyzing a dataset with vantage6 when shared patient IDs are required.

## 5. Processing Purposes

For IKNL, the purpose for processing data using the PHT is no different than for other conventional data analyses: research and statistics related to cancer.

## 6. Parties Involved

- The **Client** and the Client's organization, i.e. the researcher or party benefiting from the statistics and research gained using the PHT. The Client files the data request. Generally, the Client / Data Requesting Party qualifies as a '**Controller**'. It determines the purposes and means of the processing of personal data through PHT.

- The **PHT Service Provider and Central Service Manager**.
  The organization providing the PHT service and managing the central server. Currently, IKNL is fulfilling this role but the open source software allows any party to assume this role in the future. By design, the Central Server does not access data that can be privacy revealing. However, in a recent report (9), Pels-Rijcken argues that it cannot be proven with 100% certainty that despite these measures no identifiable data is processed. Following Pels-Rijcken, we conclude for **<u>the moment</u>** that they are **a Processor** in case of vertically partitioned data. With the right measures in place, it can be demonstrated for the horizontally partitioned data that the party is neither processor nor controller.
    - In case of horizontally partitioned data (federated learning, the central server only receives aggregated statistics – as the stations are only to accept trains that do not return patient-level data). The Stations are responsible for sharing the aggregated statistics and accepting on their data.
    - In case of vertically partitioned data (multi party computation), the central server only receives encrypted datasets. The technology ensures that the central server cannot relate the received encrypted data to any patient in the original dataset. As, within reason, individual patients cannot be identified by the central server, we believe no personal data is processed (10). However, Pels-Rijcken argue that this case cannot be ruled out that a judge would argue differently. To prevent such discussion, for vertically partitioned data, the PHT service provider would be classified as a processor.

- The hosting party utilized by the **PHT Service Provider**. Currently, this is Microsoft Azure with a cloud server in NL/ EU. The status of the hosting party depends on the status of the PHT service provider.

- The **data providers**, including IKNL.

  Generally, the data providers qualify as a '**Controller'**. They determine the purposes and means of the processing of personal data through PHT.

- The organization hosting the PHT station for the data provider, for IKNL this is Microsoft Azure.
  This organization generally qualifies as the '**Processor**'.

We assume that the trains are constructed and deployed according to the basic principle of the personal health train: no patient identifiable data is shared between parties. This is established by sharing either aggregated data (federated learning) or encrypted data (multi-party computation) with the central server.

## 6.1 A note on the Personal Health Train Report by Pels-Rijcken.

In a recent report by Pels-Rijcken (9), the status of the Personal Health Train Service Provider is discussed. Whether this party should be classified as a data controller is dependent on the implementation of the PHT and the trains allowed on the network.

The report assumes that trains are accepted that share encrypted data with the service provider. The GPPR Article 29 working group has stated its opinion (5/2014) on several encryption technologies. This working group assessed these techniques did not meet the three criteria for effective anonymization: person traceability, ability to connect data, and deductibility of personal details.

For the implementation of the Personal Health Train as discussed here, we argue that we have put measures in place that do meet these three criteria.

- The data stations are responsible for only accepting trains that do not disclose privacy sensitive data and guarantee effective anonymization. An on-going effort is to establish trust in these trains and empower organizations to review trains in a meaningful way.

- Multi-Party Computation is a novel encryption paradigm that builds upon some of the 5 techniques as reviewed by the Article 29 WG. The techniques in development at the moment are "encryption with secret key" and "homomorphic encryption". In their report, Pels-Rijcken argue that despite all advanced protective measures, the PHT service provider is required to assume the role of data controller.

- Federated learning does not assume the sharing of encrypted patient-level data. The anonymity of the data can be mathematically demonstrated.

- Additional measures are put in place to guarantee anonymity of data.
    - Use of minimal datasets – no data is to be analyzed and placed on the data station that is not strictly required for the research question to be addressed
    - Use of random selection of data. Instead of including all patients in a cohort, a random subsample can be used for analysis. In this case, a unique patient in the dataset does not need to be unique in the population.
    - Differentially privacy: calibrated randomness can be added to an algorithm or query that processes sensitive data according to the definition of differential privacy, which provides mathematical guarantees that the output of the algorithm is resistant to any form of attack that attempts to infer which individuals are present in the input data.

We believe that these measures will further drive the discussion on the role of the Personal Health Train in regard of the GDPR. We therefore consider this DPIA a living document that will be revisited on annual basis.

## 7. Processing Locations

The locations where data is processed are described in the flow chart under 3.

1. At (the data host of) the data provider
2. At the vantage6 data station – for IKNL this is Microsoft Azure (which has been certified to respect conditions defined in the GDPR)

**Using Federated Learning trains:**

- Trains are certified to only share aggregated statistics with the central server
- No processing of individual patient data takes place outside the data station
- The data providing parties are responsible for accepting trains on their stations. They will verify whether the train indeed does not share any identifiable information.

Federated learning is therefore suited for international collaborations, with data providers outside the EER. As no patient-level is shared across borders or organization, the GDPR is not applicable for as no sensitive data is processed outside the data stations. Of course, each data provider should adhere to GDPR when processing data.

**Using Multi-Party Computation trains:**

- Trains will share encrypted data with the PHT service provider
- Processing of encrypted patient-level data takes place at the PHT service provider, yet the service provider is unable to identify individual patients due to the state-of-the-art encryption techniques applied.
- MPC techniques enable privacy as no single organization can decrypt data collected at the PHT service provider.

## 8. Techniques and Methods of Data Processing Operations

Trains are implemented to provide the functionality of statistical packages that are commonly used in data analysis projects.

The PHT service provider manages and certifies the trains, while the data providers are required to accept a journey including the trains required.

- It is the responsibility of the PHT service provider to ensure that the train used in the journey is the same as specified at the moment when data providers accept the journey
- The PHT service provider will make information available to review the functionality of the trains and test them in a controlled environment (e.g. with fake/synthetic data)
- The data provider will accept the trains based on this information.

The use of vantage6 is not fundamentally different from more conventional ways of performing research as described in the Netherlands Cancer Registry DPIA.

## 9. Retention Periods

In the data request, two periods will be defined:

- The period in which the data will be made available in the PHT station
- The period the dataset will be retained at the organization as defined in the data request, as defined in the NCR DPIA

# B. Assessment of lawfulness of data processing

### 10. Legal Basis

It is important that all parties involved in a journey have a justification of lawfulness. Besides the criteria mentioned in article 6 GDPR (lawfulness) all parties involved need to also have an exception following article 9 GDPR to be able to process special categories of personal data (sensitive data).

For now, it is known that the PHT will be used in settings using sensitive data. It can be considered that for these situations article 6 under f GDPR can be invoked:

> *"processing is necessary for the purposes of the* ***legitimate interests*** *pursued by the controller or by a third party, except where such interests are overridden by the interests or fundamental rights and freedoms of the data subject which require protection of personal data, in particular where the data subject is a child."*

Mostly also article 9 sub 2 under j can be invoked:

> *"processing is necessary for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes in accordance with Article 89(1) based on Union or Member State law which shall be proportionate to the aim pursued, respect the essence of the right to data protection and provide for suitable and specific measures to safeguard the fundamental rights and the interests of the data subject."*

| | Lawfunless criteria |
|---|---|
| **IKNL** | Article 6 GDPR under f, article 9 sub 2 under j GDPR |

### 11. Special Categories of Personal Data

vantage6 aims to enable epidemiological research available in a privacy-preserving manner. The data involved may contain sensitive information.
It can contain personal data, genetic data and/or data concerning the health of individuals.

### *12.* Purpose Limitation

At IKNL, vantage6 will be used for research and statistics in cancer. For each new survey a data application will be filed as described in the process 'Maatwerk gegevensaanvragen'.

## 13. Necessity and Proportionality

Each Client (researcher) will provide IKNL with a research proposal in which both proportionality and subsidiarity are described. Specific attention will be paid to the following:

- Purpose; it must be specified, explicit and legitimate
- Basis: lawfulness of processing, prohibition of misuse
- Data minimization; adequate, relevant and limited
- Data quality: accurate and kept up to date
- Storage durations

The process 'Maatwerk gegevensaanvragen' contains an adequate monitoring of the research proposal both by members of the IKNL-department "NKR analyse' and the Çommissie van Toezicht Nederlandse Kankerregistratie'.

Furthermore the data protection officer is involved in the process of approving an data survey.

Appendix: 'Maatwerk gegevensaanvragen'

## 14. Rights of the Data Subjects

At IKNL, vantage6 will be used for research and statistics on cancer, in particular using the Netherlands Cancer Registry. The PIA for the Netherlands Cancer Registry is applicable.

Appendix: PIA NCR

| Right | Measure | Specific to PHT |
|---|---|---|
| **Transparent information, communication and modalities for the exercise of the rights of the data subject** | The hospitals distribute a folder with patient information. The agreements with the hospitals as defined in the NCR contract are listed in this folder. | No |
| **Right of access by the data subject and Right to rectification** | A process is defined to grant these rights. Engage provides access | No |
| **Right to restriction of processing.** | IKNL will process a request by a data subject | No |
| **Right to data portability** | Not applicable, GDPR article 20 clause 1 | No |
| **Right to rectification/erasure** | In case a patient does not want to be included in the NCR, they will not be registered. If a temporary or final registration already is made in the NCR, the patient with all their data are removed. The process is documented in Engage. | No |

# C. Description and assessment of the risks for the data subjects

## 15. Risks

For this analysis, we deem only the steps 5  - 10 of Fig. 8 relevant, as the other steps are not specific to the PHT. For steps 1 – 4, we refer to the "maatwerk gegevensaanvragen".

| Ref. no. | Step | Risk type | Risk |
|---|---|---|---|
| **1** | 5 | Loss of confidentiality | Unsecure file transfer to Data Station |
| **2** | 7 | loss of confidentiality | Data provider accepting a journey not reflecting the data request |
| **3** | 6,8 | Loss of confidentiality | Hack on Data Station |
| **4** | 9 | Loss of confidentiality | Use of malicious Docker image after failed certification |
| **5** | 9 | Loss of confidentiality | Use of malicious Docker image after hack on the PHT service provider |
| **6** | 9 | Loss of confidentiality | Use of Docker image of malicious train accepted by data provider |
| **7** | 9 | Loss of confidentiality | Use of very small data set such that aggregated data contains identifiable data |
| **8** | 9 | Loss of confidentiality | Authentication not sufficient allowing undesired access to other party |
| **9** | 9 | Unauthorized or unlawful disclosure and/or processing | Client (e.g. a researcher) may use data otherwise than stated in the data request (e.g. commercial application) – risk is not specific to PHT |
| **10** | all | Unauthorized or unlawful disclosure and/or processing | Interception when data is transferred from one location/system to the other. (e.g. man in the middle attack) |
| **11** | 5 | Unauthorized or unlawful disclosure and/or processing | Too much data in dataset (e.g. dob delivered rather than age) |
| **12** | n.a. | Unauthorized or unlawful disclosure and/or processing | Lack of governance structure |
| **13** | n.a. | Unauthorized or unlawful disclosure and/or processing | Patient data on cloud is not according to IKNL policy. |
| **14** | n.a. | | One of the nodes is slow or gets disconnected – research cannot be performed. |

# D. Description of measures planned

## 16. Measures

| Ref. no. | Step | Risk type | Measures | Hazard | Impact |
|---|---|---|---|---|---|
| | | | **Loss of confidentiality** | | |
| 1 | 5 | Unsecure file transfer to Data Station | The data is stored on a secured Azure server, making use of all modern web security standards including safe file transport between the IKNL and Azure servers. | unlikely | moderate |
| 2 | 7 | Data provider accepting a journey not reflecting the data request. | Will result in an unpredictable outcome or the train (algorithm) will not run on the dataset. The client responsible for the study will notice the discrepancy and take actions as the study aim cannot be achieved.<br>In the current way of working, IKNL (in the role of PHT central server manager) is responsible for the definition of the journey. The data providing organizations will review the trains before accepting the journey. As all peers (i.e. all data stations) review the journey, the implementation of the journey is not dependent on one reviewer from one organization, but is a shared effort and responsibility. | unlikely | moderate |
| 3 | 6,8 | Hack on Data Station | To use vantage6 on a data station, Docker and the vantage6 software need to be downloaded from the internet (vantage6.ai). The responsibility for downloading a correct version of the software is with the data providing organization. As it is open source, other, compatible versions yet with undesired functionality may be published on the internet. However, the source code of the installed software can always be inspected and reviewed.<br>The data is stored on a secured Azure server, making use of all modern web security standards. Trusted users review usernames and passwords<br>Future: disable accounts that are not used for 30 days. Log logins and notify Data Protection Officer when suspicious logins occur. Log files of vantage6.ai will be shared with data Station organizations to review data traffic. Authentication, encryption, and security policy will be published and reviewed by IKNL security officer. Said policy will be regularly updated and reviewed.<br>Today, data is delivered to researchers, where IKNL has limited control with respect to storing and copying sensitive data.<br>With the PHT, we address this problem but placing the NCR data on a secure server including a firewall. In PHT projects today, we use of limited datasets.<br>Log files of vantage6.ai will be shared with Node organizations to review data traffic. Authentication, encryption, and security policy will be published and reviewed by IKNL security officer. Said policy will be regularly updated and reviewed. | unlikely | moderate |

| | | | | | |
|---|---|---|---|---|---|
| 4 | 9 | Use of malicious Docker image after failed certification | In the current way of working, IKNL (in the role of PHT central server manager) is responsible for the definition of the journey, including the selection of Docker containers. The data providing organization are to define the Docker containers that are accepted on their stations.<br>If a container is accepted that is not certified, this container may conduct analyses or induce communication that is not specified. This behavior can be observed also when analyzing synthetic data. IKNL will therefore first evaluate the behavior of Docker containers on synthetic data, such that no sensitive data is exposed at the first usage of the container. | unlikely | minor |
| 5 | 9 | Use of malicious Docker image after hack on the PHT service provider | See 4<br>IKNL uses certificates and standard safety-measures on their infrastructure and monitors were applicable. | unlikely | minor |
| 6 | 9 | Use of Docker image of malicious train accepted by data provider | Each data provider (station) is responsible for their own infrastructure.<br>However, in the current way of working is IKNL responsible for the definition of the journey, including the selection of Docker containers.<br>Future: when other parties make algorithms available, the central server manager will (a) review the code by 2 data scientists, (b) publish the review on the GitHub page where the code is stored and (c) test the data communication using the algorithm on synthetic data to detect possible data leaks | unlikely | minor |
| 7 | 9 | Use of very small data set such that aggregated data contains identifiable data | Data requests need to be evaluated as they are today for “normal” requests. If data set is too small, then take corresponding measures. Measures are in DPIA NCR | unlikely | minor |
| 8 | 9 | Authentication not sufficient allowing undesired access to other party. | IKNL (as the central server manager) hosts the authorization of users and thereby the access.<br>No access is granted before all the necessary legal steps have been taken between the partners. | possible | moderate |
| | | | **Unauthorized or unlawful disclosure and/or processing** | | |
| 9 | 9 | Client/Researcher may use data otherwise than stated in the data | No data will be provided unless a signed contract is available between the partners.<br>This are the standard measures, undertaken by NKR-analyses and the legal department. | unlikely | minor |

| | | | | | |
|---|---|---|---|---|---|
| | | request (e.g. commercial application) – risk is not specific to PHT | | | |
| 10 | all | Interception when data is transferred from one location/system to the other. (e.g. man in the middle attack) | The data is stored on a secured Azure server, making use of all modern web security standards including safe file transport between the IKNL and Azure servers. The communication between station and central server is end-to-end encrypted to further ensure data protection. | unlikely | moderate |
| 11 | n.a. | Too much data in dataset (e.g. dob delivered rather than age) | Data minimization is a standard check in the processes of NKR-analyse and the Commissie van toezicht NKR. | unlikely | minor |
| 12 | n.a. | Lack of governance structure | Current measures: file for separate data requests at participating data providers and make all software open source to provide full transparency.<br>Future measure: identify (semi-)trusted third party to play the role as Central Server manager. and define contract between data providers and Central Server Manager.<br>A workflow should be defined and coordinated to execute studies with multiple data providers (stations) in order to adhere to the applicable data protection, ethics and privacy measures. | unlikely | minor |
| 13 | n.a. | Patient data on cloud is not according to IKNL policy. | IKNL has the policy not to store patient data on any cloud servers such as Azure. Although Microsoft and other providers will have state-of-the-art data protection software and measures in place, IKNL keeps data “in house”.<br>IKNL is redesigning its ICT infrastructure. This aspect will be dealt with in this process.<br>The Azure clouds used for the PHT are compliant to GDPR. | unlikely | minor |
| 14 | n.a. | One of the nodes is slow or gets disconnected – research cannot be performed. | Measures are not necessary, this will result in delay or postponing of the study. This is not different from the normal procedures when performing scientifically studies. | unlikely | moderate |

# E. References

1. Arturo Moncada-Torres, Frank Martin, Melle Sieswerda, Johan van Soest, Gijs Gelijnse. VANTAGE6: an open source priVAcy preserviNg federaTed leArninG infrastructurE for Secure Insight eXchange. AMIA Annual Symposium Proceedings, 2020 (in press).

2. IKNL. vantage6.ai – Privacy preserving federated learning [Internet]. 2019 [cited 2020 Jan 24]. Available from: https://www.vantage6.ai

3. Nina Bontje. Privacyaspecten van de Personal Health Train Aandachtspunten voor de verdere ontwikkeling. 2018.

4. Jones EM, Sheehan NA, Masca N, Wallace SE, Murtagh MJ, Burton PR. DataSHIELD – shared individual-level analysis without sharing the data: a biostatistical perspective. Norsk Epidemiologi [Internet]. 2012 Apr 13 [cited 2019 Jan 22];21(2). Available from: http://www.ntnu.no/ojs/index.php/norepid/article/view/1499

5. Jiang W, Li P, Wang S, Wu Y, Xue M, Ohno-Machado L, et al. WebGLORE: a Web service for Grid LOgistic REgression. Bioinformatics. 2013 Dec 15;29(24):3238–40.

6. Lu C-L, Wang S, Ji Z, Wu Y, Xiong L, Jiang X, et al. WebDISCO: A web service for distributed cox model learning without patient-level data sharing. Journal of the American Medical Informatics Association. 2015 Jul 9;ocv083.

7. Lee J, Sun J, Wang F, Wang S, Jun C-H, Jiang X. Privacy-Preserving Patient Similarity Learning in a Federated Environment: Development and Analysis. JMIR Medical Informatics. 2018 Apr 13;6(2):e20.

8. Chang K, Balachandar N, Lam C, Yi D, Brown J, Beers A, et al. Distributed deep learning networks among institutions for medical imaging. Journal of the American Medical Informatics Association. 2018 Aug 1;25(8):945–54.

9. Marte van Graafeiland, Nina Bontje. Toepassing van de Personal Health Train in de zorg Verdiepend onderzoek. Pels Rijcken; 2020.

10. Veeningen M, Chatterjea S, Horváth AZ, Spindler G, Boersma E, van der Spek P, et al. Enabling Analytics on Sensitive Medical Data with Secure Multi-Party Computation. Stud Health Technol Inform. 2018;247:76–80.