D8.03 "Proof-of-Principle" study: SOP HARMONY Anonymization Procedure

116026 – HARMONY

Healthcare Alliance for Resourceful Medicines Offensive against Neoplasms in Hematology

WP8 Legal, ethics and governance

Lead contributor	John Butler (49 – BAYER)
	Michel van Speybroeck (48 – JPNV-JANSSEN)
	Christiane Druml (09 – MUW Medizinische Universitaet Wien)
Other contributors	Ruben Villoria (7 – GMV Soluciones Globales Internet S.A.U.)
	Ana Heredia Casanoves (7 – GMV Soluciones Globales Internet S.A.U.)
	Klaus Wasserman (09 – MUW Medizinische Universitaet Wien)
	Gabriele Nagel (03 – UULM)
	Santiago Moralejo del Arco (01 - IBSAL)
	María Abáigar Alvarado (01 - IBSAL)
	Guillermo Sanz Santillana (02 - HULAFE

Due date	June 2018
Delivery date	November 2018
Deliverable type	OTHER
Dissemination level	PUBLIC

Description of Action	**Version**	**Date**
	V 1.6	31/10/2018

HARMONY

Table of Contents

Document History

Version	Date	Description
V1.0	26/04/2018	First draft
V1.1	02/07/2018	Review by Trusted Third Party
V1.2	17/07/2018	Review by WP3
V1.3	21/08/2018	Review IBSAL and HULAFE
V1.4	29/08/2018	Second draft
V1.5	06/09/2018	Third draft
V1.6	31/10/2018	Final version for SC review

Document references

Document
Osborne & Clarke Memo: Legal Assessment of the Anonymization Concept for the HARMONY Project
HARMONY De-Facto anonymization meeting minutes 23rd Feb 2018
HARMONY Platform Data Flow
D1.11 Data Quality Supervision Committee Rules and implementation (DQSC)
D3.09 Data Monitoring Plan

List of Acronyms

Acronym	Description
AE	Adverse Events
ALL	Acute Lymphoblastic Leukemia
AMDS	Associated Member Data Sharing agreement
AMEF	Associated Member Engagement Framework agreement
AML	Acute Myeloid Leukemia
APL	Acute Promyelocytic Leukemia
BCR	B-Cell Receptor
BMI	Body mass index
CLL	Chronic Lymphocytic Leukemia
CR1	First Complete Remission
DQSC	Data Quality Supervision Committee
DI	Direct identifiers
EFPIA	European Federation of Pharmaceutical Industries and Associations
EFS	Event Free Survival
ESAs	Erythropoiesis Stimulating Agents
FC	Flow Cytometrics
GDPR	General Data Protection Regulation
HDFS	Hadoop Distributed File System
HCT	Hematopoietic Cell Transplantation
HMs	Hematologic Malignancies
HMAs	Hypomethylating Agents
HB	Honest broker
IGVH	Immunoglobulin Variable Region Heavy Chain
ISMS	Information Security Management System
IP	Internet protocol
KDC	Key Distribution Centre
MBL	Monoclonal B Cell Lymphocytosis
MDS	Myelodysplastic Syndromes
MM	Multiple Myeloma
MRD	Minimal Residual Disease
NA	Not available
NDMM	Newly Diagnosed Multiple Myeloma
NHL	Non-Hodgkin Lymphoma
NDA	Non-Disclosure Agreement

OS	Overall Survival
PET-CT	Positron Emission Tomography - Computed Tomography
QI	Quasi-identifiers
RBC transfusion	Red Blood Cells
SCT	Stem Cell Transplantation
SFTP	SSH File Transfer Protocol
SMB3	Server Message Block 3
SOP	Standard Operational Procedure
SSH	Secure SHell
TLS/SSL security	Transport Layer Security /Secure Socket Layer
TTNT	Time to next treatment
TTP	Trusted Third Party
URLs	Universal Resource Locators
VPN	Virtual Private Network

D8.03 – "PROOF-OF-PRINCIPLE" STUDY: SOP HARMONY ANONYMIZATION PROCEDURE

1. SUMMARY:

HARMONY seeks reaching high utility from the data while guaranteeing data privacy and minimizing the risk of re-identification. The present document describes the technical procedures on the 'de-facto anonymization' process, which safeguards the security and confidentiality of patients' health information within the current legal framework and meets the needs of the HARMONY platform: not to render the data useless for research.

The purpose of this document is to guide data providers through '**in-origin' anonymization procedures**, i.e. technical measures such as removing direct identifiers at source and applying anonymization techniques to quasi-identifiers by which the original data will be changed only to the extent to which these data are still useful. The document describes the mechanism to transfer the data to a Trusted Third Party and the removal of data source identifiers. At the end of the process, the number of changes performed will be assessed to find out whether data have been sufficiently anonymized (**case-by-case de-facto anonymization and re-identification risk assessment**) and potential additional de-facto anonymization implemented in case the risk score from the previous step is exceeded. As these technical procedures are not sufficient to guarantee the due protection of patient's privacy on their own, data-flow rules and processes, in addition to several **organisational measures** (security, access permits, hosting agreement), will be implemented at a later stage to complete and ultimate a more than sufficiently safe de-facto anonymization procedure. Appropriately anonymized data is no longer linkable to an identifiable individual, which means it is no longer personal data and it does not infringe the privacy of the individuals the data concerns. Once it is no longer personal data, the GRDP does not apply[1].

The document includes:

1) an overview of the legal and ethical framework, including the list of relevant questions to be answered by the HARMONY platform for each individual hematologic malignancy (pilot and subsequent research projects: AML, ALL, MDS, CLL, NHL, MM, and childhood HMs),
2) a list of identifiers and quasi-identifiers that should be removed or treated, and how to do so through anonymization methods, including a check-list of questions for the data provider to determine whether their data is properly anonymized (upfront and residual risk analysis),
3) one last section that summarizes the additional organizational measures implemented as part of HARMONY'S data management and governance framework.

[1] "*The principles of data protection should therefore not apply to anonymous information, namely information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is no longer identifiable. This Regulation does not therefore concern the processing of such anonymous information, including for statistical or research purposes.*" Recital 26 of the REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (GDPR).

This practical document will be revised and refined with time, with the answers to the questions received, until it becomes a best practice **Guidance on anonymization and pseudonymization** (D8.05).

At the time of writing (May 2018) the EU General Data Protection Regulation (GDPR) has just come into effect (25. May 2018), replacing the DPD as HARMONY's data protection framework. With the GDPR a novel framework for processing personal data for "research in the public interest" and for "scientific research purposes" was implemented within the European Union: processing of personal data for these purposes will generally be legally allowed, provided "appropriate safeguards" are implemented, such as, e.g., pseudonymisation of personal data sets. In this regard, the right to the protection of personal data must be considered in relation to its function in society and be balanced against other fundamental rights, in accordance with the principle of proportionality[2].

Although the GDPR in general provides a comparatively research-friendly legal framework, a set of opening clauses still provides member states with flexibility to implement the GDPR in national law. These implementation processes are currently ongoing within the consortium member states, thus their implications on HARMONY's work cannot be foreseen in detail at this stage. An overview of how EU member states implement the GDPR's opening clauses into national law will be given at a later stage in HARMONY's deliverable document D8.06 "**Legal consideration of the use of medical data upon application of the General Data Protection Regulation**" which is due in M48 (December 2020).

[2] Recital 4 GDPR.

2. LEGAL AND ETHICS FRAMEWORKS:

Rapid technological developments and globalization have allowed for the use of health data on an unprecedent scale. The collection and sharing of data from increasingly numerous, available, and diverse sources (such as electronic health records, omics studies, etc.) has significantly raised in the past few years.[3] As a result, the embedding of heterogeneous data into a single unified data-pool has become a hot topic in research nowadays, and there is a growing international trend to support data-sharing initiatives[4].

In this context, HARMONY is an EU-EFPIA joint multidisciplinary Consortium aiming to **define sets of clinical outcome indicators and patient-related factors for health stakeholders**. To achieve it, HARMONY has built a **high-quality Big Data-sharing platform** that will be populated by the collection and harmonisation of data from previous clinical trials and patient data repositories. The purpose is to maximize the amount of data available to answer relevant specific questions (**HARMONY 'bench-to-bedside' research projects**), listed in section 2.1.

However, regardless of how promising Big Data Analysis may be for improving healthcare delivery and making best use of clinical studies, it also bears within itself a series of legal and ethical challenges relating confidentiality and data privacy.

HARMONY began down the path of processing clinical research data for secondary use within the previous legal framework of the EU Data Protection Directive (**Directive**95/46/EC, DPD in force up to 24 May 2018), where it was necessary either to have an explicit Informed Consent for secondary data use, or to anonymise the data.

The reach of the Informed Consent for the AML SG data in the Proof-of-Principle study was limited to the original goals of the studies for which they were required; therefore, when re-purposed beyond the scope of the original consent, the use of data required either re-consent from the patients or anonymization prior to re-use. The consent obtained at the point of data collection should not be regarded as 'once-and-for-all' and renewed consent is necessary for secondary data processing differing from the objective for which data were originally collected. This is an essential principle to guarantee confidentiality and data privacy.

However, the REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (GDPR) has come into effect during the Proof-of-Principle study's implementation process. This new European legal framework will likely facilitate legal compliance of HARMONY's data processing. The use of pseudonymised personal health data without informed consent for research purposes is permissible in the GDPR if subject to suitable and specific measures so as to protect the rights and freedoms of natural persons[5], compatible with the purposes for which the data were initially collected, the potential results can be considered being of overriding public interest[6], and/or

[3] Recital 5 GDPR.
[4] Recital 157 GDPR.
[5] Recital 54 GDPR.
[6] Recital 50 GDPR.

if effort to obtain new consent is disproportionate.

The HARMONY Ethics Advisory Board already denoted HARMONY to be of undeniable public interest, in the area of health given the fact that the project deals with very serious unmet needs in the current treatment of the haematological malignancies. Furthermore, it unites the ethical principle to give priority to the interest and welfare of the patient and the ethical duty to evaluate the effectiveness of medical treatment and protocols; to progress on improving them by helping determine more clinically relevant diagnoses and cost-effective treatments that will improve the quality of life of patients; to develop innovative pharmaceuticals; and to progress toward the so-called precision, or personalized, medicine by using the advanced analytic techniques currently offered by Big Data Analysis (mainly applied to existing non-related medical records).

HARMONY has developed a **'de-facto' Anonymization concept** within the purview of the GDPR which establishes an appropriately safeguarded pseudonymisation paradigm for processing personal data "for scientific research purposes" and/or "in the public interest". Anonymization in a legal sense does not require the data to be "fully anonymized"; i.e. redacted in a way that it is generally impossible – independent of technical and legal means as well as additional knowledge – to identify the affected individual. Rather, a de-facto anonymization is sufficient in order to exclude the qualification of the affected dataset as "personal data[7]"; i.e. sufficient anonymity is safeguarded in case identification would require an unreasonable effort.

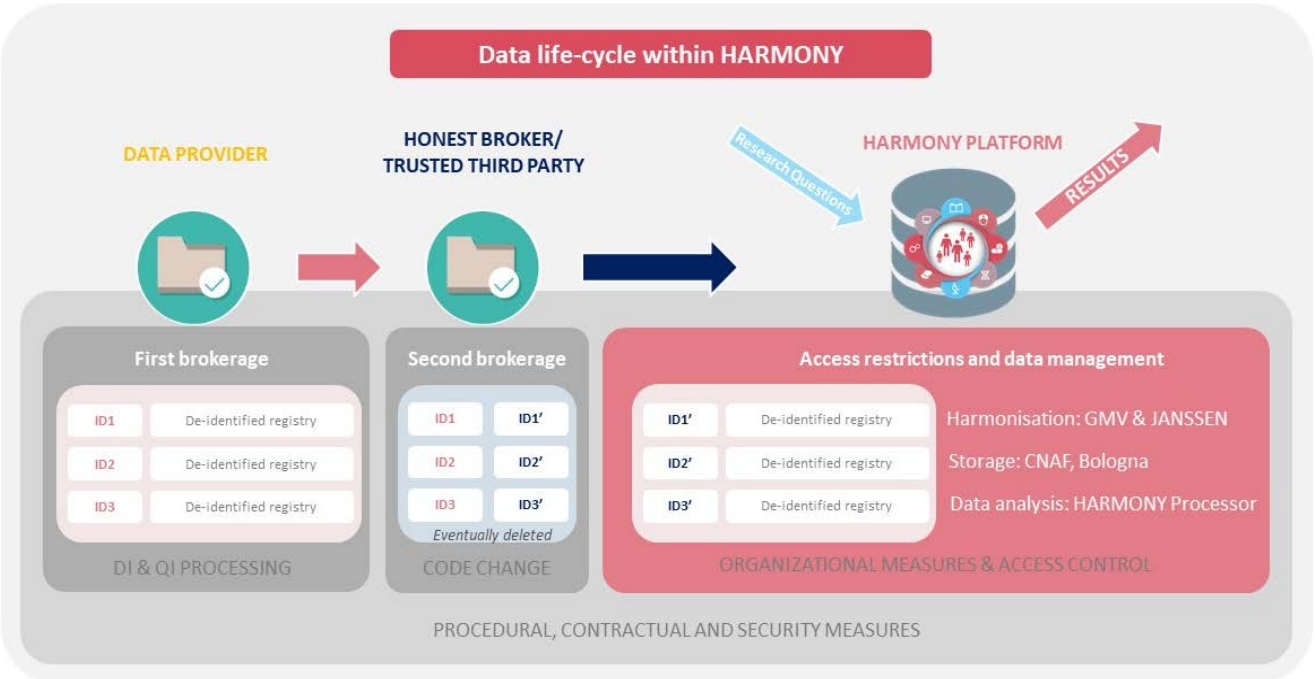

Figure 1. HARMONY's two-step pseudonymisation (coding) procedures

[7] The GDPR defines the term 'personal data' in Art.4, Definitions, 1. as "*any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person*".

The 'de-facto' Anonymization procedure consists of a safeguarded two-step pseudonymisation procedure, complemented by a third "hash" coding step, along with additional organizational, contractual, and security measures. It provides sufficient de-identification level by technical and organisational means as well as the ability to identify data records stemming from the same patients in order to be able to track disease, update and extend datasets, or to grant data subjects' rights to amendment, inter-operability or deletion of their data.

The HARMONY Anonymization Concept ensures that the intended import of data into the HARMONY Platform and their subsequent uses as envisaged within the HARMONY Project complies with applicable data protection laws on EU level including the General Data Protection Regulation (GDPR), and no means required by applicable data protection law is ignored. Under these laws, the HARMONY Anonymization Concept safeguards that the relevant data sets qualify as anonymous and not personal data and to ensure that there are no means likely reasonable to be used for identifying the data subjects to which the datasets in the HARMONY Platform relate.

Intentional re-identification is not only forbidden, but a case-by-case assessment on whether in an individual case the datasets are in fact rendered anonymous is also required. This assessment includes (1) an assessment of available technical means to reverse the anonymization technique and the foreseeable future developments in this field as well as (2) the risk of technical failures.

In addition, the HARMONY Anonymization Concept compensate possible shortcomings in a pure technical anonymization where the research purpose pursued by HARMONY in the individual scenario would be jeopardized by a further technical anonymization by way of (additional) organizational anonymization measures, as contemplated in the GDPR[8].

2.1. Examples of relevant specific questions (HARMONY 'bench-to-bedside' projects)

The following subsections summarize the first set of HM relevant questions which will form the basis for ongoing discussions on outcome definitions.

2.1.1. Acute Myeloid Leukemia (AML)

- Define patients who are suitable for intensive therapy and who benefit from SCT in CR1;
- Determine treatment impact on outcome in elderly APL patients;
- Investigation of MRD (based on molecular analyses and FC) to inform treatment in various genetically-defined AML subgroups;
- Delineation of differences in the genomic landscape between elderly and young adult well-annotated AML patients and potential impact on differing outcomes;
- Identification of a priori predictive sensitivity markers for novel therapies;
- Improved molecular characterisation of the MDS/AML overlap subtype and definition of prognostic markers and potential novel therapeutic strategies;
- Discuss the role of efficacy endpoints other than OS in patients suitable for intensive treatment and SCT;

[8] Recital 156 GDPR.

- Determine the impact of the start of treatment in very-low to intermediate risk MDS on AML progression;
- Determine the impact of immune checkpoint inhibitors on outcome in AML patients;
- Delineate the differences, if any, between responders to allogeneic transplantation and immune checkpoint inhibitors;
- Investigate the impact of erythrocyte and thrombocyte transfusions on survival endpoints in MDS/AML.

2.1.2. Acute Lymphoblastic Leukemia (ALL)

- Correlation of novel molecular risk factors with long term outcomes;
- Outcomes of patients with several types of Ph-like ALL with tyrosine kinase inhibitors;
- Impact of current therapeutic strategies on outcomes of elderly patients with ALL;
- Prognostic impact of MRD in distinct ALL molecular subtypes;
- Impact of age and ethnicity on genomic alterations in ALL;
- Evaluation of the impact of novel agents (e.g. blinatumumab, inotuzumab) regarding long term outcomes in ALL and their place in the ALL treatment algorithm;
- Identification of clinical and molecular markers of treatment refractoriness;
- Definition of MRD cut-off levels and timepoints for risk group assignment;
- Definition of clinical, age, and molecular groups that benefit the most of allogeneic SCT after CR1.

2.1.3. Myelodysplastic Syndromes (MDS)

- Identify new consensus outcomes and potential surrogates for overall survival (=> updating of International IWG 2006 criteria) ;
- Evaluate the role of newer drugs;
- Find the optimal drug/s to add to hypomethylating agents (HMAs) to improve survival advantage in higher-risk MDS;
- Find new treatments to avoid RBC transfusion dependence in lower-risk MDS with anemia (especially relevant for erythropoiesis stimulating agents [ESAs]-resistant patients) ;
- Reduce relapse risk after allogeneic hematopoietic cell transplantation (HCT), especially in very high-risk patients;
- Assess the independent prognostic role of molecular data (especially somatic mutational analysis) for refining prognosis of MDS;
- Identify those lower-risk MDS patients who may benefit from intensive therapies, including allogeneic HCT and set the best timing for starting treatment;
- Explore the potential value of the variant allele frequency of specific somatic mutations as measurement of minimal residual disease.

2.1.4. Chronic Lymphocytic Leukemia (CLL)

- Definition of a prognostic model that will clearly distinguish MBL/CLL cases that are more likely to progress to a life-threating disease requiring treatment;
- Shift toward front-line noncytotoxic regimens;
- Definition of MRD negativity as a therapeutic goal in younger and fit patients;
- Optimization of a risk-adapted therapy based upon biomarkers: cytogenetics, IGVH status etc.;
- Explore combinations of targeted agents aimed at deep remissions-> prolonged TTNT -> cure?
- Definition of a predictive model to better select treatment for patients with progressive disease, keeping in account both the host-derived and tumour-derived profiles particularly in the setting of the newer targeted and non-genotoxic treatments;
- Optimization of the sequence of treatments to improve efficacy and long-term control of the disease.

2.1.5. Non-Hodgkin Lymphoma (NHL)

- Delineate, at diagnosis, which patients are the most difficult to treat / cure by combining clinical and molecular data (from tumour and/or blood) and bring the evaluation of these parameters into routine practice;
- Define, within the different subtypes, patients that will benefit (or escape) from new targeted therapies (BCR inhibitors, epigenetic modifiers, BCL2 inhibitors, etc.) and reliably identify them;
- Further define anti-CD20 resistance, assess its potential reversibility and find ways to potentiate anti-CD20 or overcome this resistance (a critical question given the pivotal role of anti-CD20 in B-cell NHL);
- Assess the current and emerging tools (translocations, IG sequences, clonotypes, mutated circulating DNA…) to monitor MRD in NHL, to develop MRD monitoring as a clinically relevant endpoint for clinical trials in NHL;
- Enhance the standardization of PET-CT in the evaluation of (early and final) response to treatment (compare available scales, delta SUV, new approaches such as metabolic tumour volume, etc.)

2.1.6. Multiple Myeloma (MM)

- Define (long term) outcomes in pre-defined subsets of (NDMM?) patients;
- Limit to outcomes currently recorded in available data sets (e.g. EFS, OS, response rate);
- NDMM patient sub-groups of interest:
    - Transplant eligible vs ineligible
    - Patients with poor prognostic markers [FISH]
    - Age-groups (including elderly patients
    - By comorbidity
    - Subgroups of patients with early deaths (Primary resistant vs early progression vs toxic deaths)

- Rare forms, e.g. PCL;

— Describe patient outcomes;
— Define relevant cut-off values for chromosomal abnormalities/aberrations;
— Define prognostic factors for long term survival in patients with unfavourable FISH [e.g. del(17p)];
— Identify Factors for Refractoriness to treatment (e.g. IMiD, PI, High-dose, new agents);
— Characterise utility of MRD as a surrogate endpoint for OS/EFS;
— Validate R-ISS and/or ISS in EU data.

2.1.7. Pediatric HM

— Identification of clinical predictors and biological determinants of primary refractory disease (<5% patients);
— Identification of robust biomarkers of very low-risk disease, which can be treated with reduced toxicity protocols;
— Definition of the role of a variety of targeted therapies in improving the treatment and outcome of ALL;
— Evaluation of genetic biomarkers in predicting outcome after first relapse;
— Determination of the clinical and genetic risk factors for major toxicities (e.g. pancreatitis, osteonecrosis etc.) which are highly relevant in paediatric HM of children and adolescents (in particular with regard to long term sequelae).

3. FIRST BROKERAGE PSEUDONONYMIZATION BY DATA PROVIDERS

Data providers will apply two different masking technical measures that are in scope of the HARMONY Anonymization Concept in order to perform a first de-identification of the data:

- **Suppression** (i.e. elimination of data / datasets);
- **Generalization** (i.e. recoding of data into intervals, rounding, aggregation);

Each of these techniques will be applied depending on the type of variable. In HARMONY, we make a distinction between two types of variables:

- **Direct identifiers (DI)**
- **Quasi-identifiers (QI)**

In the sections 4.1. and 4.2., we will address which anonymization technique will be performed on each specific variable, as well as the specificities of each technique. Section 4.3 is designed to guide data providers in preparing ad double-checking the data before sharing it with HARMONY.

3.1. Direct Identifiers

Direct identifiers (DI) are fields that can uniquely identify individuals, such as names, Social Security numbers, email addresses, etc. DI are seldom used in any data and statistical analysis that are run on the healthcare data.

In HARMONY, DI must be supressed from the dataset by the Data Provider. Suppression means removing any value totally from an information table and replacing the attribute values with some anonymous value ("**"). Only the most important values to identify a data subject need to be supressed and replaced using the "**" value. Otherwise, the quality of the data could be drastically reduced.

Below, we can find the list of direct identifiers that must be suppressed:

- STUDY ID;
- SUBJECT ID;
- UNIQUE SUBJECT ID;
- SITE ID;
- STUDIES;
- SITE SUBJ ID;
- Staff IDs;
- Names;
- Initials;
- Telephone numbers;
- Fax numbers;
- Email addresses;
- Social Security numbers;
- Medical record numbers;
- Health plan/card numbers;
- Account numbers;
- Certificate/license numbers;

- Vehicle identifiers and serial numbers, including license plate numbers;
- Device identifiers and serial numbers;
- Web universal resource locators (URLs);
- Internet protocol (IP) address numbers;
- Biometric identifiers, including fingerprints and voiceprints;
- Full face photographic images;
- Any other unique identifying number, characteristic, or code.

With respect to **Clinical Trials**, clinical trial participant numbers should also be removed and replaced with a second set of identification numbers. As a best practice, any key linking the two sets of numbers should be destroyed[9]. Clinical Trial Investigator Information, including site name, investigator identification, and investigator affiliation should be removed or replaced with a random number. Investigator site information may also be aggregated to a national or regional level. Where appropriate, a list of sites or investigators who participates in a study can be provided, so long as individual data subjects are not linked to a particular site or investigator.

3.2. Quasi Identifiers

Concealing the name, phone number or other explicit identifiers does not ensure the security of sensitive data of an individual: Quasi Identifiers (QI) are fields that, in combination, can identify individuals. Examples of these include dates, geographic or regional location, demographic data (such as race and ethnicity), socio-economic data, anthropometric data, sensitive information, medical data, adverse events, and disease characteristics. However, unlike DI, QI are useful for data analysis. Given the fact that supressing these data would entail the loss of scientific knowledge in the results of the research studies conducted in HARMONY, QI will be treated by masking methods such as suppression and generalization (explained in sections 4.2.1. and 4.2.2.).

Following (section 4.2.3. and its subsections) is a detailed list of HARMONY QI, where the methods that Data Providers need to apply in each case are indicated.

[9] The Article 29 Working Party in Opinion 4/2007 on the concept of personal data, available at http://ec.europa.eu/justice/policies/privacy/docs/wpdocs/2007/wp136_en.pdf, concluded that controllers who possess key-coded data, but are operating within a "specific scheme" in which "re-identification is explicitly excluded and appropriate technical measures have been taken in this respect," are not engaged in "processing personal data." The Article 29 Working Party continued on to note that:
"*In other areas of research or of the same project, re-identification of the data subject may have been excluded in the design of protocols and procedure, for instance because there is no therapeutical aspects involved. For technical or other reasons, there may still be a way to find out to what persons correspond what clinical data, but the identification is not supposed or expected to take place under any circumstance, and appropriate technical measures (e.g. cryptographic, irreversible hashing) have been put in place to prevent that from happening. In this case, even if identification of certain data subjects may take place despite all those protocols and measures (due to unforeseeable circumstances such as accidental matching of qualities of the data subject that reveal his/her identity), the information processed by the original controller may not be considered to relate to identified or identifiable individuals taking account of all the means likely reasonably to be used by the controller or by any other person. Its processing may thus not be subject to the provisions of the Directive.*"

3.2.1. Generalization

Generalization consists on substituting the values of a given attribute with semantically unvarying but less particular values, or on diluting the attributes of data subjects by modifying the respective scales or order of magnitude. To this purpose, the notion of a domain (i.e., the set of values that an attribute can assume) is replaced with a set of generalized domains to hide the details of attributes, making the Quasi-Identifiers less identifying. In other words, individual records no longer exist and cannot be distinguished from other records in the same grouping.

If the value is a categorical value, it may be changed to another categorical value denoting a broader concept of the original categorical value. For example, "male" and "female" can be generalized to "person". If the value is numeric, it may be changed to a range of values. For example, the granularity of individual rates of birth can be lowered by generalizing them into a range of dates or grouped by month or year. Other numerical attributes (e.g. age, salaries, weight, height, or the dose of a medicine) can be generalized by interval values. These methods may be used when the correlation of punctual values of attributes may create quasi-identifiers.

Generalization can be applied on the attribute level (column) and in cell level.

The following table contains a non-anonymized database of patient records of a fictitious hospital in France:

Name	Age	Gender	Zip Code	Marital Status	Health Problem
Camile	29	Female	75000	Married	Cancer
Léa	24	Female	75020	Married	Viral infection
Manon	28	Female	75000	Widow	TB
Thomas	27	Male	75012	Divorced	No illness
Chloé	24	Female	75020	Single	Heart-related
Nicolas	23	Male	75012	Divorced	TB
Julien	19	Male	75020	Married	Cancer
Quentin	29	Male	75012	Married	Heart-related
Maxime	17	Male	75020	Single	Heart-related
Alexandre	19	Male	75020	Single	Viral infection

In the table below, we have replaced all the values in the 'Name' attribute and all the values in the 'Marital status' with '**' (suppression), and the values in the 'Age' attribute with a range.

Name	Age	Gender	Zip Code	Marital Status	Health Problem
**	20-30	Female	75000	**	Cancer

**	20-30	Female	75020	**	Viral infection
**	20-30	Female	75000	**	TB
**	20-30	Male	75012	**	No illness
**	20-30	Female	75020	**	Heart-related
**	20-30	Male	75012	**	TB
**	10-20	Male	75020	**	Cancer
**	20-30	Male	75012	**	Heart-related
**	10-20	Male	75020	**	Heart-related
**	10-20	Male	75020	**	Viral infection

This data has 2-anonymity with respect to the attributes 'Age', 'Gender' and 'Zip code' since for any combination of this attributes found in any row of the table there are always at least 2 rows with those exact attributes.

3.2.2. Quasi-Identifiers in HARMONY and applicable Anonymization Procedures

3.2.2.1. Dates

Only months and years may be indicated for dates related to identifying events in a data subject's life (such as birth dates, dates of death, hospital admission dates, discharge dates, health-care practitioner visit dates, and specimen collection dates). Where appropriate, a birth date may be replaced with the data subject's age at the time the information was gathered.

Age at diagnosis is an essential information that must be retained.

Should it be necessary to preserve information concerning the **temporal relationship of events** within a certain period, dates may be changed to relative time (time from date to a randomly generated reference time) Do not include random reference date in de-identified data:

a. dates may be expressed as the number of days that passed from some other event (e.g., in a clinical study, this could be the number of days since the data subject's enrolment in the clinical study).

3.2.2.2. Geographic/regional location

- Country;
- State or region;
- City;
- Street address;
- Zip code or Postcode.

Country, State/region should be maintained if no information about zip/postal codes is available.

City and **Street address** are not required and should also be deleted.

Postal codes should only be retained by using the initial three digits of the postal code only. In those case in which the zip code represents areas with populations below 20,000 persons, it should be deleted or aggregated to a larger geographic area.

3.2.2.3. Demographic data

- Sex, Gender;
- Race, ethnicity;
- Nationality;
- Household, family composition;
- Marital status;
- Pregnancy information;
- Age;
- Age at ---.

It is essential to maintain the information for **sex/ gender** and **nationality**.

If the population associated with the dataset is such that including **ethnicity** would create a risk of re-identification, ethnicity should be removed or replaced with more generalized categories, such as Clinical Data Interchange Standards Consortium's (CDISC) standard ethnicities.

Information regarding the exact **number of pregnancies** may be essential in some instances. When available, it should be maintained with exact numbers (0, 1, 2, 3), and replaced with a range for more than 3.

Information regarding **age** may be replaced with ranges where appropriate. However, all persons over the age of 89 must be grouped into a single category.

3.2.2.4. Socioeconomic data

- Name of Employer;
- Job title;
- Profession or occupation;
- Income;
- Education;
- Place of work;
- Qualifications;
- Languages spoken;

Specific **socio-economic information** may be deleted, or replaced with broad categories (for example, "post-secondary education" rather than the name of a specific education institution; "income" or specific

salaries can be replaced by salary ranges; a rare medical profession such as perinatologist can be aggregated to a more general obstetrician).

Information on the coding of the socio-economic variables must be provided in the data dictionary to ensure adequate coding (i.e. educational levels may differ across countries).

3.2.2.5. Anthropometric data

- Height;
- Weight;
- Body mass index (BMI).

Anthropometric values should be included except for extreme outlying values, which would allow singling out an individual per se. Outlier values should be deleted if they allow identification of an individual.

3.2.2.6. Sensitive information

- Drinking habit;
- Drug use;
- Finding/interventions;
- History of previous diseases;
- Laboratory value;
- Smoking habit.

Data quality may differ considerable depending on the data collection. Information on the coding must be provided in the data dictionary to ensure adequate coding. Sensitive information may be required for multiple studies and may be retained. Laboratory values are critical in hematology. History of previous diseases will help although will be rather unbalanced between registries.

3.2.2.7. Medical data

- Diagnosis;
- Medications;
- Rare health conditions;

Diagnosis, medications, and rare health conditions are core fields for HARMONY and should be included.

If the data subject's **medical history** contains information about the data subject, or the data subject's family, which could permit re-identification, then medical history should be removed or replaced with generic language (e.g., "family history of heart disease").

If the quantity of **genetic information** contained in the record could be used to match a subsequent genetic sample from the same individual to the data profile, then genetic information should be removed. Even a small set of genetic information, when combined with other factors, may be sufficient to identify a data subject. **Please note that HARMONY is using genetic data on somatic mutations of neoplastic cells that cannot identify a patient and that change over time**.

3.2.2.8. Adverse events (AE) – only for serious or severe AE

- Standard description (may not be knowable);
- Flag for AE resulting in death;
- Flag for AE being life threatening;
- Flag for AE resulting in congenital abnormalities;
- Flag for AE resulting in permanent/serious disability or incapacity;
- Flag for AE resulting in/prolonging hospitalization;

Adverse event descriptions or codes should be maintained and presented in a generalized manner that do not permit re-identification of the data subject.

3.2.2.9. Disease characteristics

- Duration of symptoms;
- Hours missed of work;
- Emergency room visits;
- Hospital stay duration;

3.2.3. Other

- Free text and verbatim statements should be removed, if they contain information which could be used to re-identify the data subject;
- Variables with only NA should be removed;
- Extreme/outlier values should be removed (e.g. ages >90, number of children, very long hospital duration).

Whether the degree of applied anonymization methods is sufficient or not depend on the overall risk profile for re-identification. The risk profile can be assessed using criteria contained in the de-facto anonymization approach. The grade of anonymity can be assessed using statistical methods on quasi-identifiers (given that all direct identifiers have been eliminated).

3.3. Upfront anonymization checklist assessment

This assessment is only to be performed for data sets that have not been already subjected to anonymization. In case your dataset has not been anonymized yet or it's unclear, and as you prepare your

data for sharing with HARMONY, use this checklist to make sure you are going on track, please check the following items:

Please confirm whether the following direct identifiers have been removed.

Aspect	Item	Yes	No	Comment
Names	Full and partial names, including initials.	☐	☐	
Geographic Subdivisions	Street addresses	☐	☐	
	City	☐	☐	
	County	☐	☐	
	State	☐	☐	
	Legislative district	☐	☐	
Contact Information	Telephone numbers	☐	☐	
	Fax numbers	☐	☐	
	Email addresses	☐	☐	
	Websites and URLs	☐	☐	
	Screen names	☐	☐	
	IP address numbers	☐	☐	
Identifying Numbers	Social Security numbers	☐	☐	
	National identifying numbers	☐	☐	
	Account numbers	☐	☐	
	Medical record numbers	☐	☐	
	Health insurance numbers	☐	☐	
	Certificate numbers	☐	☐	
	License numbers	☐	☐	
	Vehicle identification numbers	☐	☐	
	License plate numbers	☐	☐	

	Device serial codes	☐	☐	
	Internet Protocol (IP) addresses	☐	☐	
	Any other numbers capable of identifying a single person or a small number of persons	☐	☐	
Biometric Identifiers	Finger prints	☐	☐	
	Voice recordings	☐	☐	
	Pictures of identifying marks	☐	☐	
	Full-face images	☐	☐	
	Any other picture that depicts a sufficient area of the data-subject in sufficient detail to permit re-identification	☐	☐	
Verbatim quotes	Verbatim statements	☐	☐	

Please specify what has been done to the following data fields, whether they have been removed or replaced. In the second case, please indicate the technical measures implemented (as per the guidelines in sections 3.1 and 3.2 of this document). Use the "comments" cell to explain the technical measures implemented.

Aspect	Removed	Replaced	Comments
Clinical Trial information	☐	☐	
Day value in dates	☐	☐	
Ages	☐	☐	
Temporal relationship of events within a period	☐	☐	
Zip or Postal codes	☐	☐	
Demographic data	☐	☐	
Socioeconomic data	☐	☐	
Anthropometric data	☐	☐	
Sensitive Information	☐	☐	
Medical information	☐	☐	
Genetic information	☐	☐	

Adverse events	☐	☐	

If the data set contains the above data elements and you are unable to perform the redaction of the data set as suggested, a trusted third party should be engaged to perform these actions.

If the data set contains the above data elements and you are able to perform the redaction of the data set as suggested, a **residual risk analysis** should be performed – either in house or through a trusted third party (detailed instructions or risk analysis tool to be provided), in relation to three different attacker models:

- **The prosecutor scenario**: the attack aims to re-identify a specific person and relies upon pre-existing knowledge about a person known to exist in the de-identified database.
- **The Journalist scenario**: this attack also aims to re-identify an individual, but the attacker does not know for certain that the target is in the dataset. Instead, the attacker uses access to another source of public information about an individual or individuals that are also present in the de-identified dataset.
- **The Marketer scenario:** this attack involves re-identifying as many people as possible from the de-identified data even if this means some of them will be incorrectly identified.

4. SECOND BROKERAGE PSEUDONONYMIZATION BY HONEST BROKER

The present section aims to define the data flow starting at provider's servers, until it becomes part of HARMONY platform, **defined for the pilot projects and for those datasets where the data provider has confirmed that the data have been anonymized at origin.** The requirements, checks and validations that will be considered for their incorporation into HARMONY platform will also be introduced, establishing the full landscape on how information will be prepared, transferred, received, and stored. It will also lay the ground for the final process to be applied to data sources incoming in the future, eventually becoming best practices in data acquisition, transference, and inclusion.

4.1. Requirement

A series of technical and procedural requirements need to be fulfilled for an efficient data intake.

As a technical requirement, Data Providers need to provide a general description of the data, as well as a data dictionary. The information to be included in the general description and the data dictionary is detailed in the last part of the AMEF. Deliverable *D1.11 Data Quality Supervision Committee Rules and implementation (DQSC)* section *3.2.3 Evaluation Criteria* specifies which checks and validations will be applied to ensure these technical requirements are fulfilled.

As a procedural requirement, and related to the previous one, a communication channel needs to be stablished between the Data Provider and the Data Processor. To that end, a mailbox with address harmony-data@synapse-managers.com has been created, with the HB as the recipient.

4.2. Data structure transfer

Because the intake, quality check and harmonisation processes are configured in a source-by-source basis, a detailed description of the datasets is paramount in their definition.

Along with the Associated Member Engagement Framework (AMEF), the data structure needs to be provided specifying the field names, description, data types, possible values, units and catalogues, scales, guidelines, or criteria used. Also, a summary of the information contained in the dataset is required, including the data origin (e.g. patient registries, clinical trials etc.) and a brief description of the collection method.

All the information related to the dataset will reach the Data Processor via the HB, who will remove any reference to the Data Provider if present.

Upon receipt, the structure will be analysed so as to define how the data will be loaded, curated and harmonized. Should the designated person in WP3 need any assistance from the Data Provider in the definition of these processes, the same communication channel will be used.

4.3. Data delivery

Several parties are involved in the transference of the data from its original location to the platform through a data transfer area specifically set up for this purpose. These actors are in charge of preparing

the data, verifying it is 'de facto' anonymous and relocating it at specific points of the process. As a result, only high-quality, non-personal data enters the platform.

4.3.1. Data transfer infrastructure and data flow

For the safe transfer of the sources a data transfer area (Figure 2) has been created on Microsoft SharePoint online services. This area, managed by the honest broker, consists of three folders with different purposes and access permissions:

1) DATA PROVISION, gathering a folder created per data provider once the AMDS signed. A designated person per data provider will get access to its individual folder, while the honest broker will have access granted to all of them. Only individual credentials will be provided.
2) CODE CHANGE, solely accessible by the honest broker.
3) TRANSFER TO DB, accessible to both the honest broker and the data processor.

Access to the three folders is permitted only to authorized staff at the **HB**, via individual credentials, through a secure channel pre-determined by Microsoft SharePoint. Each authorized staff member is given for a limited time the minimum permissions necessary to perform their duties. The synchronization function in SharePoint has been removed to avoid potential data transfer to other IT devices. However, those who have edit permission in a specific folder can download documents from the system. To minimize risks only two staff members at the **HB** (the administrator of the system and a deputy) have such permission. They have signed Non-Disclosure clauses, and are subject to legal consequences in case of breach.

This folder structure will be setup for each data source.

Data in the transfer area flows in the order specified.

1) The data provider delivers the source in DATA PROVISION.
2) The HB transfers the data to the CODE CHANGE folder, replaces registry identifiers and leaves it in TRANSFER TO DB.
3) The data processor takes the source from TRANSFER TO DB and loads it onto the HARMONY platform.

After confirmation from the data processor that the data has successfully been loaded onto the HARMONY Platform, the source files in the DATA PROVISION and CODE CHANGE folders will be permanently deleted by the HB.

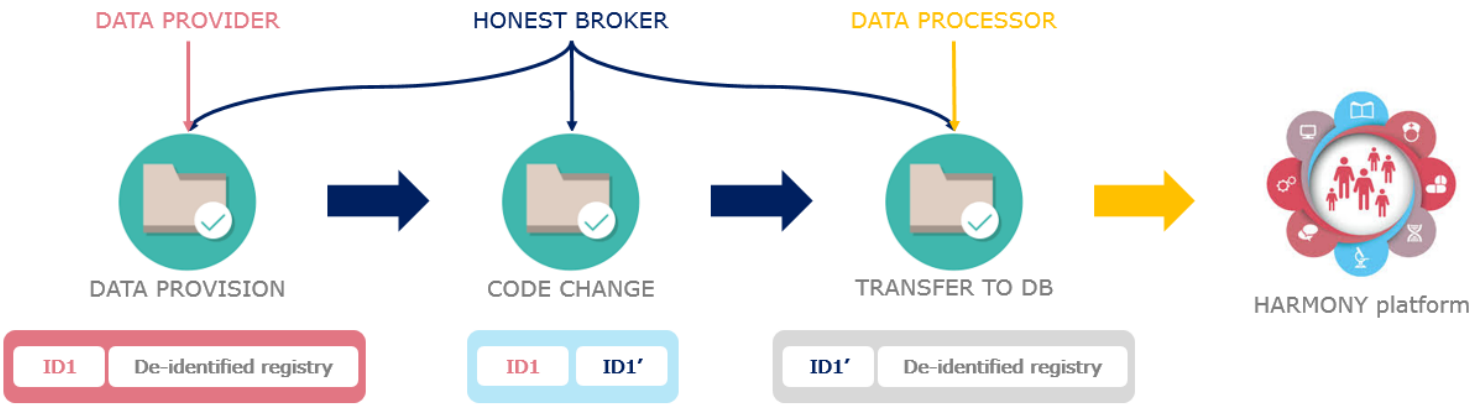

Figure 1. Data transfer SharePoint structure, permissions, and processes.

4.3.2. Responsibilities of the parties involved

Prior data delivery the **Data Provider** needs to review and sign the contracts, provide the data structure, and prepare the data according to the AMDS and the *SOP HARMONY Anonymization Procedure*. A checklist will be supplied for the provider to assess to which extent the data is anonymous.

Request access to the data transfer area by email harmony-data@synapse-managers.com has been created to this end. **HB** creates a user for each data provider and an individual folder with restricted access only for this user and manager from **HB**.

The **HB** is responsible of verifying the data is anonymised, replacing record IDs, and removing any information present in the data which relates it to its provider. Although the traceability back to the Data Provider will be kept, the **HB** will delete registry identifiers equivalences after the data brokerage.

4.4. Process validity

The European legal framework privacy and data protection has set the background for the design of the data sharing and intake processes and choosing the services on which to build the data transfer area.

In the HARMONY platform these regulatory requirements are met by implementing monitoring and securization measures as defined in the deliverable *D3.09 Data Monitoring Plan* and ISO 27001 certifying the infrastructure. Similarly, Microsoft SharePoint servers, located in Ireland and the Netherlands, have implemented security measures complying with the requirements set forth in ISO 27001, ISO 27002, ISO 27018 and the General Data Protection Regulation (GDPR) as described in their *Online Services Terms* as of July 1, 2018[10]. The fact that Microsoft will conduct security audits as well as promptly notify of any security incident, provide detailed information about it, and take reasonable steps to mitigate the effects and to minimize any damage reinforces the reliability of the services provided. Moreover, SharePoint includes an Audit Trail and Log functions that allows monitoring any access to the system.

[10] Online Services Terms available at http://www.microsoftvolumelicensing.com/DocumentSearch.aspx?Mode=3&DocumentTypeId=46

5. SECOND BROKERAGE PSEUDONONYMIZATION BY TRUSTED THIRD PARTY

The present section aims to define the data flow starting at provider's servers, until it becomes part of HARMONY platform, **defined for those datasets where the data needs further anonymization measures.**

5.1. Requirements

For the de facto anonymization with in the IMI Harmony Alliance the Institute of Epidemiology and Medical Biometry Ulm University (UULM) will serve as Trusted Third Party (TTP). The TTP will receive the dataset in disease specific formats and unique coding via a user specific secure interface. The TTP will perform quality checks according to the anonymization protocol. In case of major deviations from the protocol the dataset will be returned to the centre and updated version will be queried by the TTP. The results of the quality checks will be documented.

For datasets of sufficient quality, a Harmony ID will be generated.

The double pseudonymized data will be fetched from the TTPs SFTP server by the central HARMONY database located at Bologna. For this purpose, the Bologna centre uses its accounts (one per authorized person) managed by TTP. The Bologna centre will have access only to the harmonized data, not to the source data files.

5.2. Data structure transfer

As in section 6.2, quality check and harmonisation processes are configured in a source-by-source basis, a detailed description of the datasets is paramount in their definition.

Along with the Associated Member Engagement Framework (AMEF), the data structure needs to be provided specifying the field names, description, data types, possible values, units and catalogues, scales, guidelines, or criteria used. Also, a summary of the information contained in the dataset is required, including the data origin (e.g. patient registries, clinical trials etc.) and a brief description of the collection method.

All the information related to the dataset will reach the Data Processor via the HB, who will remove any reference to the Data Provider if present.

5.3. Data delivery

5.3.1. Data transfer infrastructure and data flow

At the TTP the datasets will be separated in match of IDS (Study ID and HARMONY ID) and medical data. After successful transmission to the central HARMONY database, the medical data will be deleted after successful transfer to the HARMONY database or at maximum after one month.

Interface 1: Transfer from study centres to the TTP

For the interface 1 a secure exchange with one account per user will be used. Data will be transferred

from Study Centres to an SFTP hosted by the TTP. Each study centre will have access only to its own uploaded data. File format for data exchange will be CSV (text based data separated by ";") encoded in UTF-8 with byte order mark.

Interface 2: Transfer to the HARMONY central database

File format for data exchange will be CSV (text based data separated by ";") encoded in UTF-8 with byte order mark.

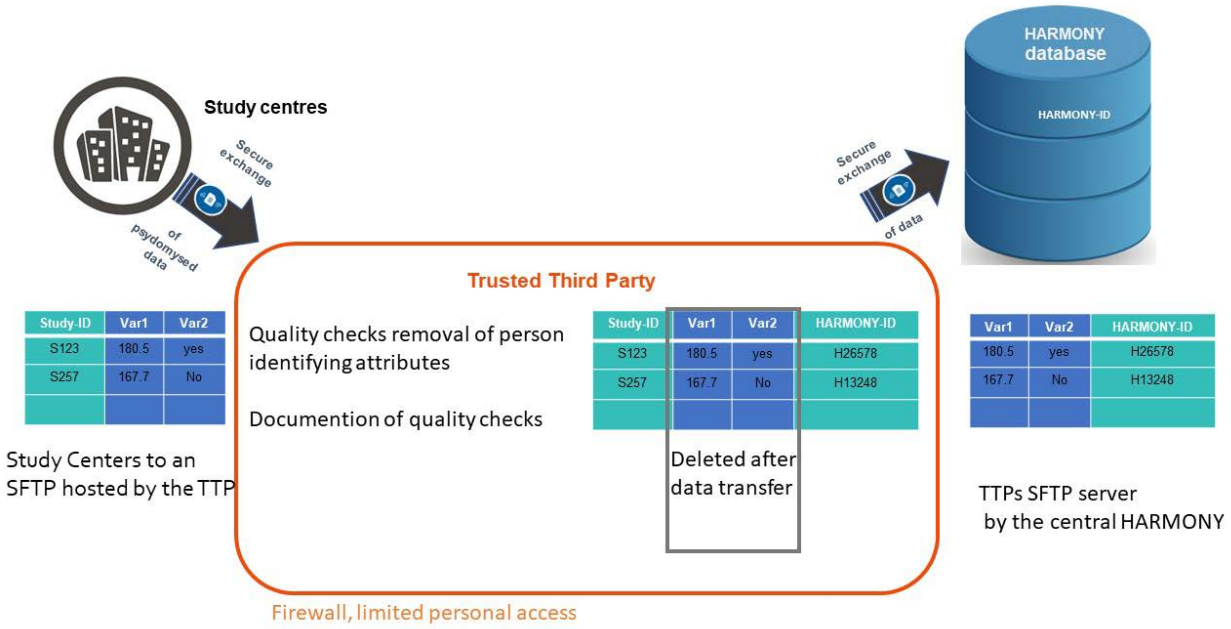

Figure 3. Trusted Third Party role, permissions, and processes

5.3.2. Responsibilities of the parties involved

Data access by cooperating centres to the SFTP will be logged. Data will be transferred to the institutes local file server for the actual data processing. The access to data on the file server is limited to the designated project members of the Institute of Epidemiology and Medical Biometry, and data will be accessed via SMB3. Encrypted data backups are performed daily.

All employees (as well as assistants) located at TTP are obliged according to the state data protection law § 6 LDSG which incorporates national and European regulations. All involved employees of TTP authenticate themselves via their personal user account using passwords. Software and hardware firewalls are used to protect against unauthorized external access.

Access to the server room at TTP, in which the electronic data are stored, is secured by a manual locking system with security locks. Various technical safety measures are available in the server rooms of the TTP: air conditioning, fire and smoke alarm systems, and nearby fire extinguisher.

6. ORGANIZATIONAL MEASURES FOR ANONYMIZATION:

The GDPR explicitly acknowledges that data anonymization can be achieved through a combination of technical and organizational measures[11]. In fact, both technical and organizational measures are ranked as equal possibilities of the controller to comply with its obligations under the GDPR and demonstrate compliance[12].

Equally, in its Opinion 4/2007 the WP29 consequently explicitly acknowledges that not only technical but also organizational measures can be a valid means to ensure anonymization[13].

In addition to the technical measures presented in Section 5, HARMONY provides organizational anonymization measures in order to (1) compensate for a possibly incomplete technical anonymization and (2) provide for additional safeguards to respond to organizational risks.

Further information about the organizational measures can be extracted from D3.09 and WP3 document on data intake.

6.1 Data Access restrictions

The HARMONY Platform provides for data access restrictions to safeguard that (1) the number of people with access to data is limited, (2) download capabilities for full data sets are inexistent and (3) data providers are established.

These data access restrictions are effective means to further ensure that the datasets in the HARMONY Platform cannot be linked back or matched with the original datasets remaining at the data provider. They ensure that no person who has access to the original dataset at the data provider (or to other datasets about the data subjects) is being granted access to the datasets in the HARMONY Platform or can download data from the HARMONY Platform and thereby minimizing possible residual risk of linkability to an extent that it is no longer reasonably likely that such linking / matching occurs.

These access restrictions include the following safeguards:

---

[11] Art. 89 para. 1 s.4 GDPR.

[12] Art. 24 para. 1 GDPR, and Recital 78 GDPR, which provides that "*the protection of the rights and freedoms of natural persons with regard to the processing of personal data require that appropriate technical and organisational measures be taken to ensure that the requirements of this Regulation are met*".

[13] The Article 29 Working Party in Opinion 4/2007 on the concept of personal data, WP 136, available at http://ec.europa.eu/justice/policies/privacy/docs/wpdocs/2007/wp136_en.pdf, states that in order to exclude identification over time "*the system should be able to […] incorporate then the appropriate technical and organizational measures*" (p.15). More explicitly, in the context of pharmaceutical research data, the WP29 acknowledges the validity and effectiveness of organizational purposes for anonymization purposes: "*Example No. 13: pharmaceutical research data. Hospitals or individual physicians transfer data from medical records of their patients to a company for the purposes of medical research. No names of the patients are used but only serial numbers attributed randomly to each clinical case, in order to ensure coherence and to avoid confusion with information on different patients. The names of patients stay exclusively in possession of the respective doctors bound by medical secrecy. The data do not contain any additional information which make identification of the patients possible by combining it. In addition, all other measures have been taken to prevent the data subjects from being identified or becoming identifiable, be it legal, technical or organizational. Under these circumstances, a Data Protection Authority may consider that no means are present in the processing performed by the pharmaceutical company, which make it likely reasonably to be used to identify the data subjects.*" (p. 15 et seq).

— Set-up **detailed and severe access management concept**. This includes:

a. Physical (access keys) and electronic (personal tag) controlled mechanisms to access the security perimeters where the HARMONY Platform is set. Staff must go through a registration process, with indication of identity (full name, employer organization), date and duration of access (start and end time), and motivation or reason of the access, in order for a specific access authorization to be granted. Record of access thought the personal electronic tags are registered and stored on a dedicated log server, administered by a staff member different from those authorized to the security perimeter and with redundant backup. Usage of the keys is also recorded.

b. Network communication protection. Machines inside the information security management system (ISMS) network have only the essential ports open, both for incoming and outgoing communication, and all the traffic is filtered by the switch firewall. The whole ISMS network is isolated as a VPN from the remaining hosting provider network aside from essential administrative services (performed only from secure terminals located inside the secure perimeter). All the communication inside, to and from the high security network is encrypted with cryptographic keys (TLS/SSL security), a firewall is implemented in the different cluster nodes along with a network authentication protocol that works based on 'tickets'. To prevent intrusion, all the systems are kept up to date with security updates, antivirus software is installed, and relevant information sources for security news have been identified and regularly checked. Access logs are stored in a dedicated log server. Staff with administrative rights on the servers do cannot have administrative right on these machines. These logs are subjects to periodic controls. Staff members receive specific training about security issues, in particular about information transfer security. The information related to changes and evolution is also collected through tasks, providing an excellent traceability mechanism.

c. Network access controls. Access to the network and the servers are permitted only to the authorized staff and registered users, who are given the absolutely minimum permissions necessary to perform their duties. Access to the system is only allowed on pre-determined ways and only through secure channels and configured with a system that guarantees that the real authorizations are consistent with the planned ones. User accounts have passwords policies and the employer organization must confirm the employment status every three months. All users must sign Non-Disclosure Agreement (NDA) and data security agreements. Additionally, all access to the high security servers are monitored and logged and the number of users is limited to who is absolutely necessary for data analysis. Access logs are stored in a dedicated log server. Staff with administrative rights on the servers do cannot have administrative right on these machines.

d. Host control access. Access to HARMONY servers is controlled and granted only to authorized entities through the need-to-know principle. Access to the program source code is also restricted and controlled. Specific procedures are present for password attribution, and user creation and management during the course of its activity. Additionally, users and entities which have accessed the network segments can always be identified. User access rights are reviewed periodically, both for normal and privileged users. This control is

repeated whenever users change their employment status, as there might be changes on their right of access to the data. Normal user access is provided once explicit authorization is granted, after they provide (i) signed copy of the access request form; (ii) copy of their personal ID; (iii) signed copy of a non-disclosure agreement; and (iv) signed copy of the "nomination as sensitive data manager" form as per law requisites. Privileged access is limited and only provided on basis of proved necessity and constantly monitored.

e. Security monitoring. Security Monitoring is provided by two Linux architecture services: SELinux and Audit. Security risk will be assessed through software means and news feed from authoritative sources. Any new vulnerability discovered will be registered into the management site, assigned to be managed by a person responsible and solved as soon as possible, and a proper risk assessment will be carried on according to the risk assessment procedures. A detailed registry of events and logs can be consulted, allowing to look up and search service logs, by keyword, host, service, or log level, without accessing the cluster machines via SSH.

f. Segregation of duties. Conflicting duties and areas of responsibility in HARMONY platform are segregated to reduce opportunities for unauthorized or unintentional modification or misuse of the information.

g. Dataset download capabilities from the HARMONY Platform are excluded.

6.2 Further Organizational Measures

6.2.1 Contractual measures

Contracts with the data providers, processor(s), and accessing entities are being conducted in order to ensure that compliance with all applicable laws, rules, regulation, ordinances and directives, including laws on the protection of personal data, and that none of them engage in re-identification activities.

In order to provide sufficient safeguards in this respect, the data providers must to apply certain anonymization techniques to ensure that the provided data have been stripped of direct identifiers and that quasi-identifiers have been adequately treated. They are obligated also not to "(i) engage in any activity to re-identify the *Contributed Data* by any means whatsoever including but not limited to singling out, linking back or matching any dataset from the HARMONY Platform with other datasets (however available to the *Data Provider*); in particular, not to match *Contributed Data* or any parts thereof with any personal or pseudonymous datasets remaining at the *Data Provider*; (ii) instruct or request the Data Processor engaged by *Data Provider* for the purpose of providing and/or anonymizing the *Contributed Data* to make available to the *Data Provider* a copy of the *Contributed Data* or any means to re-identify the *Contributed Data;* request access to *Contributed Data* contained in the HARMONY Database or to accept such access"[14].

It should be safeguarded that the Trusted Third Party, as Processor, do not act on behalf and under the instructions of HARMONY, but rather under the instructions and on behalf of the data providers. This means that a data processing agreement fulfilling the requirements of applicable data protection laws

[14] Clause 4.6 of the HARMONY Data Sharing Agreement (AMDS).

must also be concluded between the data provider as controller and the TTP as processor. According to that contract the TTP "must not engage in any activity to re-identify the Contributed Data, e.g. by way of singling out, linking back or matching any dataset provided by the Controller with other datasets (however available to the Processor). The Processor must neither transfer back to the Controller a copy of the Contributed Data or any portion of it or otherwise communicate Contributed Data to third parties other than to the HARMONY platform. Processor will also not accept deviating instructions from the Controller and even not return Contributed Data to him upon his request"[15].

Furthermore, the prohibition has also been made extensive to HARMONY Consortium members, who "shall not engage at any time during and after the term of this agreement in any activity to re-identify data by any means whatsoever including but not limited to singling out, linking back, or matching any dataset from the HARMONY Platform with other datasets (however available to the *Beneficiary*); in particular, (i) not to match any dataset from the HARMONY Platform or any parts thereof with any personal or pseudonymous datasets that are under their control, including its constituent entities, Affiliated Entities, Linked Third Parties, and Sub-contractors; (ii) instruct, or request its constituent entities, Affiliated Entities, Linked Third Parties, and Sub-contractors to engage in any activity to re-identify any dataset from the HARMONY Platform or any parts thereof (iii) extract, copy, reproduce, or duplicate the content of the database or any parts thereof from the HARMONY Platform to a database other than the HARMONY Platform"[16].

The contracts also contain effective sanctions and penalties which constitute "barriers" that reasonably prevent that a contractual entity would violate these obligations[17]. In this regard, the non-breaching Party of a data sharing agreement shall terminate it and inform the relevant Authorities[18]; the Processor shall establish a verification procedure, participated by the HARMONY Ethics Advisory Board as well as their Ethics Committee, by which it can prove compliance towards the Controller, and produce and make available on an annual basis a written report[19]; the participation in HARMONY of breaching Consortium members shall be terminated[20]. Furthermore, the termination of the respective agreements do not affect the possibility to engage liability against the breaching party.

6.2.2 Internal policies and processes

Policies, procedures, and internal guidelines which define what data use is allowed and prohibited are already developed (i.e. in the rules for the submission and approval of bench-to-bedside research questions) or foreseen as deliverables of the project. They include processes that ensure that the

[15] Clause 2.5 of the 'Agreement on the processing of personal data on behalf of a controller pursuant to art. 28 GDPR', which is signed between the data provider and the Trusted Third Party for the anonymization of data.
[16] Clause 4.5 of the HARMONY Consortium Agreement.
[17] See re. German data protection law, also Simitis/*Dammann*, BDSG, § 3 marginal 27; with regard to knowledge in the possession of a third party; see also *Arning/Forgó/Krügel*, Data protection in grid-based multicentric clinical trials: killjoy or confidence-building measure), in: Philosophical Transactions of the Royal Society A 367 (2009), 2729 (2736 et seqq.).
[18] Clause 4.7 of the HARMONY Data Sharing Agreement (AMDS).
[19] Clause 2.6 of the 'Agreement on the processing of personal data on behalf of a controller pursuant to art. 28 GDPR', which is signed between the data provider and the Trusted Third Party for the anonymization of data.
[20] Clause 4.5 of the HARMONY Consortium Agreement.

handling of the data occurs in a controlled environment; processes to response to organizational risks; the prohibition for HARMONY staff involved in the handling of data to any use of HARMONY datasets outside of what is strictly necessary for HARMONY purposes, including any transfer or communicate of the datasets. Moreover, HARMONY staff is bound to protect and maintain confidentiality.

List of Tables and Figures