H2020 – ICT-13-2018-2019

Machine Learning to Augment Shared Knowledge in Federated Privacy-Preserving Scenarios (MUSKETEER)

Grant No 824988

D2.3 Key performance indicators selection and definition

May 19

Imprint

Contractual Date of Delivery to the EC: **31 May 2019**

Author(s): **Susanna Bonura (ENG), Davide Dalle Carbonare (ENG)**

Participant(s): **Chiara Napione (COMAU), Lucrezia Morabito (COMAU), Giacomo Fecondo (FCA), Joao Correia (B3D), Petros Papachristou (HYGEIA)**

Reviewer(s): **Maria Irina Nicolae (IBM) Roberto Diaz Morales (TREE)**

Project: **Machine learning to augment shared knowledge in federated privacy-preserving scenarios (MUSKETEER)**

Work package: **WP2**

Dissemination level: **Public**

Version: **1.1**

Contacts: **Susanna Bonura – susanna.bonura@eng.it, Davide Dalle Carbonare – davide.dallecarbonare@eng.it**

Website: **www.MUSKETEER.eu**

Legal disclaimer

The project Machine Learning to Augment Shared Knowledge in Federated Privacy-Preserving Scenarios (MUSKETEER) has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824988. The sole responsibility for the content of this publication lies with the authors.

Copyright

Executive Summary

The deliverable D2.3 - Key performance indicators selection and definition, is the first result of the task T2.3 started at M1 and to be finished at M28. This task deals with the definition of KPIs related to the processing and analysis of data coming from the 2 industrial scenarios, which shall be improved through the application of MUSKETEER data platform. This task will therefore provide to the validation work package (WP7) with the expected outcomes, which shall be compared with the real results of the validation activities in order to assess the success of the project. The Evaluation Framework employed for validating the MUSKETEER platform, is based on the Goal Question Metric (GQM) method. Thus, measurement goals, questions and metrics are described to the use case implementations towards final evaluation execution by the end of the project. This document describes the planning and the definition phases, which have been achieved at this stage of the project. A revision of the KPIs and methodology will be done in M24. Starting from the final version of the D2.3, Data Collection and interpretation phases will be documented in the deliverables D7.5 and D7.6, where the description of the Smart Manufacturing and Health pilots setup and execution, together with the evaluation of the KPIs, will be reported in order to assess the usage of the MUSKETEER platform.

Document History

Version	Date	Status	Author	Comment
0.1	21 Jan 2019	For internal review	ENG	First draft TOC
0.2	28 Feb 2019	For internal review	ENG	Added Common evaluation Framework
0.3	18 March 2019	For internal review	ENG	Added Key dimensions, measurements goals and questions
0.4	10 April 2019	For internal review	COMAU, FCA	Added metrics from smart manufacturing scenario
0.5	19 April 2019	For internal review	B3D, HYGEIA	Added metrics from health scenario
0.6	26 April 2019	For internal review	ENG	Added section on related documents
0.7	3 May 2019	For internal review	ENG	Added executive summary and document structure sections
0.8	8 May 2019	For internal review	ENG	Added conclusion
0.9	10 May 2019	Ready for internal review	ENG	Overall check before internal review
1.0	27 May 2019	Final version	IBM, TREE, ENG	Feedbacks from internal reviewers
1.1	28 May 2019	Finalization	IBM	

Table of Contents

List of Figures

List of Acronyms and Abbreviations

ABBREVIATION	DEFINITION
CA	Consortium Agreement
CT	Computed Tomography
DAN	Dekanewton
DICOM	Digital Imaging and Communications in Medicine
DOW	Description of Work
GA	Grant Agreement
HL7	Health Level Seven International
J	Joule
KA	Kiloampere
MRI	Magnetic Resonance Imaging
MS	milliseconds
OEM	Original Equipment Manufacturer
PI-RADS	Prostate Imaging – Reporting and Data System
POM	Privacy Operations Modes
RSW	Resistance Spot Welding
S	second
SMC	Secure Multiparty Computation
WP	Work Package

1 Introduction

1.1 Purpose

This document defines the MUSKETEER evaluation framework and overall evaluation approach that will be implemented in WP7 according to the different MUSKETEER use cases.

Based on the Goal Question Metric method (GQM), this document is paving the way to the use case implementations towards final evaluation execution by the end of the project.

Key technical quality focuses and business priorities are identified in order to prepare the deployment of detailed appropriated evaluation questions and metrics.

This document describes the planning and the definition phases, which have been achieved at this stage of the project. A revision of the KPIs and methodology will be done in M24.

Starting from the final version of the D2.3 released in M24, Data Collection and interpretation phases will be documented in the deliverables D7.5 and D7.6, where the description of the Smart Manufacturing and Health pilots setup and execution, together with the evaluation of the KPIs, will be reported in order to assess the usage of the MUSKETEER platform.

1.2 Related Documents

MUSKETEER follows a multidimensional assessment approach, aiming at covering all the relevant aspects resulting from the project execution.

More in detail, as shown in Figure 1, the DoW [3] envisages four kinds of assessments: from the business perspective, end-users partners supported by ENG, will assess the project results on the basis of the framework and metrics identified in this document (WP2 and WP7); from a legal point of view, partners, supported by KUL, will complete a comprehensive privacy and data protection impact assessment (PCIA) that takes into account any privacy and data protection risks presented by the MUSKETEER Data Platform (WP2); technical aspects, such as scalability and computational efficiency of federated privacy-preserving machine learning algorithms, security of machine learning algorithms under the different privacy operation modes, data value extraction and monetization strategies, will be assessed on the basis of a proper framework under the leading of UC3M (WP6); finally, exploitation of the results, including market entry strategies and business models for the different industrial partners involved, will be assessed in WP8 led by IDSA.

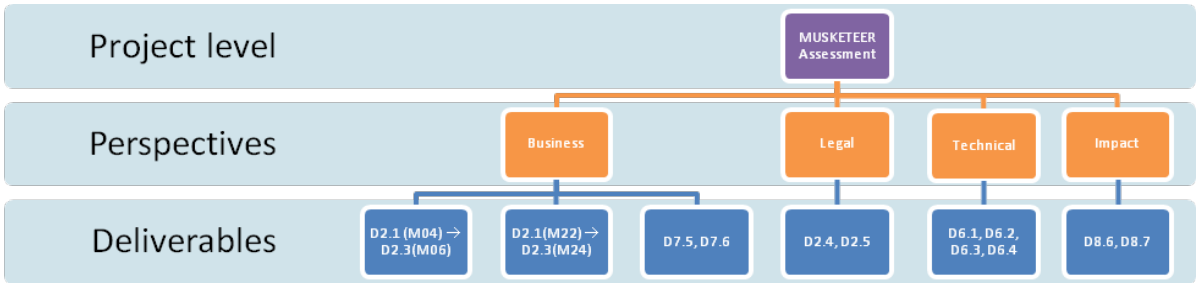

Figure 1 Deliverables related to the project assessment

Thus, the MUSKETEER assessment is covered by the following deliverables:

D2.3 - Key performance indicators selection and definition, which provides a detailed description of the technical and domain business-specific KPIs that will be used for validating the MUSKETEER data platform.

D7.5 - Use case execution and KPI evaluation in the Smart Manufacturing domain, which reports the description of the Smart Manufacturing pilot setup and execution, together with the evaluation of the KPIs in order to assess the usage of the platform (taking into account the D2.3 as input).

D7.6 - Use case execution and KPI evaluation in the Health domain, which reports the description of the Health pilot setup and execution, together with the evaluation of the KPIs in order to assess the usage of the platform (taking into account the D2.3 as input).

D2.4 - Privacy and confidentiality impact assessment report and recommendations – Initial version and D2.5 - Privacy and confidentiality impact assessment report and recommendations – Final version, which provide the findings from the PCIA completed in T2.4, and recommendations for MUSKETEER in terms of managing any privacy risks and impacts posed by the data platform to be implemented as part of a privacy-by-design approach.

D6.1 - Assessment Framework design and specification, which describes the main common evaluation framework. It will contain the design of the different tests and datasets that will be used in the evaluation, as well as the merit performance measurements to be obtained.

D6.2 - Scalability of machine learning algorithms over every POMs, which describes the scalability of every algorithm as a function of the POM, the HW available, the number of data providers and the volume of training data. It will also contain recommendations to select the best algorithm and configuration for a given privacy restriction.

D6.3 - Security of federated machine learning algorithms, which reports how confident is the accuracy of a Machine Learning algorithm when we consider attacks and detection strategies.

D6.4 - Data value extraction and monetization strategies, which describes the different data value extraction techniques and monetization strategies that can be used considering the different privacy operation modes.

D8.6 - Evaluation and impact assessment, which details the gains associated with the MUSKETEER solution, using quantitative information, and which identifies areas for further improvement and investment.

D8.7 - Business and exploitation plan, which presents MUSKETEER Business Plan for post-project exploitation of the results, including market entry strategies and business models for the different industrial partners involved.

In order to avoid overlaps among abovementioned deliverable contents, the present document will refer to the other documents when necessary.

1.3 Document Structure

The rest of the deliverable is organised as follows: Section 2 summarizes the MUSKETEER project objectives, so as described in the DoW. Section 3 introduces the MUSKETEER Evaluation Framework by drawing inspiration from Goal – Question – Metric (GQM) methodology for evaluation of processes and products. Section 4 presents measurement goals, questions and metrics which were elicited according to four dimensions: architecture, preserving operation modes, machine learning algorithms, rewarding models. In Section 5, metrics are shown from business and technical perspectives and evaluation groups are defined for each metric identified. Section 6 concludes the deliverable. It outlines the main findings of the deliverable and possible improvements that will be done in the second release of this document.

2 MUSKETEER Objectives

Based on the MUSKETEER use cases, the evaluation will demonstrate the main benefits of the MUSKETEER approach to “support the emergence of data markets and the data economy” (as outlined in the call ICT-13).

The MUSKETEER focus and overall goals will be achieved by taking into account the project objectives, as defined in MUSKETEER DoW [3]:

O1. Machine Learning over a high variety of different privacy-preserving scenarios.

O1.1. Definition of several Privacy Operation Modes (POMs) to provide compliance with the legal and confidentiality restrictions of most industrial scenarios, so to get scalable architecture design (D3.1-2) and prototype (D3.3-4) with some Privacy Preserving Modes. In order to

assess such result, it will be expected distributed efficiency (speedup/number of users) superior to 0.8 (this result will be included in report D6.2), while preserving privacy.

O1.2. Creating predictive models without directly exposing them to the data consumers (training data remains in the installations of data providers), so to obtain Federated data normalization and alignment algorithms (D4.2.1) and some of the training procedures in (D4.3.1-2, D4.4.1-2). The federated training will achieve comparable accuracy as the traditional local computing (decentralization will not affect the accuracy; this result will be included in report D6.2).

O1.3. Correct combination of different concepts of federated machine learning, differential privacy, homomorphic encryption, secure multiparty computation and distributed computing to improve the scalability of machine learning algorithms over every POM. The output will be some of the training procedures in the ML library (D4.3.1-2, D4.4.1-2), and it will be evaluated by expecting faster than current SMC privacy-preserving alternatives: PySyft, SecureML (this result will be included in report D6.2).

O1.4. Complete library of algorithms, having algorithms of different complexity levels (D4.3.1-2, D4.4.1-2). The number of implemented algorithms has to be significant. In supervised learning, the library will contain at least a classification and regression alternative of linear models, kernel methods, trees and deep neural networks. It will also include one unsupervised technique for clustering and data decomposition.

O2. Providing robustness against external and internal threats.

O2.1. Providing analysis and requirements for secure federated machine learning algorithms. We will consider vulnerabilities during training and at runtime, including the possibility of abuse from the users of the platform. Thus, the goal is having Threat model and taxonomy of the possible attacks and weaknesses for federated machine learning algorithms (D5.1).

O2.2. The POMs will be designed to allow a secure information exchange among the platform users, so to have an architecture flexible enough to handle the 8 POMs.

O2.3. Including defensive mechanisms for the federated machine learning algorithms against poisoning and evasion attacks by detecting and mitigating the effect of such attacks (D5.4 and D5.5). The defensive mechanisms will be capable of reducing the effect of poisoning (for reasonable levels of data poisoning, e.g. less than 20% of poisoning in the training dataset) and evasion attacks, compared with unsecured federated machine learning algorithms.

O2.4. Providing mechanisms to detect and mitigate the effect of abusive users in the platform trying to compromise the learning process, so to get algorithms to detect and characterize malicious users colluding to compromise the learning algorithms in the platform (D5.6, D5.7). The defensive mechanisms will be capable of mitigating colluding users' attacks for reasonable scenarios (e.g. assuming a maximum of 20% of malicious users colluding to manipulate the platform), compared with unsecured federated machine learning algorithms.

O2.5. Providing strong cyber-security against external data hackers by integrating robust and secure access and transportation protocols into the communication layers. Architecture with cyber-security mechanisms fully implemented with zero filtration of data in the validation process by surpassing.

O2.6. Developing a framework to test the security of federated machine learning against data poisoning, evasion attacks, and users' colluding attacks. This testing framework will enable the design of more secure learning algorithms and will provide an estimation of the worst-case performance of the system against different attacks with different levels of strength. They will be delivered a report and an implementation with the testing methodology to assess the security of the machine learning algorithms used in the platform against poisoning and evasion attacks (D6.3), and to evaluate the robustness of the system against malicious users (D5.6, D5.7).

O3. Enhancement of the Data Economy.

O3.1. Enhancing data providers to share their datasets thanks to the ability of creating predictive models without explicitly giving their datasets (using the FML concept), thus avoiding any possibility of personal/private information robbery (Algorithmic (D4.1) and architectural (D3.1) design). Eight different privacy operation modes will be implemented to cover the different privacy needs given in industry.

O3.2. Allowing to measure the impact of every data owner on the accuracy of the predictive models, thus allowing to monetize their contributions as a function of their real data value (Data value extraction and monetization strategies (D6.4). Different data value estimation methods (one for every POM/algorithm) will be delivered.

O3.3. European SMEs involvement (D8.5, D8.6), through more than 10 industrial diffusion events, 3 workgroups attendance, 5 workshops.

O4. Providing a standardized and extensible architecture.

O4.1. Integration with other European initiatives related with data platform, by granting the compliance with the Industrial Data Space Association reference architecture.

O4.2. Allowing interoperability with Big Data frameworks by providing portability mechanisms to load and export the predictive models from/to other platforms. The predictive models will be obtained with the ML library (D4.3.1-2, D4.4.1-2), so MUSKETEER will be capable to export the predictive models to be loaded at least into the most extended ML libraries.

O4.3. Fostering the creation of a community of developers and researchers that can extend the platform with new algorithms and attack detection mechanisms after the life of the project. Special focus will be given to Open Source Licenses. Reports with scientific dissemination (D8.1, D8.2), reports with community engagement and technology transfer (D8.2, D8.4) will be delivered.

O4.4. Fast deployment, installation and use. Architecture based on containers will ensure that applications deploy quickly, reliably, and consistently regardless of deployment environment. Software component accessible in open source repositories.

O5. Industrial demonstration of the technology advances in operational environment (TRL6)

G5.1. Demonstration that MUSKETEER will be applicable on different privacy application domains will drive the project research and developments. They will be delivered: Report with privacy confidentiality impact assessment (D2.4, D2.5), Data ownership and governance recommendations (D2.6), and correct application of MUSKETEER into two different sectors (smart manufacturing and health) will be ensured.

O5.2. Continuous monitoring and feedback process during the whole project using realistic conditions to ensure demonstration of ready-to-use technology at the end of the project (report with the technical and legal requirements (D2.1, D2.2), and industrial KPI definition (D2.3)). There will be at least one monthly meeting involving technical partners and use cases to facilitate communication and report potential problems that may arise.

O5.3. Benchmark execution, evaluation and impact assessment to ensure that the innovative technology is applicable in a wide variety of problems (report with the evaluation of every use case (D7.5, D7.6)). At least 8 correlations will be identified among the variables that characterize the welding process. The knowledge associated to these correlations will help to improve the process in time, cost, efficiency, etc. A reduction of 12% in the false alarm probability in the health use case thanks to the combination of datasets (based on use case partner estimations).

2.1 Evaluation scenarios objectives

The MUSKETEER platform will be demonstrated in two complementary scenarios. This validation phase allows developers and end-users to test the MUSKETEER functionalities under real-scenario's conditions and look for errors/bugs that need to be fixed for the MUSKETEER last version prototype that will be released by M36. At this respect two domains have been selected: Health domain (Personal data) and Smart Manufacturing domain (Industrial data). Each of the use cases is comprised by two partners: B3D and HYGEIA for Health, FCA and COMAU for Smart Manufacturing.

More in detail, with regard to the Smart Manufacturing use case, a direct impact in using MUSKETEER will be in the reduction of the manufacturing process due to:

(1) An improvement of the welding process with positive impacts both on the quality of the welding process and on the final product associated with it;

(2) A reduction in the numbers of person hours needed to configure the robots;

(3) A reduction in the robot maintenance cost.

The expected impacts in the Health use case are:

(1) Improve accuracy of AI algorithms by sharing knowledge from distinct organisations and data repositories, supporting cooperation keeping security and privacy of health data;

(2) More accurate clinical decision support tools for diagnosis and prognosis of diseases, conducting to better patient outcomes;

(3) Increasing productivity of services and more studies and patients diagnosed;

(4) More accurate clinical decision support tools for diagnosis and prognosis saving lives in emergency cases;

(5) Enable the growth of the level of research in medical imaging AI tools supported by distributed data repositories;

(6) Enable clinical practices to access medical imaging AI tools with gains of productivity and better patient outcomes;

(7) Improve Biotronics3D commercial offer, enabling partners to access its market

In both use cases, and however in other domains where the resulting assets can be directly applied, MUSKETEER will allow any organization to use our innovative Federated Machine Learning techniques to run AI algorithm solver distributed dataset of the other data provider organization in the same sector.

3 MUSKETEER Evaluation Framework

Although there are a number of comprehensive evaluation and validation methodologies in industry and academia, they often lack the goal-driven nature of businesses, thus not able to provide valuable conclusions about the real viability and sustainability of the MUSKETEER platform. The Goal Question Metric (GQM) method [1] supports such a business-driven quality improvement and validation approach quite well and has inspired the Evaluation Framework employed for validating the MUSKETEER platform.

GQM represents a systematic approach for tailoring and integrating goals with models of the software processes, products and quality perspectives of interest, based upon the specific needs of the project. The result of the application of the GQM method is the specification of a strategy targeting a particular set of issues and a set of rules for the interpretation of the measurement data. The principle behind the GQM method is that evaluation, validation and subsequent measurement should be goal-oriented.

Along the GQM method, a certain goal is defined which is refined into questions, and metrics that provide the information to answer these questions. By answering the questions, the

measured data can be analyzed to identify whether the goals have been attained. Thus, GQM defines metrics from a top-down perspective and analyses and interprets the measurement data bottom-up, as shown in Figure 2.

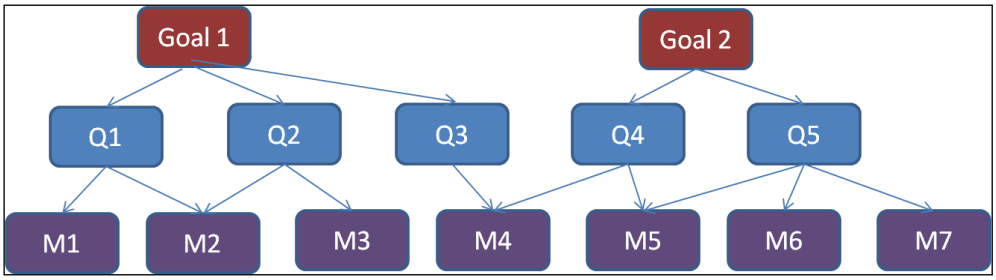

Figure 2 The GQM method

Using the GQM method, the objects of study are to be clearly identified and then validated according to a number of goals that enable focus on certain aspects of assessment. Each goal will be broken down into one or more questions that act as a vehicle for the assessment of the goal. Finally, in order to analyse and interpret the questions' results, specific metrics will be defined.

The measurement data is interpreted bottom-up. As the metrics are defined with an explicit goal in mind, the information provided by the metrics is interpreted and analysed with respect to this goal, to conclude whether or not it is attained. GQM trees of goals, questions and metrics are usually built based on the knowledge of experts.

The MUSKETEER Evaluation Framework contains four phases inspired by the GQM method (Figure 3):

- The Planning phase, during which the overall approach is defined and planned, resulting in a use case evaluation plan. The evaluation objects are defined (components, processes or resources under observation) as well as the evaluation groups (people who will participate in the evaluation process). This phase is performed to fulfil all basic requirements for conducting the validation successfully, including the definitions of actors, who will be involved, and the creation of a high-level evaluation plan.
- The Definition phase, during which the measurement scheme is defined (goal, questions and metrics are defined) and documented.
- The Data Collection phase, during which the actual data collection takes place, resulting in collected measurement and data. The data collection forms are defined, filled-in and stored.
- The Interpretation phase, during which collected data is processed with respect to the defined metrics into measurement results that provide answers to the defined questions, after which goal attainment can be evaluated.

This document describes the planning and the definition phases, which have been achieved at this stage of the project. Data collection and interpretation phases will be documented in the deliverables D7.5 and D7.6.

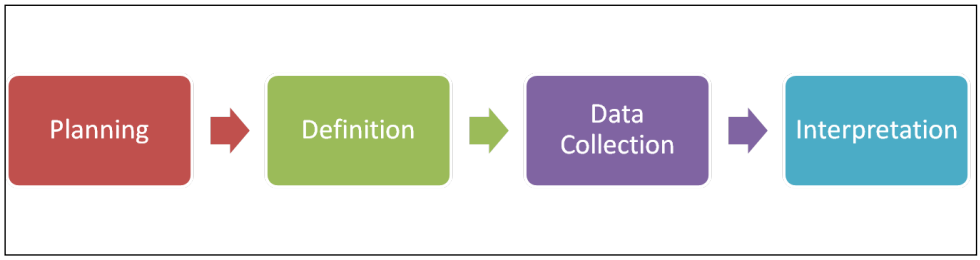

Figure 3 MUSKETEER Evaluation phases

3.1 Planning Phase

The primary objectives of this phase were to collect all required information for a successful MUSKETEER evaluation, to define the actors involved in the procedure and to prepare a high-level validation framework. This framework is supposed to act as guideline for all subsequent phases and all stakeholders involved.

Three identification steps are involved in the planning phase of the MUSKETEER evaluation:

(1) The groups of people who will participate in the evaluation process.

(2) The validation perspectives.

(3) The objects to be validated.

The following sections describe the preliminary outcome of the planning phase. However, taking into account possible updates in the prototype development (WP3), these results could be subject to further revisions until the actual validation phase performed by the end of the project (WP7).

3.1.1 MUSKETEER Evaluation Groups

On the basis of the MUSKETEER DoW, the Use cases validation and KPI evaluation will be executed in the task T7.4, where the KPI checklist resulting from this document will be used to validate the MUSKETEER platform performance, with a special focus on the business (i.e. end users) perspective, as explained in section 1.2. In addition, the KPI evaluation enables to identify the proper adjustments in order to align the results to the original requirements.

Technical assessment on confidentiality, scalability, computational efficiency, interoperability will be executed in T6.2 and T6.2 according to the framework resulting from T6.1. Privacy and data protection will be assessed in T2.4. Finally, MUSKETEER project impact assessment will be addressed in T8.3.

For the aims of the use cases validation, the evaluation groups identification takes into account the partners involved into the task T7.4: ENG, TREE, IMP, IDSA, KUL, FCA, COMAU, B3D; HYGEIA.

The MUSKETEER Evaluation Framework defines four separate evaluation groups, as shown below:

Evaluation Group	Partners
Health Evaluation Group (HEG)	B3D, HYGEIA
Smart Manufacturing Evaluation Group (SMEG)	FCA, COMAU
Development Evaluation Group (DEG)	ENG, TREE, IMP, IDSA
Legal Evaluation Group (LEG)	KUL

The domain evaluation groups (HEG and SMEG) will answer questions with a special focus on measurable business metrics. These groups based on participants from the use-case partners, will be involved mainly through surveys on the MUSKETEER platform assessment. The MUSKETEER developer group (DEG) will form the core validation group with thorough and in-depth knowledge about software quality. LEG will be called into questions specially related to POMs evaluation.

Several groups may be involved in the measurement goals assessment.

3.1.2 MUSKETEER Evaluation Perspectives

Each evaluation object defined in the next section will be validated according to two different perspectives that are defined below:

(1) The Technical Perspective, in which some aspects of the MUSKETEER architecture, Privacy preserving federated machine learning algorithms under different POMs, Federated Machine Learning Algorithms will be evaluated. Moreover, the data quality, scalability, computational efficiency and security of the MUSKETEER platform will be considered within WP6. In order to evaluate the platform from the technical perspective, the second version of this document (due to M24) will be properly enriched with metrics resulting from the deliverable D6.1.

(2) The Business Perspective, in which the response to the user needs be examined. The instrumentation used in this perspective is drawn upon usefulness and feasibility of the proposed deployment operation support.

3.1.3 Evaluation Objects

After the definition of the MUSKETEER evaluation groups and evaluation perspective, the next step is the identification and selection of appropriate evaluation objects.

The evaluation of the MUSKETEER Industrial Data Platform implies the evaluation of the following elements: *(i)* the architecture; *(ii)* the privacy operation modes (POMs); *(iii)* the machine learning techniques that will be included and the adversarial attack detection and mitigation strategies; *(iv)* the rewarding model capable to fairly monetize datasets according to the real data value.

With regard to the first item, the MUSKETEER architecture is defined to design and implement the platform on which the privacy-preserving machine learning algorithms will be instantiated. The architecture needs to ensure that the platform meets the security requirements of industrial data standards. Furthermore, scalability in the number of users, data volume and complexity of the machine learning models is required to ensure exploitability of the platform capabilities. Functionally, the platform has to provide the infrastructure and implement the services that are required to enable the distributed machine learning capabilities developed in WP4 and WP5, along with interfaces supporting the use case integration in WP7. It is worth noticing that it will be assessed from a technical and scientific point of view by means of the framework provided in WP6 according to the framework provided for this aim (task T6.1).

A key consideration in the design of the architecture will be interoperability with existing Industrial Data Space Association reference architectures [5]. The envisioned key role of internal and external connectors will be a first step in this direction. Further important design criteria are the packaging of client components as Data Apps to support the deployment in an industrial data space, as well as the ability to abstract from specific use cases in future instantiations of the platform, to make it broadly applicable across a variety of data and machine learning model types and application domains. This aspect will be evaluated in WP7.

Regarding the second item, the set of POMs will be evaluated, each one describing a potential scenario with different privacy preserving demands, but also with different computational, communication, storage and accountability features.

In addition, also the Machine Learning Algorithms will be objects to be evaluated. In fact, MUSKETEER aims to provide a high-level library of machine learning algorithms to adjust the models' parameters according to training datasets and to apply them to new data. To that end, several algorithms will be adapted to work over each POM. In order to facilitate the

integration with the most extended platforms and libraries while creating reusable models, MUSKETEER will export the predictive models to a wide variety of formats. This will facilitate the integration with other machine learning frameworks.

Finally, the rewarding model provided by MUSKETEER will be evaluated. It aims at monetizing datasets according to the adjustment of the Machine Learning model for a given task. Under these premises, the user interested in obtaining an improved model has not to pay for irrelevant data, and what is more, he/she does not need to pay in advance for data before knowing their real utility. If the final model is the result of a direct cooperation among different users (every one of them providing data to solve the task) we expect to obtain merit numbers indicating in which percentage every user data has contributed to the final model.

3.2 Definition Phase

The definition is the second phase of the MUSKETEER Evaluation Framework, and concerns all activities that should be performed to formally define a measurement scheme.

Thus, the following sections describe the preliminary results of the definition phase. However, these results are indicative as part of the overall framework and subject to further specific adjustments according to the evaluation scenarios.

As already said, MUSKETEER follows the Goal Question Metric (GQM) approach [1] in order to evaluate the effectiveness of the proposed technologies. This approach has been widely used for product and process assessment, including improvement assessment.

GQM evaluation results in the specification of a measurement system targeting a particular set of issues and a set of rules for the interpretation of the measurement data. The GQM model has three levels [2]:

(1) Conceptual level (Goal). A goal is defined for an object of measurement, for a variety of reasons, with respect to various models of quality, from various points of view, relative to a particular environment. Objects of measurement are:

   a. Products: artefacts, deliverables and documents that are produced during system lifecycle; e.g., specifications, designs, programs, and test suites.

   b. Processes: software related activities normally associated with time; e.g., specifying, designing, testing, and interviewing.

   c. Resources: items used by processes in order to produce their outputs; e.g., personnel, hardware, software, and office space.

(2) Operational level (Question). A set of questions is used to characterise the way the assessment/achievement of a specific goal is going to be performed based on some characterizing model of quality. Questions try to characterise the object of

measurement (product, process, or resource) with respect to a selected quality issue and to determine its quality from the selected viewpoint.

(3) Quantitative level (Metric). A set of data is associated with every question in order to answer it in a quantitative way. The data can be:

a. Objective, if they depend only on the object that is being measured and not on the viewpoint from which they are taken; e.g., number of versions of a document, staff hours spent on a task, and size of a program.

b. Subjective, if they depend on both the object that is being measured and the viewpoint from which they are taken; e.g., readability of a text and level of user satisfaction.

3.2.1 Key dimensions and main fields of measurements

The MUSKETEER evaluation framework and KPIs presented in this document are based on the following dimensions:

Dimension 1: MUSKETEER Architecture design, measuring the degree of adoption, integration and performance of the platform on which the privacy-preserving machine learning algorithms will be instantiated. The architecture needs to ensure that the platform meets the security requirements of industrial data standards. For this aim, the IDSA specification will be taken into account. In fact, 'the IDS Association forms the basis for a data marketplace based on European values, i.e. data privacy and security, equal opportunities through a federated design, and ensuring data sovereignty for the creator of the data and trust among participants. It forms the strategic link between the creation of data in the internet of things on the one hand side and the use of this data in machine learning (ML) and artificial intelligence (AI) algorithms on the other hand side' (https://www.internationaldataspaces.org/the-principles/). Furthermore, scalability in the number of users, data volume and complexity of the machine learning models is required to ensure exploitability of the platform capabilities. Functionally, the platform has to provide the infrastructure and implement the services that are required to enable the distributed machine learning capabilities developed in WP4 and WP5, along with interfaces supporting the use case integration in WP7.

Dimension 2: Privacy Preserving Operation Modes, measuring the degree of coverage on several scenarios. MUSKETEER must support several Privacy Operation Modes (POMs), each one describing a potential scenario with different privacy preserving demands, but also with different computational, communication, storage and accountability features. Under these modes, data never leaves the data owners' facilities, since training takes place under the Federated Machine Learning paradigm, where the model is transferred among the users, and everyone contributes by locally updating the model using their data. The resulting model is

unique, common to all the users, but in some POMs not all users get access to the trained model in unencrypted form.

Dimension 3: Machine Learning Algorithms, measuring if the set of machine learning algorithms was efficiently implemented under the assumptions of every POM. In supervised learning, the library should contain at least a classification and regression alternative of linear models, kernel methods, trees and deep neural networks. It will also include one unsupervised technique for clustering and data decomposition.

Dimension 4: Rewards model. MUSKETEER is expected to provide data value measurement/estimation such that every actor in the marketplace is rewarded according to their real contribution to the final models. A relevant marketplace needs to measure data value with more sophisticated units than the straightforward measure of data volume. The techniques to be developed will serve to guarantee that MUSKETEER will provide a rewarding model to monetize datasets according to the adjustment of the Machine Learning model for a given task. Under these premises, the user interested in obtaining an improved model has not to pay for irrelevant data, and what is more, they do not need to pay in advance for data before knowing their real utility. If the final model is the result of a direct cooperation among different users (every one of them providing data to solve the task) we expect to obtain merit numbers indicating in which percentage every user data has contributed to the final model.

4 Measurement Goals, Questions and Metrics

This section lists the metrics to the evaluation of the MUSKETEER results that will be measured in the use cases validation. The purpose is to unify the terminology and its meaning. All the project partners that use a certain metric should have the same understanding of its measurement [4].

The first step in the definition process is the definition of formal measurement goals. These validation objectives are derived from the evaluation objects and components, which are already identified in the preceding planning phase. Measurement goals have to be defined in an understandable way and with a clear structure.

The template shown in Figure 4 underpins a generic evaluation goal's purposes based on the original GQM method. The MUSKETEER measurement goals were defined accordingly both from the technical and the business perspective.

Analyse	Clear evaluation object identification
For the purpose of	Understanding, controlling or improving the object?
With respect to	The particular object quality focus
From the point of view	The concerned evaluation group(s)
In the context of	The environment in which the measurement takes place

Figure 4 GQM goal definition template

The following subsections define goal, questions, metrics by taking into account the dimensions abovementioned. They are derived from the MUSKETEER objectives as described in Section 2.

4.1 Dimension 1: MUSKETEER Architecture

4.1.1 Goals, questions, metrics

The goal G1.1 is described as follows:

G1.1	
Analyse	MUSKETEER Architecture
For the purpose of	Evaluate
With respect to	standardization and extensibility
From the view point of	Technical Perspective
In the context of	Use Cases validation (WP7)

The questions identified for the goal G1.1 are listed below.

Identifier	Questions
G1.1_Q01	Is the MUSKETEER architecture aligned with the Industrial Data Space Association reference architecture?
G1.1_Q02	Does it Allow interoperability with Machine Learning frameworks?
G1.1_Q03	Does it foster the creation of a community of developers and researchers that can extend the platform with new algorithms and attack detection mechanisms?
G1.1_Q0.4	Does it allow Fast deployment, installation and use?

The metrics identified for each question are listed below.

Identifier	KPI	Format	Method of collection and measurement
G1.1_Q01_M01	Number of artifacts aligned with IDSA reference architecture / Total number of artifacts	Decimal <=1	Online questionnaire; face-to-face interview (envisaged the involvement of IDSA in accordance with the DoW)
G1.1_Q02_M01	Number of ML libraries supported to export the predictive models /Total of the best-known ML libraries	Decimal <=1	Online questionnaire; fate-to-face interview (the best-known ML libraries will be detailed more in the second version of this document)
G1.1_Q03_M01	open source web communities' interactions	Integer	Field survey
G1.1_Q0.4_M01	Number of SW applications released as 66images	Integer	Field survey
G1.1_Q0.4_M02	Number of software components released in open source repositories	Integer	Field survey

The goal G1.2 is described as follows:

G1.2	
Analyse	MUSKETEER Architecture
For the purpose of	Evaluate
With respect to	standardization and extensibility
From the view point of	Business Perspective
In the context of	Health evaluation scenario (WP7)

The questions identified for the goal G1.2 are listed below.

Identifier	Questions
G1.2_Q01	Does it allow fast deployment and installation?
G1.2_Q02	Is it easy to use?
G1.2_Q03	Does it require special hardware locally?
G1.4_Q04	Does it allow interoperability with Medical Imaging Systems implementing DICOM, HL7 and IHE?

The metrics identified for each question are listed below.

Identifier	KPI	Format	Method of collection and measurement
G1.2_Q01_M01	(Time taken to deploy and install the MUSKETEER client) * (Number of employees involved to deploy and install the MUSKETEER client)	(HH:MM) * number of employees	Online questionnaire; face-to-face interview
G1.2_Q01_M02	(Time taken to update the MUSKETEER client) * (Number of employees involved update the MUSKETEER client)	(HH:MM) * number of employees	Online questionnaire; face-to-face interview
G1.2_Q02_M01	Time taken by one person to create a task	HH:MM	Online questionnaire; face-to-face interview
G1.2_Q02_M02	Time taken to run the training procedure associated to a given ML task	HH:MM	Online questionnaire; face-to-face interview
G1.2_Q02_M03	Time taken to select and use a trained ML model	HH:MM	Online questionnaire; face-to-face interview
G1.2_Q02_M04	Time required for training a new user	HH:MM	Online questionnaire; face-to-face interview
G1.2_Q02_M05	Number of functionalities of MUSKETEER supported by documentation (e.g. user guide) / Total number of functionalities implemented	Percentage	Verification of platform
G1.2_Q03_M01	Cost of local special equipment	Number (EUR)	Online questionnaire; face-to-face interview
G1.2_Q03_M02	Cost of setting up local special equipment	Number (EUR)	Online questionnaire; face-to-face interview
G1.2_Q04_M01	Integration profile conformance statements	Number	Verification of platform/ technical documentation

The goal G1.3 is described as follows:

G1.3	
Analyse	MUSKETEER Architecture
For the purpose of	evaluate
With respect to	standardization and extensibility
From the view point of	Business Perspective
In the context of	Smart Manufacturing evaluation scenario (WP7)

The questions identified for the goal G1.3 are listed below.

Identifier	Questions
G1.3_Q01	Does it allow fast deployment, installation and updating?
G1.3_Q02	Is it easy to use?
G1.3_Q03	Are there different visibility constraints based on user permissions?
G1.3_Q04	Is the architecture compliant with industry standard and production plant IT policies?
G1.3_Q05	Does the platform require a special hardware locally?
G1.3_Q06	Is it possible to download the model?
G1.3_Q07	Is it fast enough the training of the model?
G1.3_Q08	When a new task is launched, what is the algorithm used and its parameters?
G1.3_Q09	Is it possible to report a comment on an unexpected behavior of algorithm during a *business* user session?

The metrics identified for each question are listed below.

Identifier	KPI	Format	Method of collection and measurement
G1.3_Q01_M01	(Time taken to deploy and install the MUSKETEER client) * (Number of employees involved to deploy and install the MUSKETEER client)	(HH:MM) * number of employees	Online questionnaire; face-to-face interview
G1.3_Q01_M02	(Time taken to update the MUSKETEER client) * (Number of employees involved to update the MUSKETEER client)	(HH:MM) * number of employees	Online questionnaire; face-to-face interview

G1.3_Q02_M01	Time taken by one person to create a task	HH:MM	Online questionnaire; face-to-face interview
G1.3_Q02_M02	Time taken to run the training procedure associated to a given ML task	HH:MM	Online questionnaire; face-to-face interview
G1.3_Q02_M03	Number of screens supported by help option	Integer	Verification of platform
G1.3_Q02_M04	Time taken by one person to add a new user	HH:MM	Online questionnaire; face-to-face interview
G1.3_Q02_M05	Time taken by one person to select and use a ML model	HH:MM	Online questionnaire; face-to-face interview
G1.3_Q03_M01	Different information for different user permissions	Boolean (true/false)	Verification of platform
G1.3_Q04_M01	Compliance with Information Technology production plant policies	Boolean (true/false)	Verification of platform / technical documentation
G1.3_Q05_M01	Cost of local special equipment	Number (EUR)	Online questionnaire; face-to-face interview
G1.3_Q05_M02	Cost of setting up local special equipment	Number (EUR)	Online questionnaire; face-to-face interview
G1.3_Q06_M01	Possibility of downloading the ML model	Boolean (true/false)	Availability of a button to download the ML model
G1.3_Q07_M01	(Time taken to train the model) / (number of samples)	HH:MM	Verification of platform
G1.3_Q07_M02	Time taken to classify the current status of the equipment	HH:MM	Verification of platform
G1.3_Q07_M03	Time taken to classify the future status of the equipment	HH:MM	Verification of platform
G1.3_Q08_M01	Possibility to access information about the algorithm used and its parameters.	Boolean (true/false)	Verification of platform
G1.3_Q09_M01	Report an unexpected behavior	Boolean (true/false)	Availability of a report linked to the session

4.2 Dimension 2: Privacy Preserving Operation Modes

4.2.1 Goals, questions, metrics

The goal G2.1 is described as follows:

G2.1	
Analyse	Privacy Preserving Operation Modes
For the purpose of	evaluate
With respect to	privacy, computational overload, central storage requirements, communication requirements, data utility accountability
From the view point of	Technical Perspective
In the context of	Use Cases execution (WP7)

The questions identified for the goal G2.1 are listed below.

Identifier	Questions
G2.1_Q01	Will POMs be designed to allow a secure information exchange among platform user?
G2.1_Q02	Will POMs provide compliance with the legal and confidentiality restrictions of most industrial scenarios?
G2.1_Q03	Will the scalability of machine learning algorithms be improved over every POM, with regard to correct combination of different concepts of federated machine learning, differential privacy, homomorphic encryption, secure multiparty computation and distributed computing?

The metrics identified for each question are listed below.

Identifier	KPI	Format	Method of collection and measurement
G2.1_Q01_M01	Number of robust POMs for use cases	Integer	On field
G2.1_Q02_M01	Number of robust POMs for use cases	Integer	On field
G2.1_Q02_M02	Speedup/number of users while POM is applied	Ratio	On field
G2.1_Q03_M01	Number of training procedures implemented	Integer	On field

G2.1_Q03_M02	Speed of privacy-preserving machine learning algorithms implemented with respect to other existing solutions	Ratio	On field

The goal G2.2 is described as follows:

G2.2	
Analyse	Privacy Preserving Operation Modes
For the purpose of	evaluate
With respect to	privacy, computational overload, central storage requirements, communication requirements, data utility accountability
From the view point of	Business Perspective
In the context of	Health evaluation scenario (WP7)

The questions identified for the goal G2.2 are listed below.

Identifier	Questions
G2.2_Q01	How easy it is to verify and declare the privacy requirements?
G2.2_Q02	Due to our policy, is an adequate level of data privacy granted?
G2.2_Q03	Is it GDPR compliant?
G2.2_Q04	Is it compliant with Medical Devices' standards and regulations (MDR EU REGULATION 2017/745, EN ISO 13485:2016 and EN ISO 14971:2012)?
G2.2_Q05	Is it compliant with EN ISO/IEC 27001:2017?

The metrics identified for each question are listed below.

Identifier	KPI	Format	Method of collection and measurement
G2.2_Q01_M01	Time taken by one person to verify and declare privacy requirements	HH:MM	Online questionnaire; fate-to-face interview
G2.2_Q01_M02	Number of options expressed in natural language/ total number of steps	Decimal	Online questionnaire; fate-to-face interview
G2.2_Q02_M01	Adequate level of data privacy when sharing data with other end users	Boolean (true/false)	Online questionnaire; fate-to-face interview
G2.2_Q02_M02	Adequate level of data privacy when sharing models with other end users	Boolean (true/false)	Online questionnaire; fate-to-face interview

G2.2_Q02_M03	Adequate level of data privacy when training models	Boolean (true/false)	Online questionnaire; fate-to-face interview
G2.2_Q03_M01	Implementation of GDPR requirements	Boolean (true/false)	Platform/ technical documentation verification
G2.2_Q04_M01	Implementation of MDR, EN ISO 13485:2016 and EN ISO 14971:2012 security and privacy requirements	Boolean (true/false)	Platform/ technical documentation verification
G2.2_Q04_M01	Implementation of EN ISO/IEC 27001:2017 requirements	Boolean (true/false)	Platform/ technical documentation verification
G2.2_Q04_M02	Number of security controls implemented and verified	Number	Platform/ technical documentation verification

The goal G2.3 is described as follows:

G2.3	
Analyse	Privacy Preserving Operation Modes
For the purpose of	evaluate
With respect to	privacy, computational overload, central storage requirements, communication requirements, data utility accountability
From the view point of	Business Perspective
In the context of	Smart Manufacturing evaluation scenario (WP7)

The questions identified for the goal G2.3 are listed below.

Identifier	Questions
G2.3_Q01	How easy it is to declare the privacy requirements?
G2.3_Q02	Due to our policy, is an adequate level of data privacy granted?
G2.3_Q03	All the features of the selected POM are implemented?
G2.3_Q04	How easy it is to encrypt or decrypt data or model?
G2.3_Q05	Does my central storage support the platform requirements?
G2.3_Q06	How easy it is to verify if all the communications are working?
G2.3_Q07	Which is the maximum dimension of messages supported by the platform?

The metrics identified for each question are listed below.

Identifier	KPI	Format	Method of collection and measurement
G2.3_Q01_M01	Time taken by one person to declare privacy requirements	HH:MM	Online questionnaire; face-to-face interview
G2.3_Q01_M02	Number of options expressed in natural language/ total number of steps	Decimal	Online questionnaire; face-to-face interview
G2.3_Q02_M01	Possibility of sharing my data with other users	Boolean (true/false)	Online questionnaire; face-to-face interview
G2.3_Q02_M02	Possibility to control if other users are visualizing my data	Boolean (true/false)	Online questionnaire; face-to-face interview
G2.3_Q02_M03	Possibility to control who has the grant for visualizing my data	Boolean (true/false)	Online questionnaire; face-to-face interview
G2.3_Q03_M01	Correct receipt of my private key	Boolean (true/false)	Online questionnaire; face-to-face interview
G2.3_Q03_M02	Correct encryption of data or model using my private key	Boolean (true/false)	Real time test between end users and central node to compare data or model
G2.3_Q03_M03	Correct receipt of central node encrypted results of computation	Boolean (true/false)	Online questionnaire; face-to-face interview
G2.3_Q03_M04	Correct decrypt of central node results	Boolean (true/false)	Real time test between end users and central node to compare data or model

G2.3_Q04_M01	Time taken by one person to encrypt or decrypt data or model	HH:MM	Online questionnaire; face-to-face interview
G2.3_Q05_M01	(my central storage size) - (Storage requested by the platform)	Decimal	Direct calculation performed by the platform. Measurements: KPI has to be > 0
G2.3_Q06_M01	Possibility to verified if all the communication protocols are enabled	Boolean (true/false)	Online questionnaire; face-to-face interview
G2.3_Q07_M01	Maximum dimension of messages (sent or received) support by the platform	Decimal	Stress tests

4.3 Dimension 3: Machine Learning Algorithms

4.3.1 Goals, questions, metrics

For completeness, the G3.1 and G3.2 measurement goal descriptions are reported below. Please note that scalability, computational efficiency, security, of MUSKETEER machine learning algorithms are addressed and assessed in WP6. The main metrics identified during the tasks of WP6 will be reported in the final version of this deliverable (M24).

G3.1	
Analyse	Federated Privacy-preserving Machine Learning Algorithms
For the purpose of	evaluate
With respect to	scalability, computational efficiency
From the view point of	Technical Perspective
In the context of	evaluation scenario of WP6 (T6.2)

G3.2	
Analyse	Machine Learning Algorithms
For the purpose of	evaluate
With respect to	security
From the view point of	Technical Perspective
In the context of	evaluation scenario of WP6 (T6.3)

The goal G3.3 is described as follows:

G3.3	
Analyse	Machine Learning Algorithms
For the purpose of	evaluate
With respect to	pre-processing, normalization, data alignment, supervised and unsupervised learning
From the view point of	Business Perspective
In the context of	Health evaluation scenario (WP7)

Identifier	Questions
G3.3_Q01	Given the data on pelvis MRI exams as well as multiparametric MRI exams for male patients and a model adopted for predictions, is MUSKETEER able to improve the prediction model of the existence and grade of prostate cancer?
G3.3_Q02	Given the data on pelvis MRI exams as well as multiparametric MRI exams for male patients and a model adopted for predictions, is MUSKETEER trained model able to better identify and segment prostate cancer lesions?
G3.3_Q03	Data quality impacts on Machine Learning results, how is data quality controlled?

The metrics identified for each question are listed below.

Identifier	KPI	Format	Method of collection and measurement
G3.3_Q01_M01	[Sensitivity (true positive rate) of prediction of existence of cancer by using new trained model provided by MUSKETEER]/ [Sensitivity (true positive rate) of prediction of existence of cancer by using the model as-is]/. Formula:	decimal	Test ML models

	Sensitivity = (True Positive) / (True Positive + False Negative)		
G3.3_Q01_M02	[Specificity (true negative rate) of prediction of existence of cancer by using new trained model provided by MUSKETEER]]/ [Specificity (true negative rate) of prediction of existence of cancer by using the model as-is]. Formula: Specificity = (True Negative) / (True Negative + False Positive)	decimal	Test ML models
G3.3_Q01_M03	[Accuracy of classification of studies with PIRADS score for each Patient by using new trained model provided by MUSKETEER]/ [Accuracy of classification of studies with PIRADS score for each Patient by using the model as-is]	decimal	Test ML models
G3.3_Q02_M01	Accuracy of lesion's segmentation	Percentage	Test ML models
G3.3_Q02_M02	Accuracy of lesion's PI-RADS score classification	Percentage	Test ML models
G3.3_Q03_M01	Ratio of data errors in pre-processing, normalization, data alignment	Percentage	Tests in pre-processing, normalization, data alignment
G3.3_Q03_M02	Ratio of empty data in pre-processing, normalization, data alignment	Percentage	Tests in pre-processing, normalization, data alignment
G3.3_Q03_M03	Accuracy of ML trained model with dataset tested with external datasets	Percentage	Test ML models

The goal G3.4 is described as follows:

G3.4	
Analyse	Machine Learning Algorithms
For the purpose of	evaluate
With respect to	pre-processing, normalization, data alignment, supervised and unsupervised learning
From the view point of	Business Perspective
In the context of	Smart Manufacturing evaluation scenario (WP7)

The questions identified for the goal G3.4 are listed below.

Identifier	Questions
G3.4_Q01	Given historical data of the welding gun (current, welding time, welding force….) and a model, is MUSKETEER able to provide the improvement for grouping the dataset in such a way that objects in the same group are more similar to each other than to those in other groups?
G3.4_Q02	Given the data on the welding gun (current, welding time, welding force….) and a model, is MUSKETEER able to improve the prediction of future *n* welding parameters?
G3.4_Q03	Given the data on the welding gun (current, welding time, welding force….) and a model, is MUSKETEER able to improve the classification of quality index class?
G3.4_Q04	Are the results of the prediction interpretable? In particular, when the task is completed, is the outcome characterized also with all information concerning the context (a header containing model, algorithm, …) where it has been executed?
G3.4_Q05	Is the algorithm able to use data of different plants to extract knowledge useful for all?
G3.4_Q06	Is the ML algorithm reliable? Does it give comparable output working on the same data and in the same conditions in different sessions?
G3.4_Q07	Does the prediction provided by the model (after the MUSKETEER training) provide an improved knowledge to build a reliable relationship between spot weld issues and possible causes linked to applied welding process (parameters values,…)?

The metrics identified for each question are listed below.

Identifier	KPI	Format	Method of collection and measurement
G3.4_Q01_M01	Cluster cohesion index after the MUSKETEER training/ Cluster cohesion index before the MUSKETEER training	Double	Clustering analysis with *n* groups using all main welding parameters
G3.4_Q02_M01	Mean squared error of the difference between the predicted quality index after the MUSKETEER training and the predicted quality	Double	Test ML models

	index before the MUSKETEER training		
G3.4_Q02_M02	Recognition of a trend in parameters values (current, force, quality index, ...)	Boolean (true/false)	Trend detection algorithm
G3.4_Q03_M01	Cluster cohesion index after the MUSKETEER training/ Cluster cohesion index before the MUSKETEER training	Double	Classification of historical data in *n* groups of welding process states (suspected defect, normal...), labelled with a discretization of quality index possible values.
G3.4_Q03_M02	Accuracy of classification after the MUSKETEER training/ Accuracy of classification before the MUSKETEER training	Percentage	Test ML models
G3.4_Q03_M03	Precision after the MUSKETEER training/ Precision before the MUSKETEER training (PRECISION = TRUE POSITIVE / (TRUE POSITIVE + FALSE POSITIVE) Recall = TRUE POSITIVE / (TRUE POSITIVE + FALSE NEGATIVE)	Double	Test ML models
G3.4_Q04_M01	Time taken to one person to understand algorithm output	HH:MM	Online questionnaire; face-to-face interview
G3.4_Q04_M02	Return of the parameter which most influences the classification	Boolean (true/false)	Online questionnaire; face-to-face interview
G3.4_Q04_M03	Completeness of output header	Boolean (true/false)	Check list of fields to be characterized *Measurement*: 100% of fields have a value
G3.4_Q05_M01	Accuracy of federated model >= accuracy of local (trained with single plant's data) model	Boolean (true/false)	Test ML model

G3.4_Q06_M01	Difference of two output calculated on same input in different sessions	Double	Comparison of output *Measurement:* zero or not relevant differences detected
G3.4_Q06_M02	Ratio of data errors in pre-processing, normalization and data alignment	Percentage	Tests in pre-processing, normalization and data alignment
G3.4_Q07_M01	Provide a cause-effect relationship	Two-dimensional array whose values are boolean	Face to face interview *Measurements*: Presence of "true" values in two-dimensional array

4.4 Dimension 4: Rewarding Models

4.4.1 Goals, questions, metrics

For completeness, the G4.1 measurement goal description is reported below. It refers to models and techniques to obtain data value estimations more elaborated than the straightforward measure of data volume. The techniques will be developed in the task T6.4 and will serve to guarantee that MUSKETEER rewarding model only (or mainly) rewards those datasets relevant to adjust a good Machine Learning model (for a given task). Such rewarding models and techniques will be assessed in T6.4. The main metrics identified during the tasks of WP6 will be reported in the final version of this deliverable (M24).

G4.1	
Analyse	Rewarding model
For the purpose of	evaluate
With respect to	data value
From the view point of	Technical Perspective
In the context of	evaluation scenario of WP6 (T6.4)

5 KPIs and goals evaluation perspectives

In this section, we present the metrics for the key dimensions and the main fields of measurement introduced in Section 4, from the perspectives introduced in Section 3.1.2, by also taking into account the deliverable D2.1 as input.

5.1 Business Perspective

5.1.1 Smart Manufacturing use case

The presence and use of robots, more generally of equipment and tools, in FCA's factories is more and more pervasive and will be more and more in the years to come. High quality manufacturing processes require a high number of person-hours spent in the configuration of the robots and their posterior maintenance with routine inspections. However, reducing the number of inspections is a bad strategy that can reduce a plant's overall productive capacity by 5 to 20% since robots use to degrade until quality problems arise and it is necessary to stop the manufacturing plant.

This cost can be highly reduced with smart manufacturing thanks to the introduction of machine learning to define and update the robot settings. A predictive model, trained on historical records, can be used to alert of a possible future failure or a decrease of quality, allowing a more efficient maintenance. However, most of the times there are not enough historical records to solve these tasks collecting data from only one robot.

Since a single robot manufacturer creates instances of the same type of equipment and they can be used to perform the same operations in different sites, a potential benefit of combining the historical records of all of them can be used to improve the predictive models, with potential impacts on the product quality and plant efficiency.

Other advantages can be obtained combining historical records, not only belonging to different plants of the same company, but also to different car manufacturers. This could bring to a benefit for all the companies involved and also to the equipment supplier, which in this case is represented by COMAU.

In order to identify the effects of degraded conditions and consequent quality problems in advance of the AS IS, a possible approach is to collect and analyse all the configuration and use parameters, for example of the same class of welding guns, with the aim of searching for any correlation between the imprecision found and the conditions that generated it.

It is also necessary that the stakeholders involved share a data collection and analysis platform that is reliable and secure and that guarantees data protection. Sharing and analysis data must favour the creation of a reference model, based on Artificial Intelligence techniques, which, appropriately fed and trained, through the use of shared data, is able to guarantee the quality for that specific operation, at desired levels and with a positive impact on the final process.

The purpose of this use case is to collect and analyse the data related to the welding process, available in the various plants, in order to search, with the support of artificial intelligence technologies, any correlations among the variables that characterize the process so that a produced welding point will be of the expected quality.

Welding is the process by which two pieces of metal can be joined together thanks to a fusion of the layers. A welding gun is composed by:

- two mechanical arms, one fixed and the other which can move;
- a linear motor which allows the arm movement;
- a copper electrode at the end of each arm, which is in contact with the metal sheets to weld;
- a water-cooling system;
- a welding tray which is the controller of the current supplied for the welding.

In general, the number of metal sheets to weld varies from 2 to 3. The supplied current, the time spent on the welding process and the pressure applied by the arms on the metal sheets strictly depend on the number of layers and on their thickness.

The current is supplied by the welding tray and flows through the arms up to the pieces of metal to melt.

The spot-welding time cycle is characterized by four time-measurements: squeeze time, weld time, hold time and off time. The squeeze time represents time between pressure application and weld; the weld time represents the weld time in cycles; the hold time represents the time the pressure is maintained after weld operation is completed and off time the time in which electrodes are separated to enable the next spot.

During each welding point the electrodes are subjected to a degradation and they get dirtier. This cause a loss of quality in following welding points. For this reason, after a predefined number of points, the electrodes undergo to a dressing process, which consists on a small material removal. After some removals the electrode has to be changed. Welding data contain information of these processes by means of counter variables.

In general, RSW is based on four major factors, which most describe the welding process:

- amount of current that passes through the “work piece” [kA];
- time in which the current flows through the “work piece” [s];
- pressure that electrodes apply on the “work piece” [daN];
- the area of the electrode tip contacts with “work piece”.

The welding process can be represented as a decomposable dynamic system, as shown in the Figure 5, in Inputs, Outputs and Disturbance events.

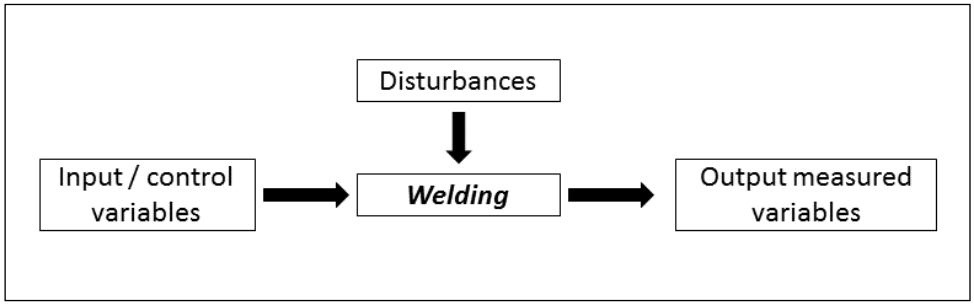

Figure 5 Welding process logical representation

As input we mean the value of the optimal parameters to be set to obtain an expected quality welding; as output we mean the measured final result, e.g. the input can be the current between the two electrodes set to obtain a right welding point; as interference factor all the boundary events that affect the expected result, like the metal sheets quality; as output the measured current effectively used by the welding gun.

In particular, as regards the inputs, the parameters involved are: welding current, running speed of the sealing element, pressure of the sealing element on the surface to be welded.

Regarding to the disturbing factors that are mainly involved in a welding process, it is necessary to take into account: mutual position of the elements to be welded, impurities of the surface to be welded, state of wear of the sealing element and stabilization of electric current used for welding. As output, we mean: evidence of detected defects and classification of the defects detected with the methods in use (e.g. visual inspection or indirect measurement by ultrasound).

Today, poor maintenance strategies can reduce a plant's overall productive capacity between 5 and 20%. Recent studies also show that unplanned downtime costs industrial manufacturers an estimated $50 billion each year. It can be difficult to determine how often a machine should be taken offline to be serviced as well as weigh the risks of lost production time against those of a potential breakdown. Machine Learning can create predictive model to improve the quality of the manufacturing processes and to detect errors and future failures, however good predictive models require combining datasets of similar equipment in different factories.

Here we can identify several of the previously defined barriers (data confidentiality, data ownership, uncertain data value, adversarial attacks) that will be avoided thanks to MUSKETEER. FCA and COMAU will share their data to identify and develop methods and techniques that allow collection and analysis of the data related to the welding process, available in the various plants, in order to search, with the support of artificial intelligence technologies, any correlations among the variables that characterize the process so that a produced welding point will be of the expected quality and the remaining time to loss of quality. The availability of the correlations sought enables the improvement of the welding process with positive impacts both on the quality of the welding process and on the final product associated with it.

In particular, the sharing and analysis of data could generate a distribution curve of the probability of error, in qualitative terms, stabilized within a set interval and such that the impact on the quality of the final product is acceptable. Moreover, the estimation of the remaining time to loss of quality can improve the maintenance organization, avoiding replacements of still functioning pieces and, on the other hand, avoiding unexpected failures.

This will have a direct impact in the reduction of the manufacturing process due to: (1) An improvement of the welding process with positive impacts both on the quality of the welding process and on the final product associated with it; (2) A reduction in the welding gun maintenance cost.

With regard to privacy and security concerns, since historical records contain information about the industrial processes of a company and used solutions. Any information leakage can potentially reveal industrial secrets about internal manufacturing processes and the problems that production teams has to face with in the plants. This information can give competitive advantage to OEM competitors and cause damages to brand. That's why we will use IDSA concepts and models to ensure confidentiality and privacy protection to the IDS ecosystem stakeholders. In addition, a factory may be hit by cyber-attacks. A data poisoning attack can produce a useless predictive model and the malfunction of the production plant. An attack can lead to a false alarm of a possible future failure and a maintenance cost increasing.

5.1.1.1 **Metrics**

All the metrics belonging to the smart manufacturing use case are listed below, together with evaluation groups involved. It is worth noticing that, as T7.4 leader, ENG will be involved in all metric evaluations.

Identifier	KPI	Format	Method of collection and measurement	Evaluation Groups
G1.3_Q01_M01	(Time taken to deploy and install the MUSKETEER client) * (Number of employees involved to deploy and install the MUSKETEER client)	(HH:MM) * number of employees	Online questionnaire; face-to-face interview	SMEG
G1.3_Q01_M02	(Time taken to update the MUSKETEER client) * (Number of employees involved to update the MUSKETEER client)	(HH:MM) * number of employees	Online questionnaire; face-to-face interview	SMEG

G1.3_Q02_M01	Time taken by one person to create a task	HH:MM	Online questionnaire; face-to-face interview	SMEG
G1.3_Q02_M02	Time taken to run the training procedure associated to a given ML task	HH:MM	Online questionnaire; face-to-face interview	SMEG
G1.3_Q02_M03	Number of screens supported by help option	Integer	Verification of platform	SMEG
G1.3_Q02_M04	Time taken by one person to add a new user	HH:MM	Online questionnaire; face-to-face interview	SMEG
G1.3_Q02_M05	Time taken by one person to select and use a ML model	HH:MM	Online questionnaire; face-to-face interview	SMEG
G1.3_Q03_M01	Different visualizations for different user permissions	Boolean (true/false)	Verification of platform	SMEG
G1.3_Q04_M01	Compliance with Information Technology production plant policies	Boolean (true/false)	Verification of platform / technical documentation	SMEG
G1.3_Q05_M01	Cost of local special equipment	Number (EUR)	Online questionnaire; face-to-face interview	SMEG
G1.3_Q05_M02	Cost of setting up local special equipment	Number (EUR)	Online questionnaire; face-to-face interview	SMEG
G1.3_Q06_M01	Possibility of downloading the ML model	Boolean (true/false)	Availability of a button to download the ML model	SMEG
G1.3_Q07_M01	(Time taken to train the model) / (number of samples)	HH:MM	Verification of platform	SMEG
G1.3_Q07_M02	Time taken to classify the current status of the equipment	HH:MM	Verification of platform	SMEG
G1.3_Q07_M03	Time taken to classify the future status of the equipment	HH:MM	Verification of platform	SMEG
G1.3_Q08_M01	Possibility to see a resume page to	Boolean (true/false)	Verification of platform	SMEG

	show what is the algorithm used and its parameters.			
G1.3_Q09_M01	Report an unexpected behavior	Boolean (true/false)	Availability of a comment area linked to the session *Measurement*: Availability of area	SMEG
G2.3_Q01_M01	Time taken by one person to declare privacy requirements	HH:MM	Online questionnaire; face-to-face interview	SMEG, LEG
G2.3_Q01_M02	Number of options expressed in natural language/ total number of steps	Decimal	Online questionnaire; face-to-face interview	SMEG, LEG
G2.3_Q02_M01	Possibility of sharing my data with other users	Boolean (true/false)	Online questionnaire; face-to-face interview	SMEG. LEG
G2.3_Q02_M02	Possibility to control if other users are visualizing my data	Boolean (true/false)	Online questionnaire; face-to-face interview	SMEG, LEG
G2.3_Q02_M03	Possibility to control who has the grant for visualizing my data	Boolean (true/false)	Online questionnaire; face-to-face interview	SMEG, LEG
G2.3_Q03_M01	Correct receipt of my private key	Boolean (true/false)	Online questionnaire; face-to-face interview	SMEG, LEG
G2.3_Q03_M02	Correct encryption of data or model using my private key	Boolean (true/false)	Real time test between end users and central node to compare data or model	SMEG, LEG
G2.3_Q03_M03	Correct receipt of central node encrypted results of computation	Boolean (true/false)	Online questionnaire; face-to-face interview	SMEG, LEG
G2.3_Q03_M04	Correct decrypt of central node results	Boolean (true/false)	Real time test between end users and central node to compare data or model	SMEG, LEG

G2.3_Q04_M01	Time taken by one person to encrypt or decrypt data or model	HH:MM	Online questionnaire; face-to-face interview	SMEG, LEG
G2.3_Q05_M01	(my central storage size) - (Storage requested by the platform)	Decimal	Direct calculation performed by the platform. Measurements: KPI has to be > 0	SMEG, LEG
G2.3_Q06_M01	Possibility to verified if all the communication protocols are enabled	Boolean (true/false)	Online questionnaire; face-to-face interview	SMEG, LEG
G2.3_Q07_M01	Maximum dimension of messages (sent or received) support by the platform	Decimal	Stress tests	SMEG, LEG
G3.4_Q01_M01	Cluster cohesion index	Double	Clustering analysis with *n* groups using all main welding parameters	SMEG, DEG, LEG
G3.4_Q02_M01	Mean squared error of the difference between the predicted quality index and the real quality index	Double	Test ML models	SMEG, DEG, LEG
G3.4_Q02_M02	Recognition of a trend in parameters values (current, force, quality index,…)	Boolean (true/false)	Trend detection algorithm	SMEG, DEG, LEG
G3.4_Q03_M01	Cluster cohesion index	Double	Classification of historical data in *n* groups of welding process states (suspected defect, normal...), labelled with a discretization of quality index possible values.	SMEG, DEG, LEG
G3.4_Q03_M02	Accuracy of classification	Percentage	Test ML models	SMEG, DEG, LEG

G3.4_Q03_M03	Precision = TRUE POSITIVE / (TRUE POSITIVE + FALSE POSITIVE) Recall = TRUE POSITIVE / (TRUE POSITIVE + FALSE NEGATIVE)	Double	Test ML models	SMEG, DEG, LEG
G3.4_Q04_M01	Time taken to one person to understand algorithm output	HH:MM	Online questionnaire; face-to-face interview	SMEG, DEG, LEG
G3.4_Q04_M02	Return of the parameter which most influences the classification	Boolean (true/false)	Online questionnaire; face-to-face interview	SMEG, DEG, LEG
G3.4_Q04_M03	Completeness of output header	Boolean (true/false)	Check list of fields to be characterized *Measurement*: 100% of fields have a value	SMEG, DEG, LEG
G3.4_Q05_M01	Accuracy of federated model >= accuracy of local (trained with single plant's data) model	Boolean (true/false)	Test ML model	SMEG, DEG, LEG
G3.4_Q06_M01	Difference of two output calculated on same input in different sessions	Double	Comparison of output *Measurement:* zero or not relevant differences detected	SMEG, DEG, LEG
G3.4_Q06_M02	Ratio of data errors in pre-processing, normalization and data alignment	Percentage	Tests in pre-processing, normalization and data alignment	SMEG, DEG, LEG
G3.4_Q06_M01	Provide a cause-effect relationship	Two-dimensional array whose values are boolean	Face to face interview *Measurements*: Presence of "true" values in two-dimensional array	SMEG, DEG, LEG

5.1.2 Health use case

Health data is a very special type of personal data that encompasses an extreme value for the person itself, considering its own health and wellbeing, and for the healthcare practitioners who should decide on the correct diagnosis and care pathways to achieve the best patient outcomes. Health data is also extremely important to the research, development and validation of new technologies, procedures and care pathways to improve the diagnosis, prognosis and treatment of diseases.

The recent years have shown important advances in Artificial Intelligence, enabled by Cloud Computing and big-data collections, with application in many different fields, and also with strong promises in the health care sector. One key element for improving AI algorithms and its results is gathering large amounts of good-quality data. In the health care sector, mainly for security and privacy reasons, but also due to some lack of interoperability and standardisation, it has been difficult to concentrate large amounts of quality data for the development of AI methodologies. Biobanks are vital source of information for fundamental and translational biomedical research aimed at the development of better predictive, preventive, personalised and participatory health care.

Multi-tenant and multi-data centre cloud solutions for medical imaging management, analysis and reporting, have been used in clinical practice for radiology and tele-radiology for a few years. They have been used by public hospitals to organise networked, collaborative reporting services, and by private practices to improve the productivity on large distributed groups and on small clinics. Vast amounts of medical imaging data are collected and reported using these cloud solutions, but each organisation accesses only its own data. Thanks to MUSKETEER this limitation will be surpassed.

The pressure for productivity is increasing due to the lack of Radiologists and the growing demand for medical imaging services. Key driving factors are the rise in prevalence of chronic diseases, technological advancements in diagnostic imaging modalities, increasing number of imaging procedures, rising awareness among the patients about early diagnosis of clinical disorders and rise in base of aging population. In addition, increasing demand from emerging countries, improved government funding towards chronic disorders, increasing investment in public and private organizations, and increasing disposable income among the population will further expected to drive the market in the coming years.

Radiology is moving toward a future in which radiologists, guided by artificial intelligence, will be able to work more closely with clinicians to provide precise therapies that offer patients an improved quality of life (according to a series of speakers at the opening press conference of the European Congress of Radiology ECR2019).

This use case intends to demonstrate the application of the Artificial Intelligence methodologies and technologies developed, enabling access to vast amounts of distributed medical imaging data to train and improve the learning algorithms, providing powerful tools to improve clinical practice. Being such a vast area, with several imaging modalities applying to different human body parts to analyse distinct conditions, we shall restrict the demonstration to one specific type of study.

The main objective will be the training of AI algorithms for support the detection of prostate cancer. B3D and HYGEIA will take a huge advantage of MUSKETEER developments to demonstrate the application of the Artificial Intelligence methodologies and technologies enabling access to vast amounts of distributed medical imaging data to train and improve the learning algorithms, providing powerful tools to improve clinical practice.

The aim is training AI algorithms for support the detection of prostate cancer. Since it is really hard to collect medical records, the benefit to collaborate sharing datasets to improve the predictive models to aid in the medical diagnosis is clear. The main barriers to be avoided with MUSKETEER are data localization, information leakage, standardisation and adversarial attacks. This project can solve these barriers.

The expected impacts of this use case are:

- improve accuracy of AI algorithms by sharing knowledge from distinct organisations and data repositories, supporting cooperation keeping security and privacy of health data;
- more accurate clinical decision support tools for diagnosis and prognosis of diseases, avoiding invasive procedures and conducting to better patient outcomes;
- faster decision support tools, enabling shorter turn-around-times, increasing productivity of services and more studies and patients diagnosed;
- faster and more accurate clinical decision support tools for diagnosis and prognosis saving lives in emergency cases;
- to enable the growth of the level of research in medical imaging AI tools supported by distributed data repositories;
- to enable clinical practices to access medical imaging AI tools with gains of productivity and better patient outcomes;
- to improve B3D commercial offer, enabling partners to access its market.

B3D provides a cloud-based Medical Imaging platform, 3Dnet™, for storage, retrieval, conditioning, fusion, analysis, presentation, interaction and reporting studies across medical imaging industry. 3Dnet Medical cloud collects and manages vast amounts of data for distributed

services of Radiology and Teleradiology, using DICOM and HL7 standards, into its secure cloud infrastructure, providing advanced visualization techniques to the Radiologists, allowing to make the right decisions and report their diagnosis, anytime and anywhere.

Hygeia is a reference organization for health care services in Greece, being the first hospital throughout Europe to carry out implantation of radioactive particles in prostate cancer. Hygeia has its own datacentre and uses 3Dnet Medical software to manage the medical im-aging data.

In terms of input data for system training reasons, Hygeia will draw all pelvis MRI exams as well as multi-parametric MRI exams for male patients. For each exam, an assessment sheet (in pdf format) is attached where lesions (regions of abnormality) are shown in images with PI-RADS score for each lesion (grading from I to V). Available histopathology reports (written as text in Greek language) from these exams will also be gathered.

All the above data is delivered as input to the software with main objective the training of AI algorithms and the potential to get precise details as an output on lesions localization and corresponding PI-RADS score.

The PI-RADS is described in the document Prostate Imaging Reporting & Data System - PI-RADS 2015 Version 2, of the American College of Radiology (ACR). A graphic presentation of the scoring system is provided by the Radiology Assistant site (Figure 6).

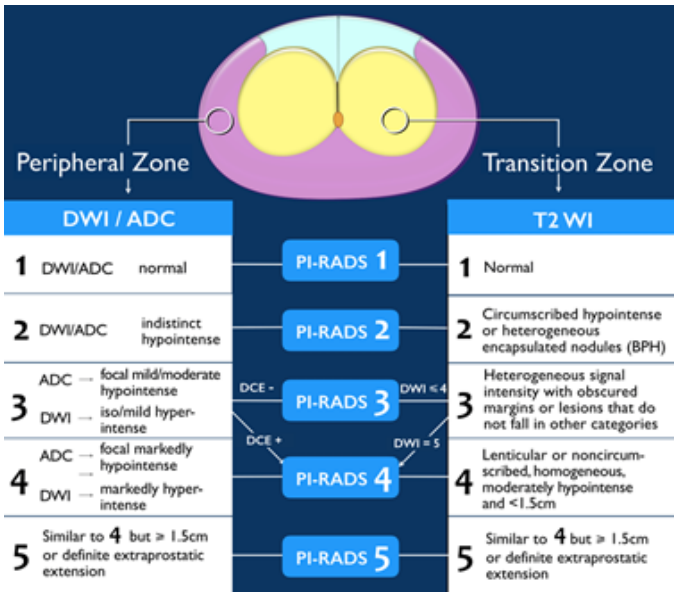

Figure 6 PI-RADS assessment (from radiologyassistant.nl)

The PI-RADS scoring system has the following grades:

- PI-RADS 1: Very low (clinically significant cancer highly unlikely)
- PI-RADS 2: Low (clinically significant cancer unlikely)

- PI-RADS 3: Intermediate (clinically significant cancer equivocal)
- PI-RADS 4: High (clinically significant cancer likely)
- PI-RADS 5: Very high (clinically significant cancer highly likely)

The image segmentation should follow the sector map used in the PI-RADS version 2 which employs 39 sectors (12 in the base, 12 in the midportion, 12 in the apex of the prostate, 2 seminal vesicles and 1 urethral sphincter).

With regard to privacy concerns, the main privacy problem in the health care scenario is the security and privacy of personal data. Machine learning algorithms can process health records to create predictive models capable to help in the medical diagnosis, these types of datasets are very valuable for research purposes.

For a single hospital it is very complicate to collect a dataset large enough to create a complex predictive model. For that reason, the benefit, in terms of predictive model accuracy, of combining datasets of different hospitals is very clear. However, having explicit consent of a patient to use his/her health records does not guarantee the protection of security and privacy and when two or more different research groups have explicit permission to use a health dataset, different barriers arise.

Data Localization barriers: To create a predictive model currently is necessary to place the dataset in a single place (same local computer or in the same cloud computing cluster). However, data localisation among different countries stems from legal rules that dictate the localisation of data for its storage or processing. Such requirements restrict the free flow of data between regions or countries.

Information Leakage barrier: Even signing a non-disclosure agreement, digital information can be easily copied and redistributed. Giving directly access to other's datasets open a door to personal data robbery. This entails severe fines for hospitals and personal damage if personal information is revealed.

Standardisation Barrier: When different hospitals create a dataset (e.g. medical images) using different medical devices (of different device manufacturers or with different calibration), then every hospital can use completely different measuring units and data standardization takes special importance.

Data Untrustworthiness Barrier: Groups may distrust the other's datasets since some partners can make a data poisoning attack in order to slow down the research of other group in a specific field.

MUSKETEER allows machine learning over datasets allocated in different locations (thus removing the data localization barrier) where the privacy preserving analytics remove any chance of information leakage and with mechanisms to provide standardisation among

different partners (based on IDSA concepts and Reference Architecture Model). In addition, the adversarial attack detection and mitigation strategies will be capable to detect data poisoning attacks and alert the other hospitals.

5.1.2.1 **Metrics**

Identifier	KPI	Format	Method of collection and measurement	Evaluation Groups
G1.2_Q01_M01	(Time taken to deploy and install the MUSKETEER client) * (Number of employees involved to deploy and install the MUSKETEER client)	(HH:MM) * number of employees	Online questionnaire; face-to-face interview	HEG
G1.2_Q01_M02	(Time taken to update the MUSKETEER client) * (Number of employees involved update the MUSKETEER client)	(HH:MM) * number of employees	Online questionnaire; face-to-face interview	HEG
G1.2_Q02_M01	Time taken by one person to create a task	HH:MM	Online questionnaire; face-to-face interview	HEG
G1.2_Q02_M02	Time taken to run the training procedure associated to a given ML task	HH:MM	Online questionnaire; face-to-face interview	HEG
G1.2_Q02_M03	Time taken to select and use a trained ML model	HH:MM	Online questionnaire; face-to-face interview	HEG
G1.2_Q02_M04	Time required for training a new user	HH:MM	Online questionnaire; face-to-face interview	HEG
G1.2_Q02_M05	Number of screens supported by help option / Total number of screens	Percentage	Verification of platform	HEG
G1.2_Q03_M01	Cost of local special equipment	Number (EUR)	Online questionnaire; face-to-face interview	HEG
G1.2_Q03_M02	Cost of setting up local special equipment	Number (EUR)	Online questionnaire; face-to-face interview	HEG
G1.2_Q04_M01	Integration profile conformance statements	Number	Verification of platform/	HEG

			technical documentation	
G2.2_Q01_M01	Time taken by one person to verify and declare privacy requirements	HH:MM	Online questionnaire; fate-to-face interview	HEG, LEG
G2.2_Q01_M02	Number of options expressed in natural language/ total number of steps	Decimal	Online questionnaire; fate-to-face interview	HEG, LEG
G2.2_Q02_M01	Adequate level of data privacy when sharing data with other end users	Boolean (true/false)	Online questionnaire; fate-to-face interview	HEG, LEG
G2.2_Q02_M02	Adequate level of data privacy when sharing models with other end users	Boolean (true/false)	Online questionnaire; fate-to-face interview	HEG, LEG
G2.2_Q02_M03	Adequate level of data privacy when training models	Boolean (true/false)	Online questionnaire; fate-to-face interview	HEG, LEG
G2.2_Q03_M01	Implementation of GDPR requirements	Boolean (true/false)	Platform/ technical documentation verification	HEG, LEG
G2.2_Q04_M01	Implementation of MDR, EN ISO 13485:2016 and EN ISO 14971:2012 security and privacy requirements	Boolean (true/false)	Platform/ technical documentation verification	HEG, LEG
G2.2_Q04_M01	Implementation of EN ISO/IEC 27001:2017 requirements	Boolean (true/false)	Platform/ technical documentation verification	HEG, LEG
G2.2_Q04_M02	Number of security controls implemented and verified	Number	Platform/ technical documentation verification	HEG, LEG
G3.3_Q01_M01	Sensitivity (true positive rate) of prediction of existence of cancer. Formula: Sensitivity = (True Positive) / (True Positive + False Negative)	percentage	Test ML models	HEG, LEG, DEG

G3.3_Q01_M02	Specificity (true negative rate) of prediction of existence of cancer. Formula: Specificity = (True Negative) / (True Negative + False Positive)	percentage	Test ML models	HEG, LEG, DEG
G3.3_Q01_M03	Accuracy of classification of studies with PIRADS score for each Patient	percentage	Test ML models	HEG, LEG, DEG
G3.3_Q02_M01	Accuracy of lesion's segmentation	Percentage	Test ML models	HEG, LEG, DEG
G3.3_Q02_M02	Accuracy of lesion's PI-RADS score classification	Percentage	Test ML models	HEG, LEG, DEG
G3.3_Q03_M01	Ratio of data errors in pre-processing, normalization, data alignment	Percentage	Tests in pre-processing, normalization, data alignment	HEG, LEG, DEG
G3.3_Q03_M02	Ratio of empty data in pre-processing, normalization, data alignment	Percentage	Tests in pre-processing, normalization, data alignment	HEG, LEG, DEG
G3.3_Q03_M03	Accuracy of ML trained model with dataset tested with external datasets	Percentage	Test ML models	HEG, LEG, DEG

5.2 Technical Perspective

As already explained, from technical point of view, some aspects of the MUSKETEER architecture, Privacy preserving federated machine learning algorithms under different POMs, Federated Machine Learning Algorithms will be evaluated in WP7 according to the MUSKETEER Evaluation framework presented in this document. Moreover, the data quality, scalability, computational efficiency and security of the MUSKETEER platform will be considered within WP6. In order to evaluate the platform from the technical perspective, the second version of this document (due to M24) will be properly enriched with metrics resulting from the deliverable D6.1.

The metrics to be evaluated from technical perspective in WP7 are listed below.

Identifier	KPI	Format	Method of collection and measurement	Evaluation Groups
G1.1_Q01_M01	Number of artifacts compliant with IDSA reference architecture / Total number of artifacts	Decimal <=1	Online questionnaire; fate-to-face interview (envisaged the involvement of IDSA in accordance with the DoW)	DEG
G1.1_Q02_M01	Number of ML libraries supported to export the predictive models /Total of the best-known ML libraries	Decimal <=1	Online questionnaire; fate-to-face interview (the best-known ML libraries will be detailed more in the second version of this document)	DEG
G1.1_Q03_M01	Open source web communities' interactions	Integer	Field survey	DEG
G1.1_Q0.4_M01	Number of SW applications released as 66images	Integer	Field survey	DEG
G1.1_Q0.4_M02	Number of software components released in open source repositories	Integer	Field survey	DEG
G2.1_Q01_M01	Number of robust POMs for use cases	Integer	On field	DEG, LEG
G2.1_Q02_M01	Number of robust POMs for use cases	Integer	On field	DEG, LEG
G2.1_Q02_M02	Speedup/number of users while POM is applied	Ratio	On field	DEG, LEG
G2.1_Q03_M01	Number of training procedures implemented	Integer	On field	DEG, LEG
G2.1_Q03_M02	Speed of privacy-preserving machine learning algorithms implemented with respect to other existing solutions	Ratio	On field	DEG, LEG

6 Conclusion

This deliverable D2.3 - Key performance indicators selection and definition, presented the MUSKETEER evaluation framework and overall evaluation approach that will be implemented in WP7 according to the different MUSKETEER use cases.

Based on the Goal Question Metric method (GQM), this document describes the first version of KPIs and it is paving the way to the use case implementations towards final evaluation execution by the end of the project. Key technical quality focuses and business priorities are identified in order to prepare the deployment of detailed appropriated evaluation questions and metrics.

This document defines the planning and the definition phases, which have been achieved at this stage of the project. A revision of the KPIs and methodology will be done in M24.

Starting from the final version of the D2.3 released in M24, Data Collection and interpretation phases will be documented in the deliverables D7.5 and D7.6, where the description of the Smart Manufacturing and Health pilots setup and execution, together with the evaluation of the KPIs, will be reported in order to assess the usage of the MUSKETEER platform.

It is worth noticing that, to be effective, metrics should be compared to established benchmarks or objectives. This provides valuable context for the values used in the metric and allows users to better act on the information they have to handle.

Thus, in the final version of this deliverable (to be released at M24), benchmarks will be provided for each metric identified. They will allow the comparison against actual values which will be measured when the MUSKETEER industrial data platform is applied to improve the quality of data providers' machine learning models (WP7).

The results of the overall evaluation of the MUSKETEER Federated Machine Learning platform and of the demonstrators will be carried out and documented in D7.3 and D7.4, focusing therefore on the correct adoption of the platform including benchmarking and assessment based on the requirements. The KPI checklist is used to validate the MUSKETEER platform. In addition, the KPI evaluation enables to identify the proper adjustments in order to align the results to the original requirements (corresponding to data collection and interpretation phases of the MUSKETEER evaluation framework described in Section 3).

7 References

[1] R. van Solingen, E. Berghout (1999). The Goal/Question/Metric Method: A Practical Guide for Quality Improvement of Software Development, McGraw-Hill
[2] V. Basili, G. Caldiera, H.D. Rombach (1994). Goal Question Metric Paradigm, Encyclopaedia of Software Engineering – 2 Volume Set.
[3] MUSKETEER Technical Annex
[4] M. Kasunic (2008). A Data Specification for Software Project Performance Measures: Results of a Collaboration on Performance Measurement.
[5] Reference Architecture Model for the Industrial Data Space, Fraunhofer-Gesellschaft, Munich, 2017