DELIVERABLE

Project Acronym: **VALIDATE**

Grant Agreement number: **101057263**
Project Title: **Validation of a Trustworthy AI-based Clinical Decision Support System for Improving Patient Outcome in Acute Stroke Treatment**

D1.1 – VALIDATE trustworthy AI framework and manual at project month 9

Revision: 1.0

Authors and Contributors	Vince I. Madai (CUB); Cathrine Bui (CUB), staff from all partners as part of WP1 activities in requirement definition			
Responsible Author	Vince I. Madai		**Email**	vince_istvan.madai@bih-charite.de
	Beneficiary	CUB	**Phone**	+49 178 725 31 05

Project co-funded by the European Commission within **HORIZON-HLTH-2021-DISEASE-04-04**		
Dissemination Level		
PU	Public, fully open	xxx
CO	Confidential, restricted under conditions set out in Model Grant Agreement	
CI	Classified, information as referred to in Commission Decision 2001/844/EC	

This project has received funding from the European Union's Horizon Europe research and innovation program under grant agreement No 101057263

Revision History, Status, Abstract, Keywords, Statement of Originality

Revision History

Revision	Date	Author	Organisation	Description
0.1	09.01.23	Vince Madai	CUB	Main structure, introduction
0.2	13.01.23	Vince Madai, Cathrine Bui	CUB	Planguage, requirements
0.3	24.01.23	Vince Madai, Cathrine Bui	CUB	requirements
0.4	30.01.23	Vince Madai	CUB	requirements
1.0	31.01.23	Vince Madai	CUB	Requirements, finalization of ethical framework and manual

Date of delivery	Contractual:	31.01.2023	Actual:	31.01.2023
Status	final ☒ /draft ☐			

Abstract (for dissemination)	This document presents the snapshot of the VALIDATE Trustworthy AI framework and manual at month 9.
Keywords	Trustworthy AI; ethics of AI; VALIDATE

Statement of originality

This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation, or both.

Table of Content

Executive Summary

AI-based prognostic tools and subsequent clinical decision support will only be approved and accepted in the clinical setting if they meet all the necessary criteria in terms of regulation, legality, ethics and robustness. Within VALIDATE, we will follow the EU guidelines for trustworthy AI. This approach ensures, for example, that (systematic) automated and manual biases are taken into account, that patient subgroups are not discriminated against, and that traceability, transparency, and safety are ensured during development, testing and validation.

However, a major challenge is that most existing frameworks - and the EU guidelines are no exception - focus on 'what' to do, but do not provide guidance on 'how' to do it in practice. As a result, to date, healthcare researchers struggle to operationalise high-level ethical principles.

VALIDATE therefore offers a methodological ethical toolbox that spans the entire project duration and includes reviews, workshops, interview studies, a z-inspection analysis and audits to provide a holistic approach to defining low-level requirements, which in turn can be mapped to the EU's high-level AI guidelines. This toolbox is embedded in a continuous interdisciplinary co-creation approach leading to a development framework for data-driven clinical AI systems.

Importantly, our framework is a living document as it reflects a co-creation approach. We are constantly revising the framework based on new findings, reviews and interdisciplinary considerations.

The project provides reviewers with an update of the ethical framework at three points in the study, reflecting the status quo of the framework at those points, namely at 9, 18 and 36 months. The document presented here is the current snapshot at month 9.

The manual, which is also part of the deliverables, is not a living document, unlike the framework. It is a shorter form of the living framework that reflects the status at the time of the deliverables and focuses solely on an overview of all requirements. The manual serves as a quick guide for daily work and as a reference book, as not every staff member can be required to constantly follow the changes and discussions in the framework. However, major changes, such as amended requirements, are included in the manual constantly.

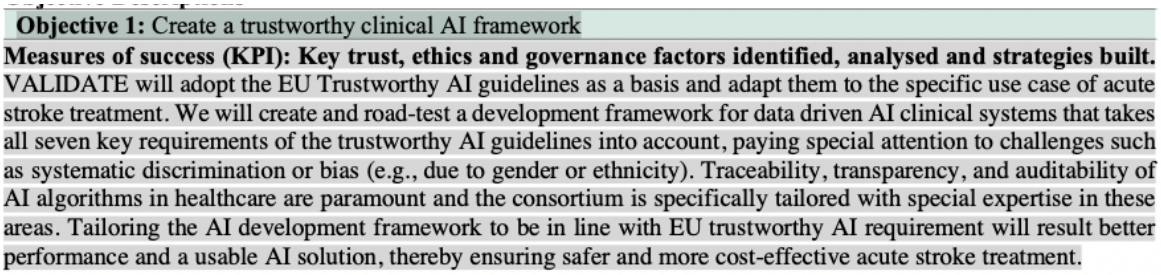
Using our ethical VALIDATE toolbox and aligning the AI development framework with the EU requirements for trustworthy AI will lead to a more powerful and usable VALIDATE AI solution, ensuring safer and more cost-effective treatment of acute stroke.

1 Introduction

Based on pre-clinical evidence and previously developed models, and an available prototype of a clinical decision support system, VALIDATE set out in this project to further develop, test, and validate an artificial intelligence (AI)-based prognostic tool for outcome prediction of acute stroke patients.

However, AI-based prognostic tools and subsequent clinical decision support will only be approved and accepted in the clinical setting if they fulfil all necessary criteria with regards to regulatory, lawfulness, ethics, and robustness.

Thus, in VALIDATE, we follow the **Trustworthy AI** guidelines of the EU's High-Level Expert Group on Artificial Intelligence (AI HLEG) as our ethical framework(figure 1).


Objective 1: Create a trustworthy clinical AI framework

Measures of success (KPI): Key trust, ethics and governance factors identified, analysed and strategies built. VALIDATE will adopt the EU Trustworthy AI guidelines as a basis and adapt them to the specific use case of acute stroke treatment. We will create and road-test a development framework for data driven AI clinical systems that takes all seven key requirements of the trustworthy AI guidelines into account, paying special attention to challenges such as systematic discrimination or bias (e.g., due to gender or ethnicity). Traceability, transparency, and auditability of AI algorithms in healthcare are paramount and the consortium is specifically tailored with special expertise in these areas. Tailoring the AI development framework to be in line with EU trustworthy AI requirement will result better performance and a usable AI solution, thereby ensuring safer and more cost-effective acute stroke treatment.

Figure 1. Objective 1: Create a trustworthy clinical AI framework

These guidelines aim to promote the development and deployment of AI systems that are safe, reliable, and respect fundamental rights and values. The guidelines include seven key principles for trustworthy AI that count as high-level norms and several sub-groups under these heading counting as mid-level norms:

1. Human agency and oversight
    1. Fundamental rights
    2. Human agency
    3. Human oversight
2. Technical robustness and safety
    1. Resilience to attack and security
    2. Fallback plan and general safety
    3. Accuracy
    4. Reliability and Reproducibility
3. Privacy and Data governance
    1. Privacy and data protection
    2. Quality and integrity of data
    3. Access to data
4. Transparency

1. Traceability
2. Explainability
3. Communication

5. Diversity, non-discrimination and fairness
    1. Avoidance of unfair bias
    2. Accessibility and universal design
    3. Stakeholder participation
6. Societal and environmental well-being
    1. Sustainable and environmentally friendly
    2. Social impact
    3. Society and Democracy
7. Accountability
    1. Auditability
    2. Minimisation and reporting of negative impacts
    3. Trade-offs
    4. Redress

Additionally, the guidelines provide recommendations for the development and deployment of AI, including the need for a governance framework and the promotion of responsible research and innovation. Overall, the EU's AI HLEG guidelines aim to ensure that AI is developed and used in a way that benefits society and respects fundamental rights and values.

However, high-level and mid-level guidelines such as the EU-guidelines above are not sufficient. As a reply to the challenges that researcher and developers are facing, not only the EU but also others have introduce trustworthy AI guidelines. There is currently a phenomenon that can only be called an inflation of ethical and trustworthy AI guidelines. Summaries of current frameworks reported numbers between 80 and almost 300 sources[1] [2] [3]. They have been written by standardisation bodies, academia and industry, (supra)national bodies and government organisations[4].

It was, however, pointed out that despite this number of guidelines on how to perform AI research ethically, there is no shortage of reports of unethical use of AI[5]. The main reason for this is that the current guidelines are very abstract and of limited practical applicability to researchers and developers of algorithms and AI systems. They are simple principle-based guidelines that are similar to *principlism*, the framework that Beauchamp and Childress developed for bioethics. A review found that 75% of major ethical guidelines only provide high-level principles with very little detail, and over 80% offer no

[1] Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S. & Floridi, L. The ethics of algorithms: Mapping the debate. *Big Data Soc.* **3**, 2053951716679679 (2016).

[2] Schönberger, D. Artificial intelligence in healthcare: a critical analysis of the legal and ethical implications. *Int. J. Law Inf. Technol.* **27**, 171–203 (2019).

[3] Jobin, A., Ienca, M. & Vayena, E. The global landscape of AI ethics guidelines. *Nat. Mach. Intell.* **1**, 389–399 (2019).

[4] Schiff, D., Biddle, J., Borenstein, J. & Laas, K. What's Next for AI Ethics, Policy, and Governance? A Global Overview. in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* 153–158 (Association for Computing Machinery, 2020).

[5] Morley, J. *et al.* Operationalising AI ethics: barriers, enablers and next steps. *AI Soc.* (2021) doi:10.1007/s00146-021-01308-8.

or the lowest level of practical insights[6]. It was also aptly observed that principles alone cannot guarantee ethical AI and that current high-level frameworks fail to address fundamental normative tensions embedded in key concepts such as fairness or privacy[7].

The majority of existing frameworks focus on "*what*" to do but do not give any guidance on "*how*" to do it in practice[8]. And those frameworks that attempt to give an answer to the "*how*" severely lack usability, i.e. they are not actionable offering very limited support on how to use them in practise.

Thus, a major challenge - that we also face in VALIDATE - is that to date healthcare researchers have little guidance that operationalizes how high-level ethical principles can be implemented in the practice of researching and using AI in healthcare.

Thus, VALIDATE provides a methodological ethical toolbox that spans the whole project time including reviews, workshops, interview studies, a z-inspection analysis, and audits to provide a holistic approach to operationalise low-level requirements that in turn can be mapped to the high-level AI guidelines of the EU.

Importantly, our framework is a living document, as it reflects a co-creation approach. We do and will constantly revise the framework based on new results, audits, task results and deliverables, and interdisciplinary reflection.

The project reviewers are presented at three time-points in the study an update about the ethical framework which reflects the status quo of the framework at those time points, namely at 9, 18, and 36 months. **This deliverable is the current snapshot at month 9.**

The manual, also part of the deliverable is, in contrast to the framework, not a living document. It is a shorter form of the living framework that reflects the status at the time points of the deliverables and focuses solely on an overview of all requirements. The manual serves as a quick guidance document for everyday work and as a reference point, as not every staff member can be required to constantly follow the changes and discussions in the framework. Major changes such as changed requirements will be reflected in the manual, however.

[6] Bélisle-Pipon, J.-C., Monteferrante, E., Roy, M.-C. & Couture, V. Artificial intelligence ethics has a black box problem. *AI Soc.* (2022) doi:10.1007/s00146-021-01380-0.

[7] Mittelstadt, B. Principles alone cannot guarantee ethical AI. *Nat. Mach. Intell.* **1**, 501–507 (2019).

[8] Morley, J., Floridi, L., Kinsey, L. & Elhalal, A. From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices. *Sci. Eng. Ethics* **26**, 2141–2168 (2020).

2 Ethical AI framework

Our ethical framework follows the structure of the the **Trustworthy AI** guidelines of the EU's High-Level Expert Group on Artificial Intelligence (AI HLEG).

These guidelines define a structure of high- and mid-level norms. High-level norms are equivalent to high-level principles, for example “Transparency”. These principles are then further defined by mid-level norms, for example “Explainability”.

Our framework extends these norms by defining low-level, practical requirements which are mapped to the high- and mid-level norms of the AI HLEG.

Our low-level requirements are defined and quantified in a certain format that is inspired by Planguage – a short form of planning language.

The use of methods for quantifying requirements is important in the field of AI ethics because what is not measured is more difficult to improve. Quantified metrics for whether a goal of auditability or trustworthiness has been met will increase transparency and accountability. When defining vague high- level principles into something quantified and measurable, it can bring to light the unsaid issues that need to be discussed. This will bring forth a concrete definition that will provide a shared language for future discussions and to a clear path of action.

2.1 Planguage

A Planning Language or its shortened term, *Planguage*, is an effective interactive tool to quantify a result from a process or work [9]. Planguage helps in defining the expected result to customers or several stakeholders by translating the text and words to measures and numbers. Tse and Kahlon described how Planguage enhances the interrelationship qualities between people, processes, and technology such as AI[10].

Planguage tools can be applied to clarify the requirements analysis of digital health project in many different fields, such as managing projects and enforcing the quality of the healthcare system[11]. It was also stated that the main problem of any IT-related projects or literature that is applied in fields such as healthcare, is the difficulties that may face the stakeholders such as health workers to understand the analysis of these IT systems or the expected results. Also, they described how Planguage was an effective tool that led a healthcare IT project to be successfully delivered and how this method enabled the project to address stakeholder values and viewpoints more visually and explicitly.

A sample of how a measurable requirement might look like in the Planguage format is the following:

% of the **[Information Type]** are available through **[Release Type]** every **[Time Interval]** to relevant **[Stakeholders]** so that they can **[Action]**.

Where the parameters are defined as:

[9] Gilb, T. (1989). A planning language (a PLanguage). *Conference Proceedings on APL as a Tool of Thought*, 169–177. https://doi.org/10.1145/75144.75168

[10] Tse, M.-C., & Kahlon, R. S. (2013). *How Planguage Measurement Metrics: Shapes System Quality*. https://www.proquest.com/openview/1145e0a0101b89a123a8c975afb64c6a/1?pq- origsite=gscholar&cbl=396494

[11] Ibid.

[Information Type]: Assumptions of the AI-tool, Limitations of the AI-tool, Operating protocols, Data, Meta-Data, Methods used for data collection, Methods used for data processing, Methods used for data labelling, Methods for developing the AI algorithm.

[Release Type]: Publication, Internal Documentation, FAQs, Translated FAQs. **[Time interval]:** Year.

[Stakeholders]: System Evaluators, Government Regulators, Auditors, Marginalized Populations, Doctors, Patients.

[Action]: Identify Errors, Conduct Effective Oversight, Understand How to Use the AI Tool. This would mean that there would be different sentence versions of that requirement when progress has been made, for example:

10% of the **operating protocols** is available through **internal documentation** every **year** to relevant **system evaluators** so that they can **identify errors**.

70% of the **limitations of the AI-tool** are available through **FAQs** every **year** to relevant **doctors** so that they can **understand how to use the AI tool**.

Note, that the definitions of these measures and numbers are somewhat subjective, they are opinion-based ideas about how something should be measured and when this measure is fulfilled. For instance, the number of "information types that should be available" is defined by the stakeholder group, it is not based on objective datapoints from sources such as wide-spread surveys.

Based on the parameterized requirement statements, we then formulate versions of the requirement that are categorized in three categories:

TOLERABLE
A minimum requirement that needs to be achieved

GOAL

An objective that the project aims to reach. If both a Tolerable and a Goal are defined, the likely project outcome will be between the two categories.

WISH

A "nice to have" outcome that we will aim to achieve but will likely be only partially achieved.

In the following the framework is ordered according to the high- and mid-level norms of the AI HLEG publication. Descriptions of these norms are copies from the URL, where the norms are described.

2.2 Human Agency and Oversight

AI systems should support human autonomy and decision-making, as prescribed by the principle of respect for human autonomy. This requires that AI systems should both act as enablers to a democratic, flourishing and equitable society by supporting the user's agency and foster fundamental rights, and allow for human oversight.

2.2.1 Fundamental rights

Like many technologies, AI systems can equally enable and hamper fundamental rights. They can benefit people for instance by helping them track their personal data, or by increasing the accessibility of education, hence supporting their right to education. However, given the reach and capacity of AI systems, they can also negatively affect fundamental rights. In situations where such risks exist, a fundamental rights impact assessment should be undertaken. This should be done prior to the system's development and include an evaluation of whether those risks can be reduced or justified as necessary in a democratic society in order to respect the rights and freedoms of others. Moreover, mechanisms should be put into place to receive external feedback regarding AI systems that potentially infringe on fundamental rights.

None of the identified requirements could be mapped to this category. We also believe that our tool does not influence any fundamental rights.

2.2.2 Human Agency

Users should be able to make informed autonomous decisions regarding AI systems. They should be given the knowledge and tools to comprehend and interact with AI systems to a satisfactory degree and, where possible, be enabled to reasonably self-assess or challenge the system. AI systems should support individuals in making better, more informed choices in accordance with their goals. AI systems can sometimes be deployed to shape and influence human behaviour through mechanisms that may be difficult to detect, since they may harness sub-conscious processes, including various forms of unfair manipulation, deception, herding and conditioning, all of which may threaten individual autonomy. The overall principle of user autonomy must be central to the system's functionality. Key to this is the right not to be subject to a decision based solely on automated processing when this produces legal effects on users or similarly significantly affects them.

Req. ID	Requirement	Parameters		Phases
HUO-01	**[X number]** of **[relevant methods]** for measuring model uncertainty are used to provide a confidence interval for predictions in the app available **[time point]**	**[X number]**	To be defined	development; testing;
		[relevant methods]	To be defined	
		[time point]	At start of the prospective study	

Tolerable

X number:To be defined of **relevant methods:To be defined** for measuring model uncertainty are used to provide a confidence interval for predictions in the app available **at start of the prospective study**.

2.2.3 Human Oversight

Human oversight helps ensuring that an AI system does not undermine human autonomy or causes other adverse effects. Oversight may be achieved through governance mechanisms such as a human-

in-the-loop (HITL), human-on-the-loop (HOTL), or human-in-command (HIC) approach. HITL refers to the capability for human intervention in every decision cycle of the system, which in many cases is neither possible nor desirable. HOTL refers to the capability for human intervention during the design cycle of the system and monitoring the system's operation. HIC refers to the capability to oversee the overall activity of the AI system (including its broader economic, societal, legal and ethical impact) and the ability to decide when and how to use the system in any particular situation. This can include the decision not to use an AI system in a particular situation, to establish levels of human discretion during the use of the system, or to ensure the ability to override a decision made by a system. Moreover, it must be ensured that public enforcers have the ability to exercise oversight in line with their mandate. Oversight mechanisms can be required in varying degrees to support other safety and control measures, depending on the AI system's application area and potential risk. All other things being equal, the less oversight a human can exercise over an AI system, the more extensive testing and stricter governance is required.

Req. ID	Requirement	Parameters		Phases
HUO-02	Human in command (HIC) governance concept available **[time point]**	**[time point]**	At the end of the project	development; testing; validation

Tolerable

Human in command (HIC) governance concept available **at the end of the project.**

2.3 Technical Robustness and Safety

A crucial component of achieving Trustworthy AI is technical robustness, which is closely linked to the principle of prevention of harm. Technical robustness requires that AI systems be developed with a preventative approach to risks and in a manner such that they reliably behave as intended while minimising unintentional and unexpected harm, and preventing unacceptable harm. This should also apply to potential changes in their operating environment or the presence of other agents (human and artificial) that may interact with the system in an adversarial manner. In addition, the physical and mental integrity of humans should be ensured.

2.3.1 Resilience to Attack and Safety

AI systems, like all software systems, should be protected against vulnerabilities that can allow them to be exploited by adversaries, e.g. hacking. Attacks may target the data (data poisoning), the model (model leakage) or the underlying infrastructure, both software and hardware. If an AI system is attacked, e.g. in adversarial attacks, the data as well as system behaviour can be changed, leading the system to make different decisions, or causing it to shut down altogether. Systems and data can also become corrupted by malicious intention or by exposure to unexpected situations. Insufficient security processes can also result in erroneous decisions or even physical harm. For AI systems to be considered secure, possible unintended applications of the AI system (e.g. dual-use applications) and potential abuse of the system by malicious actors should be taken into account, and steps should be taken to prevent and mitigate these.

Req. ID	Requirement	Parameters		Phases
TRA-02	Tool should follow **[relevant norms]** that are within our scope and resource limitations	**[relevant norms]**	ISO	development; testing;

	to prepare for compliance with the MDR **[time point]**.			validation
		[time point]	At the end of the project	

Wish

Tool should follow **ISO norms** that are within our scope and resource limitations to prepare for compliance with the MDR **at the end of the project**.

2.3.2 Fallback plan and general safety

AI systems should have safeguards that enable a fallback plan in case of problems. This can mean that AI systems switch from a statistical to rule-based procedure, or that they ask for a human operator before continuing their action. It must be ensured that the system will do what it is supposed to do without harming living beings or the environment. This includes the minimisation of unintended consequences and errors. In addition, processes to clarify and assess potential risks associated with the use of AI systems, across various application areas, should be established. The level of safety measures required depends on the magnitude of the risk posed by an AI system, which in turn depends on the system's capabilities. Where it can be foreseen that the development process or the system itself will pose particularly high risks, it is crucial for safety measures to be developed and tested proactively.

Req. ID	**Requirement**	**Parameters**		**Phases**
TRA-03	A thorough risk assessment following the MDR should be done to check and prove which risks are involved and whether they are low enough for this tool **[time point]**.	**[time point]**	When intended purpose is defined; at the end of the project	development; testing; validation

Goal

A thorough risk assessment following the MDR should be done to check and prove which risks are involved and whether they are low enough for this tool **at the end of the project**.

Wish

A thorough risk assessment following the MDR should be done to check and prove which risks are involved and whether they are low enough for this tool **at when intended purpose is defined**.

2.3.3 Accuracy

Accuracy pertains to an AI system's ability to make correct judgements, for example to correctly classify information into the proper categories, or its ability to make correct predictions, recommendations, or decisions based on data or models. An explicit and well-formed development and evaluation process can support, mitigate and correct unintended risks from inaccurate predictions. When occasional inaccurate predictions cannot be avoided, it is important that the system can indicate how likely these errors are. A high level of accuracy is especially crucial in situations where the AI system directly affects human lives.

Req. ID	Requirement	Parameters		Phases
VAL-01	Validity will be measured by doing a randomized clinical trial (RCT) to examine whether at least **[ratio]** of **[RCT study size] patients** whose doctors use the tool have a significantly improved MRS 90 compared to control group at **[time point].**	**[ratio]**	0.1%	Validation
		[RCT study size]	Number determined by estimated effect size	
		[time point]	After the project	

Wish

Validity will be measured by doing a randomized clinical trial (RCT) to examine whether at least **0.1%** of **number determined by estimated effect size patients** whose doctors use the tool have a significantly improved MRS 90 compared to control group at **after the project.**

Req. ID	Requirement	Parameters		Phases
VAL-02	% of **[RCT regulation requirements]** are addressed to prepare for an RCT **[time point]** that validates that MRS 90 patient outcome is significantly better when using the tool compared to current standards of care.	**[RCT regulation requirements]**	MDR; FDA; GDPR;(requirements)	development; testing; validation
		[time point]	After the project	

Tolerable

10% of MDR requirements are addressed for an to prepare for an RCT after the project that validates that MRS 90 patient outcome is significantly better when using the tool compared to current standards of care.

10% of GDPR requirements are addressed for an to prepare for an RCT after the project that validates that MRS 90 patient outcome is significantly better when using the tool compared to current standards of care.

0% of FDA requirements are addressed for an to prepare for an RCT after the project that validates that MRS 90 patient outcome is significantly better when using the tool compared to current standards of care.

Goal

20% of MDR requirements are addressed for an to prepare for an RCT after the project that validates that MRS 90 patient outcome is significantly better when using the tool compared to current standards of care.

20% of GDPR requirements are addressed for an to prepare for an RCT after the project that validates that MRS 90 patient outcome is significantly better when using the tool compared to current standards of care.

10% of FDA requirements are addressed for an to prepare for an RCT after the project that validates that MRS 90 patient outcome is significantly better when using the tool compared to current standards of care.

Wish

40% of MDR requirements are addressed for an to prepare for an RCT after the project that validates that MRS 90 patient outcome is significantly better when using the tool compared to current standards of care.

40% of GDPR requirements are addressed for an to prepare for an RCT after the project that validates that MRS 90 patient outcome is significantly better when using the tool compared to current standards of care.

20% of FDA requirements are addressed for an to prepare for an RCT after the project that validates that MRS 90 patient outcome is significantly better when using the tool compared to current standards of care.

2.3.4 Reliability and Reproducibility

It is critical that the results of AI systems are reproducible, as well as reliable. A reliable AI system is one that works properly with a range of inputs and in a range of situations. This is needed to scrutinise an AI system and to prevent unintended harms. Reproducibility describes whether an AI experiment exhibits the same behaviour when repeated under the same conditions. This enables scientists and policy makers to accurately describe what AI systems do. Replication files can facilitate the process of testing and reproducing behaviours.

Req. ID	Requirement	Parameters		Phases
TRA-04	The tool needs to define the intended purpose **[time point]**.	**[time point]**	As part of the demonstrator development	development; testing;

Tolerable

The tool needs to define the intended purpose **as part of the demonstrator development**.

Req. ID	Requirement	Parameters	Phases

ROB-02	% of publications authored by the consortium are open access and if relevant **[scientific methods]** to improve reproducibility.	**[scientific methods]**	(contain model cards; follow reporting guidelines; follow open code practices, provide open data access)	development; testing; validation

Tolerable

100% of publications authored by the consortium are open access and if relevant **(contain model cards; follow reporting guidelines; follow open code practices, provide open data access)** to improve reproducibility.

2.4 Privacy and data governance

Closely linked to the principle of prevention of harm is privacy, a fundamental right particularly affected by AI systems. Prevention of harm to privacy also necessitates adequate data governance that covers the quality and integrity of the data used, its relevance in light of the domain in which the AI systems will be deployed, its access protocols and the capability to process data in a manner that protects privacy.

2.4.1 Privacy and data protection

AI systems must guarantee privacy and data protection throughout a system's entire lifecycle. This includes the information initially provided by the user, as well as the information generated about the user over the course of their interaction with the system (e.g. outputs that the AI system generated for specific users or how users responded to particular recommendations). Digital records of human behaviour may allow AI systems to infer not only individuals' preferences, but also their sexual orientation, age, gender, religious or political views. To allow individuals to trust the data gathering process, it must be ensured that data collected about them will not be used to unlawfully or unfairly discriminate against them.

Req. ID	**Requirement**	**Parameters**		**Phases**
PRI-01	The AI tool complies with relevant **[privacy regulations]** **by [date]**	**[privacy regulations]**	GDPR, HIPAA, local authority privacy regulations, EU Horizon Europe grant requirements for privacy	development; testing; validation

		[date]	Start of prospective study	

Tolerables

The AI tool complies with GDPR by start of prospective study.

The AI tool complies with Horizon Europe grant requirements for privacy by start of prospective study.

The AI tool complies with local authority privacy regulations by start of prospective study.

Wish

The AI tool complies with HIPAA by start of prospective study.

Req. ID	Requirement	Parameters		Phases
PRI-02	All **[development materials]** are stored via **[practices for privacy], with practice enabled before [date]**	**[development materials]**	Research data, models, predictions, XAI results,	development; testing; validation
		[practices for privacy]	Data encryption, password protected user rights system, local protected servers, cloud environment,	
		[date]	Start of training; at 24 months into the project;	

Tolerables

Research data are stored via password protected user rights system, with practice enabled before start of training.

Research data are stored via local protected servers, with practice enabled before start of training.

Models are stored via password protected user rights system, with practice enabled before start of training.

Models are stored via local protected servers, with practice enabled before start of training.

Predictions are stored via password protected user rights system, with practice enabled before start of training.

Predictions are stored via local protected servers, with practice enabled before start of training.

xAI results are stored via password protected user rights system, with practice enabled before start of training.

xAI results are stored via local protected servers, with practice enabled before start of training.

Goal

Research data are stored via data encryption, with practice enabled at 24 months into the project.

Models are stored via data encryption, with practice enabled at 24 months into the project.

Predictions are stored via data encryption, with practice enabled at 24 months into the project.

xAI results are stored via data encryption, with practice enabled at 24 months into the project.

Wish

Research data are stored via cloud environment, with practice enabled at 24 months into the project.

Models are stored via cloud environment, with practice enabled at 24 months into the project.

Predictions are stored via cloud environment, with practice enabled at 24 months into the project.

xAI results are stored via cloud environment, with practice enabled at 24 months into the project.

Req. ID	Requirement	Parameters		Phases
PRI-04	Privacy information to answer "**[local ethics committee questions]**" to approve data collection for prospective study are needed **[time point]** in the research data management plan.	**[local ethics committee questions]**	(Why are we collecting this data? How will the data be collected? Where will data be stored? What data be collected? Will the data be shared?)	development
		[time point]	By project month 12	

Tolerable

Privacy information to answer "**Why are we collecting this data? How will the data be collected? Where will data be stored? What data be collected? Will the data be shared?** to approve data collection for prospective study are needed **by project month 12** in the research data management plan.

2.4.2 Quality and integrity of data

The quality of the data sets used is paramount to the performance of AI systems. When data is gathered, it may contain socially constructed biases, inaccuracies, errors and mistakes. This needs to be addressed prior to training with any given data set. In addition, the integrity of the data must be ensured. Feeding malicious data into an AI system may change its behaviour, particularly with self-learning systems. Processes and data sets used must be tested and documented at each step such as

planning, training, testing and deployment. This should also apply to AI systems that were not developed in-house but acquired elsewhere.

Req. ID	Requirement	Parameters	
ROB-01	% of datasets fulfill **[quality criteria]** with regards to **[quality parameters]** **[time point].**	**[quality criteria]**	To be defined
		[quality parameters]	(Missing data, errors, inaccuracies, interoperability)
		[time point]	Before final model training for demonstrator

Tolerable

100% of datasets fulfill **quality criteria:To be defined** with regards to **(Missing data, errors, inaccuracies, interoperability) before final model training for demonstrator.**

2.4.3 Access to data

In any given organisation that handles individuals' data (whether someone is a user of the system or not), data protocols governing data access should be put in place. These protocols should outline who can access data and under which circumstances. Only duly qualified personnel with the competence and need to access individual's data should be allowed to do so.

Req. ID	Requirement	Parameters		Phases
PRI-03	**[Data type]** is available for access to **[stakeholder type]** through **[process]** **[time point]**.	**[data type]**	Retrospective study data; anonymized prospective study data; data used to simulate a decision during prospective study	development; testing; validation
		[stakeholder type]	Relevant VALIDATE staff; researchers in academia: legal guardian/caregiver; patient; physician.	
		[process]	Process outlined in VALIDATE data management plan.	
		[time point].	After embargo period; at the end of the project as preparation for an RCT; for demonstrator development; during prospective study	

Tolerable

Retrospective study data is available for access to relevant VALIDATE staff through process outlined in VALIDATE data management plan for demonstrator development.

Anonymized prospective study data is available for access to researchers in academia through process outlined in VALIDATE data management plan after embargo period.

Data used to simulate a decision during prospective study is available for access to physicians through process outlined in VALIDATE data management plan during prospective study.

Goal

Data used to simulate a decision during prospective study is available for access to patients through process outlined in VALIDATE data management plan during prospective study.

Wish

Data used to simulate a decision during prospective study is available for access to legal guardians/caregivers through process outlined in VALIDATE data management plan during prospective study.

2.5 Transparency

This requirement is closely linked with the principle of explicability and encompasses transparency of elements relevant to an AI system: the data, the system and the business models.

2.5.1 Traceability

The data sets and the processes that yield the AI system's decision, including those of data gathering and data labelling as well as the algorithms used, should be documented to the best possible standard to allow for traceability and an increase in transparency. This also applies to the decisions made by the AI system. This enables identification of the reasons why an AI-decision was erroneous which, in turn, could help prevent future mistakes. Traceability facilitates auditability as well as explainability.

Req. ID	Requirement	Parameters		Phases
TRA-01	[Limitations and metadata] are available to doctors in the app during emergencies **[time point]**.	**[Limitations and metadata]**	(Data collection trail, data source, demographics, how patient-specific input affected the predictions, how model is calibrated and optimized)	development; testing; validation
		[time point]	At the end of the project	

Wish

[Data collection trail, data source, demographics, how patient-specific input affected the predictions, how model is calibrated and optimized) are available to doctors in the app during emergencies **at the end of the project**.

2.5.2 Explainability

Explainability concerns the ability to explain both the technical processes of an AI system and the related human decisions (e.g. application areas of a system). Technical explainability requires that the decisions made by an AI system can be understood and traced by human beings. Moreover, trade-offs might have to be made between enhancing a system's explainability (which may reduce its accuracy) or increasing its accuracy (at the cost of explainability). Whenever an AI system has a significant impact on people's lives, it should be possible to demand a suitable explanation of the AI system's decision-making process. Such explanation should be timely and adapted to the expertise of the stakeholder concerned (e.g. layperson, regulator or researcher). In addition, explanations of the degree to which an AI system influences and shapes the organisational decision-making process, design choices of the system, and the rationale for deploying it, should be available (hence ensuring business model transparency).

Req. ID	Requirement	Parameters	
VAL-03	% of **[explainability methods]** applied in the tool are validated by **[metric]** by **[time point].**	**[explainability methods]**	Explainability methods for tabular data; explainability methods for imaging data
		[metric]	Quantified scores; qualitative validation by users
		[time point]	Start of the prospective study

Tolerable

100% of explainability methods for tabular data applied in the tool are validated by quantified scores by the start of the prospective study.

100% of explainability methods for tabular data applied in the tool are validated by qualitative validation by users by the start of the prospective study.

100% of explainability methods for imaging data applied in the tool are validated by quantified scores by the start of the prospective study.

100% of explainability methods for imaging data applied in the tool are validated by qualitative validation by users by the start of the prospective study.

Req. ID	Requirement	Parameters		Phases
EXP-03	Explanations need to be communicated and defined in line with the **[constraints]** of the project **[time point].**	**[constraints]**	**[ethical values of the project]**; **[explainability regulations]**	development; testing
		[ethical values of the project]	Ethical framework	
		[explainability regulations]	EU AI Act; GDPR; EU-MDR	

		[time point]	Prior to start of the prospective study	

Goal

Explanations need to be communicated and defined in line with **ethical values of the project:Ethical framewok** of the project **prior to start of the prospective study.**

Explanations need to be communicated and defined in line with **explainability regulations :EU AI act;GDPR;MDR** of the project **prior to start of the prospective study.**

Req. ID	Requirement	Parameters		Phases
EXP-04	The system is not a black box achieved **[time point].**	[time point]	Prior to start of the prospective study	development; testing

Tolerable

The system is not a black box achieved **prior to start of the prospective study point.**

2.5.3 Communication

AI systems should not represent themselves as humans to users; humans have the right to be informed that they are interacting with an AI system. This entails that AI systems must be identifiable as such. In addition, the option to decide against this interaction in favour of human interaction should be provided where needed to ensure compliance with fundamental rights. Beyond this, the AI system's capabilities and limitations should be communicated to AI practitioners or end-users in a manner appropriate to the use case at hand. This could encompass communication of the AI system's level of accuracy, as well as its limitations.

Req. ID	Requirement	Parameters		Phases
EXP-01	[information] relevant for [stakeholders] to read or understand are communicated in adequate language so that it enables [stakeholders] to protect their own interests and is available [time point].	[information]	Study information; study results; user information	development; testing
		[explainee]	Patients; patient family/caregivers; users	

		[time point]	Before the prospective study start	

Tolerable

Study information relevant for **patients** to read or understand are communicated in adequate language so that it enables **patients** to protect their own interests and is available **before the prospective study start**.

Study results relevant for **patients** to read or understand are communicated in adequate language so that it enables **patients** to protect their own interests and is available **before the prospective study start**.

Study information relevant for **patient family/caregivers** to read or understand are communicated in adequate language so that it enables **patient family/caregivers** to protect their own interests and is available **before the prospective study start**.

Study results relevant for **patient family/caregivers** to read or understand are communicated in adequate language so that it enables **patient family/caregivers** to protect their own interests and is available **before the prospective study start**.

User information relevant for **users** to read or understand are communicated in adequate language so that it enables **users** to protect their own interests and is available **before the prospective study start**.

Req. ID	**Requirement**	**Parameters**		**Phases**
EXP-02	**[Relevant information]** for **[explainee groups]** need to be explained to **[explainee groups]** and translated by a science communicator when relevant.	**[Relevant information for patient family, patient]**	(Why are we using the tool? Why we don't know what the best treatment is. What are the results of the AI model? What do the results of the tool imply for them?); (Why should I use this tool? Evidence that it works. How do we know that this tool is successful? How does the model work? Which parameters/variables are input data? And which ones are deciding factors?);	development; testing

		for **[explainee groups]**	Patient/patient family/caregiver; medical users	

Tolerable

(Why are we using the tool? Why we don't know what the best treatment is. What are the results of the AI model? What do the results of the tool imply for them?) for **patient/patient family/caregiver** need to be explained to **patient/patient family/caregiver** and translated by a science communicator when relevant.

(Why should I use this tool? Evidence that it works. How do we know that this tool is successful? How does the model work?Which parameters/variables are input data? And which ones are deciding factors?) for **medical users** need to be explained to **medical users** and translated by a science communicator when relevant.

Req. ID	Requirement	Parameters		Phases
EXP-05	Explanations are specifically tailored to **[different explainee groups]** **[time point].**	**[different explainee groups]**	(users in different countries; patients/family/caregivers in different countries; patients/family/caregivers of different cultures	development; testing
		[time point]	Prior to start of the prospective study	

Tolerable

Explanations are specifically tailored to **users in different countries prior to start of the prospective study.**

Goal

Explanations are specifically tailored to **patienty/family/caregivers in different countries prior to start of the prospective study.**

Wish

Explanations are specifically tailored to **patienty/family/caregivers of different cultures prior to start of the prospective study.**

2.6 Diversity, non-discrimination and fairness

In order to achieve Trustworthy AI, we must enable inclusion and diversity throughout the entire AI system's life cycle. Besides the consideration and involvement of all affected stakeholders throughout the process, this also entails ensuring equal access through inclusive design processes as well as equal treatment. This requirement is closely linked with the principle of fairness.

2.6.1 Avoidance of unfair bias

Data sets used by AI systems (both for training and operation) may suffer from the inclusion of inadvertent historic bias, incompleteness and bad governance models. The continuation of such biases could lead to unintended (in)direct prejudice and discrimination against certain groups or people, potentially exacerbating prejudice and marginalisation. Harm can also result from the intentional exploitation of (consumer) biases or by engaging in unfair competition, such as the homogenisation of prices by means of collusion or a non-transparent market. Identifiable and discriminatory bias should be removed in the collection phase where possible. The way in which AI systems are developed (e.g. algorithms' programming) may also suffer from unfair bias. This could be counteracted by putting in place oversight processes to analyse and address the system's purpose, constraints, requirements and decisions in a clear and transparent manner. Moreover, hiring from diverse backgrounds, cultures and disciplines can ensure diversity of opinions and should be encouraged.

Req. ID	Requirement	Parameters		Phases
VAL-04	Subgroup stratified according to **[features]** OR **[intersectional feature combination]** has prediction accuracy with no significant difference to the majority population.	**[features]**	age; sex/gender; ethnicity/race; geographic location	development; testing; validation
		[intersectional feature combination]	sex/gender and ethnicity/race;	

Goal

Subgroup stratified according to **age** has prediction accuracy with no significant difference to the majority population.

Subgroup stratified according to **sex/gender** has prediction accuracy with no significant difference to the majority population.

Subgroup stratified according to **ethnicity/race** has prediction accuracy with no significant difference to the majority population.

Subgroup stratified according to **geographic location** has prediction accuracy with no significant difference to the majority population.

Subgroup stratified according to **sex/gender and ethnicity/race** has prediction accuracy with no significant difference to the majority population.

2.6.2 Accessibility and universal design

Particularly in business-to-consumer domains, systems should be user-centric and designed in a way that allows all people to use AI products or services, regardless of their age, gender, abilities or characteristics. Accessibility to this technology for persons with disabilities, which are present in all societal groups, is of particular importance. AI systems should not have a one-size-fits-all approach and should consider Universal Design principles addressing the widest possible range of users, following relevant accessibility standards. This will enable equitable access and active participation of all people in existing and emerging computer-mediated human activities and with regard to assistive technologies.

Req. ID	Requirement	Parameters	Phases

USA-01	System undergoes Ux/UI testing on **[X number] [users] [time point].**	**[X number]**	10	development; testing;
		[users]	Stroke physicians of different skill level	
		[time point]	During demonstrator development and testing	

Tolerable

System undergoes Ux/UI testing on **10 stroke physicians of different skill level during demonstrator development and testing.**

2.6.3 Stakeholder participation

In order to develop AI systems that are trustworthy, it is advisable to consult stakeholders who may directly or indirectly be affected by the system throughout its life cycle. It is beneficial to solicit regular feedback even after deployment and set up longer term mechanisms for stakeholder participation, for example by ensuring workers information, consultation and participation throughout the whole process of implementing AI systems at organisations.

Req. ID	Requirement	Parameters		Phases
INC-01	System is presented to **[X number] [stakeholders]** for feedback **[time point].**	**[X number]**	4;7;10	development; testing;
		[stakeholders]	Stroke patients and/or patient representatives from different geographical and cultural background	
		[time point]	During demonstrator development and testing	

System is presented to **4 stroke patients and/or patient representatives from different geographical and cultural backgrounds** for feedback **during demonstrator development and testing.**

Goal

System is presented to **7 stroke patients and/or patient representatives from different geographical and cultural backgrounds** for feedback **during demonstrator development and testing.**

Wish

System is presented to **10 stroke patients and/or patient representatives from different geographical and cultural backgrounds** for feedback **during demonstrator development and testing.**

2.7 Societal and environmental well-being

In line with the principles of fairness and prevention of harm, the broader society, other sentient beings and the environment should be also considered as stakeholders throughout the AI system's life cycle. Sustainability and ecological responsibility of AI systems should be encouraged, and research should be fostered into AI solutions addressing areas of global concern, such as for instance the Sustainable Development Goals. Ideally, AI systems should be used to benefit all human beings, including future generations.

2.7.1 Sustainable and environmentally friendly AI

AI systems promise to help tackling some of the most pressing societal concerns, yet it must be ensured that this occurs in the most environmentally friendly way possible. The system's development, deployment and use process, as well as its entire supply chain, should be assessed in this regard, e.g. via a critical examination of the resource usage and energy consumption during training, opting for less harmful choices. Measures securing the environmental friendliness of AI systems' entire supply chain should be encouraged.

Req. ID	Requirement	Parameters		Phases
SUS-01	% of VALIDATE machine learning engineers follow an SOP to reduce the environmental impact of model training.			development; testing; validation

Tolerable

50% of VALIDATE machine learning engineers follow an SOP to reduce the environmental impact of model training.

Goal

100% of VALIDATE machine learning engineers follow an SOP to reduce the environmental impact of model training.

2.7.2 Social impact

Ubiquitous exposure to social AI systems in all areas of our lives (be it in education, work, care or entertainment) may alter our conception of social agency, or impact our social relationships and attachment. While AI systems can be used to enhance social skills, they can equally contribute to their deterioration. This could also affect people's physical and mental wellbeing. The effects of these systems must therefore be carefully monitored and considered.

None of the identified requirements could be mapped to this subcategory.

2.7.3 Society and democracy

Beyond assessing the impact of an AI system's development, deployment and use on individuals, this impact should also be assessed from a societal perspective, taking into account its effect on institutions, democracy and society at large. The use of AI systems should be given careful consideration particularly in situations relating to the democratic process, including not only political decision-making but also electoral contexts.

None of the identified requirements could be mapped to this subcategory.

2.8 Accountability

The requirement of accountability complements the above requirements, and is closely linked to the principle of fairness. It necessitates that mechanisms be put in place to ensure responsibility and accountability for AI systems and their outcomes, both before and after their development, deployment and use.

2.8.1 Auditability

Auditability entails the enablement of the assessment of algorithms, data and design processes. This does not necessarily imply that information about business models and intellectual property related to the AI system must always be openly available. Evaluation by internal and external auditors, and the availability of such evaluation reports, can contribute to the trustworthiness of the technology. In applications affecting fundamental rights, including safety-critical applications, AI systems should be able to be independently audited.

None of the identified requirements could be mapped to this subcategory.

Auditability, however, is a main aspect of the WP1, with audits being performed throughout the project that will, in a co-creation approach, lead to framework updates via interdisciplinary collaboration.

2.8.2 Minimisation and reporting of negative impacts

Both the ability to report on actions or decisions that contribute to a certain system outcome, and to respond to the consequences of such an outcome, must be ensured. Identifying, assessing, reporting and minimising the potential negative impacts of AI systems is especially crucial for those (in)directly affected. Due protection must be available for whistle-blowers, NGOs, trade unions or other entities when reporting legitimate concerns about an AI-based system. The use of impact assessments (e.g. red teaming or forms of Algorithmic Impact Assessment) both prior to and during the development, deployment and use of AI systems can be helpful to minimise negative impact. These assessments must be proportionate to the risk that the AI systems pose.

Req. ID	Requirement	Parameters		Phases
ACC-01	Algorithmic impact assessment via **[relevant method]** performed **[time point].**	**[relevant method]**	To be determined	development; testing; validation
		[time point]	By the end of the project	

Tolerable

Algorithmic impact assessment via **relevant method:To be determined** performed **by the end of the project.**

2.8.3 Trade-offs

When implementing the above requirements, tensions may arise between them, which may lead to inevitable trade-offs. Such trade-offs should be addressed in a rational and methodological manner within the state of the art. This entails that relevant interests and values implicated by the AI system should be identified and that, if conflict arises, trade-offs should be explicitly acknowledged and evaluated in terms of their risk to ethical principles, including fundamental rights. In situations in which no ethically acceptable trade-offs can be identified, the development, deployment and use of the AI system should not proceed in that form. Any decision about which trade-off to make should be reasoned and properly documented. The decision-maker must be accountable for the manner in which the appropriate trade-off is being made, and should continually review the appropriateness of the resulting decision to ensure that necessary changes can be made to the system where needed.

None of the identified requirements could be mapped to this subcategory.

However, navigating trade-offs and tensions is one of the main overarching goals of work package one is reflected in tasks that include auditing, or the z-inspection assessment, in which identification and solution of tensions is one of the main goals.

2.8.4 Redress

When unjust adverse impact occurs, accessible mechanisms should be foreseen that ensure adequate redress. Knowing that redress is possible when things go wrong is key to ensure trust. Particular attention should be paid to vulnerable persons or groups.

None of the identified requirements could be mapped to this subcategory.

3 Manual

The VALIDATE trustworthy AI Manula is a quick reference guide for all the requirements and their category levels that were defined in the ethical framework.

3.1 Human Agency and Oversight

3.1.1 Human Agency

Tolerable

X number:To be defined of **relevant methods:To be defined** for measuring model uncertainty are used to provide a confidence interval for predictions in the app available **at start of the prospective study**.

3.1.2 Human Oversight

Tolerable

Human in command (HIC) governance concept available **at the end of the project.**

3.2 Technical Robustness and Safety

3.2.1 Resilience to Attack and Safety

Wish

Tool should follow **ISO norms** that are within our scope and resource limitations to prepare for compliance with the MDR **at the end of the project**.

3.2.2 Fallback plan and general safety

Goal

A thorough risk assessment following the MDR should be done to check and prove which risks are involved and whether they are low enough for this tool **at the end of the project**.

Wish

A thorough risk assessment following the MDR should be done to check and prove which risks are involved and whether they are low enough for this tool **at when intended purpose is defined**.

3.2.3 Accuracy

Wish

Validity will be measured by doing a randomized clinical trial (RCT) to examine whether at least **0.1%** of **number determined by estimated effect size patients** whose doctors use the tool have a significantly improved MRS 90 compared to control group at **after the project.**

Tolerable

10% of MDR requirements are addressed for an to prepare for an RCT after the project that validates that MRS 90 patient outcome is significantly better when using the tool compared to current standards of care.

10% of GDPR requirements are addressed for an to prepare for an RCT after the project that validates that MRS 90 patient outcome is significantly better when using the tool compared to current standards of care.

0% of FDA requirements are addressed for an to prepare for an RCT after the project that validates that MRS 90 patient outcome is significantly better when using the tool compared to current standards of care.

Goal

20% of MDR requirements are addressed for an to prepare for an RCT after the project that validates that MRS 90 patient outcome is significantly better when using the tool compared to current standards of care.

20% of GDPR requirements are addressed for an to prepare for an RCT after the project that validates that MRS 90 patient outcome is significantly better when using the tool compared to current standards of care.

10% of FDA requirements are addressed for an to prepare for an RCT after the project that validates that MRS 90 patient outcome is significantly better when using the tool compared to current standards of care.

Wish

40% of MDR requirements are addressed for an to prepare for an RCT after the project that validates that MRS 90 patient outcome is significantly better when using the tool compared to current standards of care.

40% of GDPR requirements are addressed for an to prepare for an RCT after the project that validates that MRS 90 patient outcome is significantly better when using the tool compared to current standards of care.

20% of FDA requirements are addressed for an to prepare for an RCT after the project that validates that MRS 90 patient outcome is significantly better when using the tool compared to current standards of care.

3.2.4 Reliability and Reproducibility

Tolerable

The tool needs to define the intended purpose **as part of the demonstrator development**.

Tolerable

100% of publications authored by the consortium are open access and if relevant **(contain model cards; follow reporting guidelines; follow open code practices, provide open data access)** to improve reproducibility.

3.3 Privacy and data governance

3.3.1 Privacy and data protection

Tolerables

The AI tool complies with GDPR by start of prospective study.

The AI tool complies with Horizon Europe grant requirements for privacy by start of prospective study.

The AI tool complies with local authority privacy regulations by start of prospective study.

Wish

The AI tool complies with HIPAA by start of prospective study.

Tolerables

Research data are stored via password protected user rights system, with practice enabled before start of training.

Research data are stored via local protected servers, with practice enabled before start of training.

Models are stored via password protected user rights system, with practice enabled before start of training.

Models are stored via local protected servers, with practice enabled before start of training.

Predictions are stored via password protected user rights system, with practice enabled before start of training.

Predictions are stored via local protected servers, with practice enabled before start of training.

xAI results are stored via password protected user rights system, with practice enabled before start of training.

xAI results are stored via local protected servers, with practice enabled before start of training.

Goal

Research data are stored via data encryption, with practice enabled at 24 months into the project.

Models are stored via data encryption, with practice enabled at 24 months into the project.

Predictions are stored via data encryption, with practice enabled at 24 months into the project.

xAI results are stored via data encryption, with practice enabled at 24 months into the project.

Wish

Research data are stored via cloud environment, with practice enabled at 24 months into the project.

Models are stored via cloud environment, with practice enabled at 24 months into the project.

Predictions are stored via cloud environment, with practice enabled at 24 months into the project.

xAI results are stored via cloud environment, with practice enabled at 24 months into the project.

Tolerable

Privacy information to answer "**Why are we collecting this data? How will the data be collected? Where will data be stored? What data be collected? Will the data be shared?** to approve data collection for prospective study are needed **by project month 12** in the research data management plan.

3.3.2 Quality and integrity of data

Tolerable

100% of datasets fulfill **quality criteria:To be defined** with regards to **(Missing data, errors, inaccuracies, interoperability) before final model training for demonstrator.**

3.3.3 Access to data

Tolerable

Retrospective study data is available for access to relevant VALIDATE staff through process outlined in VALIDATE data management plan for demonstrator development.

Anonymized prospective study data is available for access to researchers in academia through process outlined in VALIDATE data management plan after embargo period.

Data used to simulate a decision during prospective study is available for access to physicians through process outlined in VALIDATE data management plan during prospective study.

Goal

Data used to simulate a decision during prospective study is available for access to patients through process outlined in VALIDATE data management plan during prospective study.

Wish

Data used to simulate a decision during prospective study is available for access to legal guardians/caregivers through process outlined in VALIDATE data management plan during prospective study.

3.4 Transparency

3.4.1 Traceability

Wish

(Data collection trail, data source, demographics, how patient-specific input affected the predictions, how model is calibrated and optimized) are available to doctors in the app during emergencies **at the end of the project**.

3.4.2 Explainability

Tolerable

100% of explainability methods for tabular data applied in the tool are validated by quantified scores by the start of the prospective study.

100% of explainability methods for tabular data applied in the tool are validated by qualitative validation by users by the start of the prospective study.

100% of explainability methods for imaging data applied in the tool are validated by quantified scores by the start of the prospective study.

100% of explainability methods for imaging data applied in the tool are validated by qualitative validation by users by the start of the prospective study.

Goal

Explanations need to be communicated and defined in line with **ethical values of the project:Ethical framewok** of the project **prior to start of the prospective study.**

Explanations need to be communicated and defined in line with **explainability regulations :EU AI act;GDPR;MDR** of the project **prior to start of the prospective study.**

Tolerable

The system is not a black box achieved **prior to start of the prospective study point.**

3.4.3 Communication

Tolerable

Study information relevant for **patients** to read or understand are communicated in adequate language so that it enables **patients** to protect their own interests and is available **before the prospective study start**.

Study results relevant for **patients** to read or understand are communicated in adequate language so that it enables **patients** to protect their own interests and is available **before the prospective study start**.

Study information relevant for **patient family/caregivers** to read or understand are communicated in adequate language so that it enables **patient family/caregivers** to protect their own interests and is available **before the prospective study start**.

Study results relevant for **patient family/caregivers** to read or understand are communicated in adequate language so that it enables **patient family/caregivers** to protect their own interests and is available **before the prospective study start**.

User information relevant for **users** to read or understand are communicated in adequate language so that it enables **users** to protect their own interests and is available **before the prospective study start**.

Tolerable

(Why are we using the tool? Why we don't know what the best treatment is. What are the results of the AI model? What do the results of the tool imply for them?) for **patient/patient family/caregiver** need to be explained to **patient/patient family/caregiver** and translated by a science communicator when relevant.

(Why should I use this tool? Evidence that it works. How do we know that this tool is successful? How does the model work?Which parameters/variables are input data? And which ones are deciding factors?) for **medical users** need to be explained to **medical users** and translated by a science communicator when relevant.

Tolerable

Explanations are specifically tailored to **users in different countries prior to start of the prospective study.**

Goal

Explanations are specifically tailored to **patienty/family/caregivers in different countries prior to start of the prospective study.**

Wish

Explanations are specifically tailored to **patienty/family/caregivers of different cultures prior to start of the prospective study.**

3.5 Diversity, non-discrimination and fairness

Goal

Subgroup stratified according to **age** has prediction accuracy with no significant difference to the majority population.

Subgroup stratified according to **sex/gender** has prediction accuracy with no significant difference to the majority population.

Subgroup stratified according to **ethnicity/race** has prediction accuracy with no significant difference to the majority population.

Subgroup stratified according to **geographic location** has prediction accuracy with no significant difference to the majority population.

Subgroup stratified according to **sex/gender and ethnicity/race** has prediction accuracy with no significant difference to the majority population.

3.5.1 Accessibility and universal design

Tolerable

System undergoes Ux/UI testing on **10 stroke physicians of different skill level during demonstrator development and testing.**

3.5.2 Stakeholder participation

Tolerable

System is presented to **4 stroke patients and/or patient representatives from different geographical and cultural backgrounds** for feedback **during demonstrator development and testing.**

Goal

System is presented to **7 stroke patients and/or patient representatives from different geographical and cultural backgrounds** for feedback **during demonstrator development and testing.**

Wish

System is presented to **10 stroke patients and/or patient representatives from different geographical and cultural backgrounds** for feedback **during demonstrator development and testing.**

3.6 Societal and environmental well-being

3.6.1 Sustainable and environmentally friendly AI

Tolerable

50% of VALIDATE machine learning engineers follow an SOP to reduce the environmental impact of model training.

Goal

100% of VALIDATE machine learning engineers follow an SOP to reduce the environmental impact of model training.

3.7 Accountability

3.7.1 Minimisation and reporting of negative impacts

Tolerable

Algorithmic impact assessment via **relevant method:To be determined** performed **by the end of the project.**