VALIDATE Framework

Deliverable

D3.2 Ethical Plan

Project Acronym: **VALIDATE**

Project Title: **Continuous stratification for improved prevention, treatment, and rehabilitation of stroke patients using digital twins and AI**

Grant Agreement Number: **101080875**

This project has received funding from the European Union's Horizon 2022 programme under Grant Agreement No. 101080875

Authors and Contributors	Riana Minocher (CUB); Vince I. Madai (CUB)		
Responsible Author	Riana Minocher	**Email:**	riana.minocher@bih-charite.de
	Beneficiary CUB		

Work package information	**Work Package**	3	**Deliverable** D3.2 Ethical Plan
	Dissemination Level		Public

Table of contents

1 Introduction to VALIDATE framework

2 How Requirements are Organized

Requirements are structured to provide a clear view of their definitions and parameters. Below is an outline detailing how these are organized, articulated through an example of a standard requirement.

2.1 Requirement YZ-N-AAA

Description: Each requirement bears a unique ID, constructed as follows:

- **YZ**: Sub-chapter numbers reflecting the requirement's area

- **N**: Sequential number within the sub-chapter

- **AAA**: The first three letters of the chapter topic

For instance, the requirement for implementing confidence intervals in predictive modeling might state: "X of [example parameter] for measuring model uncertainty are implemented to provide a confidence interval for predictions in the app available at [time point]".

2.2 Parameters

- **[Example parameter]**: Parameters listed in brackets are adjustable by the WP 1 team during the project period.
- **[time point]**: This refers to the deadline by which a requirement should be met, typically specifying some due date.

2.3 Phases

Each requirement is mapped to a specific project phase:

- **Development**: Focuses on developing the tool's prototype until month 18.

- **Testing**: The prototype undergoes testing until month 18.

- **Validation**: Begins in month 18, marking the start of validating the developed and tested prototype.

2.4 Stakeholders and Owner

Stakeholders: Includes project staff and other relevant groups affected by or necessary for fulfilling the requirement.

Owner: Each requirement is assigned an Owner, usually a WP lead relevant to existing tasks. The Owner is responsible for task delegation and execution to meet the requirement.

2.5 Requirement Categorization

Based on the parameterized statements, requirements are classified into three categories:

- **TOLERABLE**: The minimum achievement threshold for a requirement.

- **GOAL**: Represents the objective aimed to be reached by the project. If both Tolerable and Goal are defined, the project's likely outcome will be between these two.

- **WISH**: Desirable outcomes that are aimed for but likely only partially achieved.

3 1 Human Agency and Oversight (HUM)

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.

3.1 1.1 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. Other than the requirements addressed in related sections such as privacy, we believe that our tool does not influence fundamental rights.

3.2 1.2 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.

3.2.1 Requirement 12-1-HUM

Description: In the events where there is a clear conflict between the predictions of the tool and the medical opinion of the doctor, a process protocol for epistemic authority dilemmas customized for ***physician expertise level*** can be followed. Available by ***time point***.

physician expertise level:

- beginner

- intermediate

- expert

time point:

- project end

Owner

- WP1 lead

Stakeholders

- WP 1 (ethics)

- WP 3 (design)

- WP 4 (clinical validation)

- Other: medical staff

Tolerable

- In the events where there is a clear conflict between the predictions of the tool and the medical opinion of the doctor, a process protocol for epistemic authority dilemmas customized for ***beginner; intermediate; expert*** can be followed. Available by ***project end***.

3.3 1.3 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.

3.3.1 Requirement 13-1-HUM

Description: Tool is designed with ***[capability needed]*** to enable a Human in Command (HIC) governance structure. This is implemented by ***[time point].***

capability needed:

- capability to oversee the overall activity of the AI system;

- ability to decide whether, when and how to use the system in any particular situation;

- established levels of human discretion during the use of the system where necessary;

- ability to override a decision made by a system

time point:

- project end

Owner

- WP 1 lead

Stakeholders

- WP 1 (ethics)

- WP 3 (design)

- WP 4 (clinical validation)

Tolerable

- Tool is designed with ***capability to oversee the overall activity of the AI system*** to enable a HIC governance structure. This is implemented ***by the end of the project***.

Goal

- Tool is designed with ***ability to decide whether, when and how to use the system in any particular situation*** to enable a HIC governance structure. This is implemented ***by the end of the project***.

- Tool is designed with ***established levels of human discretion during the use of the system where necessary*** to enable a HIC governance structure. This is implemented ***by the end of the project***.

- Tool is designed with ***ability to override a decision made by a system*** to enable a HIC governance structure. This is implemented ***by the end of the project***.

4 2 Technical Robustness and Safety (ROB)

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.

4.1 2.1 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.

4.1.1 Requirement 21-1-ROB

Description: Tool should follow ***[relevant aspects of norms]*** that are within our scope and resource limitations to prepare for compliance with the MDR by ***[time point]***.

relevant aspects of norms:

- ISO norms relate to the standard MDR:

- IEC 62366

- IEC 62304

- ISO 14971

- ISO norms related to MDR and recent AI advancements:

- ISO 13485

- ISO/IEC 23894:2023

- ISO/IEC 38507

time point:

- the end of the project

Phases

- development

- testing

- validation

Owner

- WP 5 lead

Stakeholders

- WP 2 (development)

- WP 3 (design)

- WP 5 (regulatory pathway to market)

Wish

- The stability of the accuracies and performance for ***when there are missing input data points*** of the tool is analyzed or monitored, available ***by the start of the validation phase***.

- Tool should follow ***relevant aspects of ISO norms related to the MDR and recent AI advancements: ISO 13485; ISO/IEC 23894:2023; ISO/IEC 38507***that are within our scope and resource limitations to prepare for compliance with the MDR by ***the end of the project***.

4.1.2 Requirement 21-2-ROB

Description: Tool should follow ***[relevant guidelines on cyber security]*** that are within our scope and resource limitations to prepare for compliance with the MDR by ***[time point]***.

relevant guidelines on cyber security:

- To be defined, needs to be researched by Owner.

time point:

- the end of the project

Phases

- development

- testing

- validation

Owner

- WP 5 lead

Stakeholders

- WP 2 (development)

- WP 3 (design)

- WP 5 (regulatory pathway to market)

Wish

- Tool should follow ***[relevant guidelines on cyber security]*** that are within our scope and resource limitations to prepare for compliance with the MDR by ***the end of the project***.

4.2 2.2 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.

4.2.1 Requirement 22-1-ROB

Description: A risk assessment following relevant aspects of ***[risk management norms]*** will be done by ***[time point]***.

risk management norms:

- ISO 14971

- ISO 23894:2023

time point:

- when intended purpose is defined

- at the end of the project

Phases

- development

- testing

- validation

Owner

- WP 5 lead

Stakeholders

- WP 2 (development)

- WP 3 (design)

- WP 5 (regulatory pathway to market)

Goal

- A risk assessment following relevant aspects of ***ISO 14971***will be done by ***at the end of the project***.

Wish

- A risk assessment following relevant aspects of ***ISO 23894:2023*** will be done by ***when intended purpose is defined***.

4.3 2.3 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.

4.3.1 Requirement 23-1-ROB

Description: 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 Modified Ranking Scale 90 (MRS 90) compared to control group at ***[time point].***

ratio:

- will be determined at the end of the prospective study

RCT study size

- number will be determined later by estimating the effect size

time point:

- after the project

Phases

- validation

Owner

- WP 4 lead

Stakeholders

- WP 4 (clinical validation)

Wish

- Validity will be measured by doing a randomized clinical trial (RCT) to examine whetherat least ***[ratio]***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*.**

4.3.2 Requirement 23-2-ROB

Description: Accuracy will be measured by ***[time point]*** by doing a prospective shadowing study to examine whether in at least ***[threshold]***of the cases, the predictions of the prototype are similar to the actual 3 months mRS according to the actual administered treatment.

threshold:

- 90%

time point:

- the end of the project

Phases

- validation

Owner

- WP 4 lead

Stakeholders

- WP 4 (clinical validation)

Goal

- Accuracy will be measured by ***the end of the project***by doing a prospective shadowing study to examine whether in at least ***90%***of the cases, the predictions of the prototype are similar to the actual 3 months mRS according to the actual administered treatment

4.4 2.4 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.

4.4.1 Requirement 24-1-ROB

Description: Tool reliably provides no significantly different performance for ***[contexts]*** at ***[time point].***

contexts:

- edge-cases

- differing local standards of care

- non-western patient minorities

time point:

- at the end of the project

Phases

- development

- testing

- validation

Owner

- WP 2 lead

Stakeholders

- WP 2 (development)

- WP 3 (design)

- WP 4 (clinical validation)

Goal

- Tool reliably provides no significantly different performance for ***edge-cases at the end of the project.***

- Tool reliably provides no significantly different performance for ***differing local standards of care at the end of the project***.

- Tool reliably provides no significantly different performance for ***non-western patient minorities at the end of the project***.

4.4.2 Requirement 24-2-ROB

Description: The stability of the accuracies and performance for ***[different risk areas]*** of the tool is analyzed or monitored, available by ***[time point].***

different risk areas [From 35-ROB:]***:***

- edge-cases

- differing local standards of care

- non-western patient minorities

- when there are missing input data points

- for sub-groups mentioned in 61-DIV

time point:

- by the start of the validation phase

Phases

- development

- testing

- validation

Owner

- WP 2 lead

Stakeholders

- WP 2 (development)

- WP 3 (design)

- WP 4 (clinical validation)

Tolerable

- The stability of the accuracies and performance for ***when there are missing input data points*** of the tool is analyzed or monitored, available ***by the start of the validation phase.***

Goal

- The stability of the accuracies and performance for ***edge-cases, differing local standards of care, non-western patient minorities***of the tool is analyzed or monitored, available ***by the validation phase.***

- The stability of the accuracies and performance for ***sub-groups mentioned in 61-DIV*** of the tool is analyzed or monitored, available ***by the validation phase***.

4.4.3 Requirement 24-3-ROB

Description: The tool has a defined intended purpose following the MDR ***[time point]***.

time point:

- as part of the prototype development

Phases

- development
- testing

Owner

- WP 3 lead

Stakeholders

- WP 2 (development)
- WP 4 (clinical validation)
- WP 5 (regulations)

Tolerable

- The tool has a defined intended purpose following the MDR ***as part of the prototype development***.

5 3 Privacy and data governance (PRI)

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.

5.1 3.1 Privacy and data protection

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.

5.1.1 Requirement 31-1-PRI

Description: The Prototype complies with relevant ***[privacy regulations]*** by ***[time point].***

privacy regulations:

- GDPR

- Health Insurance Portability and Accountability Act (HIPAA)

- National/local authority privacy regulations

time point:
- start of the prospective study

Phases

- development

- testing

- validation

Owner
- WP 5 lead

Stakeholders

- WP 1 (ethics)

- WP 2 (development)

- WP 3 (design)

- WP 5 (regulations)

Tolerable

- The prototype complies with ***GDPR*** by ***start of prospective study***.

- The prototype complies with ***Horizon Europe grant requirements for privacy*** by ***start of prospective study***.

- The prototype complies with ***national/local authority privacy regulations*** by ***start of prospective study***.

Wish

- The prototype complies with ***HIPAA*** by ***start of prospective study***.

5.1.2 Requirement 31-2-PRI

Description: All ***[development materials]*** are stored using ***[best practices for privacy]*****,** with practice enabled before ***[time point]***

development materials:

- research data

- models

- predictions

- XAI results

best practices for privacy:

- data encryption

- password protected user rights system

- local protected servers on clinical premises

time point:

- start of training

- at 24 months into the project

Phases

- development

- testing

- validation

Owner

- WP 2 lead

Stakeholders

- WP 2 (development)

- WP 3 (design)

Tolerable

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

- ***Research data*** are stored using ***local protected servers on clinical premise***s, with practice enabled ***before start of training***.

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

- Models are stored using ***local protected servers on clinical premises***, with practice enabled ***before start of training***.

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

- Predictions are stored using ***local protected servers on clinical premises***, with practice enabled ***before start of training***.

- xAI results are stored using ***password protected user rights system***, with practice enabled before start of training. xAI results are stored using local protected servers on clinical premises, with practice enabled ***before start of training***.

Goal

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

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

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

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

5.1.3 Requirement 31-3-PRI

Description: Privacy information have been collected to answer ***[local ethics committee questions]*** at ***[time point]***to approve data collection for prospective study 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?

- Who is the owner of the data?

- Who is responsible for the data?

- Who has access to the data?

- Will the data be transferred to other countries?

- Are those countries in the EU? Will the data be shared?

time point:

- by the end of the development phase

Phases

- development

Owner

- WP 4 lead

Stakeholders

- WP 2 (development)

- WP 3 (design)

- WP 5 (regulations)

Tolerable

- Privacy information have been collected 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? by the end of the development phase*** to approve data collection for prospective study in the research data management plan.

5.1.4 Requirement 31-4-PRI

Description: Consent from patients have been collected for ***[data purposes] [time point]*.**

data purposes:

- model training using the prospective data
- prospective study
- Randomized Clinical Trial
- open data-sharing required by the EU Horizon grant

time point:

- before data inclusion in CRF
- before the randomized clinical trial
- after the VALIDATE project

Phases

- development
- training
- validation

Owner

- WP 4 lead

Stakeholders

- WP 3 (design)
- WP 4 (clinical validation)
- WP 5 (regulations)
- WP 6 (patient communication)

Tolerable

- Consent from patients have been collected for ***model training using the prospective data before data inclusion in CRF***.

- Consent from patients have been collected for ***open data-sharing required by the EU Horizon grant raining before data inclusion in CRF***.

Goal

- Consent from patients have been collected for ***prospective study before data inclusion in CRF.***

Wish

- Consent from patients have been collected for ***randomized clinical trial before the randomized clinical trial, after the VALIDATE project***.

5.2 3.2 Privacy and data protection

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.

5.2.1 Requirement 32-1-PRI

Description: ***%*** of datasets fulfill ***[quality criteria]*** with regards to ***[quality parameters] [time point].***

quality criteria:

- relevant interoperability standards for labeling
- HL7 FHIR standards
- to be defined

quality parameters:

- missing data
- errors
- inaccuracies
- interoperability

time point:

- before final model training for prototype

Phases

- development
- training
- validation

Owner

- WP 2 lead

Stakeholders

- WP 1 (ethics)
- WP 2 (development)

- WP 4 (clinical validation)

Tolerable

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

Goal

- *100%* of datasets fulfill ***relevant interoperability standards for labeling*** with regards to ***interoperability before final model training for prototype.***

5.3 3.3 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.

5.3.1 Requirement 33-1-PRI

Description: ***[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
- data generated by the system or app
- meta-data or usage-data of the app/system

stakeholder type:

- relevant VALIDATE staff
- researchers
- legal guardian/caregiver
- patient
- users

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 prototype development

- during prospective study

- at the end of the project

Phases

- development

- training

- validation

Owner

- WP 3 lead

Stakeholders

- WP 2 (development)

- WP 4 (clinical validation)

- WP 5 (admin)

Tolerable

- ***Retrospective study data*** is available for access to ***relevant VALIDATE staff*** through ***process outlined in VALIDATE data management plan for prototype developmen***t.

- ***Anonymized prospective study data*** is available for access to ***researchers*** 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 ***researchers*** through ***process outlined in VALIDATE data management plan during prospective stud***y.

- ***Data generated by the system or app*** is available for access to ***users*** 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.***

- ***Meta-data or usage-data of the app/system*** is available for access to ***researchers*** through ***process outlined in VALIDATE data management plan at the end of the project***.

5.3.2 Requirement 33-2-PRI

Description: Access and use of ***[sensitive data attributes]*** is ***[safeguards]*** to protect ***[vulnerable groups]*** from discrimination and harm by ***[time point].***

sensitive data attributes:

- sex

- gender

- ethnicity

safeguards:

- logged

- only available to relevant qualified personnel with user privileges

- only available after log-in

- data is stored on local hospital premises

time point:

- by the start of the validation phase. *See page 7 for definition of validation phase.*

Phases

- development

- training

- validation

Owner

- WP 4 lead

Stakeholders

- WP 2 (development)

- WP 4 (clinical validation)

- WP 5 (admin)

- NORA (data storage)

Goal

- Access and use of ***sex; gender; ethnicity*** is ***logged*** to protect ***non-binary patients; trans patients; ethnic minorities*** from discrimination and harm ***by the start of the validation phase.***

- Access and use of ***sex; gender; ethnicit***y is ***only available to relevant qualified personnel with user privileges*** to protect ***non-binary patients; trans patients; ethnic minorities*** from discrimination and harm ***by the start of the validation phase.***

- Access and use of ***sex; gender; ethnicity*** is ***only available after log-in*** to protect ***non-binary patients; trans patients; ethnic minorities*** from discrimination and harm ***by the start of the validation phase.***

6 Transparency (TRA)

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.

6.1 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.

6.1.1 Requirement 41-1-TRA

Description: ***Limitations and metadata*** are available to doctors in the app during emergencies by ***time point***.

Parameters:

Limitations and metadata: - data collection trail - data source - demographics - how patient-specific input affected the predictions - how model is calibrated and optimized, especially whether it has a tendency for false negatives or false positives

time point:
- by the end of the project

Phases
- development - testing - validation

Owner
- WP 2 lead

Stakeholders

- WP 2 (development)

- WP 3 (design)

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 ***by the end of the project***.

6.2 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).

6.2.1 Requirement 42-1-TRA

Description: The system is able to inform about parts of the input that led to a specific outcome, achieved at ***time point***.

Parameters

time point:
-Prior to start of the prospective study

Phases
- development - testing

Owner
- WP 2 lead

Stakeholders

- WP 2 (development)

- WP 3 (design)

Tolerable

- The system is able to inform about parts of the input that led to a specific outcome achieved ***prior to start of the prospective study point***.

6.2.2 Requirement 42-2-TRA

Description: % of explanations are communicated and defined in line with the ***constraints*** of the project ***time point***.

constraints: - ethical values of the project - explainability regulations

ethical values of the project: - ethical framework

process: - process outlined in VALIDATE data management plan

explainability regulations: - EU AI Act - GDPR - EU-MDR

time point: - prior to the start of the prospective study

Phases
- development - testing

Owner
- WP 6 lead

Stakeholders

- WP 2 (development)

- WP 3 (design)

- WP 5 (regulations)

- WP 6 (patient communication)

Goal

- % of explanations are communicated and defined in line with ***ethical values of the project: Ethical framework*** of the project ***prior to start of the prospective study***.

- % of explanations are communicated and defined in line with ***explainability regulations: EU AI act;GDPR;EU-MDR*** of the project ***prior to start of the prospective study***.

6.2.3 Requirement 42-3-TRA

Description: % of ***explainability methods*** are applied in the tool and outputs are available to ***relevant users*** by ***time point***.

explainability methods: - explainability methods for tabular data - explainability methods for imaging data

relevant users: - medical personnel

time point: - start of the prospective study

Phases
- development - testing

Owner
- WP 2 lead

Stakeholders

- WP 2 (development)

- WP 3 (design)

- WP 4 (clinical validation)

Goal

- 100% of ***explainability methods for tabular data are*** applied in the tool and outputs are available to medical personnel by ***the start of the prospective study***.

- 100% of ***explainability methods for imaging data are*** applied in the tool and outputs are available to medical personnel by ***the start of the prospective study***.

6.2.4 Requiremennt 42-4-TRA

Description: % 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: - month 24 (delivery date of D2.2 which will incorporate results from T2.4 (refinement and improvement of models for stroke outcome) and T2.5 (augmentation of models with xAI))

Phases
- development - testing

Owner
- WP 2 lead

Stakeholders

- WP 2 (development)

- WP 3 (design)

Goal

- 100% of ***explainability methods for tabular data*** applied in the tool are validated by ***quantified scores*** by month 24

- 100% of ***explainability methods for tabular data*** applied in the tool are validated by ***qualitative validation by users*** by month 24

- 100% of ***explainability methods for imaging data*** applied in the tool are validated by ***quantified scores*** by month 24

- 100% of ***explainability methods for imaging data*** applied in the tool are validated by ***qualitative validation by users*** by month 24

6.3 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.

6.3.1 Requirement 43-1-TRA

Description: ***information*** relevant for ***explainee*** to read or understand are communicated in adequate language so that it enables ***explainee*** to protect their own interests and is available ***time point***.

information: - study information - study results - user information

explainee: - patients - patient family/caregivers - users

time point: - prior to start of the prospective study

Phases
- development - testing

Owner
- WP 6 lead

Stakeholders

- WP 6 (patient 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***

6.3.2 Requirement 43-2-TRA

Description: ***Relevant information*** for ***explainee groups*** are explained to ***explainee groups*** and translated by a science communicator when relevant.

relevant information: For Patient/patient family/caregiver group:

- 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 medical users group:

- Why should I use this tool?

- When should I not 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?

explainee groups: - patient/patient family/caregiver - medical users

Phases
- development - testing - validation

Owner
- WP 6 lead

Stakeholders

- Explainee groups

- WP 2 (development)

- WP 4 (clinical validation)

- WP 6 (patient communication)

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*** are explained to ***patient/patient family/caregiver*** and translated by a science communicator when relevant.

- ***(Why should I use this tool? When should I not 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*** are explained to ***medical users*** and translated by a science communicator when relevant.

6.3.3 Requirement 43-3-TRA

Description: Explanations are specifically tailored to ***different explainee groups*** by ***time point***.

different explainee groups:

- users in different countries - patients/family/caregivers in different countries - patients/family/caregivers of different cultures

time point: - prior to the start of the prospective study

Phases
- development - testing

Owner
- WP 6 lead

Stakeholders

- WP 6 (patient communication)

Tolerable

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

Goal

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

Wish

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

7 Diversity, non-discrimination, and fairness (DIV)

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.

7.1 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.

7.1.1 Requirement 51-1-DIV

Description: Subgroup stratified according to ***features*** AND ***intersectional feature combination*** has prediction accuracy with no significant difference compared to the majority population.

features: - age - sex/gender - ethnicity/race - geographic location

intersectional feature combination:
- sex/gender with ethnicity/race

Phases
- development - testing - validation

Owner
- WP 2 lead

Stakeholders

- WP 2 (development)

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.

7.2 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.

7.2.1 Requirement 52-1-DIV

Description: System undergoes UX/UI testing on ***X number user groups at time point***.

X number:
-10

user groups: - Stroke physicians of different technical skill levels - Stroke physicians of different medical skill levels

time point: - During demonstrator development and testing

Phases
- development - testing

Owner
- WP 3 lead

Stakeholders

- WP 3 (design)

- WP 4 (clinical validation)

Tolerable

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

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

7.3 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.

7.3.1 Requirement 53-1-DIV

Description: System is presented to X ***stakeholders*** for feedback ***time point***.

stakeholders: - Stroke patients and/or patient representatives from different geographical and cultural backgrounds

time point: - During demonstrator development and testing

Phases
- development - testing

Owner
- WP 3 lead

Stakeholders

- WP 3 (design)

- WP 6 (patient communication)

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***.

8 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.

8.1 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.

8.1.1 Requirement 61-1-SUS

Description: % of VALIDATE machine learning engineers follow an SOP to track the environmental impact of model training.

Phases
- development - testing

Owner
- WP 2 lead

Stakeholders

- WP 1 (ethics)

- WP 2 (development)

Tolerable

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

Goal

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

8.2 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.

8.3 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.

9 7 Accountability (ACC)

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.

9.1 7.1 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.

9.2 7.2 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.

9.2.1 Requirement 72-1-ACC

Description: Algorithmic impact assessment following ***relevant method*** is performed by ***time point***.

relevant method: - MDR - EU AI Act - To be determined

time point:

- by the end of the project

phases:

- development

- testing

- validation

Owner

- WP1 lead

Stakeholders

- WP 1 (ethics)

Tolerable

- Algorithmic impact assessment using ***MDR; EU AI Act, relevant method: To be determined*** performed ***by the end of the project.***

9.3 7.3 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.

9.4 7.4 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.

10 Work Packages