Deliverable 3.2 Ethical plan

STRATIF AI

Deliverable

D3.2 Ethical Plan

Project Acronym: **STRATIF-AI**

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** WP3	**Deliverable**	D3.2 Ethical Plan
	Dissemination Level		Public

Table of contents

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

Revision	Date	Author	Organization	Description
0.1	09.04.2024	Riana Minocher	Charité	First draft

Date of delivery		
Contractual:	30.04.2024	Actual:
Status		
final [x]		

Abstract	This document describes the plan for ethical and trustworthy design of the STRATIF-AI platform.
Keywords	ethics, trustworthy AI

Statement of originality

This deliverable contains original unpublished work, or it is clearly indicated otherwise. Proper recognition of previously published material and of the work of others has been made through appropriate citation.

Introduction

The application of artificial intelligence (AI) in healthcare poses ethical risks. Dedicated efforts are essential to ensure that AI systems in healthcare adhere to principles of lawfulness, ethics, and robustness. We have developed a plan which outlines steps to mitigate ethical risks and increase trustworthiness throughout the lifecycle of the STRATIF-AI project. We draw on interdisciplinary expertise and align our approach with three core strategies:

(1) Operationalization: Existing guidelines for AI development often provide broad, high-level principles, which are suitable for a wide range of systems. High-level principles must be translated to actionable steps that are tailored to the specific traits of individual AI systems. We leverage expertise in ethical framework development to ensure that ethical principles do not remain theoretical concepts in our project—but are instead effectively embedded into the design, implementation, and continuous management of the AI system.

(2) Co-creation: In line with existing recommendations, we have adopted a co-creation process to ensure the equitable development of solutions. As such, the content for this manual has been collaboratively developed with key stakeholders within the STRATIF-AI project, including clinicians, data scientists, medical researchers, representatives from patient organizations, and project coordinators. Engaging diverse stakeholders allows us to thoroughly assess the risks and benefits associated with specific requirements. The collaborative process fosters the creation of ethical requirements that are not only necessary and feasible but also auditable, thus contributing to the overall integrity and reliability of the AI system.

(3) Anticipation: The STRATIF-AI project introduces several novel technological solutions. Novel challenges may fall outside the scope of existing ethical guidelines, and may thus be difficult to address solely by adhering to current regulations or previously published frameworks. To proactively mitigate novel ethical risks, we pay close attention in our workshops and research to issues which may be yet unprecedented.

Overall, our ethical plan has been designed to facilitate the early identification of potential ethical risks and ultimately enhance the STRATIF-AI system's trustworthiness and societal impact.

Framework

Deliverable 3.2 presents a snapshot of the **Framework for Ethical and Trustworthy Design of the STRATIF-AI platform**.

The goal of the framework is to enumerate project-specific requirements that can help to mitigate ethical risks and increase trustworthiness throughout the STRATIF-AI project's development process (see: Introduction).

Design and Use

Our framework has been designed to be a living document[1], and, as such, will be revised following upcoming workshops, audits, and external validation. Deliverable D3.2 is a PDF document which captures a static version of the framework. The live version of the framework is is hosted online on the Charité server. The site is currently available under password protection to the STRATIF-AI consortium[2], and can be made publicly available upon its finalization.

The contents of the first version of the framework were developed based on a series of co-creation workshops (see: Workshops).

The framework will be updated dynamically. Planned updates will follow **(1) a focused stakeholder workshop series**, **(2) upcoming audits** of current framework adherence and **(3) an external Z-inspection** of the AI system.

(1) A focused stakeholder workshop series will be conducted to complete the requirements outlined in this document. More specifically, many requirements have been planned preliminarily, but still lack specific information about parameters, methods, key personnel or planned output. These details will be completed in an iterative process during a forthcoming workshop series.

(2) Audits are planned at five time points during the project, at 6-month increments, to assess adherence to the framework, potential challenges, and suggest changes to resolve issues.

(3) An external Z-inspection will be conducted to assess the trustworthiness of the STRATIF-AI platform. Z-inspection is a published and validated gold standard method by which external experts use sociotechnical scenarios to derive AI trustworthiness requirements. The results of the Z-inspection process will be integrated into the framework.

[1] a document that is continually edited and updated.
[2] username: stratifai, password: Mulch_Snout_Devotion

Regulatory Context

The Framework for Ethical and Trustworthy Design of the STRATIF-AI platform has been designed in concordance with the (1) **EU Guidelines for Trustworthy AI** and (2) **Assessment List for Trustworthy Artificial Intelligence (ALTAI)**.

The EU Guidelines for Trustworthy AI were developed on 8 April 2019 by the High-Level Expert Group (HLEG) on AI. The guidelines aim to promote three key principles to support the development and deployment of safe AI systems—i.e., lawfulness, ethics and robustness. The principles are expounded as seven key requirements (Figure 1).

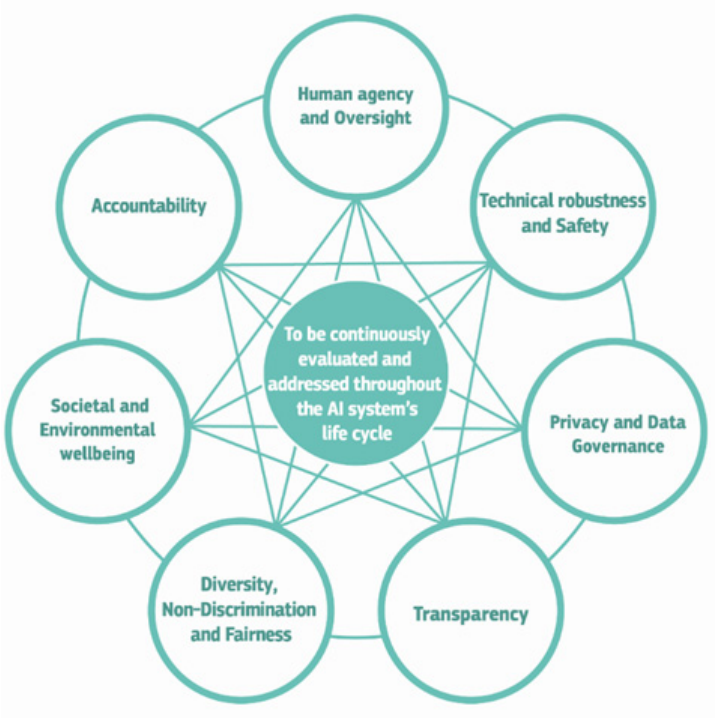

Figure 1: EU Guidelines for Trustworthy AI.

The Assessment List for Trustworthy Artificial Intelligence (ALTAI) was derived on 17 July 2020 by the High-Level Expert Group on Artificial Intelligence (AI HLEG), to aid AI developers in implementing ethical principles in practice. This tool serves as a practical instrument for assessing the compliance of AI systems with the current legal and ethical framework.

Structure

Our framework is divided into seven sections, each corresponding to one of seven requirements set forth by the EU Guidelines for Trustworthy AI:

- Human Agency and Oversight
- Technical Robustness and Safety
- Privacy and Data Governance
- Transparency
- Diversity, non-discrimination, and Fairness
- Societal and Environmental Wellbeing
- Accountability

Under each section, we provide a preliminary assessment following the questions included in the Assessment List for Trustworthy AI (ALTAI) checklist.

We then describe a set of project-specific requirements. Under each project-specific requirement, we outline a set of relevant information that is either planned or *requires revision.*

> **Note**
>
> Under each requirement, we delineate information that requires revision and will be updated following workshops/audits in *italic text*.

Requirement Template

Description We describe each requirement briefly and nominate specific ***parameters*** which need to be expounded or addressed.

parameters: We identify specific parameters within each requirement.

> **Note**
>
> When a parameter has been defined, it is delineated by plain text. If a parameter is yet to be defined, we provide *prompts which will help to define parameters in future workshops, which are delineated by italic text*.

Owner

- Each requirement has an “Owner”, who is responsible for its implementation.

Key Personnel

- Several stakeholders may serve as key personnel to contribute to or facilitate progress towards this requirement. *Named personnel* will be identified in upcoming workshops.

Schedule A preliminary schedule will be devised.

Stroke Phase We identify whether this requirement pertains to all phases of stroke (prevention, acute treatment, rehabilitation) that the STRATIF-AI tool intends to treat.

Actionable tasks

- Here we list a set of tasks

- Which will be executed in order to meet this requirement

Co-creation Workshops

Overview

A series of workshops were conducted with STRATIF-AI consortium members to define issues, requirements, and solutions regarding the design of an ethical and trustworthy AI system. The workshops were designed following feedback from a similar program, which was developed for the EU Horizon project "VALIDATE" and executed in 2023.

Schedule and participants

Workshops were conducted in March 2024 and comprised a diversity of participants from within the STRATIF-AI consortium. The planned duration of each workshop was 3.5 hours; the final workshop duration was between 3 and 3.5 hours. All workshops, with the exception of the first, were facilitated by Riana Minocher (CUB). The first workshop was facilitated by Riana Minocher (CUB) with assistance from Gabriele Pluktaite (CUB). All consortium members who are subscribed to the STRATIF-AI email list were invited to register their participation via a Google form.

Table 1: Date conducted and number of participants

Workshop number	Date conducted	Number of participants
1	05.03.2024	6
2	13.03.2024	4
3	20.03.2024	4
4	27.03.2024	2

Table 2: Details about participants in Workshop 1

Participant name	Consortium partner	Role in STRATIF-AI	Workshop number
Lucia Gregorio	TREE	Leading development of federated learning architecture	1
Elizabeth Hunter	TUD	Programming ML models for prevention	1

Participant name	Consortium partner	Role in STRATIF-AI	Workshop number
Jesper Fellenius	Z2	Development of Personal Data Vault to store patient data	1
Rachele Terragni	FINCB	Clinician	1
Giorgia Camarda	FINCB	Clinician	1
Johan Holsmater	LIU (ext)	Providing input on prevention/health as a lifestyle	1

Table 3: Details about participants in Workshop 2

Participant name	Consortium partner	Role in STRATIF-AI	Workshop number
Silvia Schiavolin	FINCB	Clinical partner/psychologist	2
Serge Timsit	BREST	Clinical studies on hemorrhage and minor stroke	2
Catalina Martínez Costa	UM	Semantic interoperability and data reuse	2
Ville Piispanen	RV	Computer simulation technology	2

Table 4: Details about participants in Workshop 3

Participant name	Consortium partner	Role in STRATIF-AI	Workshop number
Alejandro Garcia Rudolph	GUT	Cognitive and Physical rehab, predictive models, integrate digital twins into cognitive app	3
Gelu Onose		Clinician	3
Daniel Clark	SHE	Project manager D4D	3
Jerome Edmond Bickenbach	SPF	Implementation strategy for STRATIF-AI	3

Table 5: Details about participants in Workshop 4

Participant name	Consortium partner	Role in STRATIF-AI	Workshop number
T. N. Que Nguyen	TUD	Data scientist, working on model development	4

Participant name	Consortium partner	Role in STRATIF-AI	Workshop number
Mark Wright	GUT	Rehabilitation physiotherapist	4

Methodology

The first workshop functioned as a pilot session. Resources/exercises were consequently modified following the first session to accommodate feedback from participants and informal reflections. Participants received an email the day before the workshop with information and instructions on how to participate. Workshops were conducted on Microsoft Teams. Mural was used to disseminate materials and facilitate note-taking during discussions.

Each workshop was divided into three sections:

1. **Introduction and warmup**:

   - Introduction by facilitator to list workshop goals and overview
   - Participant introductions using sticky-notes in Mural
   - Creativity game: to warm up and practice using Mural

2. **Understand concepts**:

   - Free-list in answer to prompt “What ethical concerns come to mind when you think of STRATIF-AI?”
   - Discuss EU requirements meanings/associations
   - Map EU requirements to STRATIF-AI project workflow

3. **Define concrete requirements**:

   - Demo of STRATIF-AI framework site
   - Discuss requirements/solutions within groups

The phase “Understand concepts” was intended to serve as a primer for participants about various ethical issues that have been defined by governing bodies, and to begin to think about how such issues might be related to the STRATIF-AI project. Exposure to requirements and terminology was deemed necessary to guide discussion in the phase “Define concrete requirements”. In this phase, participants were expected to highlight project-specific concerns, which may pertain to their expertise, interest or own personal health history.

Output

General Reflections

The workshops were generally well-received. In Workshop 1, participants discussed topics within their expertise with enthusiasm, but some noted the insufficient time to cover all possible content. The suggestion for follow-up workshops, which could be tailored to specific stakeholder roles, was made. Workshop 2 saw broad awareness of potential risks and discussion of potential abuse of AI within healthcare. Workshop 3 faced some technical challenges, due to difficulties with using Mural effectively, but this did not prevent a serious discussion about risks, roles and responsibilities. Finally, Workshop 4 engaged participants from diverse backgrounds, fostering detailed and productive discussions, particularly about the trade-offs between data protection and model performance. Overall, the workshops provided valuable insights and highlighted the need for ongoing dialogue and clarification on roles and ethical considerations within the STRATIF-AI project.

Key Themes

1. **Data Ownership**: Participants discussed considerations regarding data ownership, re-use, quality, and bias extensively. They emphasized the importance of patient data control, particularly regarding stewardship of data in the event of death or disability, the potential to re-use data for studies, the process of obtaining informed consent for data use, and the novel challenges presented by the data harmonization and federated learning processes.
2. **Transparency**: A significant focus was placed on ensuring transparency in AI systems to empower patients and clinicians, and maintain their trust. Discussions highlighted the need to communicate prediction uncertainties effectively to patients who may not be medical or quantitative experts. Similarly, the challenge of developing transparent models which are accessible to clinicians and patients who have varying goals and experience was discussed.
3. **Patient Empowerment**: Workshops underscored the importance of patient empowerment in the AI-driven healthcare landscape. Participants emphasized the need for patients to have control over their data, and for AI systems to provide understandable and actionable insights to empower informed decision-making. The need to communicate insights without causing harm to mental or emotional wellbeing was identified.
4. **Personalized Prevention**: Workshops emphasized the significance of a user-centric approach, especially in tailoring prevention programs to diversity of end-users. The potential benefits of personalized health coaching through digital twins were explored, to drive behavioral change and contribute to broader societal wellbeing.
5. **A Common Language for Health**: Workshops emphasized the importance of integrating discussions across diverse stakeholder groups and defining a common language for health. Addressing disparities in health literacy and ensuring equitable distribution of AI platforms across communities were highlighted as key considerations.

1 Human Agency and Oversight

The need to ensure respect for human autonomy encompasses the respect for a democratic, flourishing, and equitable society; the prioritization of user agency; and the necessity of human oversight for activities/uses of the AI system. This requirement is particularly important for AI systems which guide or influence human decision-making—and is thus of central importance to the STRATIF-AI project.

Keywords: guiding, influencing, decision-making (e.g., prediction systems, predictive policing, decision-support); human-like; trust and (in)dependence

Definitions: human-in-the-loop (HITL): capability for human intervention in every decision cycle of the system human-on-the-loop (HOTL): capability for human intervention during the design cycle of the system and monitoring the system's operation human-in-command (HIC): 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 AI system in any particular situation—e.g., override decision or discontinue use

1.1 Human agency and autonomy

ALTAI assessment

STRATIF-AI is intended to guide decisions made by human end-users—i.e., both doctors and patients—, which are in turn projected to impact end-users (patients' health and autonomy, as well as doctors' autonomy) and society in general. To ensure STRATIF-AI does not generate confusion about the extent to which decision-making is guided by algorithms and the allocation of epistemic authority, a plan to disseminate information in an accessible manner to a diversity of end-users will be necessary. The risk of interpretation of a digital twin or "personalized health coach' ' as a human rather than AI system is generally believed to be low, but a risk assessment needs to be made to rule out this possibility. The risk of addiction or unhealthy attachment to the system are again deemed low, but measures to mitigate or minimize the risk of such activities need to be developed. The risk of manipulation needs to be investigated.

1.1.1 Define epistemic authority

Description In the events where there is a ***discrepancy*** 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*** must be described.

discrepancy: difference in recommended course of treatment; difference in output *from existing medical device*; ***physician expertise level:*** early; proficient; expert

Owner

- WP4
- WP5
- WP6

Key Personnel

- medical staff (TBD)
- design staff (TBD)

Schedule At any point, before end of project.

Stroke Phase EACH (3)

Actionable tasks

- appoint key personnel
- define discrepancy
- clarify if comparison of AI and clinician will be conducted in clinical study
- prepare process protocol outline
- define physician expertise levels
- completion of process protocol
- feedback from medical staff
- publication/research design about epistemic authority

1.1.2 Communicate prediction uncertainty

Description A ***methodology*** for measuring and communicating quantitative prediction uncertainty to ***end-users*** must be established and ***validated***.

valid methodology: *TBD*; *set of ways in which predictions will be shown in platform*
end-users: medical professionals, patients ***validated:*** validated in clinical setting; validated in external clinical setting; validated through feedback from experts

Owner

- WP2

Key Personnel

- medical staff (TBD)
- technical staff
- design staff

Schedule At any point, before end of project.

Stroke Phase EACH (3)

Actionable tasks

- appoint key personnel
- define methods to communicate prediction interval
- validation of methods
- feedback from medical staff
- feedback from patients and patient representatives
- development of front-end communication for patients with specific conditions, age, education
- publication of validation and methodology

1.1.3 Foster patient autonomy

Description A standardized ***procedure*** for explaining ***risks*** of oversight customized for patient ***sub-groups*** must be designed.

procedure: video; written material; training protocol for medical staff;
risks: procedures for epistemic authority; HIC governance structure; AI system overview
subgroups:: *speaking different languages; different education levels; different age groups; degree of disability*

Schedule At any point, before end of project.

Stroke Phase EACH (3)

Owner

- WP7

Key Personnel

- patient representatives
- technical staff (TBD)
- design staff (TBD)

Actionable tasks

- appoint key personnel
- draft documentation
- feedback from patients/patient representatives
- translation of documentation
- adaptation in simple and complex format
- adaptation for youth
- adaptation for visually impaired
- adaptation for mentally impaired
- assessment by disability experts
- additional feedback from patients/patient representatives

1.1.4 Establish degree of trust

Description Degree of ***trust*** in the STRATIF-AI platform, determined by ***end-users***, should be quantified.

trust: *operational definitions of trust* ***end-users:*** medical professionals; patients with stroke; patients at risk of stroke; patients without stroke

Schedule At any point, before end of project.

Stroke Phase ALL

Owner

- WP3

Key Personnel

- patient representatives
- medical staff

Actionable tasks

- appoint key personnel
- define degrees of trust or conception of trust
- design questionnaire/interview to survey trust of system
- pre-register survey
- execute survey with key personnel
- report outcomes
- translate outcomes for end-users in platform
- disseminate outcomes in STRATIF-AI pilot studies and solicit feedback
- publication on trust in digital twins

1.2 Human oversight

ALTAI assessment

STRATIF-AI must be constructed as a Human-in-Command system. Medical professionals need to be trained to exercise oversight according to a process protocol that has been validated. A system to detect and report misuse or adverse effects of the system must be constructed. STRATIF-AI is a self-learning system; autonomy in predictions needs to be specified to ensure human oversight/control over the type of prediction generated.

1.2.1 Establish HIC governance

Description Tool is designed with ***capabilities*** to enable a Human in Command (HIC) governance structure.

capbilities: oversee the overall activity of the AI system; ability to decide whether, when and how to use the system in any particular situation; ability to override a decision made by a system

Owner

- WP3

Key Personnel

- WP7
- design staff (TBD)

Schedule At any point, before end of project.

Stroke Phase ALL

Actionable tasks

- appoint key personnel
- create governance structure draft
- disseminate structure
- feedback from medical staff and implementation staff
- publication/research design about epistemic authority

1.2.2 Institute reporting feedback loop

Description A system for ***end-users*** to make reports about ***discrepancies*** should be established within the platform and ***periodically*** be evaluated.

end-users: medical professionals, patients
discrepancies: disagreement between system recommendation and end-user decision; end-user discomfort with system recommendation; indication of errors
periodically: *TBD*

Stroke Phase ALL

Owner

- WP2

Key Personnel

- design staff
- technical staff

Actionable tasks

- appoint key personnel
- determine evaluation schedule
- design and validate system

1.2.3 Manage patient expectations

Description In the events where there is a ***discrepancy*** between the predictions of the tool and the perspective or opinion of the patient, a process protocol for managing expectations, customized for patients of different ***subgroups*** must be described.

discrepancy: *acceptable threshold for difference* ***subgroups:*** *stroke risk category; age groups; education levels; degree of disability*

Owner

- WP4
- WP5
- WP6

Key Personnel

- medical staff

Schedule At any point, before end of project.

Stroke Phase EACH (3)

Actionable tasks

- completion of process protocol
- report number of discrepancies
- feedback on process protocol from medical staff
- feedback on process protocol from patient representatives

> **Note**
>
> This requirement is closely related to 1.1.1; 1.1.2; and 1.2.2—this requirement is tailored to end-users who are patients. Requirement 1.2.2 will enable reporting discrepancies for patients, and the execution of 1.1.2 will inform the process protocol for this requiremnt.

1.2.4 Enable autonomy of end-users

Description ***End-users*** should have sufficient control over platform ***parameters*** to foster independent use and trust in the system predictions.

end-users: medical staff, patients
parameters: predictive model type; *data to be used for predictions; data to be ommitted for predictions;*

Owner

- WP2

Key Personnel

- medical staff (TBD)
- technical staff (TBD)
- design staff (TBD)

Schedule At any point, before end of project.

Stroke Phase EACH (3)

Actionable tasks

- appoint key personnel
- determine parameters which can be controlled at end-use
- design interface to select parameters
- feedback from medical staff
- feedback from patients

2 Technical Robustness and Safety

AI systems must be technically robust to prevent harm, unintentional or otherwise, to human dignity, mental and physical integrity. The system must be guard against vulnerability to attack by third parties which could compromise its integrity or alter its behaviour. Safety procedures to enact in the event of errors or risks must be instituted. The accuracy of the system is of critical importance where the case for AI-assisted medical decision-making. Finally, reproducibility and reliability help to ensure that the performance and behaviour of AI system can be critically assessed.

Keywords: accuracy, AI Bias, AI System, AI Reliability, AI Reproducibility, (Low) Confidence Score, Continual Learning, Data Poisoning, Model Evasion, Model Inversion, Pen Test, Red-team

2.1 Resilience to attack and security

ALTAI Assessment

In the event of technical faults, misuse or defect, STRATIF-AI could have damaging effects to human safety. The risk of vulnerability to cyber-attacks will be evaluated to ensure the system is made compliant with security standards under the Cybersecurity Act in Europe. Vulnerability to data poisoning, model evasion, model inversion will be addressed throughout model validation and testing. Measures to ensure robustness to attacks after the system has been deployed will be designed. The expected timeframe of validity for security measures need to be defined.

2.1.1 Define trade-offs in Federated Learning

Description Measures to assess whether federated learning exposes patients to ***risks*** must be established, assessed and reported.

risks: model inversion (information in parameters to reconstruct data); data leakage; data poisoning

Owner

- WP1

Key Personnel

- technical staff (TBD)

Schedule During model training.

Stroke Phase ALL

Actionable tasks

- appoint key personnel
- define list of risks and parties
- document measures to assess risks
- verify metrics are sufficient to dissuade concerns about vulnerability
- publish validation study

2.1.2 Comply with cybersecurity law

Description The system must be made compliant with the ***relevant cybersecurity legislation***.

relevant cybersecurity legislation: EU; country-specific; global

Owner

- WP7

Key Personnel

- technical staff (TBD)

Schedule At any point, before end of project.

Stroke Phase ALL

Actionable tasks

- appoint key personnel
- define relevant legislation/ compliance categories

2.1.3 Establish emergency protocols

Description In the event of a breach to security, a process protocol based on the ***severity*** and ***type*** of attack must be followed.

severity: *define metrics to measure security breach, and scale of severity* ***type:*** data poisoning; model evasion; *define types*

Owner

- WP2

Key Personnel

- technical staff (TBD)
- WP7

Schedule At any point, before end of project.

Stroke Phase ALL

Actionable tasks

- appoint key personnel
- define severity levels for attack
- define parameters for process protocols
- external validation of process protocols

2.2 General safety

ALTAI Assessment

The risks associated with the AI system for each use-case will be assessed through pilot and clinical studies. Procedures to continuously measure and access risks will be devised, validated and implemented. A plan to inform end-users of existing and potential risks will be disseminated. STRATIF-AI is a novel concept and technology; possible threats and vulnerability to misuse will additionally be explored through a systematic literature review of ethical dilemmas regarding the use of digital twins in healthcare. Stability and reliability of the system will continually be assessed throughout model development. A plan to regularly evaluate the system, upon changes to technical infrastructure, will also be devised.

2.2.1 Assess risk of use

Description An independent ***risk assessment*** of use of the STRATIF-AI platform for different ***end-users*** must be carried out.

risk assessment: *a standardized risk assessment must be designed* ***end-users:*** patients; patients with stroke; patients at risk of stroke; patients without stroke; *language; education; age; disability; medical professionals; doctors; physiotherapists; psychologists; care-workers; proficiency-levels; learner; proficient; expert*

Owner

- WP4
- WP5
- WP6

Key Personnel

- medical staff (TBD)
- patients/patient representatives (TBD)
- WP3

Schedule At any point, before end of project.

Stroke Phase EACH (3)

Actionable tasks

- appoint key personnel
- define end-users and risk assessment framework
- develop protocol for clinical study/pilot study
- conduct risk assessment
- publish risk assessment

2.2.2 Implement risk-assessment feedback loop

Description A system to assess ***risk*** for ***different types of end-users*** should be established within the platform and ***periodically*** be evaluated.

risk: threat to safety; threat to validity; risk of misuse; issues with reliability; *additional risks?*
end-users: medical professionals; patients ***periodically:*** *TBD*

Stroke Phase ALL

Owner

- WP2

Key Personnel

- design staff
- technical staff
- WP3

Actionable tasks

- appoint key personnel
- determine evaluation schedule
- design and validate system

> **Note**
>
> A system for feedback reporting, with both automated and non-automated components, must be designed and integrated into the STRATIF-AI platform, to address several requirements together, including this one.

2.2.3 Communicate risks to patients

Description A standardized procedure for explaining ***technical risks*** of the system customized for patients of different ***subgroups*** must be designed.

technical risks: vulnerability to misuse; technical robustness metrics ***patients:*** language capabilities; educational experience; age groups; degree/type of disability;

Schedule At any point, before end of project.

Stroke Phase EACH (3)

Owner

- WP7

Key Personnel

- patient representatives
- technical staff (TBD)
- design staff (TBD)
- WP3

Actionable tasks

- appoint key personnel
- define technical risks
- draft documentation
- feedback from patients/patient representatives
- translation of documentation
- adaptation in simple and complex format
- adaptation for youth
- adaptation for visually impaired
- adaptation for mentally impaired
- assessment by disability experts
- additional feedback from patients/patient representatives

> **Note**
>
> To address several requirements, including this one, a comprehensive set of patient subgroups will need to be defined, which should include all potentially vulnerable or marginalized patient profiles.

2.2.4 Process protocol for technical updates

Description A process protocol for assessing ***system functions*** following any technical developments or changes to infrastructure should be designed and followed.

system functions: TBD; validity; risks to safety; evaluation metrics

Schedule At any point, before end of project.

Stroke Phase ALL

Owner

- WP2

Key Personnel

- technical staff (TBD)
- design staff (TBD)

Actionable tasks

- appoint key personnel
- define regular system function metrics
- design process protocol
- feedback from experts/ external review
- dissemination of protocol

2.3 Accuracy

ALTAI Assessment

Effective and safe use of the STRATIF-AI platform hinges upon a high level of accuracy. As such, various measures to ensure data is of high quality, representative, and continuously monitored for accuracy will be put in place. The external and internal validity of the system will similarly be established and periodically reviewed. Levels of accuracy and validity established for the STRATIF-AI platform will clearly be communicated to end-users.

2.3.1 Evaluate data accuracy within Federated Learning

Description The accuracy of the Federated Learning system under different ***scenarios*** must be analyzed and monitored ***periodically***.

scenarios: *missing data; data interoperability; heterogeneity of data; updates to data types*
periodically: *TBD*

Schedule At any point, before end of project.

Stroke Phase ALL

Owner

- WP1

Key Personnel

- technical staff (TBD)

Actionable tasks

- appoint key personnel
- define scenarios to assess validity
- define periodic intervals for assessment
- design system

2.3.2 Establish internal validity

Description The behaviour of the tool, based on a set of core ***performance metrics*** will not differ under different ***contexts***.

performance metrics: *sensitivity; specificity; area-under-curve; error reporting frequency*
data contexts: *clinical settings*

Schedule At any point, before end of project.

Stroke Phase EACH (3)

Owner

- WP2

Key Personnel

- technical staff (TBD)
- WP3
- patients/patient representatives

Actionable tasks

- verify whether this is planned in clinical studies

- appoint key personnel
- define performance metrics
- define data contexts to assess
- design study protocol
- feedback from patient representatives
- validate in pilot study
- publish validation

2.3.3 Establish external validity

Description The ***output*** of the models should not differ across end-users of different ***subgroups***.

output: *prediction types* ***subgroups:*** *patient subgroups; geography; ethnicity; gender; socio-economic background; health history; disability;*

Schedule At any point, before end of project.

Stroke Phase EACH (3)

Owner

- WP2

Key Personnel

- technical staff (TBD)
- WP3
- WP7
- patient representatives

Actionable tasks

- appoint key personnel
- define patient subgroups
- define performance metrics
- design study for external validation
- feedback on study design
- publication of external validity

2.4 Reliability and reproducibility

ALTAI Assessment

Reliability and reproducibility will play a crucial role along the entire STRATIF-AI workflow. Reproducibility metrics will be defined and evaluated periodically. Methods of verification of reproducibility/reliability

will clearly be documented. A procedure to track errors and trace their source will be developed. A continual learning system to adapt system performance to evolving medical guidelines or diverse medical principles will also be instated.

2.4.1 Track system errors

Description A system to track, trace, and report ***errors*** within the platform must be implemented.

errors error type; TBD

Schedule At any point, before end of project.

Stroke Phase ALL

Owner

- WP2

Key Personnel

- technical staff (TBD)
- design staff (TBD)

Actionable tasks

- appoint key personnel
- define error reporting metrics
- design error reporting system
- verification by technical experts
- publication (internal) of system
- publication of reports

2.4.2 Ensure continuous adaptation

Description Measures to ensure that the system is able to adapt to ***evolving treatment guidelines*** while maintaining ***predictive accuracy*** will be taken.

This involves addressing data scarcity challenges and handling variations in healthcare practices, particularly in countries with different approaches to guideline adherence.

evolving treatment guidelines: variation in healthcare practice; countries with different guidelines; changing regulations; ***predictive accuracy:*** *acceptable range TBD*

Schedule At any point, before end of project.

Stroke Phase EACH (3)

Owner

- WP7

Key Personnel

- technical staff (TBD)
- design staff (TBD)
- WP3

Actionable tasks

- appoint key personnel
- define boundaries for predictive accuracy/acceptable range
- define evolving treatment guidelines
- develop process protocol for evaluation

> **Note**
>
> Several requirements, including this one, refer to the *predictive accuracy* of the output; this parameter will be defined through the implementation of requirement 1.1.2 Communicate prediction uncertainty.

3 Privacy and Data Governance

To prevent undue harm to society, the fundamental right to privacy must be protected. Data-driven AI systems like STRATIF-AI may pose substantial risks to individuals' privacy, and as such, a comprehensive data governance protocol is necessary to maintain integrity and privacy of data.

Relevant resources: https://gdpr.eu/data-protection-impact-assessment-template/. https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation- gdpr/accountability-and-governance/data-protection-officers/.

3.1 Privacy

ALTAI Assessment

In the event of data leakage or misuse, STRATIF-AI could have significant impacts on the rights to privacy, physical, mental and moral integrity, and data protection. A process protocol to report violations to privacy or data integrity must be developed and implemented.

3.1.1 Investigate right to privacy within Federated Learning

Description The Federated Learning platform transfers model parameters rather than data; ***vulnerabilities*** of the system for data privacy must be ***quantified***.

vulnerabilities: potential to reconstruct training data; glean private information; model inversion
quantified: *levels of security to be identified*

Owner

- WP1

Key Personnel

- WP2
- modelling staff (TBD)

Schedule At any point, before end of project.

Stroke Phase ALL

Actionable tasks

- appoint key personnel
- identify potential vulnerabilities
- identify metrics to quantify scale/severity of vulnerability
- assess performance
- publish results

3.1.2 Allocate data usage and access rights

Description Design for different ***data types*** detailing ***access*** and ***safeguards*** for ***vulnerable patient groups*** must be made available within the Data Management Plan.

data types: training data; Personal Data Vault
access: rights to access; location of storage; potential re-use; encryption
safeguards: stewardship upon death; stewardship in event of disability; stewardship of carers/wards
vulnerable patient groups: *speaking different languages; age groups; ethnicity; disability*

Owner

- WP1

Key Personnel

- WP7 (TBD)
- patient representatives (TBD)
- WP4 (TBD)
- WP5 (TBD)
- WP6 (TBD)

Schedule At any point, before end of project.

> **Note**
>
> The STRATIF-AI project includes a Data Management Plan (DMP) as a recurring deliverable. This requirement will be addressed within the scope of upcoming DMPs.

Stroke Phase EACH (3)

Actionable tasks

- appoint key personnel
- review current data management plan and details
- identify patient subgroups; access types; safeguards
- feedback on DMP from patient representatives
- feedback from medical professionals
- make plan publicly and internally available/within apps/platform

3.1.3 Implement feedback system

Description A system to flag violations to privacy or data integrity must be implemented within the platform and a corresponding process protocol drawn up.

Owner

- WP2

Key Personnel

- design staff
- WP1 (TBD)

Schedule At any point, before end of project.

Stroke Phase ALL

Actionable tasks

- appoint key personnel
- design system
- draft process protocol
- feedback on prototype by end-users

3.2 Data governance

ALTAI Assessment

STRATIF-AI is trained using personal healthcare records and will be continually informed by private and securely transferred data within the Personal Data Vault. A Data Protection Impact Assessment will be executed, relevant Data Protection Officers will be appointed, and the requirements under General Data Protection Regulation will be adhered to. In addition, mechanisms to ensure oversight of data processing, data transformation and data harmonization within the Federated Learning environment will be drafted. Data minimization possibilities will need to be investigated. The right to withdraw consent, right to object, and right to be forgotten will be revised and communicated to end-users. The AI system will be aligned with relevant standards and widely adopted protocols within the relevant regulatory framework.

3.2.1 Assess data quality

Description A quality control system to ensure different ***data types*** fulfill ***quality critera*** will be implemented and made available to ***end-users*** before use for model training or prediction.

data type: training data; Personal Data Vault
quality criteria: robust to missing data; robust to inaccuracies; robust to input errors; *relevant thresholds*
end-users: patients; medical professionals

Owner

- WP1

Key Personnel

- design staff (TBD)
- technical staff (TBD)
- WP1 (TBD)

Schedule At any point, before end of project.

Stroke Phase ALL

Actionable tasks

- appoint key personnel
- define quality control metrics and acceptable thresholds
- update any existing process protocols
- produce quantitative assessments for existing data/prototype data
- translate assessment for end-users

3.2.2 Standardize Federated Learning data procedures

Description The ***procedures*** involved in manipulating training data for Federated Learning across project ***partners*** must be standardized.

Standardise the complex procedures and protocols involved in federated learning across project organizations and countries. Ensure that data recording, preprocessing, evaluation, forecasting, validation, and ML monitoring processes are normalized and follow similar guidelines, even in countries with potentially different regulatory frameworks. Establish shared schemas and protocols to facilitate effective federated learning outcomes.

procedures: pre-processing; evaluation; forecasting; validation; ML monitoring
partners: *each hospital/data contributor*

Owner

- WP1

Key Personnel

- technical staff (TBD)
- WP4
- WP5
- WP6

Schedule At any point, before end of project.

Stroke Phase ALL

Actionable tasks

- appoint key personnel
- define set of procedures
- identify partner set and any potential need for discrepancies/specific operationalization
- standardize procedures through iterative process with partners
- audit adherence
- report adherence and publish procedures

3.2.3 Standardize data harmonization procedures

Description The ***parameters*** of the harmonization process for making data interoperable must be defined and be reviewed ***periodically***.

parameters: drift detection; retraining procedure;
periodically: *TBD*

Owner

- WP1

Key Personnel

- technical staff (TBD)
- WP4
- WP5
- WP6

Schedule At any point, before end of project.

Stroke Phase ALL

Actionable tasks

- appoint key personnel
- define parameters for harmonization process
- create SOP for harmonization so this information is available to partners
- feedback from partners

> **Note**
>
> During co-creation workshops, a lack of clarity about the data harmonization process when the STRATIF-AI platform is operational emerged. How the harmonization process will be governed after the model training procedure is complete must be clarified within this requirement.

3.2.4 Acquire informed consent

Description A process protocol for obtaining consent must be drafted, which outlines ***access rights***, ***restrictions*** once consent has been provided, and is tailored to specific ***patient subgroups***.

access rights: TBD; which physicians/medical staff will have access; stewardship in event of disability;
patient subgroups: language; age; education; disability
restrictions: TBD; withdraw consent; right to object; right to be forgotten; how will models be retrained if consent is withdrawn;

Owner

- WP4
- WP5
- WP6

Key Personnel

- WP3
- WP1 (TBD)
- medical staff
- patient representatives

Schedule At any point, before end of project.

Stroke Phase EACH (3)

Actionable tasks

- appoint key personnel
- draft consent forms
- define parameters for forms
- tailor to each stroke phase
- feedback from patient representatives

3.2.5 Adhere to GDPR

Description ***Relevant requirements*** under General Data Protection Regulation must be adhered to. ***Relevant requirements:*** *the applicability of requirements must be translated for the STRATIF-AI platform and included in the DMP*

Owner

- WP1

Key Personnel

- *TBD*

Schedule At any point, before end of project.

Stroke Phase ALL

Actionable tasks

- check if this is part of DMP
- create additional document providing translation for patients/end-users

3.2.6 Investigate data minimization

Description The ***costs*** of collecting intimate data from patients must be balanced against violation of patient trust, and validated by appropriate ***methodology***.

costs: *perceived harm; erosion of trust* ***methodology:*** *cost-benefit analysis; qualitative judgement; Delphi consensus;*

Owner

- WP3

Key Personnel

- WP4 (TBD)
- WP5 (TBD)
- WP6 (TBD)
- patient representatives

Schedule At any point, before end of project.

Stroke Phase EACH (3)

Actionable tasks

- incorporate in trust study
- research about potential data minimization
- elicit patient/expert feedback on costs/benefits of highly invasive data

4 Transparency

A key requirement for ethical AI is transparency. Transparency includes traceability of the AI system and its decisions, i.e., all information and procedures which led to the system's predictions. Such traceability is essential to enable the identification of errors and improvement of future results. Traceability is closely linked to explainability, which is the ability to explain the technical processes and the human oversight procedures in place within the AI system. STRATIF-AI is a complex CDSS involving numerous technological solutions and a HIC governance process—the limits of which should be clearly communicated to end-users. High explainability may come at the cost of system accuracy: if increases in accuracy necessitate a high degree of complexity. The degree to which AI predictions affect human decisions should be balanced with the ability to explain the rationale behind such decisions. Finally, the the degree to which human decisions depend on AI predictions, and the underlying limitations of the system and its predictions, must be communicated in an interpretable manner to a diversity end-users.

Keywords:
AI System, end-user, explicability, lifecycle, subject, traceability, workflow of the model

4.1 Traceability

ALTAI Assessment

Traceability is difficult to achieve for complex data-driven AI systems. STRATIF-AI will be engineered with measures to enable traceability of the system as far as possible. The ability to trace back data, data quality, and decision rules of the system will be quantified, assessed, and reported to end-users. The decisions of the system and the associated "quality" of such decisions will be continuously monitored.

4.1.1 Enable system traceability

Description The ability to trace back key ***system parameters*** will be ***assessed*** systematically and translated for a diversity of ***end-users***.

system parameters: data source; data demographics; patient-specific data; model parameters; model calibration metrics;
assessed: *a system within the platform to report data* ***end-users:*** medical professionals; patients;

Owner

- WP2

Key Personnel

- technical staff (TBD)
- medical staff (TBD)
- patient representatives

Schedule At any point, before end of project.

Stroke Phase EACH (3)

Actionable tasks

- appoint key personnel
- define system parameters that should be identified
- design platform/system to assess/enable info
- feedback from end-users on system and params
- translation of outputs
- feedback from end-users on interpretability of output
- publication of traceability system

4.1.2 Quality control of predictions

Description The ***decisions*** of the system must be ***systematically compared*** with ***human feedback*** to infer prediction ***quality***.

decisions: predictions; prediction intervals; data output
systematically compared: *design of study comparison* ***human feedback:*** *discrepancy with medical decision; patient discomfort;*
quality: *construct a quality score*

Owner

- WP2

Key Personnel

- WP3
- technical staff (TBD)
- medical staff (TBD)
- patient representatives

Schedule At any point, before end of project.

Stroke Phase EACH (3)

Actionable tasks

- appoint key personnel
- define decisions to be recorded and feedback format
- design platform within app to do so
- construct quality score

- study design and preregistration
- publication of quality of decisions/ update to system

> **Note**
>
> This requirement is closely connected with requirements under Human Agency and Oversight.

4.2 Explainability

ALTAI Assessment

STRATIF-AI is a clinical decision support-system. Both doctors and medical professionals must be provided with a reasonable understanding of the decisions made by the system.

4.2.1 Provide decision-making parameters to end-users

Description Information about the relevant ***parameters*** that led to an algorithmic decision will be validated and available in a ***simplified format*** to ***end-users***.

simplified format: video; written material; ***parameters:*** modelling parameters; training data parameters; Data Vault parameters;
end users: medical professionals; patients;

Owner

- WP2

Key Personnel

- technical staff (TBD)
- design staff
- medical staff (TBD)
- patient representatives

Schedule At any point, before end of project.

Stroke Phase EACH (3)

Actionable tasks

- appoint key personnel
- define parameters
- interview/feedback from key personnel
- create in-app design
- feedback on prototype in pilot study

4.2.2 Monitor explainability

Description Establishing the explainability of the STRATIF-AI platform could serve as a blueprint for digital-twin based/personalized medicine projects. The ***satisfaction*** of ***end-users*** with the explainability of the platform over the course of its use should be continuously monitored.

satisfaction: TBD; means to assess suitability of explanatory materials/information; combined with feedback reporting loop;
end-users: TBD; medical professionals; experience levels; patients; education; age; language; disability;

Owner

- WP2

Key Personnel

- WP7
- technical staff (TBD)
- medical staff (TBD)
- patient representatives

Schedule At any point, before end of project.

Stroke Phase EACH (3)

Actionable tasks

- appoint key personnel
- decide on study or in-app reporting system
- define satisfaction criteria
- design and validate system
- combine with feedback reporting loop?

4.2.3 Assess costs of model explainability

Description The costs of explainability for the model parameters and decision-rules will be traded off against their accuracy. During model development and training such ***costs*** should be systematically assessed and recorded.

costs: TBD; format to quantify explainability versus accuracy tradeoff;

Owner

- WP2

Key Personnel

- WP3
- technical staff (TBD)

Schedule At any point, before end of project.

Stroke Phase ALL

Actionable tasks

- appoint key personnel
- define explainability metrics
- design system to record/assess
- is this part of existing modelling protocol
- publication of explainability/accuracy tradeoff assessment
- translation of materials for end-users

4.3 Communication

ALTAI Assessment

There is little risk of the AI system being interpreted as a human. However, the HIC governance concept must be clearly communicated to end-users. The benefits, technical limitations, potential risks must additionally be communicated to end-users in a transparent and explainable manner. A comprehensive training protocol for end-users—particularly medical personnel—must be developed in collaboration with the system developers. ### Standardize training materials

Description The recommendations for ***proper use*** of the STRATIF-AI must be ***translated*** for a diversity ***end-users***.

proper use: TBD; technical limitations; potential risks; error rates; evidence-based health benefits;
translated: TBD; written manual; visual (presentation); flyer; video
end-users: TBD; medical professionals; experience levels; patients; education; age; language; disability;

Owner

- WP7

Key Personnel

- technical staff (TBD)
- medical staff (TBD)
- patient representatives

Schedule At any point, before end of project.

Stroke Phase EACH (3)

Actionable tasks

- appoint key personnel
- define proper use parameters
- interview/feedback from key personnel

- make prototype written and flyer materials
- outsource video/audio materials
- translation for disability, language, education, age
- publication of materials

4.3.1 Tailor explainability material

Description An overview of STRATIF-AI system ***functions*** will be communicated in a diversity of ***formats***, designed specifically for different groups of ***end-users***.

functions: TBD; modelling framework; modelling maps; decision-making process; training data; private data vault; security protocols; human oversight; evidence of efficacy; rationale for using tool;
formats: TBD; written manual; visual (presentation); flyer; video
end-users: TBD; medical professionals; experience levels; patients; education; age; language; disability;

Owner

- WP3

Key Personnel

- WP7
- technical staff (TBD)
- medical staff (TBD)
- patient representatives

Schedule At any point, before end of project.

Stroke Phase EACH (3)

Actionable tasks

- appoint key personnel
- define parameters
- interview/feedback from key personnel
- refine content
- make prototype written and flyer materials
- outsource video/audio materials
- translation for disability, language, education, age
- publication of materials

> **Note**
>
> Requirements 4.3.1 and 4.3.2 are closely related. Requirement 4.3.1 involves the creation of training materials for use of the platform, i.e., serving as the guide for use, while Requirement 4.3.2 will deliver the explainability materials, i.e., serving as a background or primer on the AI system.

5 Diversity, Non-discrimination and Fairness

To ensure the trustworthiness of the STRATIF-AI platform, priority should be given to inclusion and diversity at every stage of the project lifecycle. Efforts to address historical biases, incompleteness, and governance models to mitigate the risk of unintended prejudice and discrimination should be made. Moreover, beginning with the identification and mitigation of bias from the outset is essential. Additionally, the design of STRATIF-AI systems should adhere to a user-centric approach, facilitating accessibility for all individuals, irrespective of demographic factors. This will include ensuring accessibility for persons with disabilities across diverse societal groups.

Keywords:
AI bias, AI system, AI designer, AI developer, accessibility, assistive technology, end-user, fairness, subject, universal design, use case

5.1 Avoidance of unfair bias

ALTAI Assessment

To mitigate inadvertent creation of, or perpetuation of, biased outcomes within the AI system, input data selection and algorithmic design need to be scrutinized for bias. Throughout development, consideration will be given to the diversity and representativeness of end-users and subjects within training data. Testing protocols will be tailored to rectify potential biases, to ensure integrity and impartiality of the platform across user demographics and application scenarios. Educational initiatives aimed at cultivating awareness among AI designers and developers regarding bias and its ethical ramifications will be initiated, fostering a culture of accountability. Mechanisms for identifying and resolving issues pertaining to bias or discrimination within the STRATIF-AI platform will be established within clear communication channels. Discourse with impacted communities, particularly at-risk patient subgroups, will be conducted to ensure that the STRATIF-AI project consortium's conception of fairness resonates with diverse perspectives. A quantitative analytical framework will be embraced to measure and evaluate the efficacy the applied definition of fairness, and institute mechanisms to achieve as such.

5.1.1 Represent diversity in training data

Description Inclusive representation of patients of specific (under-represented) ***subgroups*** in training data will be assessed against an acceptable ***threshold*** and rectified or reported accordingly.

subgroups: patients; patients with stroke; patients at risk of stroke; patients without stroke; *language; education; age; disability; ethnicity; gender; migration status; socio-economic status;* ***threshold:*** *reasonable representation; if zero; report in prediction uncertainty*

Owner

- WP4
- WP5
- WP6

Key Personnel

- medical staff (TBD)
- patients/patient representatives (TBD)
- WP3

Schedule At any point, before end of project.

Stroke Phase EACH (3)

Actionable tasks

- appoint key personnel
- define at-risk patient subgroups
- define acceptable thresholds
- feedback from patient representatives/experts
- assess training data
- report in publication
- note challenge of oversimplified categorization

5.1.2 Monitor bias

Description A ***system*** to ***periodically*** monitor bias in different project ***stages*** must be implemented.

system: TBD; within app for end-users; during model development;
periodically: TBD;
stages: training data; modelling; federated learning;

Owner

- WP2

Key Personnel

- technical staff (TBD)
- WP4
- WP5
- WP6
- WP3

Schedule At any point, before end of project.

Stroke Phase ALL

Actionable tasks

- appoint key personnel
- create bias testing pipelines
- define parameters
- send pipelines to necessary spaces (e.g., clinical partners)
- pilot bias testing in preliminary study
- report output
- publication of tool/pipeline

5.1.3 Assess prediction bias

Description System ***outputs*** will be assessed to ensure there is no ***difference*** between patient ***sub-groups***.

outputs: model predictions; discrepancy
difference: *quantitative threshold; acceptable margin of error;*
subgroups *gender; sex; ethnicity; education; health history; disability; socio-economic background*

Owner

- WP2

Key Personnel

- technical staff (TBD)
- WP3

Schedule At any point, before end of project.

Stroke Phase ALL

Actionable tasks

- appoint key personnel
- define parameters model outputs
- define patient subgroups
- define thresholds
- feedback on thresholds
- publication of results

5.1.4 Implement bias reporting system

Description A system for ***end-users*** to make reports about ***bias*** should be established within the platform and ***periodically*** be evaluated.

end-users: medical professionals, patients
bias: *any discrimination; process protocol for what constitutes bias; response;*
periodically: *TBD*

Owner

- WP2

Key Personnel

- technical staff (TBD)
- design staff

Schedule At any point, before end of project.

Stroke Phase ALL

Actionable tasks

- appoint key personnel
- determine evaluation schedule
- outline process protocol for updates upon substantial bias
- design and validate system

5.1.5 Cultivate ethical awareness

Description A ***strategy*** to ensure awareness about ethical issues and accountability for the system, tailored to specific ***stakeholders***, will be designed and executed.

strategy: workshop series; training programme; accountability declaration;
stakeholders: developers; doctors; clinicians; care-workers; patients; family members; technical designers; project managers;

Owner

- WP3

Key Personnel

- all relevant leads

Schedule At any point, before end of project.

Stroke Phase ALL

Actionable tasks

- write up workshop paper
- create training package materials
- interview study with experts
- design accountability declaration for project partners
- publish memorandum

5.1.6 Assess fairness across use-setting

Description The fairness of system ***outputs*** across different ***use-settings*** needs to be quantitatively evaluated to minimize risk of unfair prediction values for specific locations/services.

outputs: *prediction intervals; prediction types; accessibility; speed;*
use-settings: hospitals; homes; resource differences; medical equipment; medical staff;

Owner

- WP7

Key Personnel

- WP4
- WP5
- WP6
- WP3

Schedule At any point, before end of project.

Stroke Phase ALL

Actionable tasks

- appoint key personnel
- design study to assess fairness across contexts
- define parameters to study
- pre-register study
- pilot with prototype of tool or collect data within clinical trials
- publication of results

5.1.7 Assess and build trust within vulnerable patient groups

Description Degree of ***trust*** in the STRATIF-AI platform, determined by ***end-users***, will be quantified. An assessment of trust within specifically vulnerable patient and care-work ***subgroups*** will be assessed and measures to respond to discrepancies between subgroups devised.

trust: *operational definition of trust* ***end-users:*** medical professionals; patients with stroke; patients at risk of stroke; patients without stroke
subgroups: *language; education; age; disability; ethnicity; gender; migration status; socio-economic status;*

Schedule At any point, before end of project.

Stroke Phase ALL

Owner

- WP3

Key Personnel

- patient representatives
- medical staff

Actionable tasks

- appoint key personnel
- define degrees of trust or conception of trust
- design questionnaire/interview to survey trust of system
- pre-register survey
- execute survey with key personnel
- report outcomes
- assess how outcomes vary within patient subgroups
- disseminate outcomes in STRATIF-AI pilot studies and solicit feedback
- publication on trust and diversity in digital twins

> **Note**
>
> This requirement will constitute a part of the trust study designed to fulfill Requirement 1.1.4 Establish degree of trust.

5.1.8 Implement fairness reporting system

Description A system to report whether the system ***outputs*** are evaluated as ***feasible*** must be implemented within the platform for ***end-users***, and the results quantitatively assessed.

outputs: prediction intervals; recommendations; rehab demos;*
feasible: comprehensible in language; feasible to varying levels of resource access; feasible within time budgets;
end-users: patients; medical staff

Schedule At any point, before end of project.

Stroke Phase EACH (3)

Owner

- WP7

Key Personnel

- patient representatives
- medical staff
- design staff (TBD)

Actionable tasks

- appoint key personnel

- define parameters output and feasibility
- prototype system in apps/platform
- pilot study
- report results
- publication

5.2 Accessibility and universal design

ALTAI Assessment

Measures to prioritize inclusivity and accessibility of STRATIF-AI to a wide and diverse user-base will be taken. Assessment will be conducted to gauge the usability of the AI system's user interface for individuals with special needs, disabilities, or those at risk of exclusion. Efforts will be made to guarantee that information about the AI system, as well as its user interface, remains accessible and usable for users of assistive technologies, particularly those relevant for patients of stroke. Furthermore, Universal Design principles will be integrated into various stages of planning and development, where applicable, to enhance accessibility for all users. The potential impact of the AI system on end-users and/or subjects will be carefully assessed, including any disproportionate effects on specific groups.

5.2.1 Utilize inclusive design principles

Description The platform must be designed, tested, and trained to account for ***needs*** of vulnerable ***subgroups*** of ***end-users***.

needs: TBD*
end-users: medical staff; patients;
subgroups: *language; ethnicity; socio-economic background; health history; trust;*

Schedule At any point, before end of project.

Stroke Phase EACH (3)

Owner

- WP3

Key Personnel

- patient representatives
- medical staff
- design staff (TBD)

Actionable tasks

- appoint key personnel
- this will move to a different owner after preliminary assessment
- define set of needs that should be accessible

- identify different patient subgroups
- conduct interviews/solicit patient feedback
- design protocol to evaluate prototype design
- move WP owner to design staff
- pilot design and solicit feedback

5.2.2 Assess financial risk of unfair design

Description The system ***outputs*** must be critically evaluated to avoid ***risks*** of financial misuse in different ***clinical settings***.

outputs: *TBD; prediction intervals; discrepancy between clinician and prediction; discrepancy between patient and clinician;*
risks:* TBD; insurance system; financial overhead; costs of stay;
clinical settings: TBD; hospitals; rehab centers;

Schedule At any point, before end of project.

Stroke Phase EACH (3)

Owner

- WP7

Key Personnel

- patient representatives
- medical staff
- design staff (TBD)
- technical staff (TBD)

Actionable tasks

- appoint key personnel
- define parameters
- perform evaluation
- e.g., via experimental study/ pilot/ lit review
- update outputs
- institute necessary data protections

5.3 Stakeholder participation

ALTAI Assessment

STRATIF-AI is a project with a significant diversity in stakeholders. Consistent feedback and co-creation will be maintained with stakeholders during early stages of project development. Mechanisms

to ensure sustained stakeholder involvement after the implementation of the platform will also be designed.

5.3.1 Incorporate patient perspectives

Description Feedback and perspectives about system ***functions*** from representatives of patient ***subgroups*** will be solicited ***periodically*** and incorporated into system development.

subgroups: gender; disability; ethnicity; socio-economic status;
periodically: *TBD*
functions: *prediction intervals; design; overall trust; use;*

Schedule At any point, before end of project.

Stroke Phase EACH (3)

Owner

- WP3

Key Personnel

- patient representatives

Actionable tasks

- appoint key personnel
- define parameters
- design study or format
- design a reproducible questionnaire that can easily be answered
- pass to WP4, WP5, WP6
- design process protocol for responding to feedback
- design procedure for disseminating feedback
- publication on patient perspectives

5.3.2 Foster end-user feedback

Description A system for ***end-users*** to provide ***feedback*** should be established within the platform and ***periodically*** be evaluated.

end-users: medical professionals, patients;
discrepancies: *format of feedback;*
periodically: *TBD;*

Stroke Phase ALL

Owner

- WP2

Key Personnel

- design staff
- technical staff

Actionable tasks

- appoint key personnel
- determine evaluation schedule
- design and validate system

6 Societal and Environmental Well-being

The impacts of the AI system on the wellbeing of society must be considered throughout the AI system's life cycle. A system which is encountered in aspects of daily life—e.g., healthcare—has the potential to negatively impact mental health and emotional wellbeing. Such effects must be carefully considered and mitigated. Additonally, the sustainability of the AI system under development must be a point of discussion, such that the development and utilization of an AI system benefits all human beings, including future generations. Finally, the system must not undermine democratic processes or pose a threat to democratic society.

6.1 Environmental Wellbeing

AI systems should be as environmentally friendly as possible. The development, deployment and use-processes must therefore be assessed in terms of energy consumption and resource demands, in order to minimize waste. During the development of the STRATIF-AI project, the potential of negative harm on the environment needs to be systematically considered. Measuring the environmental impact through energy consumed and carbon emissions will be built into the development pipeline. Measures to reduce these impacts will be subsequently explored.

6.1.1 Identify environmental impact of model development

Description The ***environmental impacts*** of ***system development processes*** need to be measured or estimated.

environmental impacts: carbon emissions; energy used; ***system development processes***: model training; *other process with footprint?*

Owner

- WP2

Key Personnel

- technical staff

Schedule At any point, before end of project.

Stroke Phase ALL

Actionable tasks

- appoint key personnel
- define parameters (environmental and possibly costly processes)
- define method to estimate costs
- report costs to consortium/in documentation

6.1.2 Reduce environmental impact

Description A SOP which outlines acceptable ***thresholds*** and a resulting process protocol must be developed to mitigate environmental impacts.

thresholds: *limit for carbon emissions for model training; limit for time spent model training*

Owner

- WP2

Key Personnel

- technical staff
- WP3

Schedule At any point, before end of project.

Stroke Phase ALL

Actionable tasks

- appoint key personnel
- define parameters/ thresholds and procedures
- draft SOP (WP3)
- develop SOP for modelling team
- publish SOP

6.2 Impact on Work and Skills

ALTAI Assessment

STRATIF-AI will be designed with the objective of enabling a governance system conducive to human-in-command and prioritizing the epistemic authority of physicians. However, as a clinical decision-support system, STRATIF-AI will potentially have significant impacts on human work and work arrangements, necessitating a thorough assessment through an implementation strategy. While medical professionals will play a central role in the development pipeline of STRATIF-AI, broader inclusion and consultation with potentially affected workers and their representatives will be carried out. Measures to minimize the system's impact on human work will be implemented, following an in-depth investigation of STRATIF-AI's potential risks regarding the de-skilling of doctors or the alteration of training procedures and experiential learning.

6.2.1 Investigate impact on work arrangements

Description Assess concerns of potentially affected ***workers*** to inform implementation strategy.

workers: doctors; nurses; insurance personnel; training personnel; rehabilitation specialists; physiotherapists; *any affected working persons involved in the care of stroke patients*

Owner

- WP3

Key Personnel

- WP4
- WP5
- WP6
- WP7

Schedule At any point, before end of project.

Stroke Phase EACH (3)

Actionable tasks

- appoint key personnel
- define full set of at-risk workers
- conduct workshops with key personnel
- design interview to elicit concerns from workers
- write process protocol for implementation strategy
- feed into HIC concept

> **Note**
>
> In the initial co-creation workshop series, participants expressed concerns that integrating the STRATIF-AI platform into clinical practice too early and without warning may limit the natural experience-building process that physicians generally undergo.

6.3 Impact on Society at large or Democracy

ALTAI Assessment

Potential ramifications on society at large and democracy will be considered throughout the STRATIF-AI project cycle. To assess the societal impact of the AI system's utilization beyond those to immediate end-users and subjects, we will solicit feedback throughout planned pilot and clinical trial studies. A key aspect of the STRATIF-AI platform is developing an infrastructure to empower citizens to monitor their overall health and wellbeing while screening for risk of stroke—which will contribute to the benefit of society at large. The STRATIF-AI platform is not at risk of undermining democratic principles or adversely affecting democracy.

6.3.1 Solicit citizen feedback

Description Include diverse ***stakeholders*** in discussions at relevant ***stages*** of the STRATIF-AI pipeline.

stakeholders: citizens; hospital workers; *persons who may be overlooked in stakeholder discussions*
stages: pilot study; interview study (WP3); *any upcoming implementation discussions*

Owner

- WP3

Key Personnel

- all WP leads

Schedule At any point, before end of project.

Stroke Phase ALL

Actionable tasks

- appoint key personnel/ WP leads
- define parameters, i.e. stakeholders who may be overlooked
- clarify potential procedures within WP tasks to incorporate discussions
- assess input from stakeholders
- disseminate information

7 Accountability

The principle of accountability highlights the need to establish mechanisms to guarantee responsibility throughout the lifecycle of AI systems. This principle intersects closely with risk management practices, which involve transparently identifying and mitigating risks that can be comprehended and audited by external parties. In instances where unjust or adverse impacts arise, accessible mechanisms for accountability must be established, to facilitate the opportunity for redress to affected parties.

Keywords:
accountability, AI Ethics Review Board, redress by design

7.1 Auditability

ALTAI Assessment

The auditability of STRATIF-AI's workflow and system lifecycle is a fundamental aspect of our ethical plan. By creating a living document, we aim to establish a transparent and traceable log of all measures taken to adhere to ethical principles and legal requirements. To ensure the AI system can be audited by independent third parties, the ethical plan will be made available or disseminated accordingly.

7.1.1 Facilitate auditability

Description ***Means*** to facilitate external audits of STRATIF-AI processes will be instituted and validated.

Means:* A publicly available and version-controlled ethical plan; *any additional implementation strategies?*

Owner

- WP3

Key Personnel

- WP7
- WP2

Schedule At any point, before end of project.

Stroke Phase ALL

Actionable tasks

- appoint key personnel
- define additional outputs?
- discuss site prototype and any legal barriers
- solicit feedback on prototype
- create sub-page to report important parameters
- define important parameters
- e.g. traceability of the development process, the sourcing of training data and the logging of the AI system's processes, outcomes, positive and negative impact
- make pages public

7.2 Tradeoffs/ Risk-management

ALTAI Assessment

Ensuring both the ability to report on actions or decisions contributing to the outcomes of AI systems and to address the consequences of such outcomes is paramount. It is essential to identify, assess, document, and mitigate potential negative impacts of AI systems, particularly for those directly or indirectly affected. Safeguards must be in place to protect whistleblowers, NGOs, trade unions, and other entities reporting legitimate concerns about AI systems. As part of WP3 within the STRATIF-AI project, we will conduct an external Z-inspection assessment, which is a gold-standard and validated methodology to assess the ethical impacts of an AI tool. A comprehensive risk training protocol will be designed and adhere to any legal protocols. Our ethical plan has been designed in concordance with the Assessment List for Trustworthy AI (ALTAI), and, as such, will be continually evaluated throughout the AI system's lifecycle.

While implementing these requirements, tensions may arise, necessitating inevitable trade-offs. We will identify relevant interests and values implicated by the AI system and explicitly evaluate trade-offs in terms of their risk to safety and ethical principles, including fundamental rights. These decisions will be thoroughly documented throughout our ethical audit process. Additionally, a a process for third parties (e.g. suppliers, end-users, subjects, distributors/vendors or workers) to report potential vulnerabilities, risks or biases in the AI system, and a following risk management protocol will be instituted. Accessible mechanisms to ensure adequate redress will be established.

7.2.1 Institute external reporting system

Description A system for ***third-parties*** to make reports about ***risks*** should be established within the platform and ***periodically*** be evaluated.

third-parties: citizens; hospital workers; developers; *any contributors of technology or data*
risks: vulnerabilities; errors; concerns
periodically: *TBD*

Stroke Phase ALL

Owner

- WP2

Key Personnel

- design staff
- technical staff
- WP3

Actionable tasks

- appoint key personnel
- define all parameters
- define link to feedback reporting loop
- implement/design procedure
- feedback from third parties to validate
- test system

7.2.2 Collect post-prediction feedback

Description A system for ***end-users*** to provide feedback about ***medical applicability*** should be established within the platform. This is closely related to the requirement under Section Human Agency and Oversight; Institute reporting feedback loop.

end-users: medical professionals
medical applicability: effectiveness; relevance; alignment with evolving medical knowledge and practices;

Stroke Phase EACH (3)

Owner

- WP4
- WP5
- WP6

Key Personnel

- medical staff
- design staff
- technical staff

Actionable tasks

- appoint key personnel
- medical professionals to define key parameters
- prototype for system/information parameters
- implementation by WP2/ design staff
- design and validate system

7.2.3 Conduct budget impact analysis

Description The economic justification of the implementation of the STRATIF-AI platform for different ***stakeholders*** must be made via cost-benefit analyses.

stakeholders: hospitals; insurance companies; patients; rehab centers; *any affected institutional parties?*

Stroke Phase EACH (3)

Owner

- WP7

Key Personnel

- medical staff

Actionable tasks

- appoint key personnel
- define parameters
- design budget impact/ cost-benefit analysis study
- conduct analysis
- report results
- publication/dissemination of results