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** Confidential / Restricted / **Public**

Table of contents

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

Table 1: **Revision History**

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This deliverable contains original unpublished work, or it is clearly indicated otherwise.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 integration of artificial intelligence (AI) in healthcare presents several ethical challenges, which may have implications for the protection of fundamental human rights. Dedicated efforts are essential to ensure that AI systems in healthcare adhere to principles of lawfulness, ethics, and robustness. To address these challenges within the STRATIF-AI project, we develop a plan which outlines steps to mitigate ethical risks and increase trustworthiness throughout the AI development process. To construct this plan, we draw on interdisciplinary expertise and utilize three core strategies:

(1) Operationalization: Existing guidelines or soft-law documents tend to be rather high-level or generally applicable, which makes them appropriate for a diversity of AI systems. As a consequence, overarching principles must be contextualized into actionable steps, which are specifically tailored to the characteristics of a particular AI system. We utilize expertise on ethical framework building to ensure that legal guidelines do not remain theoretical concepts, but are rather effectively integrated into the design, deployment, and ongoing management of the AI system. For the STRATIF-AI project, this will enable identification of potential ethical risks early in the project lifecycle, to ultimately enhancing its trustworthiness and societal benefit.

(2) Co-creation: To ensure the equitable development of solution, content for this manual is being co-created with relevant stakeholders within the STRATIF-AI project, i.e., clinicians, data scientists, medical researchers, patient organization representatives and project coordinators. We thus assess the risks and benefits of various requirements to ensure ethics and trustworthiness, to create necessary, feasible and auditable ethical requirements.

(3) Anticipation: STRATIF-AI introduces several novel technological solutions, which may not fall within the scope of previously enumerated ethical guidelines. A hybrid machine-learning and mechanistic modelling framework will be utilized to simulate patient specific responses to changes in behaviour, diet, exercise and medication. Models will be trained on diverse sources of electronic health data, collected from 6 clinical studies, within a federated learning platform. Data will be uploaded and stored in a Personal Data Vault—to ensure patient control—and subjected to a semantic harmonization process to make themminteroperable. Such a multi-faceted technology is yet unprecedented and could indeed spur new ethical dilemmas. Solutions based solely on existing regulations will be reactive rather than prescriptive, having to rely on the deployment and failure of technology to guide decision-making. To pre-empt such changes, we are conducting a scoping review of the ethical considerations in the field of digital twin technology in health-care.

Framework

The framework for ethical and trustworthy design of the STRATIF-AI platform has been designed in concordance with the EU Guidelines for Trustworthy AI. Contents of the manual have been developed based on a series of co-creation workshops, preparation of a systematic review, and previous ethical frameworks developed for the EU-Horizon blueprint project "VALIDATE". The contents will be updated following audits of the STRATIF-AI consortium's adherence to the framework, and an external Z-inspection of the ethical solution.

Regulatory Context

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.

Our manual was derived by following the self-assessment checklist—Assessment List for Trustworthy AI (ALTAI)—developed by the HLEG in 2020. The ALTAI checklist is intended to help organizations identify key elements and concepts to design ethical AI systems, and serves as a practical instrument for assessing the compliance of AI systems with the EU's ethical principles.

Structure

The first draft of the framework has been structured in concordance with the EU Guidelines for Trustworthy AI. Each section corresponds to one of seven requirements. Each section first describes the

preliminary assessment following the Assessment List for Trustworthy AI (ALTAI checklist). The section then lists the set of suggested requirements for the STRATIF-AI project.

The framework is hosted online, on the Charité server, and has been made available under password protection to the STRATIF-AI consortium. The framework in a html format manual will be used by consortium members ("owners") throughout the audit process to fill in information, report progress on actionable tasks, and provide updates on requirements.

Requirement Template

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

parameters: Based loosely on the concept "Planguage", we identify specific parameters within each requirement to delineate progress.

Owner

- Each requirement has an "Owner", who is responsible for organizing its implementation.

Key Personnel

- Several stakeholders may serve as key personnel to contribute to or facilitate progress towards this requirement.

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
- These are a set of tasks which must be executed
- In order to meet this requirement

Co-creation Workshops

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 VALIDATE project and executed in 2023.

Methodology

Four workshops were conducted in March 2024 and comprised a diversity of participants from within the consortium. Workshops were conducted online on Microsoft Teams, and with the use of the online whiteboard Mural. The full workshop report and associated documents are available online ().

Output

Reflections on workshop series.

Description of general themes.

Utility of requirements.

Fundamental Rights Impact Assessment

The protection of people's fundamental rights is the term in the European Union used to refer to the human rights set forth within the EU Treaties, the Charter of Fundamental Rights and Internal Human Rights Law. These rights encompass dignity and non-discrimination, data protection and privacy. As recommended by policy-makers, this manual begins with a broad assessment of fundamental rights. The relevant solutions are linked below in response to each assessment question.

1. Does the AI system potentially negatively discriminate against people on the basis of any of the following grounds (non-exhaustively): sex, race, colour, ethnic or social origin, genetic features, language, religion or belief, political or any other opinion, membership of a national minority, property, birth, disability, age or sexual orientation?

- list bias testing plans for development, deployment, and use
- list processes to address bias

2. Does the AI system respect the rights of the child, for example with respect to child protection and taking the child's best interests into account?

3. Does the AI system protect personal data relating to individuals in line with GDPR?

- data protection impact assessment, including necessity/proportionality of processing operations in relation to their purpose
- risks, safeguards, security measures for personal data

4. Does the AI system respect the freedom of expression and information and/or freedom of assembly and association?

- which mechanisms are in place/testing/monitoring systems

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 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*** must be described.

physician expertise level: learner; proficient; expert

Owner

- WP4
- WP5
- WP6

Key Personnel

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

Schedule At any point, before end of project.

Stroke Phase ALL

Actionable tasks

- appoint key personnel
- clarify when comparison of AI and clinician will be conducted
- 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 ***valid methodology*** for measuring and communicating quantitative prediction uncertainty to ***end-users*** must be established.

valid methodology: TBD
end-users: medical professionals, patients

Owner

- WP2

Key Personnel

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

Schedule At any point, before end of project.

Stroke Phase ALL

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 ***patients*** must be designed.

risks:: procedures for epistemic authority; HIC governance structure; AI system
patients:: language; education; age; disability

Schedule At any point, before end of project.

Stroke Phase ALL

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.

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 clear conflict between the predictions of the tool and the perspective or opinion of the patient, a process protocol for managing expectations, customized for ***patients*** must be described.

patients:: with stroke; at risk of stroke; without stroke; age; education; disability

Owner

- WP4
- WP5
- WP6

Key Personnel

- medical staff

Schedule At any point, before end of project.

Stroke Phase ALL

Actionable tasks

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

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: metrics or summary statistics about training data and real-time data; predictive model type (TBD)

Owner

- WP2

Key Personnel

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

Schedule At any point, before end of project.

Stroke Phase ALL

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*** by various ***parties*** must be established, assessed and reported.

risks: model inversion (information in parameters to reconstruct data); data leakage; data poisoning
parties: internal (consortium); beneficiaries (collaborators); external

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*** of the attack must be followed.

severity: TBD; data poisoning; model evasion;

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.

end-users: TBD; 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 ALL

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 ***levels*** of ***risk*** should be established within the platform and ***periodically*** be evaluated.

levels: TBD; stakeholder risk levels; patients
risk: TBD; threat to safety; validity; misuse; reliability; metrics
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

2.2.3 Communicate risks to patients

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

technical risks: TBD; vulnerability to misuse; technical robustness metrics ***patients:*** language; education; age; disability

Schedule At any point, before end of project.

Stroke Phase ALL

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

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: TBD; 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 ***data contexts***, validated within a randomized clinical trial design or prospective follow-up study design.

performance metrics: discrepancies between physician and system; discrepancies/discomfort of patient with system; other technical error reports; TBD
data contexts: type of stroke; patient demography; age; education; socio-economic background; ethnicity; TBD

Schedule At any point, before end of project.

Stroke Phase EACH (prevention, acute, monitoring)

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 generalizability of modelling ***predictions*** across a ***diversity*** of end-users must be established.

predictions: TBD; core metrics to assess behaviour
diversity: patient subgroups; geography; ethnicity; gender; socio-economic background; health history; disability; TBD

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

2.3.4 Disseminate metrics of accuracy and validity

Description A standardized procedure for disseminating established metrics of ***accuracy*** and ***validity*** of the system customized for ***patients*** must be designed.

accuracy: metrics TBD
validity: metrics TBD
patients: language; education; age; disability

Schedule At any point, before end of project.

Stroke Phase ALL

Owner

- WP7

Key Personnel

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

Actionable tasks

- appoint key personnel
- define metrics of accuracy and validity
- feedback on metric interpretability from patients/representatives
- adaptation in formats
- publication of initial metrics (internal/otherwise)

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 norms/guidelines; TBD
predictive accuracy: boundaries of accuracy; 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

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.

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: TBD; potential to reconstruct training data; glean private information; model inversion
quantified: TBD; levels of security 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 Data management plan for different ***data types*** detailing ***access*** and ***safeguards*** for ***vulnerable*** patient groups must be made available.

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
vulnerabilities: language; age; 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.

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*** fulfil ***quality critera*** will be implemented and made available to ***end-users*** before use for model training or prediction.

data type: datasets; training data; Personal Data Vault
quality criteria: TBD; missing data; inaccuracies; 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: TBD; pre-processing; evaluation; forecasting; validation; ML monitoring
partners: TBD; 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: TBD; drift detection; retraining procedure; will harmonization happen every time new data is added; how will the process be governed?
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

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 will be adhered to.

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 Data collected in clinical studies and throughout the planned use of the STRATIF-AI platform will be highly intimate. The costs of collecting invasive data collected need to be balanced against the risks to patients' trust/privacy.

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: TBD; data collection; data source; data demographics; patient-specific data; model parameters; model calibration metrics;
assessed: TBD; 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: TBD; predictions; prediction intervals; data output
systematically compared: TBD; design of study comparison;
human feedback: TBD; discrepancy with medical decision; patient discomfort; related to req 1.1.;
quality: TBD; 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

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 available in a simplified format to ***end-users***.

parameters: TBD; modelling parameters; training data summary; own data key parameters;
end users: TBD; medical professionals; experience levels; patients; education; age; language; disability;

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, recorded and disseminated to ***end-users***.

costs: TBD; format to quantify explainability versus accuracy tradeoff;
end-users: TBD; medical professionals; experience levels; patients; education; age; language; disability;

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.

4.3.1 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

4.3.2 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

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: TBD; patients; patients with stroke; patients at risk of stroke; patients without stroke; language; education; age; disability; ethnicity; gender; migration status; socio-economic status;
threshold: TBD; 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 ***subgroups***.

outputs: TBD; model predictions; divergence between clinician prediction and model prediction;
difference: TBD; quantitative threshold; acceptable margin of error;
subgroups TBD; 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 Assess efficiency/diversity tradeoff

Description ***Methods*** to weigh risk of producing predictions which are not ***generalizable*** against the need to represent diverse data and communicated to end-users.

methods: TBD; modelling metrics;
generalizable: TBD; acceptable prediction interval/agreement with medical professional

Owner

- WP2

Key Personnel

- technical staff (TBD)
- WP3

Schedule At any point, before end of project.

Stroke Phase ALL

Actionable tasks

- appoint key personnel
- create testing metrics/methods
- define problem of generalizability/ acceptable threshold
- publication of results

5.1.5 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: TBD; 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

5.1.6 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: workshops; training materials; 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

5.1.7 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: TBD; 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

5.1.8 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.

end-users: medical professionals; patients with stroke; patients at risk of stroke; patients without stroke
subgroups: TBD; 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

5.1.9 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: *TBD; prediction intervals; recommendations; rehab demos;*
feasible:* TBD; language; comprehension; resource access; time;
end-users: patients; medical staff

Schedule At any point, before end of project.

Stroke Phase ALL

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: TBD; 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: TBD; gender; disability; ethnicity; socio-economic status;
periodically: TBD;
functions: TBD; 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: TBD; 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 Your Book Title

7 Your Book Title

7.1 Subtitle or any additional information

Author: Your Name

Date: April 15, 2024

References