PRECISE4Q

PREDICTIVE MODELLING IN STROKE

DELIVERABLE

Project Acronym: PRECISE4Q

Grant Agreement number: **777107**

Project Title: Personalised Medicine by Predictive Modelling in Stroke for better Quality of Life

Personalised Medicine by Predictive Modelling in Stroke for better Quality of Life

Revision: 1.0

Authors and Contributors	Sara Kijewski (ETH Zurich), Julia Amann (ETH Zurich), Alessandro Blasimme, Kelly Ormond (ETH Zurich), Effy Vayena (ETH Zurich)			
Responsible Author	Sara Kijewski		**Email**	Sara.kijewski@hest.ethz.ch
	Beneficiary	ETH Zurich	**Phone**	+41 44 632 46 19

Project co-funded by the European Commission within **H2020-SC1-2016-2017/SC1-PM-17-2017**		
Dissemination Level		
PU	Public, fully open	X
CO	Confidential, restricted under conditions set out in Model Grant Agreement	
CI	Classified, information as referred to in Commission Decision 2001/844/EC	

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

Revision History

Revision	Date	Author	Organisation	Description
0.1	05.03.2022	J. Amann	ETH Zurich	Preliminary outline
0.2	04.04.2022	J. Amann	ETH Zurich	First draft
0.3	11.04.2022	J. Amann	ETH Zurich	Feedback KO
0.4	18.10.2022	S. Kijewski	ETH Zurich	Second draft
0.5	19.10.2022	S. Kijewski	ETH Zurich	Feedback KO
0.6	21.10.2022	S. Kijewski	ETH Zurich	Feedback KO and AB
0.7	22.10.2022	S. Kijewski	ETH Zurich	Feedback EV
1.0	23.10.2022	S. Kijewski	ETH Zurich	Sent for internal review EMP, CUB
1.0	30.10.2022	S. Kijewski	ETH Zurich	Submission

Date of delivery	Contractual:	28.10.2022	Actual:	31.10.2022
Status	final ☒ /draft □			

Abstract (for dissemination)	This document describes the activities conducted in relation to T1.4, T1.5, T1.6, and T1.7 (WP1). Based on the activities carried out as part of T1.6, we refined the PRECISE4Q Reflective Framework. In its final version, the framework consists of ten sub-sections across development and deployment. The framework was revised based on feedback from the pilot survey and the expert workshop series, presented in this document. The document concludes by highlighting some of the limitations and provides recommendations for future applications of the framework.
Keywords	Ethics, AI, autonomy, justice, data protection and privacy, disclosure, responsibility, empathy, explainability

Statement of originality

This deliverable draws to a large extent on the following manuscript, which is currently in preparation: *Amann et. al Co-devising an ethical framework for medical AI in stroke medicine* In the aforementioned article, the framework development is presented in a scientific format, here we present it as a deliverable.

Table of Contents

Executive Summary

PRECISE4Q aims to minimize the burden of stroke for individuals and society through multi-dimensional predictive modeling. This deliverable contributes to this aim by providing a reflective ethical framework that should guide the development of these technologies and their translation into the clinical context beyond the termination of the project. The reflective framework provides consortium partners with a tool they can use to identify pertinent ethical issues and take appropriate measures. The reflective framework is rooted in normative considerations, builds on existing ethical frameworks, considers the lived experience, attitudes, values, and expectations of prospective users, beneficiaries, and developers. It has undergone two rounds of revisions based on the consortium partners' feedback and a public symposium to ensure its utility, usability, and overall fit for purpose.

This is the final deliverable of WP1, it is based on activities carried out in relation to T1.4, T1.5, T1.6, and T1.7 (WP1). Specifically, in this deliverable, we will summarize two key areas:
First, based on the insights gathered through the various stages of the project and the pilot evaluation carried out as part of T1.6 and which will be summarized here, we refined the PRECISE4Q reflective ethical framework. We present the final (revised) version, which consists of ten sub-sections across development and deployment. We also present preliminary analysis of the observational findings from the three expert workshops held with consortium partners to 'test' the reflective framework for preventative, acute and rehab settings. Second, based on the aforementioned evaluation processes, this deliverable concludes with a set of recommendations for the adoption of the reflective framework beyond the specific context of PRECISE4Q.

1 Rationale and overall objective of the deliverable

In recent years, a plethora of ethical principles, frameworks, and guidelines have been issued by the public and private sector to ensure the ethical development and use of AI-based technologies in healthcare research and practice. However, to date these valuable tools are rarely adopted in practice. This make it difficult to determine whether they are fit for purpose and meet the needs of AI researchers and developers to support them in ethical decision making. To address this translational gap from guidance to practice, this deliverable presents a pilot test of the *PRECISE4Q Reflective Framework for Big Data Health Research* using three concrete use cases.

The overall purpose of the framework is to ensure that the PRECISE4Q research activities and resulting tools are reconcilable with core ethical values that should guide all clinical research and practice. The framework is intended to guide further development of the PRECISE4Q technologies and their translation into the clinical context beyond the termination of the project. The initial reflective framework was developed as part of D1.7. It is rooted in normative considerations, builds on existing ethical frameworks, but also takes the lived experience, attitudes, values, and expectations of prospective end-users (i.e., clinicians), beneficiaries (i.e., patients and their families), and developers (i.e., researchers) of AI-powered clinical decision support systems into account.

The reflective framework has undergone two rounds of revisions based on the consortium partners' feedback to ensure its utility, usability, and overall fit for purpose. The present deliverable summarizes the observational findings from the workshops, the evaluation survey and revision of the framework and provides recommendations for the adoption beyond the specific context of the PRECISE4Q project.

As part of the deliverable, The PRECISE4Q activities in ethics were presented and discussed in a public symposium organized by the Health Ethics & Policy Lab, ETHZ Zurich.

2 The revised PRECISE4Q Reflective Framework

Phase	Theme	The PRECISE4Q Reflective Framework for Big Data Health Research
Development	Purpose of the tool	- In what phase of stroke is the tool to be used (prevention, acute, rehabilitation or reintegration)? - What is the specific problem the tool aims to address? - Who is the primary user group? - Are there any secondary end-user groups? - What is the intended purpose of the tool in clinical practice? - How might the tool be used by clinicians and how may this shape their professional role perceptions? - Is there a risk of inappropriate use and how might this risk be mitigated?
	Data quality and representativeness	- How has the data been obtained, and which ethical principles were considered in this process? - What do we know about the data quality and its representativeness for the target population? - What measures are in place to ensure data quality and representativeness (e.g., who might be under or overrepresented)? - What consequences may data characteristics have on the performance of the model for these population(s)?
	Explainability	- What kind of information on the tool will be available to end-users? - Are models explainable and if so, is there an impact on predictive performance? - If available, are explanations tailored to the needs of end-users? - How may the information end-users have or lack impact their interaction with the tool?
	Usability and user experience	- Have prospective end-users been involved in the development process and if so, how has their input shaped the tool? - If prospective end-users were not involved in the development, what consequences may this have on the tool and its adoption in clinical practice? - Have usability and user experience been assessed, and if so, how?
	Clinical validation	- How is the tool validated? - What does clinical validation mean to developers, what does it mean to clinicians and patients? - What impact may clinical validation have on clinicians' and patients' trust? - What impact may clinical validation have on clinicians' perceived responsibility and accountability?
Deployment	Disclosure of AI	- How much information can and should be disclosed to the patient? - How much do clinicians need to know about the tool and its application to fulfil their role? - What impact may predictive health information with disclosure of AI have on patient autonomy, trust, and the doctor-patient relationship (e.g., shared decision-making)? - What about the impact of disclosure on vulnerable populations (e.g., socially disadvantaged groups, stigmatized groups, groups with lower health literacy skills)?
	Responsibility	- How is responsibility/liability addressed? - Is there a risk of deskilling? - What is the developers' responsibility? - What impact may incorrect decisions caused by the tool have on clinicians' moral responsibility?
	Empathy	- How may the tool impact clinicians' empathy towards patients? - How can patient values, beliefs, and preferences be incorporated into the decision-making process? - Might the tool replace human contact in the clinical encounter and if so, what consequences may this have for patients and clinicians?
	Privacy & Data Protection	- Given that stroke prevention takes place before any symptoms occur, how can health benefits and privacy be balanced? - Should there be different privacy standards for the different phases of stroke (prevention, acute, rehabilitation, reintegration)? - Which mechanisms would need to be in place to ensure patient privacy? - What might be the consequences of failing to ensure patient privacy?
	Monitoring & Evaluation	- What should process and impact monitoring and evaluation look like along the patient journey and life cycle of the technology? - Who is responsible for conducting continuous monitoring and evaluation? - What might be the consequences of failing to conduct continuous monitoring and evaluation?

3 Existing ethical frameworks for medical AI

The PRECISE4Q reflective framework builds on existing ethical frameworks and guidance, most notably it is informed by the WHO guidance on Ethics & Governance of Artificial Intelligence for Health, the Assessment List for Trustworthy Artificial Intelligence (ALTAI) introduced by the High-Level Expert Group on AI (AI HLEG), UNESCO's Recommendation on the Ethics of Artificial Intelligence and the Organization for Economic Co-operation and Development (OECD) Principles of AI. The below sections provide a brief overview of these two frameworks.

3.1 WHO guidance on Ethics & Governance of Artificial Intelligence for Health

Introduced in 2021, the WHO guidance on Ethics & Governance of Artificial Intelligence for Health identifies the ethical challenges and risks of using artificial intelligence in healthcare and outlines six consensus principles to ensure that AI works to the public benefit of all countries. The purpose of these principles is to guide all stakeholders that develop, deploy, and evaluate AI for health, including health care personnel, developers, policymakers, health system administrators, and governments. The principles are:

1. **Protecting human autonomy**
2. **Promoting human well-being and safety and the public interest.**
3. **Ensuring transparency, explainability and intelligibility**
4. **Fostering responsibility and accountability**
5. **Ensuring inclusiveness and equity.**
6. **Promoting AI that is responsive and sustainable**

The guidance underscores the ethical challenges related to health data. Data quality and privacy preservation represent two major concerns that may inhibit the effective use of health data to develop AI systems. Regarding data quality, the guidance highlights the perils of under- or over-representation in the data, undermining the representativeness of the data. Safeguarding privacy is presented as another pressing ethical challenge arising from the use of health data.

The report also identifies legal, regulatory, and non-legal measures aimed at promoting the ethical use of AI for health and to avoid its misuse to undermine human rights and legal obligations. It reviews governance frameworks and provides specific advice for implementation of the guidance for three stakeholders groups: AI technology developers, ministries of health, and health-care providers.

3.2 Assessment List for Trustworthy Artificial Intelligence (ALTAI)

In 2020, the High-Level Expert Group on AI put forward Ethics Guidelines for Trustworthy AI complemented by the Assessment List for Trustworthy Artificial Intelligence (ALTAI). The Guidelines identify four ethical principles (Respect for human autonomy, Prevention of harm, Fairness, Explicability) and seven requirements that organizations should adhere to, in order to achieve trustworthy AI. The overall objective of ALTAI is to help organizations carry out self-evaluations to determine whether an AI system that is being developed, deployed, procured, or used, complies with the seven requirements of Trustworthy AI, specified by the *EU Ethics Guidelines for Trustworthy AI*. A set of questions for each of the seven requirements serves as an operationalization of the requirements and aims to guide organizations through the assessment process. The seven requirements are:

1. **Human agency and oversight**
   fundamental rights, human agency, and human oversight
2. **Technical robustness and safety**
   resilience to attack and security, fall back plan and general safety, accuracy, reliability, and reproducibility
3. **Privacy and data governance**
   respect for privacy, quality and integrity of data, and access to data
4. **Transparency**
   traceability, explainability and communication
5. **Diversity, non-discrimination, and fairness**
   avoidance of unfair bias, accessibility and universal design, and stakeholder participation
6. **Societal and environmental wellbeing**
   sustainability and environmental friendliness, social impact, society, and democracy
7. **Accountability**
   auditability, minimization and reporting of negative impact, trade-offs, and redress

The ALTAI checklist, which is also available as a prototype web application, can support organizations in understanding what Trustworthy AI is and what (unforeseen) risks an AI system may entail. In doing so, it raises awareness of the potential impact of AI on society, the environment, and various stakeholder groups (e.g., patients, clinicians). It can also aid organizations to determine whether meaningful and appropriate measures are or need to be put in place to ensure adherence to the seven requirements for trustworthy AI. If applied rigorously, ALTAI can help to promote responsible and sustainable AI innovation in Europe to ensure that AI-based technologies benefit, empower, and protect individual interests as well as the common good of society.

3.3 Recommendation on the Ethics of Artificial Intelligence

In 2021, the 41st General Conference of the United Nations Educational, Scientific and Cultural Organization (UNESCO) adopted the “Recommendation on the Ethics of Artificial Intelligence”. This recommendation was the first global standard-setting instrument, seeking to advance inclusive and transparent governance across multiple disciplines involving a multitude of stakeholders. They were developed with a key focus on human dignity The ten principles that were identified were:

1. **Proportionality and doing no harm**

proportionality of AI method to aim, no infringement of foundational values and human rights, appropriateness to context and grounded in scientific evidence

2. **Safety and security**
   avoidance, prevention and elimination of safety risks and security risks throughout system lifecycle
3. **Fairness and non-discrimination**
   social justice, fairness, and non-discrimination, benefits of AI available and accessible to all, minimization, and avoidance of reinforcement of bias throughout lifecycle
4. **Sustainability**
   human, social, cultural, economic, and environmental dimensions
5. **Privacy and data protection**
   respect, protect and promotion of privacy, data protection frameworks and governance mechanisms, privacy impact assessments
6. **Human oversight and determination**
   ethical and legal responsibility at any stage of the system lifecycle, human accountability
7. **Transparency and explainability**
   intelligibility, humans informed on decisions by or based on algorithms,
8. **Responsibility and accountability**
   ethical responsibility and liability for decisions and actions, oversight, impact assessment, audit and due diligence mechanisms, auditability and traceability of AI systems and their working
9. **Public awareness and literacy**
   promotion of public awareness and understanding of AI technologies and value of data, approach grounded in impact on human rights and access, the environment and the ecosystem
10. **Multi-stakeholder governance and collaboration**
    respect of international law of national sovereignty in use of data, inclusive governance approach, open standards and interoperability

The Recommendation identifies health as a focus area and specifically promotes AI systems that promote health and the protection of life. Further, it argues for special attention to AI-based solutions for prediction, detection and treatment with regard to oversight and mitigation of bias, privacy and data protection requirements, informed consent for data use and analysis, human agency, and ethics approval. Finally, it puts forward that interactions with AI systems in the health-sphere should be easily identifiable and refusable.

3.4 The OECD Principles of AI

Adopted in 2019, the Organization for Economic Co-operation and Development (OECD) Principles of AI represented the first intergovernmental policy guidelines. The aim of these guidelines is to guide governments in the promotion of innovative and trustworthy AI by shaping governance preferences in member states and beyond. The recommendation presents the five following principles values-based principles for trustworthy AI:

1. **Inclusive growth, sustainable development and well-being**
2. **Human-centred values and fairness**
3. **Transparency and explainability**
4. **Robustness, security and safety**

5. **Accountability**

3.5 Shortcomings of existing ethical frameworks for medical AI

There is no shortage of ethical guidelines and frameworks for medical AI, including the four aforementioned examples. While a convergence around human-centered principles and human rights can be observed, seeking to promote safe development and deployment of these technologies, a major shortcoming of existing ethical guidelines is that they often fail to be rigorously applied in practice [1]. This may be attributed to the fact that adherence to ethical frameworks is voluntary – contrary to the adherence to regulatory guidelines, which is usually overseen by dedicated legal departments within organizations. Another commonly raised point of criticism is that ethical frameworks are often conceived without direct input from the intended target audiences and AI beneficiaries [2]. As a result, guidelines may not be context-specific and therefore not applicable or compatible with existing workflows and processes and can thus not easily be embedded. It is therefore not uncommon, that if at all, ethical assessments are carried out toward the end of product development with little reflection or stakeholder engagement.

Some authors have gone as far as to argued that AI developers are not able or incentivized to successfully translate ethical principles and guidelines into practice [1, 3, 4]. A recent study, for instance, found that ethics are frequently ignored in software start-up like environments due to a lack of practical guidance that help translate abstract ethical principles into actionable recommendations [5]. We share these concerns and consider it unlikely that it is feasible to produce a one-size fits all ethics checklist that will fit every context and technology. This is, why rather than providing a deterministic "ethics checklist", we consider tools to promote ethical reflection among AI researchers and developers to be key when it comes to embedding ethics into AI development.

A first step towards ensuring that ethical guidelines for medical AI are applicable and responsive to different stakeholders' needs is to directly involve different stakeholder groups in the development and in the governance of medical AI [6-9]. From an intuitive and normative point of view, involving those involved in the development of medical AI and those most directly affected by it seems both reasonable and adequate. There is also empirical evidence highlighting the advantages of stakeholder involvement, showing that stakeholder engagement in developing AI ethics guidelines can lead to more comprehensive ethical guidance with greater applicability [2]. However, as a recent review showed, only 38% of AI ethics guidelines reviewed, reported some form of stakeholder engagement with even fewer documents providing detailed information on the engagement process [2].

4 Methodology

We followed a multi-stage participatory approach to devise a first version of the P4Q Reflective Framework, which was previously detailed in D1.7. As part of the current deliverable (T1.6), we carried out a second round of internal workshops, one with each of the three task forces (Prevention, Acute, Rehabilitation/Reintegration) to pilot test the P4Q Reflective Framework for each phase (development and deployment). Workshops were held virtually via Zoom from March to May 2022 with three to seven participants for each session. Each workshop lasted 1.5 hours.

The central aim of the workshops was to explore the consortium partners' experiences in applying the P4Q Reflective Framework for big data and AI health research, and to evaluate the Clinical Decision Support System's fit for purpose. In preparation for the workshop, participants received P4Q Reflective framework in pdf format and were asked to familiarize themselves with it. The workshops began with a brief introduction provided by the workshop lead (JA) followed by a joint discussion on each question item of the framework. The workshop lead also addressed all unclear concepts or questions throughout the workshop. Where needed, the workshop lead attempted to stimulate response by rephrasing the question or providing food for thought.

Workshops were recorded, and these recordings were auto-transcribed using the Zoom transcribe-function; unstructured field notes were also taken during each workshop. All workshop recordings and transcripts were carefully reviewed by the responsible author of this deliverable (SK) and a research assistant who was neither involved in the framework development nor its application during the workshop. Upon reviewing the recordings and transcripts, the research assistant produced an independent observation report, one for each of the three workshops. After the workshop, all participants were asked to complete a short evaluation survey, to assess their overall experience applying the framework, as well as challenges, risks, and suggestions for improvement (see Table 1 in the Appendix). Survey responses were analyzed both quantitatively and qualitatively.

Feedback and suggestions were incorporated into a revised version of the framework. The framework was also discussed in a public symposium organized by the Health Ethics and Policy Lab at ETH Zurich. Figure 1 presents an overview of the methodological approach.

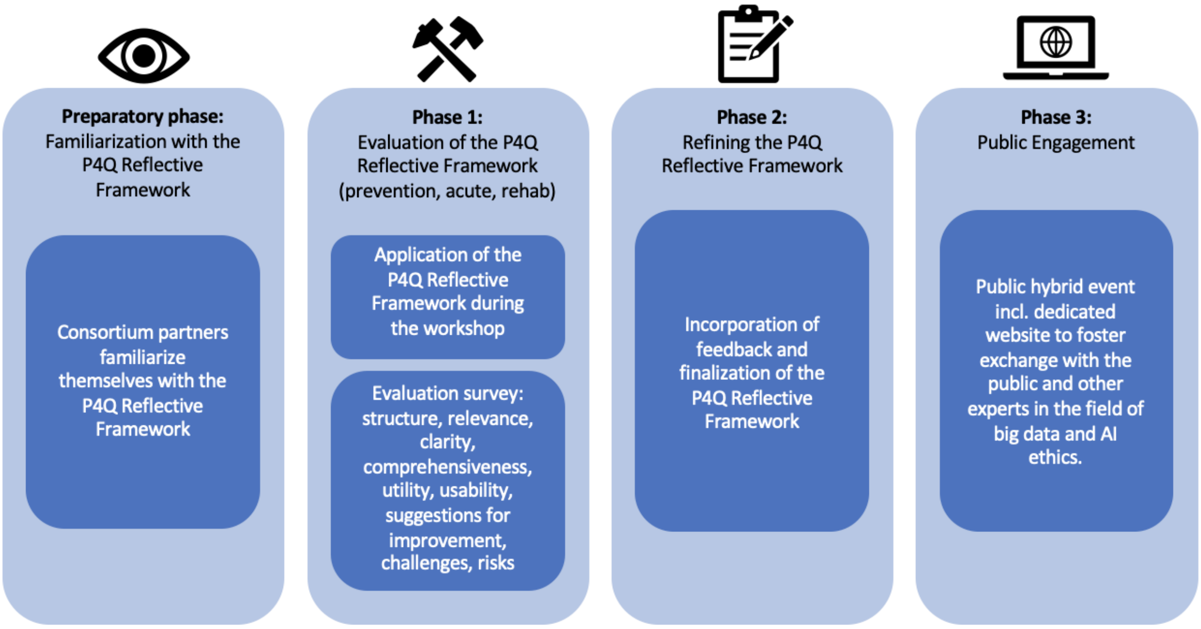
Preparatory phase:
Familiarization with the P4Q Reflective Framework
Consortium partners familiarize themselves with the P4Q Reflective Framework
Phase 1:
Evaluation of the P4Q Reflective Framework (prevention, acute, rehab)
Application of the P4Q Reflective Framework during the workshop
Evaluation survey: structure, relevance, clarity, comprehensiveness, utility, usability, suggestions for improvement, challenges, risks
Phase 2:
Refining the P4Q Reflective Framework
Incorporation of feedback and finalization of the P4Q Reflective Framework
Phase 3:
Public Engagement
Public hybrid event incl. dedicated website to foster exchange with the public and other experts in the field of big data and AI ethics.

Figure 1 Methodology

5 Findings

5.1 Observational findings from the framework application

During the three separate workshops, the task force members examined the suitability of the CDSS in the three different phases of stroke by applying the P4Q Reflective Framework. The following three sub-chapters for each of the task forces will first describe the environment/setting in which the discussions took place, before a discussion and analysis of the reflections on the framework.

The framework was successful igniting discussions in all the three task forces. The tone of the conversation in all of them was friendly and highly supportive. The frequency of engagement in the discussion varied strongly across the participants. Whereas some participants shared their views continuously, others only expressed themselves occasionally. Although applying the same ethical framework, the discussions within the groups interestingly diverged in quite different directions.

5.1.1 Development

As a first step, the participants were asked to reflect upon the problem that the tool aimed to address. The Prevention task force identified the morbidity of stroke as a key problem. Concerning the primary user group, a distinction between citizens and patients was quickly drawn, categorizing the latter as the primary end-user group and patients as the secondary group. The Acute task force devoted attention to the importance of the creation of trust and reliability for the end-user and problems at different levels that the tool aims to solve. The participants did, however, not specify what the end-user can expect in this regard. The members of the Rehabilitation and Reintegration task force group immediately began to discuss the solutions which are made possible through the tool, thereby reflecting on the “aims” rather than the “problems”.

Regarding the intended purpose of the tool in clinical practice, members of the Prevention task force quickly mentioned the potential improvement of the patient-caregiver relationship. Whereas it first was argued in the group that it the tool would merely add ‘add another layer’ to what the doctor does in stroke prevention, the discussion shifted toward a more differentiated view on the impact on this relationship, considering potential variations among clinicians. Related questions that were not resolved included the questions on how to get informed consent, what happens if individuals *do not provide informed consent,* how to tackle patients who refuse medical treatment (check) or do not wish to receive information, and how clinicians should use the tool). The Acute task force narrowed the purpose of the tool to recommendation of a treatment based on the individual features of the patient and guidance of the best possible outcome within reasonable time.

With regard to the use of the tool by clinicians and its impact on their professional role perceptions, the Acute task force members argued that the tool should be integrated in the workflow. Further specification on how this should be done would have been useful since it is widely known that clinicians work under a high level of pressure, which can make the integration of new process workflow more difficult. During the discussion, the focus shifted to the impact of the tool on clinicians. One of the participants argued that the tool can strengthen and support clinicians' decision-making. Another participant then argued that the person making the decision (e.g. for a surgery) and the one executing the decision (e.g. the surgeon) is not necessarily always the same person. It would have been valuable to further discuss the latter statement and how it was linked to the first claim. In the case of the Rehabilitation and Reintegration task force, the implementation of the tool and its impact on the professional role perceptions of clinicians, one of the participants explained that the tool is developed closely together with clinicians and that it therefore will improve clinicians' decision-making. Another participant added that one advantage of the tool is that bias driven by clinicians can be minimized through long-term experience with the program. It would have been useful to know how decisions could be improved, but also how the bias caused by the health care professional is related to the bias of the clinical decision support system.

The question of possible inappropriate uses of the tool and how to minimize such actions sparked a discussion in the Acute task force on the consequences of a use of the CDSS when it transitions from serving as a support to functioning as a decision-making tool. The members of the task force argued that experienced clinicians with extensive knowledge on the issue-matter would be able to question the authority of the tool. However, we note that they did not discuss in-depth how to prevent less experienced clinicians from overly relying on the tool in-depth. Also where clinicians are required to describe the process of decision-making in a statement, the risk remains that less experienced clinicians first make decisions with the help of the tool before producing a fitting explanation. Participants in the Prevention workshop highlighted the right not to know and possible biased findings due to the algorithm. Related to this, one of the workshop participants asked who would have access to the data. This was, however, not discussed any further by the workshop participants. Rehab and Reintegration workshop participants reflected on the necessity of clinicians to learn how to work with the tool and how explainability is related to this. With this, a part of the concerning the model's explainability was touched upon. The second part of this question concerning the impact of explainability on performance remained open.

Moving on to questions on data quality and representativeness, it was confirmed by members across all three task forces, without any further explanations, that the data had been obtained in an ethical manner. It would have been of interest to know what it means that the data has been obtained in an ethical manner, how the data had been obtained and which ethical principles that had been considered in the process. Proceeding to data quality and representativeness for the target population, participants in the Rehabilitation and Reintegration task force pointed to at the extensive efforts made to clean and acquire a deeper understanding of the data, including among other things to detect biases (e.g., age). Further, the participant added that bias also can represent an advantage. It would have been helpful to have this statement further explained. The discussion in the two other task forces on these issues were shaped by the project goal being developing proofs of concept. In both of these workshops, it was stated and agreed on that the level of data quality was appropriate for the development of the tool. One participant more specifically noted that proof of concepts can be biased and that discussions on the impact of data characteristics on performance and potential misuses were therefore theoretical.

Following the discussions on the representativeness of the data, the participants in the Acute task force initiated a discussion around technical devices currently used in the clinical setting and the importance of ensuring that the variety of their outputs is reflected in the training data. They further noted that this also extends to the representation of a diversity of treatment options in the data if the tool seeks to support treatment decisions. Patients, as key stakeholders, were missing in this discussion as was how this data might impact patients. What if a new type of treatment is released on the market which the tool “does not know about”? Will there be updates? Do updates require new validation? One of the members of the Rehabilitation and Reintegration task force underscored the importance of not drawing general statements from the possible results provided by the tool. Regrettably the participant did not further elaborate on the possible consequences for the population, e.g., the patients.

A further important group of questions treated the topics transparency and explainability. The Prevention workshop participants discussed the role of clinicians in the decision on how much information a patient should receive, how explainable a model must be and what information and to what extent information is made available for the end-user. One of the participants argued that this should be done following an evidence-based approach, however, that this would imply further efforts and costs for the AI-tool. The task force members did not completely clear whether it would have an impact of the performance of the tool.

Concluding the first part of the workshop, the groups discussed the meaning and impact of clinical validation of the tool. In the Prevention task force, the claim that clinical validation is less relevant for patients and clinicians represents one such issue was presented. One of the task force members explained that these stakeholder groups are generally interested in the outcome, and if the outcome is good, then the validation will not be paramount. Members of the Acute task force discussed the impact of clinical validation on clinicians’ trust and responsibility. They put forward the argument that clinical validation does not impact on trust or responsibility, since it would just be another information tool in the clinicians’ everyday work and compared the new tool with a MRI. From an observer perspective, it would have been valuable to hear more about the rationale for this comparison. Interestingly, an opposing argument was presented in the Prevention task force, where it was emphasized that validation would promote trust in clinicians, which would drive clinicians' willingness to take the responsibility of using the tool.

5.1.2 Deployment

The second part of the workshop on the deployment of tool began with interesting aspects on the disclosure of AI. In this part of the workshop, the practical aspects of the tool weighed heavily in the discussions. The introductory question concerned how much information can and should be disclosed to the patient. This discussion is anchored in the fundamental debate on whether one should have the right not to know and how comparable the disclosure of information is to in the case of other tools. Drawing the same comparison between existing tools and the AI-tool as earlier by members of the Acute task force, members of the Prevention task force argued that AI-systems do not necessitate more disclosure than for example imaging methods or laboratory tests. Again, further elaboration on the differences between AI and imaging methods and how they shape opinions could have been discussed further. In the Acute workshop, the participants did not provide a final answer to any of these two questions; however, the consensus was rather the opposite. Members of this task forced emphasized that artificial intelligence has a stronger impact on individual decision-making than other clinical tools and therefore perhaps should be treated differently.

Considering the question of how trust, patient autonomy and the doctor-patient-relationship may be impacted by predictive health information, one of the members of the Prevention task force argued that there is "simple math behind" AI. For the observers, this appeared to contradict the hitherto discussion of AI as a complex concept, from which the questions on disclosure arise. The result of using such systems on the doctor-patient trust-relationship remained uncleared. The question of the consequences for the relationship is also linked to the question of deskilling and the substitution of clinicians through programs. If the patient can use the tool, then the clinician or some functions of him/her becomes replaced as the patient to a large extent can inform himself. Members of the Rehabilitation and Reintegration discussed that patients are likely to be happy about personalized treatments and increases patient autonomy and that this will positively affect the doctor-patient-relationship. Further, deskilling was by some stated to be an inevitable consequence of a good tool. With this, also the question of whether human contact can be replaced was addressed.

Proceeding to questions of responsibility, members of the Prevention task force discussed how responsibility and liability depends on whether the clinicians and patients use the tool together or if the patient applies it independently. The members of the Acute task force chose to focus on the comparison of the tool to MRI while discussing what impact incorrect decisions by the tool may have on clinicians' moral responsibility. The participants did not consider the clinician to such a responsibility. Similarly, the clinician was not assigned responsibility in the case of monitoring and evaluation of the tool. Concerns related to patients and clinicians were not mentioned. This was also the case during the discussion of the consequences of lacking privacy, where participants shared their views from a business perspective and not from a patient and clinician perspective. This was however the perspective taken by the Rehabilitation and Reintegration task force which placed patients at the center of their discussion. The Rehabilitation and Reintegration task force members also agreed that privacy should be considered in all phases of the PRECISE4Q project, however, did not reach a consensus the responsibility of the developer

A final issue discussed among the members of the Acute task force was the issue of balancing health and privacy. Here the question of privacy risks, since the tool is consent-based, was raised. A more thorough discussion of the implications of consent and its scope would have been

5.2 Evaluation of the PRECISE4Q Reflective Framework

Figure 2 presents workshop participants’ evaluation of the PRECISE4Q Reflective Framework.

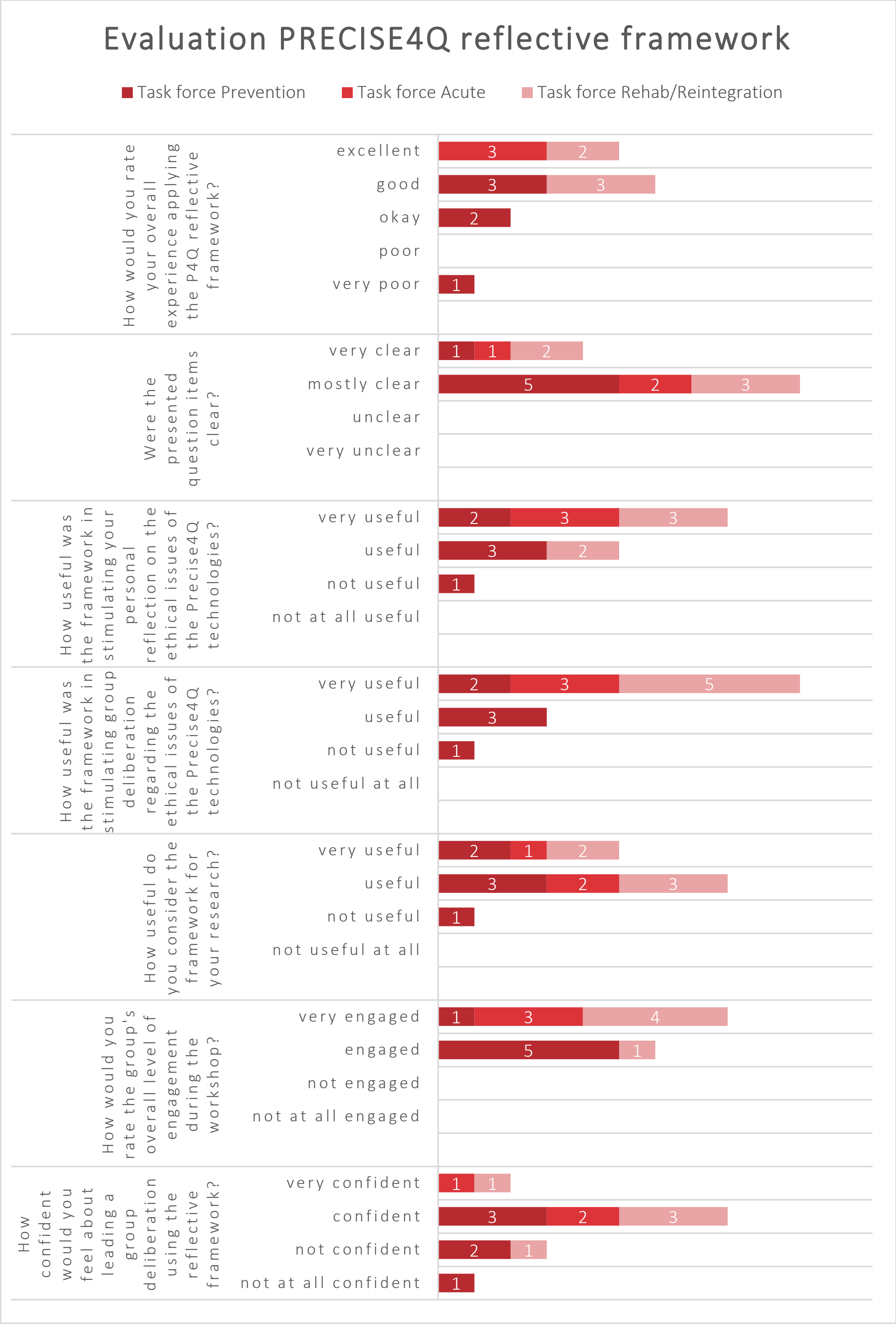

Figure 2 Evaluation PRECISE4Q Reflective Framework

Overall, most participants rated their experience applying the PRECISE4Q Reflective Framework as good or excellent (n=11). The three participants who rated their experience with the framework to be less than good reported that they particularly found it challenging to understand the meaning of certain concepts used in the questions, that they were unfamiliar with certain concepts, or a good enough understanding of the whole project (beyond the part they individually worked on).

As represented in Figure 2, all participants evaluated the presented question items as very or mostly clear (n=14) and all but one participant reported the framework to be useful to stimulate personal reflection and group reflection on ethical issues related to PRECISE4Q and its related technologies (n=13). Similarly, all but one participant (n=13) assessed the framework to be useful or very useful for their research. All participants reported that the group was very engaged (n=8) or engaged (n=6). Finally, most participant (n=10) responded that they felt confident about leading a group deliberation using the reflective framework.

Table 1 presents an overview of what participants liked, as well as of the perceived risks and challenges, and suggestions for improvement. The participants liked the framework's structure, flow and the content, how it stimulated interdisciplinary discussion it stimulated, the self- and group reflection it promoted, and the guidance it provided. The risks and challenges identified by the participants with regard to the application of the framework evolve around five themes: difficulties in understanding and interpreting the questions, lack of understanding and/or difficulties of engagement (including English language proficiency), group composition, limited applicability, effectiveness. In a final step, the workshop participants proposed improvements of the framework. These suggestions could be divided into four categories: structure and content, procedure, group composition, and scope of the framework.

Finally, with regard to coverage of ethical aspects, a majority of the participants reported that the project covers most of the relevant issues. A few participants pointed to missing aspects. These include gender and ethnic representation as separate topics, comparative perspectives across different groups (e.g., the implications of less well-off individuals' lower or higher probability to be processed by AI systems in comparisons with wealthier population groups), the ethics of prediction and consequences for patients, and aspects related to products that already are on the market.

Table 1 Summary of respondent feedback on the application of the P4Q Reflective Framework

What participants liked
Structure, flow, and content
• The appropriate, well-chosen, and well ordered themes and guiding questions to guide the group reflection • Appreciated the aspects about the highly relevant, but often neglected section on empathy in the Deployment
Interdisciplinary discussion
• Brought ideas together and supported gaining more knowledge of use among different stakeholders • Elicited interesting discussion
Self-reflection
• Promoted reflection on issues that frequently are neglected • Represented a tool to stimulate self-reflection on own work and its potential impact
Guidance
• It helped to thoroughly check if important principles were implemented as well as facilitated thinking about future steps and considerations.
Risks & Challenges
Understanding and interpretations of questions
• Some need of question clarification, explanation, and guidance on how to interpret them
Engagement
• Challenging to partake due to a lack of understanding of the project and /or specific information Difficult to participate due to lower English proficiency
Interdisciplinary, multi-stakeholder perspectives
• Challenging to discuss multi-disciplinary perspectives and to put oneself in the shoes of other stakeholders (e.g., patients and clinicians) • Risk of neglecting other stakeholders' views (i.e., patients and clinicians) among members of a homogeneous project group
Limited applicability
• Difficult to apply in complex projects where multiple datasets and research groups are involved • Questionable generalization of results to similar contexts
Effectiveness
• Risk of framework becoming a box-ticking exercise • Remaining risk of not discovering errors in data etc.
Suggestions for improvement
Structure and content
• Clarification of certain questions • Creation of scenarios to aid reflection and provide resources for next steps where problematic issues are identified • Division of framework into further sub-categories next to 'development' and 'deployment'
Procedure
• Having participants work through the framework prior to the workshop • Offering the possibility to comment in questionnaire while on Zoom or use of interactive tools to stimulate participation • Provision of more time
Group composition
• Inclusion of a variety of stakeholders in the discussion groups
Scope of the framework
• Expansion of the scope of the framework

5.3 Discussion

Despite the occasional need for further explanations on certain questions, the workshops and the evaluation demonstrate that the PRECISE4Q Ethical Framework was successful in stimulating discussion around ethical challenges related to the development and deployment of the tool.

In two of the three task forces, a change in the nature of the task force members' participation in discussion was observed over the course of the workshop. In the prevention taskforce, question responses were simpler and more limited earlier in the workshop, moving towards a more open discussion characterized by more extensive reflection on the project at the end of the workshop. In the case of the acute taskforce, the identification and discussion of possibly problematic features of the tool also increased towards the later phase of the workshop. Although a part of the explanation for this change may be that the beginning and the end of the workshop correspond with the two phases development and deployment, it is highly likely the application of the framework question by question has promoted deliberative discourse within the task forces.

The workshops also revealed some of the limitations of the framework in its current form. First, although the depth of the discussion in all of the taskforces was adequate, deeper reflection could have occasionally been stimulated through further questions by the workshop lead in order to cover even more aspects that are relevant for the development and deployment of ethically-sound tools. Second, despite participants' efforts to take the position of patients and clinicians, the homogeneous composition of the task forces undermined the variety of stakeholder perspectives presented in the group discussions and at times neglected the perspectives of patients and clinicians. In the acute taskforce, for example, merely legal issues were discussed in relation to the consequences of failing to conduct monitoring and evaluation. Patients (or clinicians) ere not mentioned. Similarly, when discussing the consequences of lacking privacy, participants focused only on the business perspective, and not from a patient and clinician perspective. This was particularly noted in the workshops of the rehabilitation and reintegration taskforce. Finally, related to the aforementioned issue, the discussions were at times characterized by a high level of agreement, with few opposing arguments. In the prevention workshop, for example, this was particularly pronounced in the early phase of the workshop with critical reflection on other participants statements increasing over time. A similar picture was evident in the acute taskforce where few contradictory arguments were presented when an opinion was expressed by one of the participants. The two latter issues, disregarding groups of stakeholders and high levels of agreement may inhibit the consideration of a variety perspectives in the development and deployment of a tool.

6 The Revision of the P4Q Reflective framework

Both the observational findings and the survey evaluation support the notion that the P4Q Reflective Framework is successful in stimulating ethical reflection. They findings, however, also reveal certain limitations of the framework. The following sections describe how we revised the framework to improve its applicability and ability to promote ethical deliberation and mitigate limitations of the framework. Further, it provides recommendation for future applications of the framework.

The pilot test of the framework led to minor modifications in item wording and order. Further, based on the participant feedback and workshop experience, we reformulated the question item concerning obtaining data in an ethical manner to promote broader reflections ethical data. Similarly, we further specified the question of explainability and its impact on performance. We Refrained from the further division of the framework into categories or sub-categories in order to keep the clear structure of the structure.

Based on our experience from the workshops and the participant feedback, we have several recommendations for future applications of the P4Q Framework. The first issue concerns allowing the participants to complete the framework by themselves prior to and discussions or focused workshops that apply the framework. This step may aid overcoming challenges related to language proficiency as well as differences in background or conceptual knowledge. Participants not only mentioned difficulties expressing themselves due to language but also due to a lacking familiarity with ethical concepts or the meaning of concepts such as “de-skilling” and “explainability” in the context of AI. Allowing future workshop participants to prepare can bring all participants to a “knowledge base-level” before the beginning of group discussions, making them more fruitful and complete.

Second, we recommend creating scenarios around which the participants can discuss. Without altering the general structure of the framework, scenarios can facilitate discussion as they are adaptable to the specific context in which the framework will be applied. Arguably it will particularly aid participants who may not have a full project overview or extensive knowledge on the ethics of AI in their reflection, and thereby enhance group reflection.

Finally, we recommend ensuring the representation of different stakeholder groups (e.g. clinicians and patients) in any future application of the framework. There are several ways this might be approached. While including external stakeholders could broaden and deepen the discussion and ethical deliberation, future applications of the framework could also be carried out as part of a role play, where participants are assigned and prepared to play different roles (e.g., clinician, patient, family member, developer) which they are asked to enact when applying the framework as a group. A role play exercise may also help to foster empathy and understanding for the needs and priorities of others [10].

7 Limitations

The work presented here should be considered in light of some limitations. Specifically, the pilot test of the framework was carried out within three project-internal workshops and a post-workshop evaluation survey; it did not include any external stakeholders (e.g., clinicians, stroke survivors, family members) due to feasibility concerns in a practical setting.

Another limitation may be seen in the fact that the workshop and the development of the reflective framework were led by the same investigators at ETH. This may have led to a positive bias in the evaluations due to social desirability bias. We tried to mitigate this risk by making the survey anonymous and by asking for participants' honest feedback and suggestions for improvement. Leading both the framework development and its application may have also led to an observer bias, in that we experienced participants to be more engaged with the framework than they actually were. We aimed to mitigate this risk by carefully reviewing workshops recording together with a research assistant who was neither involved in the framework development nor its application during the workshop. Upon reviewing the recordings, the research intern produced an independent observation report which was then compared to the field notes taken by the workshop lead and fed into the subsequent analysis.

8 Outlook

It is our hope that the framework presented here will be helpful in guiding the final stages of the project and will also find adoption beyond the project's completion.

When applying to framework to other AI-based medical technologies beyond PRECISE4Q, we advise consortium partners to involve relevant stakeholders beyond the development team wherever possible (e.g., patients, clinicians, informal caregivers). Should it not be feasible to bring together a diverse group for the assessment, the applications of the framework could also be carried out as part of a role play, where participants are assigned and prepared to play different roles (e.g., clinician, patient, family member, developer) which they are asked to enact when applying the framework as a group.

9 References

1. Morley, J., et al., *From what to how: an initial review of publicly available AI ethics tools, methods and research to translate principles into practices*, in *Ethics, Governance, and Policies in Artificial Intelligence*. 2021, Springer. p. 153-183.
2. Bélisle-Pipon, J.-C., et al., *Artificial intelligence ethics has a black box problem.* AI & SOCIETY, 2022: p. 1-16.
3. Mittelstadt, B., *Principles alone cannot guarantee ethical AI.* Nature Machine Intelligence, 2019. **1**(11): p. 501-507.
4. McNamara, A., J. Smith, and E. Murphy-Hill. *Does ACM's code of ethics change ethical decision making in software development?* in *Proceedings of the 2018 26th ACM joint meeting on european software engineering conference and symposium on the foundations of software engineering*. 2018.
5. Vakkuri, V., et al. *"This is Just a Prototype": How Ethics Are Ignored in Software Startup-Like Environments*. in *International Conference on Agile Software Development*. 2020. Springer, Cham.
6. Vayena, E., A. Blasimme, and I.G. Cohen, *Machine learning in medicine: Addressing ethical challenges.* PLoS medicine, 2018. **15**(11): p. e1002689.
7. Vayena, E. and A. Blasimme, *Towards Adaptive Governance in Big Data Health Research*, in *The Cambridge Handbook of Health Research Regulation*. 2021, Cambridge University Press. p. 257-265.
8. Blasimme, A. and E. Vayena, *The Ethics of AI in biomedical research, patient care and public health*, in *Oxford Handbook of Ethics of Artificial Intelligence*. 2020, Oxford University Press. p. 703-718.
9. Vayena, E. and A. Blasimme, *Health research with big data: time for systemic oversight.* Journal of Law, Medicine & Ethics, 2018. **46**(1): p. 119-129.
10. Bearman, M., et al., *Learning empathy through simulation: a systematic literature review.* Simulation in healthcare, 2015. **10**(5): p. 308-319.

10 Appendix

Table I: Evaluation survey – PRECISE4Q Reflective Framework
Q1: How would you rate your overall experience applying the P4Q reflective framework? Excellent Good Okay Poor Very poor
Q2: Were the presented question items clear? Very clear Mostly clear Unclear Very unclear
Q3: How useful was the framework in stimulating your personal reflection on the ethical issues of the PRECISE4Q technologies? Very useful Useful Not useful Not at all useful
Q4: How useful was the framework in stimulating group deliberation regarding the ethical issues of the PRECISE4Q technologies? Very useful Useful Not useful Not useful at all
Q5: How useful do you consider the framework for your research? Very useful Useful Not useful Not useful at all
Q6: How would you rate the group's overall level of engagement during the workshop? Very engaged Engaged Not engaged Not at all engaged
Q7: How confident would you feel about leading a group deliberation using the reflective framework? Very confident Confident Not confident Not at all confident
The following questions were open-ended questions: Q8: What did you find particularly challenging about applying the framework? Q9: In your view, does the Framework cover all relevant ethical aspects or are there aspects missing? Q10: Where do you see potential risks of the P4Q Reflective Framework? Q11: How confident do you feel about leading a group deliberation using the reflective framework? Q12: How could the P4Q Reflective Framework be improved? Q13: Do you have any other comments, suggestions, or feedback?