

Working Group on Clinical Evaluation of AI for Health (WG-CE): Workshop

1) Introduction to DEL07.4

(Agenda – setting the stage)

2) Discussion of DEL07.4 & next steps

(Agenda – plenary deliverable)

Eva Weicken, Fraunhofer HHI, Germany

Email: eva.weicken@hhi.fraunhofer.de

1) Introduction to DEL07.4

- WG-CE objectives:
 - Build a **community of collaboration** around **clinical evaluation** of AI for health
 - Develop **guideline documentation** for use by researchers, clinicians/patients, developers, and policy-makers
- DEL07.4 key **output document** of WG-CE
- **Basis for collaboration** in subgroups

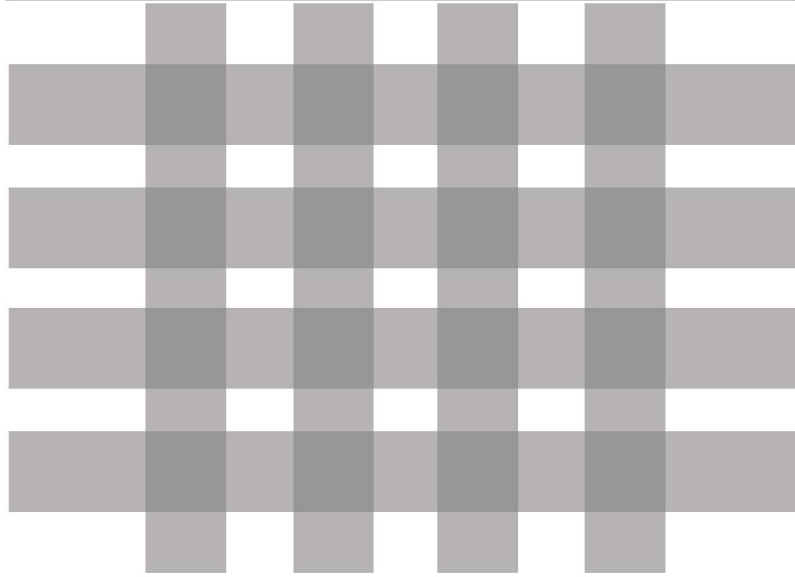
FG-AI4H topic groups & working groups

FG-AI4H working groups

- Data and AI solution assessment methods (WG-DAISAM)
- Data and AI solution handling (WG-DASH)
- Ethical considerations on AI4H (WG-Ethics)
- Regulatory considerations on AI4H (WG-RC)
- Clinical evaluation (WG-CE)

FG-AI4H topic groups

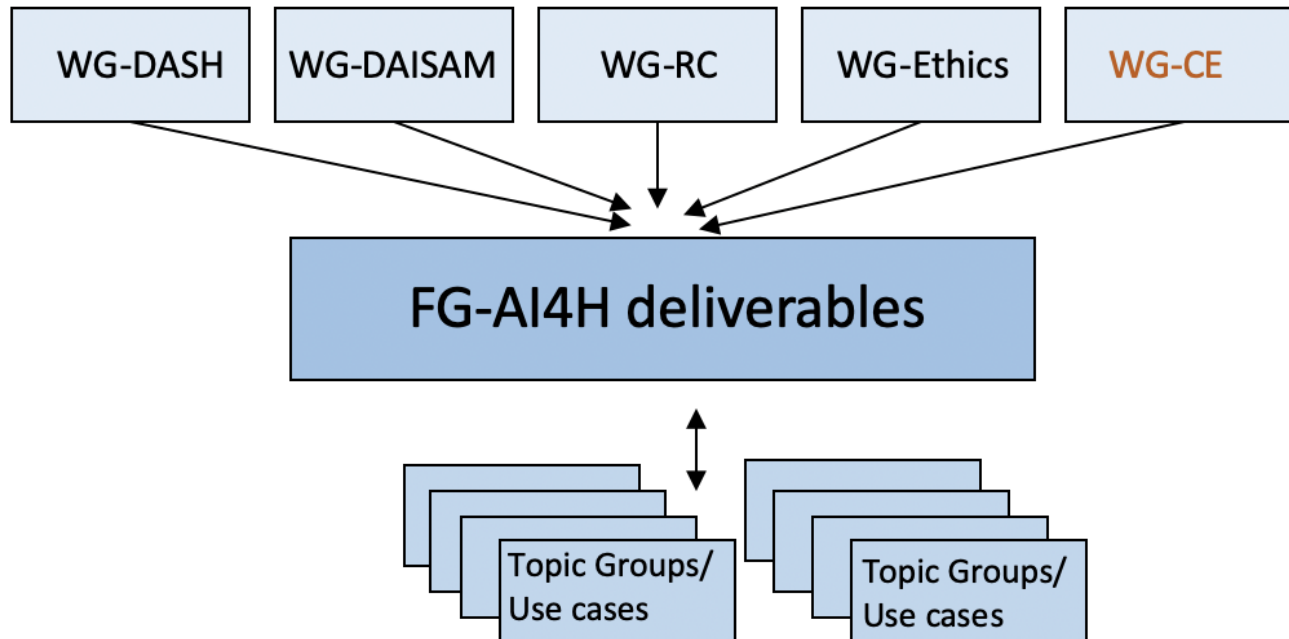
- | | |
|------------------|-----------------|
| TG-Cardio | TG-TB |
| TG-Derma | TG-DiagnosticCT |
| TG-Falls | TG-Bacteria |
| TG-Histo | TG-Dental |
| TG-Neuro | TG-Diabetes |
| TG-Outbreaks | TG-Endoscopy |
| TG-Ophthalmology | TG-FakeMed |
| TG-Psy | TG-Malaria |
| TG-Snake | TG-MCH* |
| TG-Symptom | TG-Radiology |
| | TG-MSK+ |



*MCH: maternal and child health
 +MSK: musculoskeletal medicine

Combining features from both groups is the “Ad Hoc Group on Digital Technologies for COVID Health Emergencies” (AHG-DT4HE)

FG-AI4H deliverables



No.	Deliverables categories
00	Overview of the FG-AI4H deliverables
01	AI4H ethics considerations
02	AI4H regulatory best practices
03	AI4H requirements specification
04	AI software life cycle specification
05	Data specification
06	AI training best practices specification
07	AI4H evaluation considerations
08	AI4H scale-up and adoption
09	AI4H applications and platforms
10	AI4H use cases: Topic description documents
11	Open source reference implementation

DEL07.4

- Part of DEL07 “**AI4H evaluation considerations**” (umbrella)

N°	Deliverable title
DEL07.1	AI for health evaluation process description
DEL07.2	AI technical test specification
DEL07.3	Data and AI assessment methods reference
DEL07.4	Clinical evaluation of AI for health
DEL07.5	Assessment platform

- **Output document** of WG-CE
- **Collaboration** with **other WGs and TGs** of the FG-AI4H

Specific issues

- Phases of evaluation
- Efficacy and comparative efficacy
- Safety
- Generalisability/bias and inclusiveness
- Evaluation adaptive/learning models
- Reporting of evaluation (following EQUATOR)
- **Clinically meaningful endpoints**
- **Post-deployment surveillance (overlap with regulation)**
- **Specific considerations for low- and middle-income settings**
- Collaboration and engagement

Current evaluation frameworks

- Draw on current evaluation frameworks
 - EQUATOR Network: CONSORT AI & SPIRIT AI reporting guidelines
 - IMDRF SaMD: Clinical Evaluation
 - Strong examples: Digital health scorecard, Model facts labels

2) Discussion of DEL07.4 & next steps

- **Status quo DEL07.4** (presentation in [Sharepoint](#))
- Remember to **become a focus group member**
 - Register on FG-AI4H [website](#)
 - Join [mailing list](#)

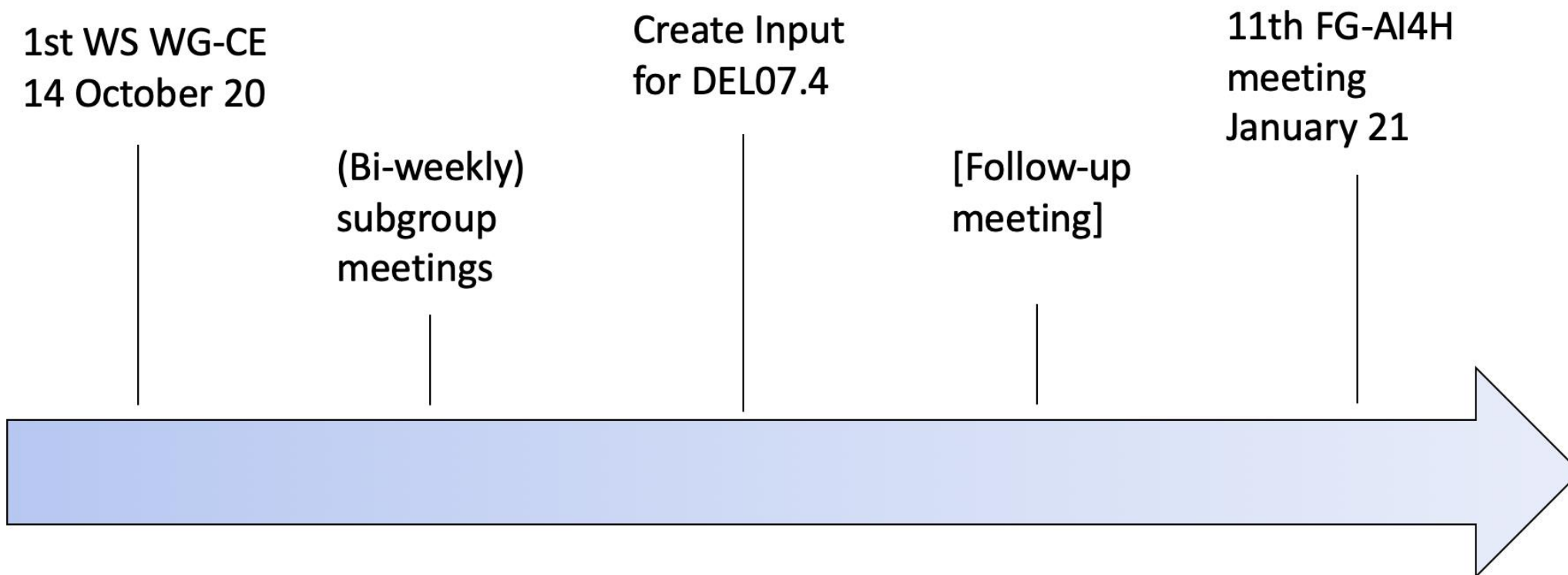
Subgroups

- **Subgroups:**
 - **Pre Deployment Clinical Evaluation**
 - **Post Deployment Clinical Evaluation**
- **Volunteer** to join subgroups (assign to **spreadsheet [Link](#)**)

Working in subgroups

- Subgroups have **members of the FG-AI4H** (co-chairs) to help coordinate & keep track of progress
- Subgroups define **own internal way of work** (e.g., in shorter biweekly follow-up meetings)
- Exchange about progress within subgroups by co-chairs

Next steps



Pre Deployment Clinical Evaluation



Analytical Validation

- Require a specific statement about the use case
- External validation in a dataset that is representative of the population in which the use is intended
- Understand situations or groups in which performance may be lower

User validation

- How and where does the tool fit in the workflow?
- Does the interaction of the user affect the performance of the tool?

Clinical studies (safety and efficacy)

- Different levels of tool/integration into the workflow/level of autonomy may require different levels of study
- Retrospective studies – prospective cohort studies – interventional/comparative studies (stepped wedge/RCT)

Cost effectiveness

Post Deployment Clinical Evaluation



Medical interventions - Risk vs Benefit

SaMD Regulation - Intended use, intended users and claimed benefits

Need robust evidence for claimed benefits to balance against potential (mitigated) risks

The higher the risk, the bigger the burden of evidence

Evidence of effectiveness - Performance, Safety and Impact

Impact - Patients, Clinical workflow and Health systems/Population

How much can be achieved with independent dataset benchmarking?

How much needs to be addressed by other means (eg prospective on site clinical studies)

Limitations of benchmarking (eg proxy to real clinical evidence)

Limitations of “real” clinical evidence - Slow. AI systems evolving rapidly