General Study Protocol

Project AIR

1. **Rationale**
   Currently more than 150 artificial intelligence (AI) based software products are on the European market for radiology[1]. An overlap exists in use cases being addressed by the different vendors. For radiology departments it becomes increasingly difficult disseminating which software is optimal for their situation. Performance of the software is one of the factors taken into consideration in the procurement process. However, there is a lack of public evidence regarding the performance of these products. A recent study showed that only 31 out of 100 CE marked products that were evaluated had peer reviewed evidence demonstrating diagnostic accuracy efficacy[2]. If case studies are available, usually different datasets are used for validation making it difficult to compare products and vendors.

2. **Objective**
   The aim of Project AIR is to increase transparency in the market of AI for radiology. Project AIR develops a physical and digital infrastructure to perform validation studies demonstrating the stand-alone performance of commercially available (CE marked or FDA cleared) AI based software for radiology use cases on a hidden, multicenter dataset.

   The process is repeatable when algorithms are updated or new vendors join. Results are made publicly available via www.Grand-Challenge.org[3] and linked to in www.AIforRadiology.com[1]. The aim is to address all use cases that have multiple AI solutions (Appendix A shows an example list).

3. **Study design**
   Project AIR is an ongoing retrospective study in which commercial AI products are being validated on a dataset from multiple centers. The process is split up in several phases as shown in Figure 1.

   PROPOSE: Study protocols are shared publicly on www.AIforRadiology.com/project-air and are open for feedback from both vendors as clinical institutes for a period of at least one month.

   JOIN: Feedback is processed and vendors and hospitals decide to join the finalized proposal. Contracts are signed.

   GATHER: Clinical data (hospitals) are collected. Algorithms (vendors) are temporarily made available to the research team.

PERFORM: Reader Studies are set up on Grand Challenge. Radiologists read the data to provide context to the stand-alone software performance. Algorithms are validated on the gathered data.

PUBLISH: Results are analysed and published on Grand Challenge. A link to these results will be added on AIforRadiology.

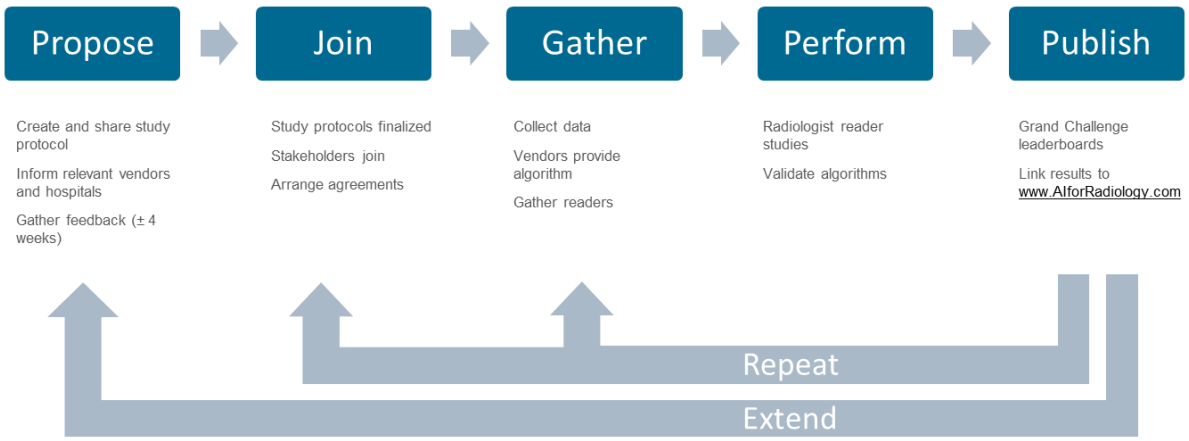

Figure 1. Overview of the study process

4. Data providers

4.1. Handling and storage of data

4.1.1. Upload

The data provider and data receiver sign a Data Transfer Agreement (DTA) before data sharing takes place. Data is securely shared through SURFdrive by SURF (collaborative organisation for ICT in Dutch education and research). The data provider is responsible for the anonymization of the data. Imaging data is uploaded in the DICOM format. Clinical data is uploaded through an excel file for which a specific template is provided for each use case.

4.1.2. Usage

Data is kept hidden to ensure the independent and fair validation of the algorithms. A small subset (aim of three images per data provider, per use case) to be determined by the dataprovider are made public under the CC-BY license. These images will be used to:

- Check the algorithm results with the expected results by the vendor;
- Serve as example images (showing the images, reference standard and algorithm output);
- Illustrate in public outings (publications, presentations, etc).

4.1.3. Storage

Data is stored for ten years after the DTA has been signed with the purpose of this study. When data is not deemed relevant anymore after this date for validation purposes (clinical data distribution has changed, because of new scanners, sequences, protocols etc.), we strive to make the data available to the public under the CC-BY license. More information can be found in the DTA which is available upon request.

4.2. Publication

The Project AIR Working Group is co-author of the scientific publication and data providers can opt to join. Read more in 7.1.

5. Vendor participation

5.1. Eligibility

Algorithms to be validated in Project AIR should be CE marked and/or FDA cleared. The task being validated should be part of the intended use according to the clearance.

5.2. Handling and storage of algorithms

5.2.1. Upload

To ensure a fair process, the test data will not be shared with the vendors. The vendors are asked to make their algorithm temporarily available to evaluate the validation set. Technically, multiple options exist and feasibility and preferences are discussed with the participating vendors individually. We prefer to make a test installation on a temporary AWS virtual machine.

5.2.2. Storage

Algorithms will be stored until processing of the data has taken place with a maximum of a year after upload. In case software has been updated, vendors may provide a new version of the algorithm (if CE marked/FDA cleared) at a maximum rate of once per year. Results are updated accordingly.

5.3. Checkpoint

For each new dataset being evaluated a small, public subset of the data (aim of three images per data provider, per use case) are made available and shared with the vendor to ensure the outcomes from the virtual machine or docker container are as expected. The images from the subset will also function as demo cases and examples for publication.

5.4. Publication

Results are presented in a transparent manner, naming the product and vendors. Vendors are included in the acknowledgement section of the scientific publication. Results are made available through Grand Challenge. Read more in section 7.

5.5. Ethical and legal considerations

Algorithms are only used for the purpose of this study. Algorithms are not duplicated or shared with third parties.

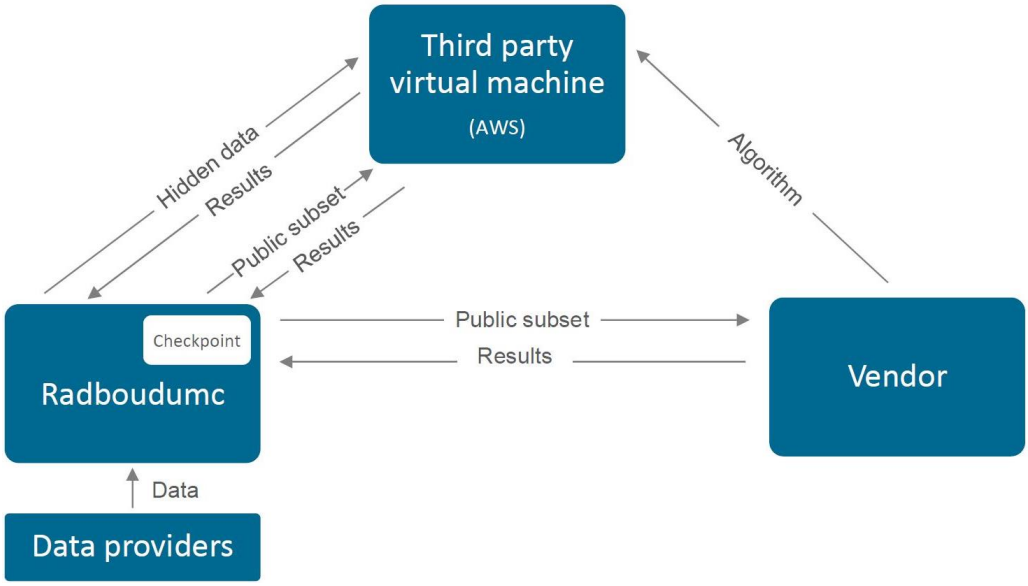

Figure 2. Overview of the technical process

6. Reader studies

6.1. Participation

Reader studies provide context to the performance of the AI algorithms. Radiologists (resident, general radiologist, specialized radiologist) are invited to read (a subset of) the studies using the infrastructure and workstations of Grand Challenge. To try out a demonstration reader study for lung nodule detection, please click here[4]. The residents and radiologist are asked to submit the number of years of experience, the country of work, role (resident, general radiologist, specialized radiologist), specialization (general, breast, cardiovascular, chest, emergency, abdomen, head and neck, musculoskeletal, neuro, pediatric, intervention) and whether he/she would like to be part of the Project AIR Working Group. The aim is to have a minimum of ten readers per use case. Radiologists that were directly involved in the data collection for a specific use case can not join as a reader for that use case. They are invited to become a reader for one of the other use cases.

6.2. Publication

Reader study results are published on Grand Challenge in an aggregated and/or anonymous form. Readers get insights into their own performance in relation to the reference standard. The Project AIR Working Group is co-author of the paper and readers can opt to join. Read more in 7.1.

6.3. Ethical and legal considerations

Readers have to confirm that they view the data only in context of this study and in no condition save or duplicate the data or have right to the data.

7. Publication of results

7.1. Scientific publication

A single article will be submitted to a peer-reviewed journal about the first set of use-cases and vendors to demonstrate the framework and process. The initial results will be presented. The extensive and most up to date results will be available online (7.2). With the expansion of use cases, papers may be published on the results of use cases specifically. The Project AIR Working Group is co-author of the scientific publications. Participating vendors are mentioned in the acknowledgements.

7.2. Online publication of results

Results and subresults are shared as a leaderboard on Grand Challenge. On AIforRadiology, the applicable products get a referring link to the leaderboard. Products that were validated through this procedure appear higher up in the listing (in order).

References

1. AI for Radiology, Radboud university medical center, www.AIforRadiology.com
2. van Leeuwen, K.G., Schalekamp, S., Rutten, M.J.C.M. et al. Artificial intelligence in radiology: 100 commercially available products and their scientific evidence. Eur Radiol (2021). https://doi.org/10.1007/s00330-021-07892-z
3. Grand Challenge, Radboud university medical center, www.Grand-Challenge.org
4. Demonstration reader study, Grand-Challenge, https://grand-challenge.org/reader-studies/project-air-demo-lung-nodule-detection-x-ray/

Appendix A

A non-exhaustive list of isolated AI tasks commercially offered by multiple vendors.

BREAST
Breast density mammography
Breast cancer detection mammography (2D)
Breast cancer detection tomography (3D)

NEURO
LVO detection CTA
ICH detection CT
Brain MS lesion quantification MR
Brain volume/region quantification MR

CHEST
Lung nodules frontal x-ray
Abnormality detection chest X-ray
Pneumothorax detection chest X-ray
Atelectasis detection chest X-ray
Pleural effusion detection chest X-ray
Tuberculosis detection chest X-ray
Covid-19 detection chest X-ray
Covid-19 detection chest CT
Lung nodule detection chest CT

MSK
Bone age hand radiographs
Cervical spine fractures CT
Knee Osteoarthritis XR
Fracture detection XR

ABDOMEN
Prostate volume quantification MR