Appendix 5

Project summary:

Introduction to artificial intelligence

"Artificial Intelligence" or simply "AI" is a broad term for computer software programmes that aim to imitate human intelligence by learning and making decisions. Generally, these software programmes or "algorithms" consists of a set of instructions to perform a specific task. However, some tasks, which may be easy for humans to do, are very difficult to write as instructions for computers to complete; a classic example being deciding if an image shows a cat or a dog. This is where AI algorithms can be useful as, rather than writing step-by-step instructions for the task, these algorithms are "trained" by being given an input (such as a picture of a cat) and being told the correct output ("this is a picture of a cat"). For each training image the algorithm's own output is compared with the correct output and small adjustments are made to its internal calculations to minimise the difference between the two. With a large amount of good quality training data these algorithms learn to complete a task well without ever being explicitly told how to do so.

In radiology, AI algorithms are trained in this manner to detect disease using thousands of medical images which a radiologist has marked as containing cancer or not. A new set of images, which weren't used for training, can then be presented to test the algorithm's ability to detect cancer.

Our research

Our research will systematically test how well AI algorithms, which have been developed by different groups and companies, perform in the context of breast cancer screening, with a view to improving breast cancer screening.

Breast screening (using mammogram x-rays) is conducted as part of a national programme in the UK for women aged between 47 and 70 years. With women older than 70 being able to self-refer. At present, the screening mammograms are read by two expert readers (radiologists and radiographers), and around 70-80% of the cancers are detected. Screening and early detection of breast cancer is known to improve outcomes for the women affected however it is a labour intensive process. It is believed that AI algorithms have the potential to work alongside expert readers to ease the workload in the breast screening programme and make it more effective by:

- Improving the accuracy of reading mammograms and cancer detection
- Acting as one of the two readers to reduce the manpower requirements
- Making reading more efficient by marking and prioritising mammograms which it believes are more likely to contain cancer
- Identifying women with dense breast tissue, for whom mammography is less sensitive, to benefit from additional imaging methods such as ultrasound or MRI.

To be able to use AI algorithms within the NHS we require regular monitoring and data security infrastructure. This requires a partnership between the NHS, academics and

23.10.20 v1.4
IRAS 258761

commercial companies to develop, test and regularly monitor as well as update systems ensuring both patient safety and patient data safety.

Our database

As part of this research we plan to build a database containing mammograms from approximately 150,000 women to be representative of the UK population. The mammograms will be obtained from the Cambridge and Norwich Breast Screening Programmes and the database will form an essential resource for testing AI algorithms as well as being valuable to the creation and development of new AI algorithms. This is because both training and testing benefit from large numbers of varied, high quality images. We will initially collect the existing mammograms of women who attended screening between 2011 and 2019 and, during the course of our research, we will continue to collect new screening mammograms until 2022.

First, a key database, of **non-anonymised** data, will be created and stored at the hospital (Cambridge University Hospitals NHS Foundation Trust / Norfolk and Norwich University Hospital NHS Foundation Trust) securely within an NHS IT system. This key database will contain information that is identifiable information, which is normally collected as part of routine care (date of birth, NHS number, hospital number, screening date, screening site, image accession number, episode record ID and screening number). These identifiers allow us to find the mammogram images and clinical information relating to the image, for each woman. A new unique trial ID will also be created and added, to identify each case. The management of this database will follow the security and data management rules for confidential information within the health sector. Access to this database will only be allowed by specific members of the research team who have been granted permission for access. This database will not be used for testing and will remain at the hospital site.

Second, a final testing database containing only **de-identified** information will be created and stored at the university (Cambridge). The de-identified information includes the mammogram images and clinical information relating to the image for each woman; details of previous breast biopsies, treatment (e.g. surgery or radiotherapy), and genetic risk factors alongside the unique trial ID. Data will be de-identified according to national standards such that all the information that could be used to identify someone (name, address, date of birth, NHS number, etc.) will be removed. It is this de-identified database that will then be used for testing.

Our AI plan

Many AI algorithms have been developed by universities and computer software experts following recent technological advances. Using our proposed final testing database of mammograms, the most advanced of these AI algorithms will be tested on data that is representative of the UK breast screening population.

Our work will also aim to highlight the performance and limitations of each AI algorithm, which is vital for clinicians to understand how to use them effectively in future real-world practice.

Appendix 5

Our data plan

The de-identified final testing database for research will be stored at University of Cambridge. All our studies testing the AI algorithms and the analysis of the results will take place there as well. The database will be governed by a Database Access Committee who will oversee all uses of this database.

In certain circumstances, limited amounts of de-identified data will be released to academics and commercial companies but with strict restrictions on how data can be accessed and used. This could occur for purposes such as adapting an algorithm to function with the type of mammogram which will be used in testing, allowing for more fair comparison between algorithms which were originally trained on different data.

The national ethics boards, the authorities which approved this study, have concluded that consent from individual women is not required for their mammograms to be de-identified and included in the creation of an image database governed by strict information governance policies.

Information posters will be displayed in screening centres and on screening vans that are taking part in this research to provide the research team's contact details. Women who do not want their data used for this research can contact the team to discuss opting-out.

This research is part of a Cancer Research UK programme grant for 'Risk adaptive breast cancer screening: a tailored imaging approach'.

Contact details:
Phone: 01223 348937
Email: braid.trial@nhs.net

Address:
Department of Radiology
University of Cambridge School of Clinical Medicine
Box 218 Cambridge Biomedical Campus
Cambridge
BB2 0QQ

Website: https://radiology.medschl.cam.ac.uk/research/research-themes/breast-imaging/cambridge-cohort/

23.10.20 v1.4
IRAS 258761