A method to benchmark AI algorithms for triage in breast screening programmes

N.R. Payne[1], J. Rothwell[1], Z. Deng[1], E. Atakpa[2], K. Freeman[2], S. Taylor-Phillips[2], and F.J. Gilbert[1,3]

1: Department of Radiology, University of Cambridge, Cambridge, UK
2: Division of Health Sciences, University of Warwick, Warwickshire, UK
3: Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK

np520@cam.ac.uk

Purpose

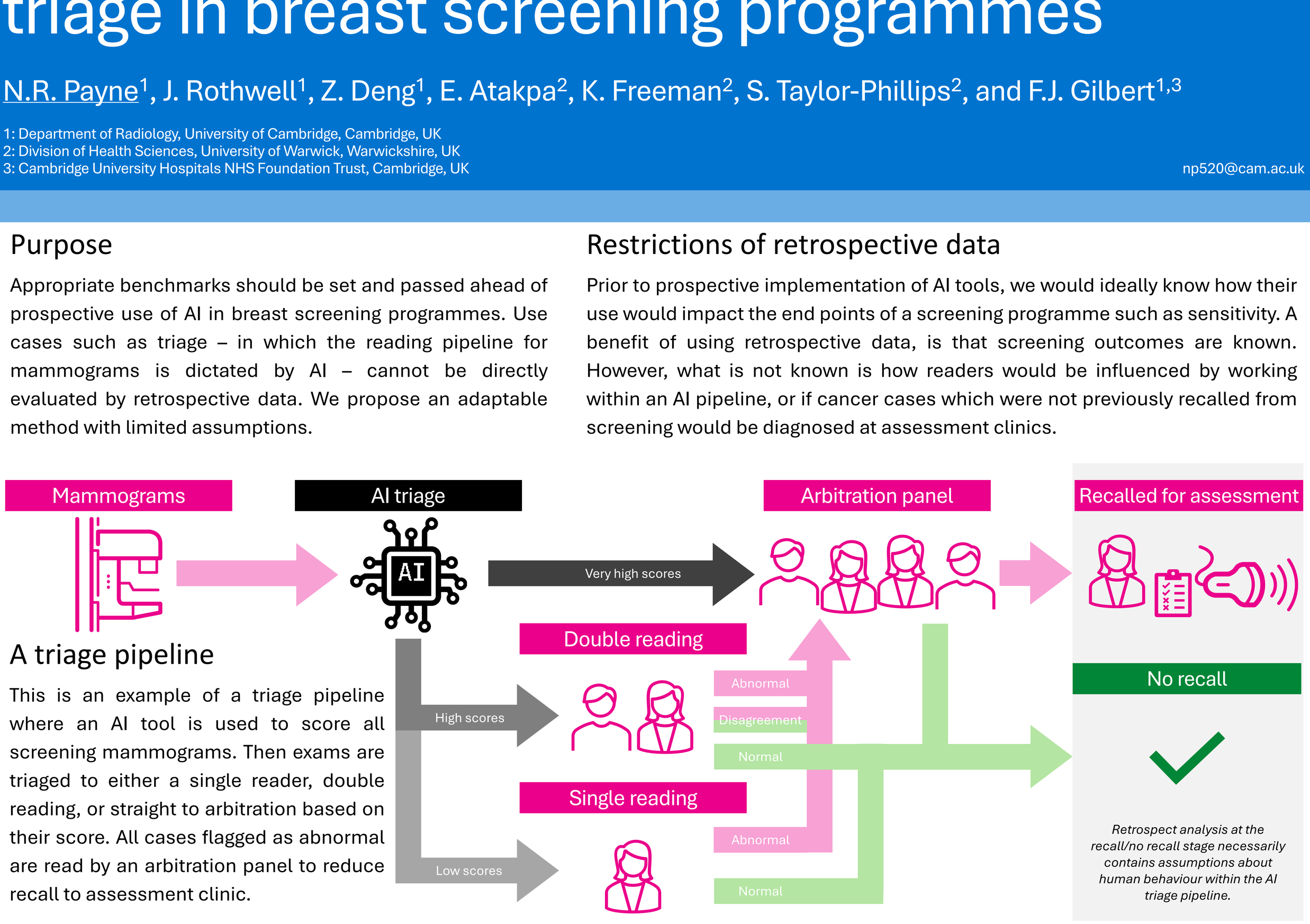
Appropriate benchmarks should be set and passed ahead of prospective use of AI in breast screening programmes. Use cases such as triage – in which the reading pipeline for mammograms is dictated by AI – cannot be directly evaluated by retrospective data. We propose an adaptable method with limited assumptions.

Restrictions of retrospective data

Prior to prospective implementation of AI tools, we would ideally know how their use would impact the end points of a screening programme such as sensitivity. A benefit of using retrospective data, is that screening outcomes are known. However, what is not known is how readers would be influenced by working within an AI pipeline, or if cancer cases which were not previously recalled from screening would be diagnosed at assessment clinics.

A triage pipeline

This is an example of a triage pipeline where an AI tool is used to score all screening mammograms. Then exams are triaged to either a single reader, double reading, or straight to arbitration based on their score. All cases flagged as abnormal are read by an arbitration panel to reduce recall to assessment clinic.

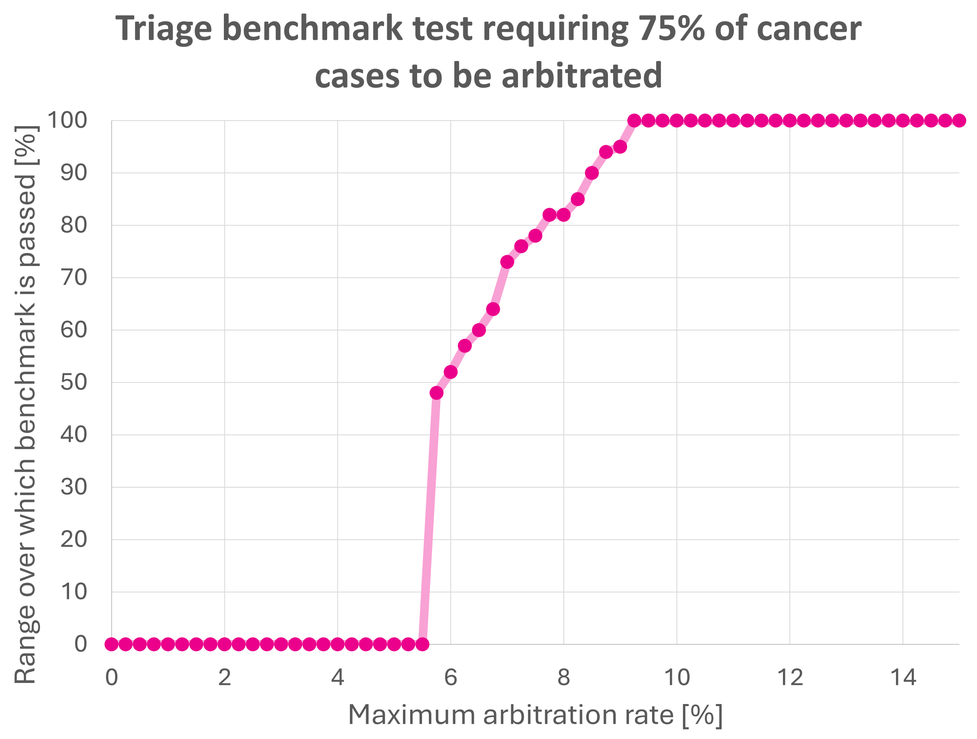


Reduce assumptions

As we cannot know which cancers would be diagnosed at assessment, we instead restrict analysis to the set of exams sent to arbitration and can set a targeted percentage of cancer cases to be included. For example, we could move from a 67% sensitivity target to 75% of cancer cases being arbitrated. To account for the range of influence an AI score could have on a reader at a given arbitration rate, we can use a weighted sum to vary between readers ignoring AI (i.e. their original decision) to complete deferral (i.e. the AI decision). A benchmark pass rate can then be found.

Materials and methods

Lunit INSIGHT v1.1.7 was used to score 73,652 GE mammograms including 860 cancer cases (590 screen detected and 270 interval cancers). The top 1% were sent to arbitration, the lowest 70% to a single reader, and the remainder were double-read. A benchmark was defined as 75% of cancer cases being sent to arbitration while maintaining a maximum arbitration rate of 10%. Reader opinions were converted to a score (normal = 0, abnormal = 1) – these were averaged for double-reading. The AI score was normalised to fall between 0 and 1. The combined score was computed as a weighted sum across the full range of AI influence and the proportion of this range for which the benchmark was passed was calculated at each maximum arbitration rate.



Results

The target of referring 75% of all cancer cases to arbitration met the benchmark across the full influence range at a maximum arbitration rate of 10%, giving high confidence that this would also be found in practice. Conversely, the benchmark was met over only half the range at 6% ma

Conclusion

Benchmarking can be aligned with intended use case but must be adapted to the constraints of retrospective data.

UNIVERSITY OF CAMBRIDGE
WARWICK THE UNIVERSITY OF WARWICK
edith AI in Breast Screening
FUNDED BY NIHR | National Institute for Health and Care Research