EST. 1597
GRESHAM COLLEGE

The AI Revolution in Cancer Imaging
Dr Richard Sidebottom
9 January 2024

A note to the readers: I have used chat GPT 4 to help compile this accompanying paper from the script for the lecture (which I wrote using help and suggestions from only the natural intelligence of my colleagues). After several unsatisfactory attempts, the prompts I used were: [perhaps we will try one section at a time: my script for explaining the amazing technology used for medical imaging needs to be rewritten slightly more concisely and in a less conversational, more formal scientific style. It is: …] and [Please rewrite the script for this next section using the same approach as last prompt: …]

I used several versions of the chatGPT output for some sections, and have combined and edited these, but it has required less revision than I expected, largely to remove or reduce unnecessary conclusions and hyperbole. If there are errors, these belong to me.

In addition, I would like to explain that the examples I have used concentrate largely on breast imaging, this is because I am familiar with work in this field, and also because it is one of the areas where development and deployment of AI systems is most advanced. Most of the issues explored in the context of breast imaging are also applicable to other areas of medical imaging.

Richard Sidebottom

The amazing technology used for medical imaging

Medical imaging has significantly evolved since the late 19th century, emerging from rudimentary experiments to become a cornerstone in healthcare delivery. The inception of this field is marked by the development of radiography, as exemplified by Wilhelm Roentgen's pioneering X-ray image of his wife's hand. These initial radiographs were simple projections captured on photographic plates. Over the course of the 20th century, plain X-ray radiography has been, and remains, a fundamental, cost-effective diagnostic tool in medicine due to its speed and simplicity.

A major advancement came in the 1970s with Geoffrey Hounsfield’s invention of computed tomography (CT) scanning. This technique utilized a series of narrow beam X-ray exposures, allowing for the reconstruction of cross-sectional imaging. The evolution of CT technology has been continuous, with modern scanners capable of quickly producing detailed volumetric images, now absolutely essential for comprehensive medical care.

Further diversification in medical imaging was achieved with the development of ultrasound technology. This modality uses high-frequency sound waves emitted from a transducer, with images formed from the echoes reflected back from body tissues. Ultrasound is valuable in oncology and breast imaging particularly, due to its complementarity to X-ray mammography and utility in image-guided interventions. An example of its advanced application is in shear wave elastography, which provides insights into tissue stiffness, aiding in the differentiation between malignant and benign masses.

Magnetic Resonance Imaging (MRI), developed in the late 1970s by Peter Mansfield, represents perhaps the most incredible imaging technology, or in fact any technology yet developed in any domain. MRI employs a potent superconducting magnet, operating at temperatures near absolute zero, to align protons in the body. Subsequent radiofrequency pulses are used to disturb this alignment, with the resulting signals processed to create detailed images of various tissue properties. This technology generates extensive data, such as in breast MRI, where two thousand images per examination are typical. Advanced MRI applications, like MR tractography, allow for visualization of neural pathways, further underscoring the versatility of this modality.

The field of nuclear medicine has also evolved, particularly with the advent of Positron Emission Tomography (PET). PET utilizes radioisotopes emitting positrons, the antimatter counterparts of electrons, generated in cyclotrons. We inject this substance which is typically bound to a sugar analog molecule designed to accumulate in highly active cells, including cancer cells. The annihilation events between positrons and electrons produce gamma rays detectable by the scanner. PET imaging, often combined with CT (PET-CT), offers both anatomical and functional information, and is now widely used for cancer staging.

These advancements in medical imaging technologies have provided an unprecedented depth of anatomical and functional insights. However, the resultant large volumes of complex data pose significant challenges in interpretation and utilization. This is where the field of 'radiomics' - the extraction of quantifiable features from imaging that may go unnoticed by human observers - becomes pertinent. As AI technology advances, it holds the potential to harness these vast data sets, enabling a more comprehensive understanding and application in clinical practice.

Defining AI and its application in breast imaging

The UK National AI Strategy (2021) defines AI to be:

‘Machines that perform tasks normally requiring human intelligence, especially when the machines learn from data how to do those tasks’

Stuart Russell who is a professor at Berkley, and who’s BBC Reith lectures I highly recommend says:

‘AI is about building machines that do the right thing, that act in ways that can be expected to achieve their objectives’. He also highlights the critical importance of setting appropriate goals for AI systems, cautioning against the risks of mis-specifying these objectives.

Artificial intelligence, a term coined in the 1950s, encompasses a wide range of technologies aimed at emulating human-like intelligence in machines. Machine learning, a subset of AI, involves the development of algorithms that adapt and learn from data, without being explicitly programmed for each task. Neural networks, inspired by the human brain's information processing system, form the backbone of many AI applications. Among these, deep learning, characterized by multiple layers in neural networks, has been the focal point of recent advancements in the field.

Breast screening is by no means perfect. In the UK and most of Europe we use double reading, where each set of mammograms are read by two different readers and if there is a disagreement between them the case is decided by a third reader, or group of readers. Despite our best efforts, about 3 out of 4 patients that we recall for further assessment because of concerns about the appearance of the mammograms do not have cancer – they are false alarms, and about 1/3 of cancers in screened women are not detected at screening.

This room for improvement together with high volume, well ordered clinical data where the objective are a little easier to specify than in many areas, has made breast screening an attractive area to try to apply computer analysis.

Early attempts to apply computer analysis in this area involved collaboration between radiologists and computer scientists. They focused on creating mathematical models to identify key features in mammograms, such as shape, size, and density. These features were then analysed using advanced statistical methods and classical machine learning techniques, forming the basis of early Computer-Aided Detection (CAD) systems.

The adoption of this previous generation of CAD systems was widespread in the United States but very limited in Europe, primarily due to their inability to match human specificity in diagnosis. A US study[i] published in 2015 by Professor Lehman and colleagues, involving over half a million cases, concluded that CAD systems did not improve the diagnostic accuracy of mammography. This finding underscored the need for more advanced and accurate AI systems in breast imaging, supporting the exploration of more sophisticated AI techniques involving deep learning.

Understanding artificial neural networks and deep learning in medical imaging

Recent advancements in medical imaging analysis are primarily based on artificial neural networks and deep learning. These networks are inspired by biological neurons, which receive multiple inputs, perform simple processing of the inputs, and send outputs to other neurons. In a similar manner, an artificial neuron receives inputs, processes them, and passes on outputs to subsequent neurons.

Each connection in an artificial neural network (ANN) is assigned a 'weight', which influences the strength of its contribution to the output. The network's ability to learn and produce accurate outputs depends on adjusting these weights.

ANNs consist of multiple layers: the input layer, where data is received (such as a mammogram image); hidden layers, where the data undergoes processing; and the output layer, where the final classification (like cancerous or non-cancerous) is determined. Initially, the network's weights might be set randomly, leading to non-specific outputs. The network is trained through a process known as supervised learning, where it is fed labelled examples (the 'ground truth') and learns to minimize the error between its predictions and the actual data.

Backpropagation is a key technique used to optimize the weights across the network, incrementally improving the accuracy of the output. As the network learns, certain pathways and features become dominant in predicting outcomes, leading to a probabilistic classification, perhaps similar to the nuanced judgments made by human radiologists.

A specific type of ANN used in image classification tasks is the Convolutional Neural Network (CNN). CNNs are inspired by the structure and function of the human visual system. They perform mathematical operations known as convolutions or filtering, where filters detect specific features or patterns in the input image. These filters create activation maps indicating the presence of these features, which are then passed through multiple layers of the network.

Each layer of a CNN contains multiple filters, creating a stack of feature maps that undergo further analysis in subsequent layers. As the network is trained, the weights describing these filters are optimized to identify key features that enhance classification accuracy. The complexity and abstraction of the information increases has the information travels through the network.

CNNs consist of various layer types, including convolutional layers for feature extraction and fully connected layers for output classification. In medical image classification, these networks can have over a hundred layers with millions of connections and weights, making them far more versatile and comprehensive than the earlier generation of CAD systems.

Training Convolutional Neural Networks for Medical Imaging

The effectiveness of these networks largely depends on the quantity and quality of the data they are trained on. Research in both medical and non-medical settings indicates that the larger the dataset, the better the network's performance. Datasets used in the context of breast imaging analysis typically contain tens or hundreds of thousands, or even millions, of images.

The training data is typically divided into three distinct parts: training, validation, and test sets. The training set is used repetitively by the network to adjust and set its weights. This process involves the network learning from the provided data, constantly updating its internal parameters to reduce errors in output.

The validation set serves a different purpose. It consists of data that the network has not been trained on, used to evaluate the network's performance and its ability to generalize. The insights gained from the validation set are crucial for monitoring the learning progress of the model and making necessary adjustments to the network architecture. This might include changing the number or arrangement of hidden layers to optimize performance.

The test set, on the other hand, is analogous to a 'final exam' for the network. It is a collection of data completely independent from the training and development sets and is used only to assess the final performance of the model.

For the training to be effective, the data must not only be abundant but also reliable. The outcomes that the model is intended to predict or assist with must be accurately represented in the training data. Any biases present in the validation and test sets can significantly impact the network's performance, leading to skewed or inaccurate results. Therefore, it's essential that the test set mirrors real-world scenarios as closely as possible to ensure the network's applicability in practical settings.

Evaluating the performance of AI systems in breast screening: progress and research

To understand the performance of AI systems, particularly in breast screening, both studies with retrospective data and prospective clinical trials are valuable. A useful paper using retrospective data is the 2020 study by Salim et al[ii], which performed a comparison of breast screening AI systems using receiver operating characteristic (ROC) curves. These curves are a standard tool for assessing the sensitivity (true positive rate) and specificity (true negative rate) of diagnostic tests. The ROC curve plots sensitivity against one minus specificity. The ideal scenario is a curve that rises sharply towards the top left corner, indicating high sensitivity and specificity. The curve is shaped by adjusting the decision threshold, the probability at which a case is classified as positive. The shape of these curves provides insights into the trade-offs between sensitivity and specificity as the decision threshold is varied. Lowering this threshold improves sensitivity but may reduce specificity, leading to more false positives.Such curves are useful in comparing the performance of different AI algorithms provided precisely the same cases are used in testing. If we want to benchmark them against human readers then a single operating point must be prespecified and chosen, because we don’t get a continuous output for human decision making, so cannot plot a continuous curve. In the Salim et al study, the best-performing AI algorithm exhibited performance comparable to the first and second human readers but fell short of the consensus opinion, which represents the standard of care. McKinney et al[iii] (also published in 2020), showed similar findings, showing that while AI can perform at a level akin to an expert human reader, it hadn't yet reached the collective accuracy of multiple experts.

These studies used retrospective data. This approach has several strengths, including access to large volumes of data with the real clinical results of what happened both in terms of radiologist interpretation, and then crucially in terms of longer term outcomes, both for interval cancers that present symptomatically and for cancers detected at the next round of screening. Retrospective studies are also used to model, within limitations how alternative screening strategies may alter screening performance and efficiency[iv].

Prospective clinical trials are also required because when relying on retrospective data, it is inherently a look at older data, and therefore may not represent true system performance from screening images that are being acquired today. Prospective studies will also capture the results of the uncertain interaction between human readers and AI systems, crucial to demonstrating safe use.

Larger prospective studies are now underway and producing encouraging results[v]. The Masai trial[vi] is in progress in Sweden using AI supported screening where only women in the top 10% risk of cancer (as determined by the AI) require double reading and the lower 90% can be single read. All readers have access to the AI results. Their published preliminary results indicate an increase in cancer detection rates without an increase in false positives, and a significant reduction in overall screen reading volume. Importantly this trial also demonstrated very high acceptance amongst the screened population.

Beyond breast screening, AI's potential in radiology is being explored in various other applications, such as predicting valvular disease from chest X-rays[vii]. A website by Radboud University[viii] lists numerous AI interpretation products that have received CE marking, covering a wide range of radiological applications. Many of these products focus on improving efficiency in familiar tasks, while others venture into new territories like quantitative risk prediction.

The emergence of large language models in AI and their potential in healthcare

The landscape of artificial intelligence (AI) in healthcare may be about to accelerate again with the advent of large language models based on transformer networks. Introduced by Google researchers in 2017, these models represent a leap in AI capabilities, initially demonstrated in language translation tasks. However, their applicability extends far beyond, and are likely to have implications in healthcare.

Large language models operate on the principle of self-supervised learning, which was initially applied to text but has now expanded to multimodal data, including imaging. This learning approach involves the model predicting the next item in a sequence - a technique that leverages vast amounts of internet text for training. By masking a word in a sentence and predicting it based on context, these models refine their learned representation of language and concepts.

The architecture of these models is complex, consisting of numerous layers that contribute to a highly abstract representation of data. One key aspect of these models is the attention mechanism, which allows them to focus on relevant parts of the input data, thereby understanding the context and nuances of language and other inputs.

For instance, in a simple two-dimensional representation, the model might closely associate the words 'cat' and 'dog' due to their frequent contextual relationship. Extending this into multiple dimensions allows for a more nuanced understanding, where relationships between various concepts can be represented in a multi-dimensional space.

These foundational models, exemplified by systems like GPT-4, have shown remarkable abilities in language processing, and also demonstrate emergent skills where they perform tasks they were not explicitly trained for. However, they are not without limitations, such as potential biases and the creation of factually incorrect content, while still appearing credible.

The potential of these technologies in healthcare is particularly intriguing. They could play a pivotal role in precision medicine, offering insights that extend from risk stratification in screening processes to tailored treatment plans. By integrating data from various sources, including family history, general healthcare information, genomics, and medical imaging, these models could significantly enhance diagnostic accuracy and treatment efficacy.

In decision-making for treatment pathways, the 'guess what comes next' approach of these models aligns closely with clinical decision-making processes. For example, integrating radiological and pathological imaging with electronic health records could lead to more accurate predictions of treatment outcomes, potential side effects, and tailoring post-treatment surveillance.

These systems offer us very significant challenges to work out if they can be made safe and useful, particularly in understanding and managing the complexity and biases of these systems. However their apparent ability to synthesize vast amounts of data and provide contextually rich insights is obviously worth exploring.

Reflecting on the challenges and potential of AI in healthcare

Ethical Imperative: "Do No Harm"

Ensuring the reliability of AI systems before their widespread use is critical. This involves rigorous testing through both retrospective data analysis and prospective clinical trials. Continuous performance monitoring is essential to detect and mitigate any adverse outcomes. For instance, when changes in mammography machines affect AI performance, recalibration or additional training of the system may be required. This poses a challenge in scenarios where new technology is introduced without historical data for AI training, particularly in resource-constrained settings like the NHS. The potential over-reliance on these systems, and the risks posed by IT failures, further complicate their integration into healthcare.

Human-Machine Interaction and the "Black Box" Issue

The interaction between humans and AI systems in medical decision-making brings forth the "black box" problem, where the reasoning behind AI decisions is often not transparent. Research is ongoing to make these processes more interpretable and increasing familiarity with using these system may mitigate this issue. At the moment it appears that the way to achieve optimal performance involves a synergistic approach where AI assists human clinicians. However, if AI systems continue to improve, there may be ethical dilemmas if human intervention in AI decisions on average, lead to suboptimal patient outcomes.

Information Overload and Patient Autonomy

The advent of AI could lead to an overload of information for patients, particularly regarding individual risk and prognosis. This abundance of data can be overwhelming and requires sensitive handling, akin to discussions in genetics clinics.

Security, Control, and Funding in AI

The broader implications of AI in society include issues of data security, centralization of control, and funding. Often, AI systems are developed using anonymized patient data, and the resulting products are controlled by private entities. This raises questions about data ownership, privacy, and the equitable distribution of AI's benefits. Given that these systems are built on data contributed by the population, there is an argument[ix] for more democratic control and oversight of AI technologies in healthcare.

Conclusion:

The revolution alluded to in the title, is underway but not yet very tangible in its impact in the clinic. We can see that the tools now beginning to be deployed look like they will have real clinical benefit. However, the implementation of these technologies in clinical settings is still in its infancy, and a cautious approach is warranted. The speed and scope of recent AI developments are startling, however the approach to medical research is necessarily much more cautious. Robust evidence from trials and practical experience from early deployments will need to accumulate before we will be ready to rely on totally novel methods. If the new generations of AI technologies are able to deliver anything like their promise, then we might just see entirely new approaches to some aspects of healthcare.

[i] Lehman CD, Wellman RD, Buist DS, Kerlikowske K, Tosteson AN, Miglioretti DL; Breast Cancer Surveillance Consortium. Diagnostic Accuracy of Digital Screening Mammography With and Without Computer-Aided Detection. JAMA Intern Med. 2015 Nov;175(11):1828-37. doi: 10.1001/jamainternmed.2015.5231. PMID: 26414882; PMCID: PMC4836172.

[ii] Salim M, Wåhlin E, Dembrower K, Azavedo E, Foukakis T, Liu Y, Smith K, Eklund M, Strand F. External Evaluation of 3 Commercial Artificial Intelligence Algorithms for Independent Assessment of Screening Mammograms. JAMA Oncol. 2020 Oct 1;6(10):1581-1588. doi: 10.1001/jamaoncol.2020.3321. PMID: 32852536; PMCID: PMC7453345.Salim and Strand paper JAMA2020

[iii] McKinney SM, Sieniek M, Godbole V, Godwin J, Antropova N, Ashrafian H, Back T, Chesus M, Corrado GS, Darzi A, Etemadi M, Garcia-Vicente F, Gilbert FJ, Halling-Brown M, Hassabis D, Jansen S, Karthikesalingam A, Kelly CJ, King D, Ledsam JR, Melnick D, Mostofi H, Peng L, Reicher JJ, Romera-Paredes B, Sidebottom R, Suleyman M, Tse D, Young KC, De Fauw J, Shetty S. International evaluation of an AI system for breast cancer screening. Nature. 2020 Jan;577(7788):89-94. doi: 10.1038/s41586-019-1799-6. Epub 2020 Jan 1. Erratum in: Nature. 2020 Oct;586(7829):E19. PMID: 31894144.

[iv] Hickman SE, Payne NR, Black RT, Huang Y, Priest AN, Hudson S, Kasmai B, Juette A, Nanaa M, Aniq MI, Sienko A, Gilbert FJ. Mammography Breast Cancer Screening Triage Using Deep Learning: A UK Retrospective Study. Radiology. 2023 Nov;309(2):e231173. doi: 10.1148/radiol.231173. PMID: 37987665.

[v] Dembrower K, Crippa A, Colón E, Eklund M, Strand F; ScreenTrustCAD Trial Consortium. Artificial intelligence for breast cancer detection in screening mammography in Sweden: a prospective, population-based, paired-reader, non-inferiority study. Lancet Digit Health. 2023 Oct;5(10):e703-e711. doi: 10.1016/S2589-7500(23)00153-X. Epub 2023 Sep 8. Erratum in: Lancet Digit Health. 2023 Oct;5(10):e646. PMID: 37690911.

[vi] Lång K, Josefsson V, Larsson AM, Larsson S, Högberg C, Sartor H, Hofvind S, Andersson I, Rosso A. Artificial intelligence-supported screen reading versus standard double reading in the Mammography Screening with Artificial Intelligence trial (MASAI): a clinical safety analysis of a randomised, controlled, non-inferiority, single-blinded, screening accuracy study. Lancet Oncol. 2023 Aug;24(8):936-944. doi: 10.1016/S1470-2045(23)00298-X. PMID: 37541274.

[vii] Ueda D, Matsumoto T, Ehara S, Yamamoto A, Walston SL, Ito A, Shimono T, Shiba M, Takeshita T, Fukuda D, Miki Y. Artificial intelligence-based model to classify cardiac functions from chest radiographs: a multi-institutional, retrospective model development and validation study. Lancet Digit Health. 2023 Aug;5(8):e525-e533. doi: 10.1016/S2589-7500(23)00107-3. Epub 2023 Jul 6. PMID: 37422342.

[viii] https://grand-challenge.org/aiforradiology/

[ix] Sidebottom R, Lyburn I, Brady M, Vinnicombe S. Fair shares: building and benefiting from healthcare AI with mutually beneficial structures and development partnerships. Br J Cancer. 2021 Oct;125(9):1181-1184. doi: 10.1038/s41416-021-01454-2. Epub 2021 Jul 14. PMID: 34262148; PMCID: PMC8548298.