22HLT05 MAIBAI

Publishable Summary for 22HLT05 MAIBAI
Developing a metrological framework for assessment of image-based Artificial Intelligence systems for disease detection

Overview

Image-based artificial intelligence (AI) systems for disease detection are increasingly being developed, and it is vital that these tools are robust and effective in heterogeneous clinical settings. To date, performance has been assessed in an ad hoc manner, as there are no approved guidelines for evaluation. Most studies have methodological weaknesses and results that are not comparable. A standardised and impartial framework for performance, generalisability, and suitability assessment of AI tools could address such needs, enabling a more efficient, reliable, and reproducible validation of image-based AI systems for disease detection. This project will use breast cancer screening as the exemplar, benchmarking AI tools on a large real-world database of mammographic images, with the aim of designing a metrological framework for AI assessment in diagnostics.

Need

AI has the potential to revolutionize healthcare systems, playing a key role in future clinical decision-making. The exponential increase in healthcare data over the last decade, as well as the fast-paced technology developments, have resulted in promising novel AI approaches for diagnostic applications and risk prediction. However, the adoption of AI in clinical settings remains limited, mostly due to i) limited data quality, structure, and interoperability across heterogeneous clinical centres and electronic health records, ii) absence of robust validation procedures, iii) distrust of predictions and decisions generated by AI systems, and iv) lack of harmonised government proposals and consensus guidelines on steps for their adoption. One barrier that continues to prevent the use of AI tools in clinical practice is that the training and benchmarking of AI algorithms requires large datasets that cover the full range of both patient variability and measurement conditions, or bias can be introduced. Poorly trained systems could lead to misdiagnosis and lack of generalisability. To overcome this barrier, large medical imaging databases and infrastructures for data collection, sharing and custodianship are therefore needed. Focusing on radiological imaging, the evaluation and testing of AI products require a high-quality set of images and associated clinical information on the screened population. However, there is a strong heterogeneity of clinical data, in terms of quality, volume, patient demographics, imaging equipment and acquisition/processing procedures. Hence, there is a need for categorisation and integration of data based on clinically relevant subgroups and image acquisition key factors. From a more technical point of view, a clear methodology to benchmark the quality of predictive AI models (with relevant associated metrics) is essential if AI is going to be used in disease detection and risk prediction. In such applications, where high-stake decisions are frequently taken, lack of interpretability can undermine trust in AI models, which are typically seen as “black boxes”. In the realm of clinical practice, the availability of interpretation methods for explainable and traceable AI tools is therefore strongly needed.

Thorough and consistent validation of image-based AI systems for disease detection is fundamental, not only in terms of sensitivity and specificity, but also for the implementation in the clinical environment, where a

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or EURAMET. Neither the European Union nor the granting authority can be held responsible for them.

The project has received funding from the European Partnership on Metrology, co-financed from the European Union’s Horizon Europe Research and Innovation Programme and by the Participating States.

European Partnership Co-funded by the European Union

METROLOGY PARTNERSHIP EURAMET

technical integration and interoperability of AI tools would be required, making indispensable the design of a global, standardised, and impartial AI assessment framework.

Objectives

The overall objective of this project is to develop a metrological framework necessary to support standardisation in image-based AI systems for disease detection. Using breast screening as an exemplar, the performance of explainable and traceable AI tools will be analysed on a large real-world database of mammographic images, informing the design of the standardised assessment framework.

The specific objectives of this project are:

1. To develop a technical infrastructure to be able to query and extract the relevant data from medical imaging databases, such as the OPTIMAM Mammographic Image Database (OMI-DB). To establish a methodology for centralised and common metadata indexes, creating an open-source middleware for translation of metadata from different imaging databases.
2. To identify image acquisition key factors and population subgroups and use these to categorise the clinical data into subsets, determining where there is sufficient data for training and validation of AI tools for disease screening. To develop a methodology to generate synthetic data derived from physics-informed models and data augmentation techniques based on measurement knowledge.
3. To use explainable and traceable AI tools for disease screening, providing the capability to train and retrain the tools as necessary. To benchmark the AI tools in terms of prediction performance, robustness, fairness and uncertainty quantification, under at least three scenarios, including low versus high image quality data, validation for specific patient demographics, presence of machine-based artefacts and noise sources. To develop and validate methods for the explainability and interpretability of the trained AI tools.
4. To provide an AI validation toolbox for diagnostic imaging, to summarise the performance testing evaluations and to give recommendations for the assessment of explainable and traceable AI tools for disease screening, with a focus on understanding their generalisability and sensitivity to varying populations, manufacturers, image processing, and acquisition techniques. To use the recommendations to design a global, standardised, and impartial AI assessment framework.
5. To facilitate the take up of the technology and measurement infrastructure developed in the project by the measurement supply chain, standards developing organisations (British Standards Institution, ISO/IEC JTC 1/SC 42 – Artificial intelligence), and end users (e.g., clinical stakeholders, manufacturers of medical and healthcare products, regulators).

Progress beyond the state of the art and results

Technical infrastructure for interacting with medical imaging databases (Objective 1)

In order to evaluate the generalisability of an AI tool, the medical imaging dataset used should cover the range of populations and imaging technologies that the AI tool will encounter clinically. There are currently relatively few large-scale medical imaging databases. Each of these databases contains images from a single country, along with the imaging technology and data storage used in that country. This can make it challenging for end users to identify data suitable for their task, as the data are stored using different methodologies. This project will overcome the above issues by establishing a methodology for centralised and common metadata indices. This will allow high-level investigation of disparate imaging datasets across Europe using a common open-source middleware layer.

Differentiation of clinically relevant subgroups in data and generation of derived data (Objective 2)

For the training and validation of AI tools in a healthcare context, it is important to have enough clinically relevant subgroups to ensure the tools are generalizable and unbiased. Often, the subgroups can be unbalanced and there is missing data for key clinical and/or measurement factors. This project will develop a methodology to differentiate the subgroups based on key factors using existing clinical imaging data and will also develop methodologies to generate derived data to supplement the missing subgroups. A comprehensive strategy will be implemented, involving *in silico* modelling (i.e., computer simulations of biophysical imaging), data augmentation techniques and machine-learning (ML) based approaches.

Assessment of AI-based tools for breast cancer screening with mammography (Objective 3)

For breast cancer screening, which is the exemplar that this project will work on, AI tools are increasingly being developed to detect disease and make risk predictions from mammographic images, showing very promising results, but there are still methodological concerns around the assessment of these tools that mean there is still a lack of wide-scale adoption in the clinic. This project will define prediction tasks and associated performance metrics of high clinical relevance for breast cancer screening and procure relevant AI models for the selected prediction tasks that can be trained and retrained on selected datasets in a systematic manner. In this way, a thorough evaluation of the AI tools can be carried out, providing specific working examples for the assessment framework development.

Interpretation methods for explainable and traceable AI tools (Objective 3)

A clear methodology to benchmark the quality of predictive AI models that inform clinically relevant decisions is essential for AI to be used in clinical settings. This project will define and implement a taxonomy of metrics covering relevant dimensions of model quality that, together, ensure that the model is fit for purpose and trustworthy. Quantitative performance metrics will be calculated, and model quality will be assessed in terms of accuracy or prediction performance, ability to maintain high accuracy in a wide range of scenarios (robustness) and for diverse patient populations of interest (fairness), ability to produce understandable and interpretable predictions (explainability) and, potentially, ability to deliver well-calibrated confidence/credibility intervals of the predictions (uncertainty quantification).

In applications where important decisions are frequently taken, like in healthcare, lack of interpretability in predictive models can undermine trust in those models, which are typically seen as “black boxes”. The use of so-called “explainable AI” (XAI) tools is therefore strongly encouraged, when developing clinical decision-support systems. However, formal requirements for XAI methods are lacking and existing approaches are insufficiently validated, preventing them from being useful for actual quality assurance purposes as needed by medical applications. To overcome this limitation, this project will develop a framework to define a notation of correctness of XAI methods and to formally verify the correctness of XAI tools for mammography, using carefully selected ground-truth data and directly interfacing with radiologists.

Design of global, standardised and impartial AI assessment frameworks (Objective 4)

The challenges associated with the safe implementation and deployment of AI in healthcare are clear, and government initiatives and guidelines demonstrate that it is a priority worldwide. The USA Food and Drug Administration (FDA) recently published an *AI-based Software as a Medical Device* action plan, where supporting regulatory science effort on the development of a methodology for the evaluation and improvement of AI algorithms is one of their main goals. The European Joint Research Centre (JRC) recently published a *Science for Policy Report on AI in Medicine and Healthcare*, where they highlight the lack of tests, benchmarking, and evaluation processes as one of the issues delaying the implementation of AI systems in healthcare. Overall, guidelines on the regulation of AI technologies include high-level directions, but not specific guidance on the practical steps in AI evaluation and in the development of reliable assessment frameworks. This project will develop a standardised and impartial assessment framework that will enable more efficient, reliable, and reproducible validation of image-based AI systems for disease detection, using breast cancer screening with mammographic images as the exemplar. Specifically, this project will provide an accessible and operational AI validation toolbox for diagnostic imaging, and give, based on the performance testing evaluations, recommendations for the assessment of explainable and traceable AI tools for disease screening, in line with the current regulatory guidelines.

Outcomes and impact

Outcomes for industrial and other user communities

The standardised and impartial assessment framework developed in this project will enable more efficient, reliable, and reproducible validation of image-based AI systems for disease detection. This in turn will result in the scalability of AI systems for disease detection, enabling the reliable and safe use of AI in healthcare. More broadly, such a framework will provide a way for the AI-tech and digitalization industry to safely develop and implement their products, and for health service providers to build the trust needed to use them, giving a level playing field for competition and innovation in this key new technological arena. The consolidation of AI tools for medical application, as pursued by this project, will boost healthcare technologies, pushing the integration of diagnostic equipment with AI platforms or the use of AI-based software as a standalone device for both detection and diagnosis, facilitating clinical analysis automation, without the need for data science expertise.

In the case of breast cancer screening, the provision of a standardised platform for assessing the generalisability of AI-based tools will enable earlier implementation of these into clinical use. This could address issues with delays in reading images due to highly-skilled staff shortages, which are expected to worsen over the next few years. Furthermore, this project will support the development of explainable and traceable AI tools for mammogram analysis, and also generate methods for data curation, data processing and data augmentation of the existing mammographic databases, providing validated methodologies to the medical community working in breast screening. Their utility can also impact on other AI assessment tasks in healthcare, provided datasets with similar features (size, population coverage, longitudinal nature, etc.) exist.

Outcomes for the metrology and scientific communities

The complexity of healthcare, compounded by the user- and context-dependent nature of AI applications, calls for a multifaceted approach beyond the traditional evaluation of AI. This project will bring together data science domain-specific expertise, *in silico* modelling, clinical research, and access to medical data. These will be combined with metrological expertise in uncertainty quantification, key comparison data evaluation and standardization.

Metrology institutes have a key interest in the explainability of AI and in ensuring confidence in data, evidenced through their own work programmes related to AI and their engagement with AI programmes at a national level. The commitment of National Metrology Institutes (NMIs) in analytical and statistical models as well as AI-based digital technologies for healthcare is indeed well-aligned with the strategies of the European Metrology Network for Mathematics and Statistics (EMN Mathmet), of which most of the NMIs involved in this project are members. The advances in AI and digitalization tool application in society sectors, like healthcare, are strongly recommended by the metrological and scientific communities, as addressed by the EURAMET's 2030 Strategy and clearly stated in the White Paper on Artificial Intelligence, published in 2020 by the European Commission. This project will further develop the capability of the NMI community working collaboratively at the European level, by developing and validating methods for the interpretation of the behaviour of AI tools and trained networks, which is crucial within the remit of achieving traceable and explainable AI. Moreover, this project will contribute to the building of an Open Access culture, through the implementation of the FAIR digital data principles (findability, accessibility, interoperability and reusability). This scientific attitude will also impact on the promotion of the SI Digital Framework, as recommended by the International Committee for Weights and Measures (CIPM), within the International Bureau of Weights and Measures (BIPM).

Outcomes for relevant standards

The work developed in this project will provide specific guidance on the practical steps for AI evaluation in healthcare and will be in line with the programmes of high-level standards bodies, such as the ISO/IEC JTC 1/SC 42 – Artificial intelligence. The designed AI assessment framework will be applicable globally, have an impact therefore at the European level in the context of regulatory guidelines and address one of the European Commission's priorities defined for 2019-2024 "A Europe fit for the digital age", which envisages as an action for healthcare the use of AI tools. Additionally, multi-country efforts in harmonisation and standardisation, as foreseen by this project, can have more commercial leverage on potential AI vendors, providers and end users than any single country. The relevance of such outcomes for regulatory bodies is well evident also through the recent actions of the International Organisation for Standardisation (ISO), which signed in April 2022 the Joint Statement of Intent "On the digital transformation in the international scientific and quality infrastructure", previously signed by the BIPM, the International Organisation of Legal Metrology (OIML), the International Measurement Confederation (IMEKO), the International Science Council (ISC) and its Committee on Data (CODATA).

Longer-term economic, social and environmental impacts

Benefits from this project are expected in different economic sectors, including healthcare and medical device industries, as well as companies specialised in digital technologies and data platforms. The growing datasets of patient health-related digital information, increasing demand for personalised medicine, and the rising demand for reducing care expenses are some of the major driving forces of the healthcare market growth. Looking specifically at the diagnostic application of AI, the market size was valued at € 576.3 million in 2021 and is projected to grow at a CAGR of 26.3% from 2022 to 2030.

By developing a robust infrastructure for the management of cancer screening data, and by providing high-quality training data and explainable AI-based tools for clinical interpretation, this project could concur to the future scalability of low-cost screening programmes. This will directly impact on the quality of life, with the possibility to treat diseases at early stages, improve the monitoring of their evolution and address therapy strategies in a more personalised way. As longer-term effects, increased survival rate and reduced costs for national health systems and the society are expected, considering that the total cost for cancer treatment and its management in Europe is € 199 billion/year. Furthermore, the World Health Organisation (WHO) has estimated that by 2030 the world will be short of 9.9 million of doctors, nurses and midwives, adding further urgency to address the challenge of already overburdened health systems. Supporting the widespread adoption of AI could help alleviate capacity shortfalls, and increase the viability of screening programmes in the developing world.

List of publications

-

This list is also available here: https://www.euramet.org/repository/research-publications-repository-link/

Project start date and duration:	1 September 2023, 36 months	
Coordinator: Alessanda Manzin, INRIM Tel: +390113919825 E-mail: a.manzin@inrim.it Project website address: https://maibaiproject.eu/		
Internal Beneficiaries: 1. INRIM, Italy 2. CMI, Czechia 3. DFM, Denmark 4. IMBiH, Bosnia and Herzegovina 5. IPQ, Portugal 6. PTB, Germany	External Beneficiaries: 7. FhG, Germany 8. ISS, Italy 9. LRCB, Netherlands 10. UL, Slovenia	
Associated Partners: 11. NPL, United Kingdom, 12. RSFT, United Kingdom		