ABSTRACT

Background In recent years, AI has made significant advancements in medical diagnosis and prognosis. However, the incorporation of AI into clinical practice is still challenging and under-appreciated. We aim to demonstrate a possible vertical integration approach to close the loop for AI-ready radiology.

Method This study highlights the importance of two-way communication for AI-assisted radiology. As a key part of the methodology, it demonstrates the integration of AI systems into clinical practice with structured reports and AI visualization, giving more insight into the AI system. By integrating cooperative lifelong learning into the AI system, we ensure the long-term effectiveness of the AI system, while keeping the radiologist in the loop.

Results We demonstrate the use of lifelong learning for AI systems by incorporating AI visualization and structured reports. We evaluate Memory Aware-Synapses and Rehearsal approach and find that both approaches work in practice. Furthermore, we see the advantage of lifelong learning algorithms that do not require the storing or maintaining of samples from previous datasets.

Conclusion In conclusion, incorporating AI into the clinical routine of radiology requires a two-way communication approach and seamless integration of the AI system, which we achieve with structured reports and visualization of the insight gained by the model. Closing the loop for radiology leads to successful integration, enabling lifelong learning for the AI system, which is crucial for sustainable long-term performance.

Key Points:
- The integration of AI systems into the clinical routine with structured reports and AI visualization.
- Two-way communication between AI and radiologists is necessary to enable AI that keeps the radiologist in the loop.
- Closing the loop enables lifelong learning, which is crucial for long-term, high-performing AI in radiology.
Introduction

The introduction of deep learning models in clinical radiology is evolving more gradually than anticipated due to issues involving effectiveness, regulatory concerns, and the difficulty in establishing sustainable business models [1, 2]. Furthermore, the development of artificial intelligence-driven radiology requires trust and interpretability of AI systems for radiologists and patients [3–5]. Deep learning and AI-driven radiology have the potential to aid radiotherapy planning [6] and tumor detection [7], among other use cases. These technologies promise to improve the accuracy and efficiency of many critical processes, leading to better patient outcomes.

For AI systems to become more relevant for radiology in practice, AI support needs to become seamless to reduce the time spent per case. Recent developments in AI try to elevate the tight time constraints that radiologists face when reviewing cases, which lead to missed findings and long turnaround times [8]. Considering this situation, we see the need for vertical integration of images of AI-driven decisions and machine-readable structured reports to support decision-making with AI systems.

Most current research bypasses complex technical integration into real-world applications by simplifying assumptions or working only with immaculate data that have been carefully cleaned and homogenous datasets selected for the research. In real-world applications, the data gathered in hospitals is heterogeneous due to differences between hospital infrastructures, different devices, or other inconsistencies. As a result of such domain shifts, many AI systems suffer from worse performance in practice as the AI system ages [9, 10] or is deployed in new environments [11–13]. Common domain shifts occur between different institutions due to changes in populations or devices. The data also shifts over time due to updates to reconstruction algorithms and acquisition protocols. Neglecting technical integration problems while benchmarking assistive technology in isolation leads to silent failures of AI systems [9–13].

A leading approach to address this problematic setting is the concept of lifelong learning [10, 13, 14], also known as continual learning. We first developed an AI system to the best of our knowledge and continuously updated the system with new data. This concept has the advantage that it can handle data until it becomes inaccessible, e.g., due to GDPR [15], or adapt to newly available data distributions, e.g., in the case of COVID-19 [16]. An example of how these advantages can be leveraged to provide fast adaptability to unpredictable events, such as future pandemics, is by allowing individuals to altruistically share their relevant data without compromising their privacy. As the AI system matures, individuals may want to redact any shared data they had initially provided. This can be done without impacting the overall performance of the AI system.

In order to integrate lifelong learning into the clinical workflow, we must engage our AI systems in cooperative lifelong learning with radiologists. For successful collaboration, the AI system needs to receive information from radiologists through a machine-readable format instead of unstructured text. The radiologist, similarly, must receive human-readable insight from the AI system. A machine-readable report uses a structure that complies with guidelines and grasps relevant information without incurring additional costs. While the benefits of lifelong learning are evident [14], current medical device regulations hinder its applicability.

To collect structured reports efficiently, we support the radiologist with images of the insight gained by the AI system [17]. This insight needs to integrate seamlessly into the clinical workflow of a radiologist in order to foster trust and enhance the effectiveness of the cooperation. The integration of AI systems with structured reports, lifelong learning, and AI images enables realistic studies regarding AI that keeps the radiologist in the loop.

As an example of such a workflow, we demonstrate a possible use for diagnosing pulmonary embolism (PE) from CT scans. This article shows the integration of visualization techniques for AI systems with structured reports and lifelong learning. The system gives reliable insight into the model's predictions for radiologists.
while collecting crucial information and increasing comprehensi-

Materials and Methods

In order to achieve the complex vertical integration of our system,

AI Visualization

One of the predominant directions for interpretable AI is the

Structured Reporting

Structured reporting is a method of describing medical images in

Fuchs M et al. Closing the loop... Fortschr Röntgenstr | © 2023. Thieme. All rights reserved.
into radiology is the ability to improve the interpretability of deep learning models.

In this section, we describe how we design our scheme and adapt it to better serve the integration of AI systems into the clinical workflow. First, we develop the structured reporting template for PE localization according to guidelines and our radiologists [29, 30]. The basic template covers all mandatory topics and questions to answer regarding PE. This includes a structure to classify the PE location according to different levels: central, lobar, segmental (seg.), and subsegmental (subseg.). Furthermore, we divide each side of the lung into three segments consisting of upper, middle, and lower for the left and right sides. Finally, a bifurcation can be labelled separately, resulting in 21 location labels. We support this with an easy-to-fill design in the OHIF viewer [31] while leaving room to capture additional patient information. The different images are in the left bar. Our design allows for the communication of insight from the AI system to the clinician while being able to switch back to the original CT scan. By filling the report with more information in a structured way, the radiologist allows us to feed the AI system with new knowledge.

Additionally, our template includes fields for location and measurement of probability for all labels, which open up when a respective field is selected. The added transparency, provided by probabilities and visualization, increases the perceived reliability and trust in the AI system. However, the template can only cover some possible information in a structured way. Creating a comprehensive template that encompasses every potential finding in a chest CT scan is not possible since new unknown diseases are going to emerge (e.g., COVID-19). Consequently, it is crucial to determine an effective approach to address rare findings, such as atypical pathologies (e.g., Castleman’s disease), foreign bodies (dental work, misplaced catheters), or anatomical variations (Azygos lobe, Horseshoe lung).

Therefore, we still offer the possibility of adding relevant observations in plain text. The unstructured text is essential for maintaining the daily clinical workflow, especially for sporadically occurring cases, as the information would otherwise be lost. However, text is initially hard to process for any AI system. Future AI systems could integrate this lost information with the help of advanced large language models. An example of a fictional patient being opened for reporting is shown in Fig. 2.

Integration and Workflow for Lifelong Learning

To integrate the AI images into the radiology workflow, we run our model automatically on thorax CT examinations and generate possible predictions for PE location. We then extract the images and forward them as an additional modality. An overview of the integration is illustrated in Fig. 3.

Reporting on a workstation is quickly started by clicking on the selected study, which opens the reporting platform. The platform
consists of two parts. First, the study is displayed on the left half of the screen. The clinician can choose between the standard CT scan or the CT with the image of the prediction results in the left bar. Second, the template is opened on the right side, enabling structured reporting for the selected task according to appropriate guidelines. The AI system pre-fills the reporting template for the learned tasks. Additionally, it offers further information, such as predicted probability and images of the relevant slices for the prediction. The radiologist can use this information to accelerate the report’s completion and avoid missing findings, e.g., in the case of multiple PEs.

After the radiologist completes the report, the new ground truth annotation can be fed back to the AI model for further improvement. This functionality opens up the possibility of adapting the AI system over time, which prevents the deterioration of model performance as time passes and data distribution changes [9, 10]. This loss in performance often goes unnoticed, as deep learning models report high confidence even for low-quality predictions. This is denoted as silent failure. Allowing for model adaptation also makes it possible for the model to work in new hospital environments [9, 10, 16].

However, training models continuously introduces new risks that must be cautiously handled. For starters, care must be taken to avoid catastrophic forgetting [14]. In this phenomenon, the model’s performance with respect to data from older distributions deteriorates significantly. The goal is to train a model that produces high-quality predictions for all data sources seen during the training process.

We address these two challenges, namely the inherent heterogeneity of data and catastrophic forgetting, in the following fashion.
For the first problem, our structured reporting template provides a solution of structuring the output data for the AI system. While CT images vary from hospital to hospital as well as device to device, they are denoted in HU [32] and basic interoperability is given by the DICOM standard. However, this still might lead to a performance drop when the distribution of data shifts quite significantly, as seen for our red hospital in ▶ Fig. 4.

To achieve resiliency against catastrophic forgetting, we explore different popular continual learning methods. The first is Rehearsal [33], where we interleave samples (20%) from the previous datasets into the present training. This approach produces good empirical results. However, GDPR guidelines [11] often do not allow the storage of patient studies, and even if they do, the studies could become unavailable later due to the right to forget. We compare these results to Memory Aware Synapses (MAS) [34]. In contrast to Rehearsal, MAS identifies the most important model parameters and prevents them from changing too much from their initial state. Therefore, MAS would be preferable under GDPR regulations.

Results

In our analysis, we take the typical approach of starting with a state-of-the-art model pre-trained on a large, public, and heterogeneous dataset, namely the RSNA pulmonary embolism CT dataset [22] and the challenge-winning model [23]. We further collect cohorts from two German university clinics with our structured reporting template, generating annotations on a sample-wise level. The radiologist found the system easy-to-use and approved the transparency images. We examine the situation where we first fine-tune the model using the first clinical data, then the second clinical data. We aim to obtain a model that performs well across all test sets and training orders. While the RSNA dataset enables us to obtain a pre-trained model for the AI system, we adapt the final classifier to the structure of the reporting template.

Our results are displayed in ▶ Fig. 4. Each clinic training set consists of 694 samples, while the red clinic has 13.83% PE-positive results and the blue clinic has 30.55%. Each clinic’s test dataset contains 86 examples, which we evaluate for PE detection and location in the form of localization labels consisting of central, lobar, seg., and subseg. embolisms. We first evaluate the latest dataset training based on the test set in red and then the previously trained clinic in blue.

The results show how the simple rehearsal approach to training the AI system leads to consistent performance with respect to PE detection (avg. accuracy of 75.85%) for both datasets. The method causes a slight loss in ability to adapt to the data distribution of the latest clinic, e.g., rehearsal top row in red. Overall, the rehearsal method performs well for detecting and localizing pulmonary embolisms. However, it requires the storage of samples, which may become problematic with GDPR standards. On the other hand, MAS can keep up for the most part (avg. accuracy of 75.28%) while reducing catastrophic forgetting and maintaining the ability to adapt to the new clinic’s data.

Discussion

Reflecting on the results of our study, we find that the rehearsal approach achieves consistent performance on both datasets but lacks the ability to adapt to new data distributions. MAS instead performs similarly well in terms of reducing catastrophic forgetting and is more adaptable to new data. These findings are consistent with previous studies on continual learning [34]. Furthermore, rehearsal requires storage of samples and may not comply with GDPR standards, while MAS does not have these issues. Overall, both methods have their advantages and limitations.

Despite the urgent need for lifelong learning AI systems for radiology, the current legislative guidelines for medical devices do not yet provide an acceptable framework for quick and effective model updating. This requires the AI system provider to obtain approval whenever the system is deemed outdated [35, 36]. In practice, model updates – which incorporate new knowledge on acquisition practices and changes in the population – only become effective after a lengthy reverification cycle. Also, local fine-tuning steps where the model is adapted to specific characteristics of the data on-site become untenable.

We firmly believe that lifelong learning should be accompanied by monitoring of the model performance and annotations that are made. However, the current regulatory framework needs to include this significant opportunity for promoting the safe and effective use of AI. A potential solution to the current situation is a pre-certification approach [37, 38] that consists of a change control plan and predefined development and monitoring practices for the manufacturer to develop and update their devices safely and effectively rather than approving each individual update. However, we see two factors that must be improved upon for such an approach to succeed. Firstly, a closer collaboration is needed between regulators, device manufacturers, academic researchers, and other stakeholders to develop new strategies and guidelines for lifelong learning medical devices. The second important factor is increased transparency in AI systems, data acquisition, and the evaluation process. Transparency can be achieved by providing interpretable explanations and open access to the codebase, which would allow for building trust and understanding with experts and other stakeholders without regard to their background knowledge. As current plans by the FDA suggest [39], it would be risky to leave the design and evaluation of the development and monitoring practices solely to one party. When measurements and models can be updated at the same time, the risk for metric manipulation would be high to better suit marketing strategies or avoid direct comparisons with competitors. This highlights the importance of accountability and comparison of similar medical devices.

We show with our results that lifelong learning is beneficial – and indeed needed – for maintaining high predictive standards through the product lifecycle. Structured reporting allows the seamless integration of expert feedback into the learning loop. Giving radiologists insight into the AI system in the form of appropriate images provides a second layer of reliability to increase trust in our system.
Fig. 4 Comparison of Memory Aware Synapses (MAS) [34] and Rehearsal [33] approaches for lifelong learning. We first pre-trained our AI system to predict PEs using the public RSNA dataset. Then we deployed the AI system in a first clinic, followed by a second clinic and vice versa. We evaluated the latest dataset training using the test set in red and the previously trained clinic in blue. On average, Rehearsal achieves an accuracy (Acc.) for PE detection of 75.85% and MAS 75.28%. However, the performance between the two clinics varies. As we show at the top, this is not due to a bias regarding sex and is rather due to different PE occurrence rates and sizes.

Conclusion

We present an easy-to-integrate workflow for lifelong learning that leverages advances in structured reporting and interpretability. Our approach builds on the vertical integration of AI-ready radiology with a deep learning system, which requires two-way communication between both parties. We incorporate reliability measurements, namely prediction probabilities for labels and visualization to deliver dependable insight into the predictions of an AI system. Cooperating radiologists found our approach to be an easy-to-use system that facilitates lifelong learning. Furthermore, we discuss potential regulatory changes to improve the applicability of lifelong learning algorithms. Finally, we advocate for better integration of AI in radiology departments and closer collaboration between AI systems and clinicians.

Funding

Bundesministerium für Gesundheit (EVA-KI [ZMV1–2520DAT03A])

Conflict of Interest

The authors of this manuscript declare relationships with the following companies: Phillip Matthies is an Employee of Smart Reporting GmbH.

Acknowledgements

This work was supported by the Bundesministerium für Gesundheit (BMG) with the grant [ZMV1–2520DAT03A]. Further, we would like to thank Zhangjie Yang, Mayur Arvind, Melda Eski and Jonathan Stieber for their involvement in earlier project stages.

References


Fuchs M et al. Closing the loop... Fortschr Röntgenstr | © 2023. Thieme. All rights reserved.
[26] Szeliski R. Computer vision: algorithms and applications; Springer Nature; 2022