Re-defining Radiology Quality Assurance (QA) - Artificial Intelligence (AI)-Based QA by Restricted Investigation of Unequal Scores (AQUARIUS)

Axel Wismüller\textsuperscript{a,b,c,d}, Larry Stockmaster\textsuperscript{a}, and M. Ali Vosoughi\textsuperscript{b}

\textsuperscript{a}Department of Imaging Sciences, University of Rochester, NY, USA
\textsuperscript{b}Department of Electrical and Computer Engineering, University of Rochester, NY, USA
\textsuperscript{c}Department of Biomedical Engineering, University of Rochester, NY, USA
\textsuperscript{d}Faculty of Medicine and Institute of Clinical Radiology, Ludwig Maximilian University, Munich, Germany

\textbf{ABSTRACT}

There is an urgent need for streamlining radiology Quality Assurance (QA) programs to make them better and faster. Here, we present a novel approach, Artificial Intelligence (AI)-Based Quality Assurance by Restricted Investigation of Unequal Scores (AQUARIUS), for re-defining radiology QA, which reduces human effort by up to several orders of magnitude over existing approaches. AQUARIUS typically includes automatic comparison of AI-based image analysis with natural language processing (NLP) on radiology reports. Only the usually small subset of cases with discordant reads is subsequently reviewed by human experts. To demonstrate the clinical applicability of AQUARIUS, we performed a clinical QA study on Intracranial Hemorrhage (ICH) detection in 1936 head CT scans from a large academic hospital. Immediately following image acquisition, scans were automatically analyzed for ICH using a commercially available software (Aidoc, Tel Aviv, Israel). Cases rated positive for ICH by AI (ICH-AI+) were automatically flagged in radiologists’ reading worklists, where flagging was randomly switched off with probability 50%. Using AQUARIUS with NLP on final radiology reports and targeted expert neuroradiology review of only 29 discordantly classified cases reduced the human QA effort by 98.5%, where we found a total of six non-reported true ICH+ cases, with radiologists’ missed ICH detection rates of 0.52\% and 2.5\% for flagged and non-flagged cases, respectively. We conclude that AQUARIUS, by combining AI-based image analysis with NLP-based pre-selection of cases for targeted human expert review, can efficiently identify missed findings in radiology studies and significantly expedite radiology QA programs in a hybrid human-machine interoperability approach.

Further author information: (Send correspondence to Axel Wismüller)

Axel Wismüller: E-mail: axel.wismueller@gmail.com
1. MOTIVATION

The provision of high-quality health care is the goal of all medical services. In the case of radiology quality assurance (QA) programs, patient selection, the conduct of the examination, and the interpretation of the results can all have an impact on the achievement of this goal, such as by improving the diagnostic information content, reducing radiation exposure, reducing medical costs, and improving departmental management [1]. QA programs thus contribute to the provision of high-quality health care.

Specifically, QA programs in radiology are of immense value in improving patient safety, optimizing the clinical usefulness of medical imaging exams, and increasing financial profits of radiology practices by lowering operational costs. Yet, contemporary approaches to radiology QA, such as widely used methods of randomly selecting cases for human review, suffer from limited effectiveness and efficiency. Hence, there is an urgent need for streamlining radiology QA programs to make them better and faster. Here, we present a novel hybrid human-machine interoperability approach, Artificial Intelligence (AI)-Based Quality Assurance by Restricted Investigation of Unequal Scores (AQUARIUS), for re-defining radiology QA, which has the potential of reducing human QA effort by several orders of magnitude over existing approaches.

In the following, we will explain the basic idea of AQUARIUS, describe its clinical application within a retrospective clinical trial for QA in AI-based detection of Intracranial Hemorrhage (ICH) in emergent care head CT scans, and will report initial radiology QA results for this clinical application.

This work is embedded in our group’s endeavor to expedite artificial intelligence in biomedical imaging by means of advanced pattern recognition and machine learning methods for computational radiology and radiomics, e.g., [2–4].

2. METHODOLOGIC CONCEPT

2.1 Shortcomings of Current Approaches to Radiology QA

The potential of conventional approaches to radiology QA is limited and subject to manifold shortcomings. A typical scenario for contemporary QA-related analysis would be to randomly select cases for human expert review, where medical imaging studies are compared with the findings described in the associated radiology reports. Although the scope of the analysis can be restricted to specific imaging study subsets, such as originating from given pre-defined radiology practices, scanners, imaging modalities, radiology exam types, patient origin
(inpatient, outpatient, or emergency department), specific referring physicians, etc., the principal challenge remains the same, namely a lack of focus by not pre-selecting cases with a high ‘pre-test’ probability of discordant interpretation based on image content. A classic example are ‘peer-review’ approaches, such as the American College of Radiology (ACR) RADPEERTM program [5], which restricts QA to cases with existing prior imaging studies. This program starts from the assumption that if, during interpretation of a new examination, there are prior images of the same area of interest, the interpreting radiologist will typically form an opinion of the previous interpretation while interpreting the new study. If the opinion of the previous interpretation is scored, a peer review event has occurred, where the reviewer scores the previous interpretation using a standardized rating scale. This approach requires significant human effort, excludes studies with no prior exams, does not distinguish between high and low ‘pre-test’ probability for discordant reads, so that only a tiny subset of inaccurate interpretations may be detected, given the usually low probability of clinically relevant discordant interpretations. In practice, most of the reviewed studies will be rated as ACR category 1 signifying a concordant read. This involves the issue of vigilance fatigue for rare events, e.g., [6], as known from non-medical domains, such as policing [7] or airport security, where the underlying probability of alert-inducing events is usually low. Case selection is usually based on a random approach, such as automatically requesting peer review for a certain number of cases (typically two) at the beginning of a radiologist’s shift, which will preclude the majority of inaccurate interpretations from being detected in the peer review process. In addition, such a peer-review approach does not only require human intervention by the original peer-reviewing radiologist, but also significant time and effort by a second human reviewer that performs a secondary review of cases with peer-review scores indicating discordant reads between current and prior studies. In conclusion, despite requiring a significant amount of human expert intervention time and effort, the vast majority of inaccurate radiology interpretations will remain undetected under the currently used radiology QA approaches. Hence, there is an urgent need to make the radiology QA process better and faster, with the goal to both improve patient safety and reduce costs.

2.2 Artificial Intelligence (AI)-Based Quality Assurance by Restricted Investigation of Unequal Scores (AQUARIUS)

AQUARIUS consists of two steps: In a first step, AI-based case analysis is compared to a ‘gold standard’ given by an automatically retrieved estimate of human case assessment. In a second step, cases with “unequal scores”, i.e., discordant assessments between AI-based and human interpretation, are reviewed by human experts. A typical diagnostic radiology implementation of AQUARIUS would include automatic comparison of AI-based image analysis results with natural language processing (NLP) results on final radiology reports. Only the usually small subset of cases with discordant reads would then be manually reviewed by dedicated subspecialty-trained expert
radiologists. By restricting human effort to only those cases with a high ‘pre-test’ probability for missed or over-called findings, AQUARIUS optimizes the use of human and fiscal resources spent on radiology QA programs, thus demonstrating a significant potential for both ‘saving lives’ and ‘saving money’ simultaneously. We emphasize that multiple organizational and technical challenges have to be met simultaneously for implementing a successful AQUARIUS study design. These challenges include, but are not limited to, (i) acquiring Institutional Review Board approval for automatic data retrieval and radiologists’ participation in such QA initiatives, (ii) aligning the interests of diverse stakeholders, including radiologists, hospital administrators, AI solution vendors, and regulatory bodies, such as the ACR, and (iii) significant interface programming efforts for automatically retrieving heterogeneous information from multiple radiology and hospital IT systems, such as AI-based image analysis results and NLP-based classification of radiology reports for relevant clinical imaging findings.

3. CLINICAL APPLICATION: INTRACRANIAL HEMORRHAGE DETECTION TRIAL

Clinical Relevance: To demonstrate the applicability of our novel hybrid human-AI quality assurance (QA) approach using the AQUARIUS framework, we performed a study for evaluating radiologists’ performance on accurately reporting intracranial hemorrhage (ICH) in emergent care setting head CT scans. Timely detection of ICH on medical imaging studies is critical, because delayed therapeutic interventions in emergency settings may be detrimental for patient outcome. For example, in hemorrhagic stroke, the American Heart Association (AHA)/American Stroke Association (ASA) 2018 guidelines for early management emphasize the time-dependent benefit of tissue plasminogen activator (tPA) therapy [8]. Patients with ischemic stroke must be identified as soon as possible so that tPA can be administered within 3-4.5 hours of symptom onset [8]. A non-contrast head CT scan, which must be negative for hemorrhage, is a required diagnostic test before tPA can be administered in a timely manner [9–11]. The clinical outcomes of patients receiving tPA within a designated time limit are closely linked to time of treatment [11,12]. Besides hemorrhagic stroke, there are numerous other clinical conditions, in which ICH will lead to rapid increase of intracranial pressure, where early detection and treatment of ICH can significantly reduce patient morbidity and mortality.

Data and Methods: A total of 1936 consecutive non-contrast emergency-setting head CT scans from 2 CT scanners at a large academic hospital were prospectively acquired over 47 consecutive days. Immediately following image acquisition, scans were automatically analyzed for ICH using commercially available software (Aidoc, Tel Aviv, Israel). Cases rated positive for ICH by AI (ICH-AI+) were automatically flagged in radiologists’ reading worklists, where flagging was randomly switched off with probability 50%, see e.g. [13]. For ICH-AI+ cases,
radiologists’ missed ICH detection rates (ratio of number of missed ICH cases and number of all true ICH-AI+ cases) was calculated and compared between flagged and non-flagged ICH-AI+ cases, where images of all ICH-AI+ cases with ICH-negative results detected by natural language processing (NLP) of final radiology reports (ICH-AI+NLP- cases) were re-analyzed by an experienced neuroradiologist to identify true ICH+ cases missed by original radiology readings.

4. RESULTS

Among all 1936 CT scans, 381 ICH-AI+ cases were found, of which 190 cases were flagged. A total of 29 ICH-AI+NLP- cases were found, where 6 had been reported ICH+ by the radiology report. Of the remaining non-reported 23 ICH-AI+ cases, neuroradiology expert review identified 6 non-reported true ICH+ cases, where 5 cases were non-flagged, and only 1 case was flagged. This yields radiologists’ missed ICH detection rates of 0.52% and 2.5% for flagged and nonflagged cases, respectively.

5. CONCLUSIONS

We present a novel hybrid human-machine interoperability approach, Artificial Intelligence (AI)-Based Quality Assurance by Restricted Investigation of Unequal Scores (AQUARIUS), for re-defining radiology QA, which has the potential of reducing human QA effort by several orders of magnitude over existing approaches, making radiology QA programs better and faster. AQUARIUS consists of two steps: In a first step, AI-based case analysis is compared to a ‘gold standard’ given by an automatically retrieved estimate of human case assessment. In a second step, cases with “unequal scores”, i.e., discordant assessments between AI-based and human interpretation, are reviewed by human experts. In our initial clinical study on detecting QA for detecting Intracranial Hemorrhage (ICH) on emergent noncontrast head CT scans, our results suggest that flagging ICH-AI+ cases on radiologists’ worklists may decrease the rate of missed true ICH+ cases, although this was not found statistically significant based on small numbers of discordantly reported cases. Yet, our novel method of human expert review of ICH-AI+NLP- cases successfully revealed true ICH+ cases missed by original radiology reports with minimum human effort, and can therefore provide valuable contributions to radiology QA programs as an efficient hybrid human-AI approach. We conclude that, by combining AI image analysis with NLP-based pre-selection of cases for targeted human expert review, our AQUARIUS approach can efficiently identify missed findings in radiology reports and thus expedite radiology QA programs.

ACKNOWLEDGMENTS

This research was partially funded by the American College of Radiology (ACR) Innovation Award “AI-PROBE: A Novel Prospective Randomized Clinical Trial Approach for Investigating the Clinical Usefulness of Artificial Intelligence in Radiology” (PI: Axel Wismüller) and an Ernest J. Del Monte Institute for Neuroscience Award from the Harry T. Mangurian Jr. Foundation (PI: Axel Wismüller). This work was conducted as a Practice Quality Improvement (PQI) project related to American Board of Radiology (ABR) Maintenance of Certificate (MOC) for A.W.
REFERENCES

[1] Davies, E. R., “Quality assurance in diagnostic radiology, world health organisation, eastbourne (1982) quality assurance in nuclear medicine, world health organisation, switzerland (1982), 70 pp.” (1982).

[2] Wismüller, A., Dsouza, A. M., Vosoughi, M. A., and Abidin, A., “Large-scale nonlinear Granger causality for inferring directed dependence from short multivariate time-series data,” Scientific reports 11(1), 1–11 (2021).

[3] Wismüller, A., Dsouza, A. M., Vosoughi, M. A., and Z. Abidin, A., “Network connectivity analysis in complex systems using large-scale non-linear Granger causality (lsNGC),” International Society for Optics and Photonics (2021).

[4] Wismüller, A. and Vosoughi, M. A., “Classification of schizophrenia from functional MRI using large-scale extended Granger causality,” International Society for Optics and Photonics (2021).

[5] “RADPEER.” https://www.acr.org/Clinical-Resources/RADPEER. Accessed: 2022-03-07.

[6] Freudenburg, W. R., “Perceived risk, real risk: Social science and the art of probabilistic risk assessment,” Science 242(4875), 44–49 (1988).

[7] Krause, M., “Vigilance fatigue in policing: A critical threat to public safety and officer well-being,” FBI L. Enforcement Bull. 81, 3 (2012).

[8] Powers, W. J., Rabinstein, A. A., Ackerson, T., Adeoye, O. M., Bambakidis, N. C., Becker, K., Biller, J., Brown, M., Demaerschalk, B. M., Hoh, B., et al., “2018 guidelines for the early management of patients with acute ischemic stroke: a guideline for healthcare professionals from the American Heart Association/American Stroke Association,” Stroke 49(3), e46–e99 (2018).

[9] Lees, K. R., Bluhmki, E., Von Kummer, R., Brott, T. G., Toni, D., Grotta, J. C., Albers, G. W., Kaste, M., Marler, J. R., Hamilton, S. A., et al., “Time to treatment with intravenous alteplase and outcome in stroke: an updated pooled analysis of ecass, atlantis, ninds, and epithet trials,” The Lancet 375(9727), 1695–1703 (2010).

[10] Walter, S., Kostopoulos, P., Haass, A., Keller, I., Lesmeister, M., Schlechtriemen, T., Roth, C., Papanagiotou, P., Grunwald, I., Schumacher, H., et al., “Diagnosis and treatment of patients with stroke in a mobile stroke unit versus in hospital: a randomised controlled trial,” The Lancet Neurology 11(5), 397–404 (2012).

[11] Katzan, I. L., Hammer, M. D., Hixson, E. D., Furlan, A. J., Abou-Chebl, A., and Nadzam, D. M., “Utilization of intravenous tissue plasminogen activator for acute ischemic stroke,” Archives of neurology 61(3), 346–350 (2004).

[12] Mazighi, M., Chaudhry, S. A., Ribo, M., Khatri, P., Skoloudik, D., Mokin, M., Labreuche, J., Meseguer, E., Yeatts, S. D., Siddiqui, A. H., et al., “Impact of onset-to-reperfusion time on stroke mortality: a collaborative pooled analysis,” Circulation 127(19), 1980–1985 (2013).

[13] Wismüller, A. and Stockmaster, L., “A prospective randomized clinical trial for measuring radiology study reporting time on Artificial Intelligence-based detection of intracranial hemorrhage in emergent care head ct,” in [Medical Imaging 2020: Biomedical Applications in Molecular, Structural, and Functional Imaging], 11317, 113170M, International Society for Optics and Photonics (2020).