June 30, 2020

Dockets Management Staff (HFA-305)
Food and Drug Administration
5630 Fishers Lane, Rm. 1061
Rockville, MD 20852

Subject: (Docket No. FDA-2019-N-5592) "Public Workshop - Evolving Role of Artificial Intelligence in Radiological Imaging;" Comments of the American College of Radiology

The American College of Radiology (ACR)[1] and Radiological Society of North America (RSNA®)[2] appreciate the opportunity to comment on the proceedings of the U.S. Food and Drug Administration (FDA) public workshop regarding the "Evolving Role of Artificial Intelligence in Radiological Imaging" held February 25-26, 2020 (Docket No. FDA-2019-N-5592).

The FDA defined autonomous radiology AI as "software in which AI/ML is being used to automate some portion of the radiological imaging workflow (e.g. detection, diagnosis, reporting)," and differentiated these solutions from the kind of "augmented intelligence" innovations currently on the market. While certain applications of autonomous artificial intelligence (AI) could soon become useful to physicians and health systems in caring for their patients, we have some concerns with the approaches suggested at the workshop by a number of researcher/developer presentations with respect to FDA authorization pathways for autonomously functioning AI algorithms in medical imaging. While we understand the desire among industry and others to swiftly advance autonomous AI, our organizations strongly believe it is premature for the FDA to consider approval or clearance of algorithms that are designed to provide autonomous image interpretation independent of physician expert confirmation and oversight because of the present inability to provide reasonable assurance of safety and effectiveness. To date, there is a lack of comprehensive research-based criteria for ensuring algorithms are generalizable and a considerable body of published research to suggest that they often perform poorly across heterogeneous patient populations. In light of the broad heterogeneity in imaging equipment and image acquisition protocols and the lack of a definable mechanism to ensure the longitudinal performance of the algorithm, we are concerned that autonomously functioning algorithms of the varieties discussed at the workshop would pose a significant risk to patient safety.

Some have touted the recently FDA-authorized AI tool for autonomous detection of diabetic retinopathy (IDx-DR) as an example how autonomous AI could work in medical imaging; however, we do not agree this is an apt analogy. That algorithm is designed to assist physicians who are not experts in fundoscopic examinations in making a referral to those that are. The output of the IDx-DR algorithm is to recommend ophthalmologic referral for additional assessment, not to recommend treatment. In contrast, the

[1] The **American College of Radiology (ACR)** is a professional association representing over 40,000 diagnostic radiologists, interventional radiologists, nuclear medicine physicians, radiation oncologists, and medical physicists.

[2] The **Radiological Society of North America (RSNA)** is a non-profit organization representing radiologists, radiation oncologists, medical physicists and related scientists with over 52,000 members from 153 countries around the world.

potential output of autonomously functioning AI algorithm in radiological imaging would be to use an AI algorithm to potentially bypass the physician experts in image interpretation and refer patients to physicians who are not experts in medical imaging for treatment based on the results of the algorithm. Because there are no physicians with imaging expertise reviewing the data, potentially overruling or modifying the outputs, applying appropriate medical judgement and providing additional context and/or identifying secondary findings, in our opinion, autonomously functioning algorithms for imaging examinations pose a much greater risk for patient safety if the algorithm fails (both overdiagnoses and underdiagnoses) than if a product such as IDx-DR fails.

Current Use of AI in Medical Imaging

In order to evaluate the safety, effectiveness, and usefulness of autonomously functioning AI in radiological imaging, it is important to understand how AI is currently being used by the medical imaging community. To date, no autonomously functioning radiology AI algorithms have gone to market. As such AI in radiological imaging currently may only be marketed to assist with specific components of an imaging specialist's workflow. In this context, the responsibility for the patients' imaging care ultimately rests solely with the interpreting physician, and AI is just another tool at the expert's disposal.

In a recent survey of its members, ACR found that only 30% of radiologists are using AI in their clinical practice, and many of them are using AI exclusively in research applications; as such, the penetrance of even FDA-cleared AI solutions is limited. For the 30% that do use AI, of the currently cleared FDA algorithms, ACR found that its members are using only 1.2 algorithms per radiologist. The medical imaging algorithms currently on the market are being used to assist physicians in image interpretation, examination prioritization and select administrative tasks and are not intended to render final image interpretations or to automate patient care. Many of these algorithms have been cleared without comprehensive investigation by the developers into the generalizability of the algorithm across heterogeneous patient populations and do not account for variability in radiological equipment and variety of imaging protocols used. For these reasons, it is not surprising that 93% of radiologists using AI in our survey said that the results of AI in their practices are inconsistent, and 95% said they would not use the AI algorithms without physician overread.

Processes for Safe and Effective AI Use in Medical Imaging

Prior to any consideration of autonomously functioning AI in radiological imaging, the regulatory pathways for all healthcare AI should be improved to provide reasonable assurance of product safety and effectiveness. Many of the lower risk imaging AI algorithms currently on the market and much of the AI research supporting their development relied on single institution data for premarket testing. **AI algorithms should be required by FDA to undergo testing using multi-site heterogeneous data sets to ensure a minimum level of generalizability across diverse patient populations as well as variable imaging equipment and imaging protocols. Additionally, the FDA should have post-market oversight mechanisms that ensure algorithms function as expected longitudinally.** Imaging equipment and protocols change rapidly, and the AI algorithms have to maintain an ability to function effectively in this changing environment. As such, rigorous post-market assessment is critical to ensure that patient safety is maintained. We urge FDA to develop requirements for continuous monitoring of all algorithms used in clinical practice. The performance of the algorithms should be monitored by the interpreting physicians and data regarding patient demographics, type of equipment and imaging protocols tailored to the algorithm should be collected and assessed as a condition of approval/clearance. Furthermore, developers should provide clear labeling regarding what equipment and protocols will be supported and advise that algorithm use be limited to only the devices and protocols that were studied during the validation process. Before AI is approved for autonomous use, more stringent review processes should

be instituted to provide reasonable assurance of safety and effectiveness. Algorithms that automate key aspects of medical care without the direct oversight of physician experts are at substantially higher risk to patients than algorithms that merely provide decision support or administrative assistance. The autonomous radiology AI functionalities discussed at the workshop would be of such high risk that general and special controls would be insufficient to provide reasonable assurance of product safety and effectiveness. As such, these types of medical devices would be most appropriately classified as Class III, requiring the premarket approval application process. Additionally, we note that SaMD for automated diagnoses involving critical scenarios or conditions would be defined under the International Medical Device Regulators Forum (IMDRF) risk categorization framework as Category IV, the highest impact category. Given the implications of that category, FDA should adopt a rigorous pre-market approval process that ensures patient safety across heterogeneous patient populations and equipment types and develop continuous post-market monitoring requirements, perhaps leveraging trusted third-party registries, to protect patients and the public. Additionally, FDA should work closely with radiologist organizations to identify practical use cases for autonomous AI to ensure the agency's resources are being directed at solutions most likely to be practical and in demand.

Additional Concerns Regarding Continuously Learning/Adaptive Algorithms

The FDA's 2019 discussion paper, *Proposed Regulatory Framework for Modifications to AI/ML-Based SaMD*, stated that AI/ML-based SaMDs exist on multiple spectrums categorized by risk to patients and also by "locked" to "continuously learning." This comment submission assumes that modifications to the types of autonomous AI algorithms described at the workshop would be on the locked side of that spectrum. However, we do have additional concerns regarding continuously adaptive AI algorithms when physician-experts are not intended to provide immediate oversight and overrule capabilities. We believe that without the safeguards provided by direct physician-expert oversight during each use **in addition** to a total product life cycle (TPLC) regulatory approach with comprehensive, real-time monitoring of deployed products, it may be infeasible for FDA to ensure the safety and effectiveness of continuously adaptive algorithms.

Autonomously Functioning AI in Clinical Practice

During the Workshop, two potential uses of autonomously functioning AI algorithms were considered:

1. Algorithms for identification of "normal" radiological examinations, and
2. Algorithms for "ruling out" critical diseases

<u>Normal Examinations</u>:

In considering using AI to identify normal examinations, presenters discussed using AI for screening mammography to identify "normal" examinations that would not be reviewed by a radiologist. We are not confident that these examinations can be safely excluded from radiologist interpretation while maintaining the current level of patient safety; and certainly not without rigorous pre-market evaluation and post-market monitoring of the algorithms using data from diverse patient demographics and equipment from different vendors. FDA's regulatory paradigm should reflect the variety of ways screening mammography is performed including 2D and 3D (digital breast tomosynthesis) techniques; each algorithm will have to produce effective outcomes in all of these settings. For radiologist interpretations, the false negative rate is approximately 0.8 /1000 examinations (Lehman, et.al.; Radiology April 2017); FDA should require autonomously functioning AI to perform at that level prior to being considered for autonomous interpretation. A sample size of 195,537 over-reads would be required to accurately estimate that proportion of false negatives with 95% confidence for each algorithm. More importantly, the FDA must have a robust method of evaluating the performance of the algorithm in real world settings. The prevalence of breast cancer in screening populations can be as few

as 2 cases of breast cancer for 1000 patients. If an algorithm ceases to function properly, without radiologists overreading the examinations, thousands of patients might be screened before algorithm failure is recognized. As such, post-market surveillance with enough radiologist overreads within an AI registry is critical to maintain patient safety at current levels. The number would be variable and dependent on the prevalence of the disease processes to be detected. When the goal is to detect disease with a very low prevalence, independent validation and monitoring using a large number of cases to ensure longitudinal performance of the algorithm is at least equal to that pf physician experts. In the case of breast cancer, which has a prevalence in screening populations as low as 0.2%, a sample size of 766,756 examinations would be necessary for monitoring to ensure adequate performance of the algorithm, clearly a number that cannot achieved by a typical site. As such institution of autonomously functioning AI in screening populations would require significant changes to current workflows in imaging care potentially including centralization of data within registries for secondary interpretation to ensure longitudinal performance of the algorithms. The FDA should impose a rigid statistical approach to ensure enough cases are continuously and longitudinally evaluated by physicians so that potential algorithm failure can be detected promptly, and patients harm averted.

Presentations at the public workshop did not clearly define what a normal examination looks like. In the context of mammography, the FDA-mandated BI-RADS reporting system is not just "positive" or "negative" system and the various BIRADS scores, including the benign categories, have nuanced clinical implications requiring physician judgement for proper assignment. While detection of breast cancer is the primary goal of mammographic screening, a number of benign processes, including breast calcifications and soft tissue masses, can mimic breast cancer and will therefore need to be identified by the AI algorithm. Notably, the number of these "benign" diseases is limited in the breast as compared to other organ systems. If "identification of normal examinations" is expanded to other more complex body systems, AI algorithms would need to accurately recognize normal anatomic variants and exclude potentially thousands of other processes. AI algorithms must be tested against the myriad of disease processes that radiologists exclude during radiographic interpretation as many of these produce subtle findings that may go unrecognized by an algorithm unless it was trained to find them.

Many have suggested, algorithms may function best when used in conjunction with radiologists' interpretations rather than stand alone. In a paper published in March 2020 about the performance of AI in screening mammography, Shaffter, et al conclude that, *"While no single AI algorithm outperformed radiologists, an ensemble of AI algorithms combined with radiologist assessment in a single-reader screening environment improved overall accuracy. This study underscores the potential of using machine for enhancing screening mammography interpretation"* [JAMA Network Open. 2020;3(3):e200265. doi:10.1001/jamanetworkopen.2020.0265]. We agree with this study's conclusion and believe it is applicable across a broad array of machine learning applications in radiology.

<u>Product Labelling Should Clearly Delineate the Limitations of Rule-Out Algorithms and Recommend Additional Follow-up of "Negative" Results</u>

Workshop presenters used "detection of intracranial hemorrhage" (ICH) as an example of how an autonomously functioning algorithm might work in clinical practice. We agree that detection of intracranial hemorrhage is a vital function for the use of computed tomography (CT) in emergency settings and that current experience shows that a number of AI algorithms are able to reliably detect ICH in clinical practice. Currently these algorithms are being used in the clinical setting for worklist prioritization with subsequent interpretation by radiologists for validation and disease characterization. While some suggest that absence of hemorrhage might be a suitable "final" diagnosis, we disagree. To ensure appropriate patient care in the "rule-out" setting, all cases would still need radiologist

interpretation to exclude the myriad of non-hemorrhagic diagnoses (tumor, stroke, demyelinating disease and infection) that could mimic symptoms and signs of ICH. The same concerns also apply to other "rule-out" scenarios such as pulmonary thromboembolic disease (PTE) or fracture detection. As an example, the number of chest diseases that could mimic signs and symptoms of PTE – including critical diseases such as aortic dissection or pneumothorax – is extensive; ruling out PTE can in no way be considered a diagnosis in and of itself. Similarly, as was discussed at the Workshop, fracture detection algorithms may be able to identify subtle fractures, but until they are able to identify other diseases including infection, stress reaction and neoplasm (some of which can be quite subtle), they cannot function autonomously, without radiologist overread, and maintain current patient safety. Rule-out algorithms and automated interpretations (unless part of a series of automated algorithms) would be expected to miss incidental/secondary findings that could otherwise be discovered by a human reader, and in many circumstances would be the standard of care for a radiologist to identify.

The same rigorous validation processes and post-market surveillance also apply for rule-out scenarios. Validation must be across a heterogenous patient group and variety of machines and imaging protocols. Rule-out algorithms must also have extensive product labeling letting users know that the algorithm has only been validated for a specific disease or diseases (e.g. PTE) and is only intended for use with a specific set of equipment and protocols (scanner manufacturer, slice thickness and radiation dose). Also, end-users must be warned that other clinically significant diseases that could mimic signs and symptoms of the target disease processes will not be detected.

Finally, radiologists do more than recognize and characterize disease process during their interpretation. Physician training includes information processing and contextual integration of data in order to render a medical judgment. Physicians are able to provide these contextual medical judgements; machines are not. An example would be machine detection of a potential but equivocal small intracranial hemorrhage at a rural hospital. If there is no radiologist oversight of the algorithm to say this is unlikely to be a significant finding, a local physician without expertise in imaging might initiate an emergency helicopter transfer to a tertiary care facility where the ultimate care would likely only be observation and repeat examinations. If a radiologist who understands the down-stream consequences of overdiagnosis is able to intervene, unnecessary care might be avoided. Taking out the human with oversight/overrule capabilities eliminates the ability for the physician expert to contribute to ensuring the safety/performance of the algorithm. We have yet to learn if there is any overall degradation in care or radiologist performance when radiologists are exclusively exposed to data flagged as abnormal by automated AI. Concerns exist regarding bias that may be introduced by the AI that would hasten a satisfaction of search by a radiologist and conceivably overlook findings that were also undetected by the AI. Research into these potential deleterious outcomes of AI is necessary before autonomous AI can be trusted in routine clinical practice.

Potential Uses of Autonomous AI in Radiological Care

Our organizations believe that there are some potential early uses for autonomously functioning AI in clinical practice. AI is poised to have tremendous clinical benefits in population health management. Diseases with important clinical ramifications such as pulmonary emphysema, hepatic steatosis, high body mass index, osteoporosis and many others can be identified and quantified by AI algorithms in patients undergoing imaging for other reasons such as trauma or inflammatory diseases. Although these diseases are frequently identified by radiologists interpreting the examinations, most often the results are buried in the reports and the presence of these diseases never makes it to the patient's medical record problem list for future evaluation. We believe that algorithms that can identify and quantify these disease processes and then transmit the information to the patient's care team via the EHR could

have significant positive impact on population health management analogous to autonomous detection of diabetic retinopathy by IDx-DR. We believe these types of autonomously functioning AI will enhance patient care and potentially save lives if treatment can be instituted early. While rigorous validation and monitoring would still be necessary, in this setting patient safety is maintained as the output is not specifically designed to impact contemporaneous treatment for the primary reason imaging was performed but rather to facilitate referral for additional care those patients at risk of consequences from these important chronic disease processes.

Summary

In summary, the ACR and RSNA believe it is unlikely FDA could provide reasonable assurance of the safety and effectiveness of autonomous AI in radiology patient care without more rigorous testing, surveillance, and other oversight mechanisms throughout the total product life cycle. Before developing pathways for the authorization of autonomously functioning AI in radiological imaging, the FDA should first wait until current AI algorithms have a broader penetrance in the marketplace so that their efficacy and safety in a "supervised" manner can be documented which could then inform decision making regarding the premarket approval and post-market surveillance process requisite for autonomously functioning AI. If the goal of autonomous AI is to remove the physician from the image interpretation, then the public must be assured that the algorithm will be as safe and effective as the physicians it replaces, which includes the ability to incorporate available context and identify secondary findings that would typically be identified during physician interpretation. We believe this level of safety is a long way off, and while AI is poised to assist physicians in the care of their patients, autonomously functioning AI algorithms should not be implemented at this time. The value that human interpretation with independent medical judgement brings to patient care cannot currently be replaced. It would be best for FDA to focus its regulatory resources on solutions that address areas of clinical value to radiologists and their patients. Algorithms that assist physicians in population health management by incidentally detecting and quantifying potentially undiagnosed chronic diseases would be an excellent way to begin bringing autonomous AI into radiological care.

Thank you for your consideration of these comments. For more information, please contact Gloria Romanelli, JD, ACR Senior Director of Senior Director, Legislative and Regulatory Relations, or Michael Peters, ACR Director of Legislative and Regulatory Affairs, at (202) 223-1670 or mpeters@acr.org.

Sincerely,

Howard B. Fleishon, MD, MMM, FACR
Chair, Board of Chancellors
American College of Radiology

Bruce G. Haffty, MD
Chair, Board of Directors
Radiological Society of North America