ACR®
AMERICAN COLLEGE OF RADIOLOGY
QUALITY IS OUR IMAGE

June 3, 2019

Submitted via Regulations.gov

The Honorable Norman E. Sharpless MD
Acting Commissioner
Food and Drug Administration
10903 New Hampshire Ave
Silver Spring, MD 20993

Subject: (FDA-2019-N-1185) FDA Discussion Paper: Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning-Based Software as a Medical Device; Comments of the American College of Radiology

Dear Commissioner Sharpless:

The American College of Radiology (ACR)—a professional association representing over 38,000 diagnostic radiologists, interventional radiologists, radiation oncologists, nuclear medicine physicians, and medical physicists—appreciates the opportunity to comment on the discussion paper from the U.S. Food and Drug Administration (FDA) titled, "Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD)" (FDA-2019-N-1185). The following comments were compiled by leaders of the ACR Commission on Informatics and Data Science Institute.

General Comments

The proposed regulatory framework outlined for public comment in the discussion paper represents a novel, progressive approach towards AI/ML-based SaMD. If implemented as envisioned, the discussed regulatory framework would directly address several key issues related to AI/ML software, including some of the most critical (and challenging) regulatory issues facing this new technology. By encouraging the development of a 'predetermined change control plan' that anticipates the evolution of these machine learning solutions as part of a pre-marketing submission, the FDA may enable a future where machine learning solutions adapt and improve during real-world usage.

In general, the ACR agrees with the approach of allowing developers to include the addition of the 'SaMD Pre-Specifications' (SPS) to define potential scope of intended changes and an 'Algorithm Change Protocol' (ACP) to control the risks of anticipated modifications. It is consistent with our perspective that evolution of this technology must be enabled for the good of patients, but only while ensuring the safety and effectiveness of these devices. Implementation of Good Machine Learning Practices (GMLP) is critical, as understanding quality in the context of dynamic AI/ML-enabled systems is a work in progress. Finally, given the dynamic nature of machine learning models, ensuring safety and efficacy throughout the lifecycle of AI/ML software is absolutely essential.

The FDA’s discussed framework for regulating AI/ML-based SaMD is encouraging, and we believe—with refinement prior to finalization, and collaboration with the physician community during implementation—the agency’s proposed approach will further bolster development of this important technology. Additionally, we would like to underscore several important high-level considerations:

- **Enabling the broader imaging community to play a central part in assessing AI/ML models is a critical component of the success of this proposal’s implementation. This is true for both changing and unchanging AI/ML models to ensure safety and model performance throughout the total product lifecycle**. Effective improvements in the model will rely on effective surveillance incorporating inputs from a broad set of clinical users. We know, for instance, that local differences in disease prevalence, image acquisition, and patient presentation can decrease model performance. Thus, it is our view that solutions that rely on AI/ML must take this variability into account, engage a broad set of radiologists to ensure performance, and be validated using data from multiple sites prior to deployment for widespread clinical practice. The clinical imaging community must be enabled broadly to provide feedback to the models as the work in the real world for the proposed approach to take hold. This is true both of models that are in the process of being improved as well as static models whose performance may change over time. This will require a commitment to, and investments in, tools and education for the radiology community to assess model performance locally.

- **We believe that model development should be informed by requirements that are developed by clinical experts**. One barrier to success for many model developers is understanding the clinical requirements for these models in the real world. As we expect these models to increase in complexity with additional data inputs as specified in the proposal, we suggest that model developers will be most successful working closely with clinicians to define model requirements. For example, the ACR has produced a set of standard “use cases” for imaging AI that can serve as a basis for the development of these models.

- **A high level of traceability and visibility into training data, as well as performance assessments in different scenarios, will be most valuable for clinicians and engineers to drive awareness about where models may fail and how to improve them.** Much of the performance of deep learning models is predicated upon the training data that was used to create the model. In addition, clear visibility into the way that training data was annotated, which data elements for used, and visibility to metadata, is critical for “debugging” models in the clinical environment. Tools and training to allow this kind of traceability, as part of GMLP, will be vital. Imaging registries, such as those held at the ACR, may serve as a foundation for such an approach.

Specific Comments

AI/ML-SaMD Modification Types

In the FDA's proposed regulatory framework, changes to the type of inputs used in an AI/ML model, expansion of training data sets with no change in intended use, and modifications to the software's intended use, are all treated differently. The ACR agrees with this strategy, as a change to the indication of use of software should prompt a higher level of scrutiny than augmenting a model with an approved indication for use.

Several other considerations may be in order--for example, outputs for the models may also change without a change in the intended use. For example: the addition of a saliency map for a pulmonary nodule detector may constitute a change in the output of the model, but not a change in the indication. This may nevertheless change the way the device is used, and should be considered.

In addition, the difference between “modifications related to performance, with new change to the use or new input type” and “modifications related to inputs, with no changed the use“ may require additional detail for some edge cases. For example, in the pulmonary nodule use case, it should be considered whether the addition of image data developed with new image reconstruction methods, or low-dose acquisitions, would constitute an additional input data or categorically different data. We would encourage the FDA to engage the clinical imaging community to help address these border cases.

Finally, as currently posed, software embedded in a medical device (SiMD)—including software employing ML/AI for tasks such as reconstruction, as well as segmentation or classification tasks on raw/non-reconstructed data—is outside the scope of SaMD. However, as AI/ML models become more integrated into new scanners and other hardware medical devices, FDA should consider applying same regulatory strategy to AI/ML-enabled SiMD.

Good Machine Learning Practices

Within the discussed framework, the emphasis on demonstrating a culture of high-quality engineering is appreciated. GMLP should include diligent model training, ensuring model generalizability, and rigorous clinical validation.

Data traceability is a critical component that may require additional attention. This is the practice of documenting the origins and processing of any data used in model development and is commonplace in other AI-based industries, such as the autonomous vehicle industry. Without a formal documentation process, it may be impossible to reproduce models that were built from extremely large and unstructured data sets. This would hamper future model development and make investigating model errors almost impossible. The FDA could support traceability by defining the minimum information required for each data point when creating medical AI models.

This is further compounded if model iteration proceeds to a site-specific development of models. It will be key to approach this problem by using specifically-crafted software and databasing to make sure every model can be traced back not just to its architecture and the way that it was parameterized, but to each individual component of training data. This way, failures can be addressed by augmenting specific slices of the data, either by collecting new data or synthetically generating representative data.

Such solutions do not currently exist within AI/ML models within healthcare. Empowering this capability should be a goal for an effective GMLP. Undoubtedly, developing the tools needed to enable this will require collaboration across the industry, as well as stakeholders in the clinical and regulatory communities. This may be empowered by connectivity to registries which can hold parts of the source training data, such as those developed /maintained by the ACR.

SaMD Pre-Specifications and Algorithm Change Protocol

The discussion paper outlines a general approach to ensuring safety and effectiveness while improving model performance through the ACP, which details the data and procedures to be followed so that the modification safely achieves its goals. FDA suggests that the ACP should include processes for new data collection and curation, model retraining methodology, evaluation criteria, and a strategy for updating software.

Iteration and retraining models may have unintended consequences, including decreasing performance in some locations and disease types while increasing performance in others.

Thus, it will be important for changes to models that are performed globally to be assessed locally. Engagement of the clinical community for this task is of the utmost importance--having a set of clinicians involved in the assessment and validation of the retreat models is the only way to provide reasonable assurance of safety and efficacy across multiple practice sites nationally.

Data access and portability may present a challenge, and with increasing data privacy concerns and greater recognition of the value of data, there are significant hurdles in sharing data. This may be true even between a model's manufacturer and its end users. Therefore, the FDA should request that manufacturers outline in the ACP plans for continuing data access. This could be facilitated by leveraging an accredited model deployment platform which handles tasks (such as anonymization and data security) and is trusted by end users—an example is the ACR's AI-LAB.

While the responsibility for performance improvement lies predominantly with the manufacturer, FDA should anticipate the likelihood that deployment sites may optimize models to their own local data and user preferences. Such sites should work with the manufacturer to continually test and optimize their local version of the model. In this scenario, the manufacturer should be required to provide clear technical guidance for model updates and periodically measure and report population-level performance, perhaps against standardized datasets. The concept outlined here is analogous to a resident adapting to local clinical presentations and protocols while periodically undergoing central board exams to ensure a basic level of clinical safety.

One solution would be to leverage parts of the licensing and certification strategies developed to maintain quality performance on scanners via the existing TRIAD system provided by the ACR. Through this mechanism, one might use a distributed method to certify the AI models locally whenever a change is performed. This would ensure the performance of the models not just with a standard test set, but also that as data in the field continues to evolve, models maintain their capabilities over time.

Premarket Review

The ACR generally agrees with the discussed approach of premarket review augmented by the predetermined change control plan. In general, a focused review may include competition of models with additional members of and already established patient population within the model. It may also represent the expansion of data across different scanner types or different kinds of reconstructions as they emerge, given that these evolutions may not be anticipated for inclusion within an SPS. However, we would reiterate that assessment of these models should be performed by broad set of clinical evaluators--preferably by testing the model against a gold standard data set and data collected by broad set of end-users, as described in the section on transparency and real-world performance monitoring.

To that end, the ACR Data Science Institute offers imaging AI developers an independent evaluation and validation service (Certify-AI) as a means of demonstrating performance to regulators and end-users. By combining a clearly defined use case with a well-curated 'ground truth' reference standard dataset, Certify-AI independently evaluates algorithm performance to safeguard against non-generalizable results in routine clinical practice. Certify-AI datasets are created with cases from multiple institutions and include cases that span the range of known sources of variability. Evaluation of the algorithm entails defining the appropriate statistical performance metrics and minimal acceptable criteria for formal statistical testing.

Transparency and Real-World Performance Monitoring

The discussed framework emphasizes that accurate reporting of a model's performance and approved indications must occur consistently and in an appropriate and accessible form for the model's users. This is especially important because user behavior often adapts to a product's characteristics. For example, users may learn to ignore a model which generates many false-positive alarms. However, if this model's specificity is subsequently improved and users are not properly notified, potentially serious alarms could be ignored despite a well-functioning model.

Thus, we believe that real-world performance monitoring of AI/ML models depends on continued validation and certification of models in the context of the various environments in which the models are used. The clinical environments in which the models are deployed undergo continuous evolution. Even in scenarios where AI/ML models remain static, the data on which they are expected to perform changes. Therefore, this technology must be validated and certified on an ongoing basis.

To promote transparency, end-users should be supplied with characteristics of the population in which the training data was performed and information explaining how the training data has changed during any interval updates. This data should be readily available and easily understood at the point of care, so that relevance of the model for individual patients can be assessed if necessary. This is another critical reason for developing strong traceability for models to training data as part of GMLP. Individual site data (as assessed by site-specific testing), as well as global metrics of performance (which can be generated through certification strategies such as utilizing ACR's TRIAD), should be readily accessible.

To that end, the ACR Data Science Institute provides monitoring of algorithm performance in clinical practice by capturing real-world data during clinical use in a clinical data registry (Assess-AI). Using data collected in the registry, Assess-AI combines specific information related to an algorithm's effectiveness reported by radiologists at the point of care, as well as specific metadata related to the exam as specified in the defined use case.

Thank you for your consideration of these comments. As always, the American College of Radiology welcomes further collaboration and dialog with FDA staff regarding the proposed regulatory framework in the discussion paper and any related issues. For additional information, please contact Gloria Romanelli, JD, ACR Senior Director of Legislative and Regulatory Relations, at gromanelli@acr.org; and Michael Peters, ACR Director of Legislative and Regulatory Affairs, at mpeters@acr.org | (202) 223-1670.

Sincerely,

Geraldine B. McGinty, MD, MBA, FACR
Chair, Board of Chancellors
American College of Radiology