**Question & Answers**

**Question from Alice Geaney: Will the FDA be defining consensus standards for Paige Prostate device product code?**

**Answer from Peter Yang:** A guidance document might be one way to provide such information. However, for the different kinds of applications there would be a lot of details needed. My thinking is less a general guidance, but rather taking the individual device function into account to assure that it is analytically and clinically valid for its purpose.

**Question from Joy Kavanagh: Peter you very rightly recommend early engagement with FDA via pre-submission process. can you comment on current wait time for pre-subs, i.e. has this returned to normal availability similar to pre-COVID or still limited?**

**Answer from Peter Yang:** This is currently changing. There are still some limitations. For in vitro diagnostics, as of June 1st, 2022, the device review group is accepting non-COVID pre-submissions. However, while they might be accepted, non-COVID pre-submissions might be reviewed on a slower timeline.

**Question from Alice Geaney: Could the FDA also talk to defining the image viewer in the intended purpose?**

**Answer from Peter Yang:** The pathology group will likely have some level of control over the entire pipeline and the related requirements.

**Answer from Brandon Gallas:** The viewer opens the image and provides the pixels to the software as a medical device (SaMD). The viewer likely also displays the SaMD outputs. As such, the viewer is important to the successful clinical implementation of the SaMD. I believe that we currently expect that the viewer will be defined in the product labeling. The viewer is also part of the regulatory device (scanner, image management system, and viewer). It is possible that standards in this area could change our thinking, as could evidence and arguments about exact pixel-level equivalence across viewers.

**Question from Steph Hess: Why was the approach as an aide afterwards after initial diagnosis and not a guide ahead of final diagnosis?**

**From Samuel Chen: As I read the document, Paige's software performs only AFTER a pathologist independently gives his/her diagnosis 1st. Is there a particular reason for that?**

**Answer from Emre Gültürk:** We consider this a brand-new industry. The first products are geared towards more conservative settings until enough post-market data has been produced to assure safety. In addition, adjunct settings enable production of post-market data.

**Answer from Brandon Gallas:** Indications for use are the manufacturers responsibility. Manufacturers are responsible for picking that path. The indications for use determine how to validate the device. There can be conversations between the FDA and the manufacturer about the appropriate validation for different indications for use. Indications drive the study.

**Answer from Emre Gültürk:** An autonomous AI algorithm. Keeping a device at a Class II level – will be sufficient to start collecting data.

**Question from Jennifer Picarsic: Is there a recommended sample size going forward for other AI submissions (including for non-cancer AI detection)**

**Answer from Brandon Gallas:** Every disease and task can have a different baseline performance, the performance without the AI/device. This baseline performance, estimates of pathologist variability, and the study design are the key components that impact the uncertainty (confidence intervals) in a clinical study (a study comparing pathologist performance with and without the AI device). Please check out the literature on multi-reader multi-case (MRMC) analysis methods that account for the variability from the readers and the cases. Such an accounting is appropriate for sizing a study to satisfy the performance goals. Additionally, it is not just a statistical question. The risk of the disease and therapies, of course, plays a role, and the number and characteristics of the readers and cases need to represent the populations of interest. There is no blanket answer.

**Answer from Peter Yang:** If we were to provide general recommendations or standards that would apply for any application, we would likely end up advocating for a conservative approach, which would not be suitably flexible for a particular application.

**Question from Jansen Seeheult:** Does the FDA do a sample size calculation? E.g. to take into the claim and precision?

**Answer from Brandon Gallas:** In our research we develop methods to do pilot studies to size pivotal studies. Are there expectations for sample size?

The size is determined by the manufacturer. Assuming the manufacturer meets with FDA before launching their study in a pre-submission meeting, they will discuss and justify the study size. This should be based on statistical methods and the estimation goals. Please check out the literature on “Multi-reader multi-case analyses” (MRMC), which are mostly arising from the radiology world.

**Question from James Harrison: If the local verification with 10% positive cases yields different accuracy from the vendor data with about 50% positive cases, what does that mean? The vendor can correctly claim that a significant difference in accuracy is expected with changes in class balance, but that doesn't provide local guidance on whether a particular performance level represent verification or clinical acceptability of performance.**

**Answer from Jansen Seheult:** Good question. Your accuracy verification study should mirror your patient population. If you fail to verify the vendor's claim, then there is additional work needed to understand why the algorithm doesn't perform as stated: is it failing in one patient population? We use the vendor's claim as a starting point, but of course, the laboratory director may choose more stringent criteria. This is a very complex topic that is challenging to cover in 5 minutes. Importantly, if you have an imbalanced sample, you may instead want to verify the sensitivity/ specificity claims directly instead of a naïve accuracy claim since the prevalence of the entity of interest does affect the sample size for verifying sensitivity/ specificity.

**Answer from Brandon Gallas:** Accuracy is not a general metric we prefer at the FDA, because it pools the two different classes (disease + non-disease). It obfuscates the performance in both. You cannot look at accuracy by itself. We always ask for sensitivity and specificity. In the radiology world we look at the ROC curve (and AUC, the area under ROC curve). ROC analysis shows the sensitivity and specificity pair across the full range of decision thresholds. We find this approach to be more robust and yield more definitive results compared to sensitivity and specificity, which are affected by the decision-threshold variability from the radiologists.

**Question from Arun Nemani: What is the FDA's guidance for models that ideally should be retrained periodically to ensure robustness. Does this require a formal review upon every retrain?**

**Answer from Peter Yang:** You are fundamentally changing the performance of the model. Some will know that I’m referring to a guidance that talks about making modifications to your device and when a new submission is required. We are working on updated guidance. We prefer stepwise – and detailed explanations what you are trying to accomplish. Again, it depends on the specific intended use.

**Answer from Brandon Gallas:** There are two guidance documents when you make changes to your device or your software device.

Links to guidance on when to come back to the agency for changes

<https://www.fda.gov/media/99812/download>

<https://www.fda.gov/media/99785/download>

**Question from Arun Nemani:** **Does the FDA also have requirements for model monitoring in production. For example, for performance monitoring when a model is deployed.**

**Answer from Peter Yang:** The concern would be that the model changes over time. You would not be able to ensure safety and effectiveness over time. The agency has been deliberating through the conceptual framework for this.

**Answer from Brandon Gallas:** The agency is currently working on a framework that has been proposed for managing models that change. Currently there are three devices that allow continuous model updates. Devices that have continuous learning capabilities will need to have special controls that need to accompany such a device submission.

**Question from James Harrison: The primary concern in monitoring is that the data changes over time, not the model. Even a static model needs performance monitoring in case the underlying data shifts or drifts.**

Answer from Joe Lennerz: Monitoring of model performance would likely be part of the ongoing performance assessment (under CLIA; e.g. proficiency testing)

**Question from James Harrison: One question is whether the package insert should contain target accuracy values that can be used directly for comparison with local verification results. For that, the FDA data set for performance calculation would need to represent "typical" case mixes.**

**Answer from Brandon Gallas:** The decision summary is the place and opportunity for describing the studies conducted to support device clearance, granting, or approval. However, the decision summary is really a reflection of the information that FDA used to decide to permit marketing of the device. It can serve as a useful comparison to support future submissions to FDA, but is not necessarily written beyond that purpose. If you want more information than you find in the decision summary, you should engage the FDA through the Q-sub process. You can ask for more information or complain that more information should be provided.

* FDA CDRH, “Requests for Feedback and Meetings  for Medical Device Submissions: The Q-Submission Program.” FDA, 2019. Accessed: Dec. 08, 2021. [Online]. Available: <https://www.fda.gov/regulatory-information/search-fda-guidance-documents/requests-feedback-and-meetings-medical-device-submissions-q-submission-program>

**Question from James Harrison: Or perhaps establish a range of case mixes across which the performance of the system is acceptable. Probably should use "class balance" rather than "case mix" in my comments. ROC, sensitivity, and specificity are still susceptible to class balance issues. I used "accuracy" in a general sense, not naive accuracy. The question still remains how differing class imbalances should be handled.**

**Answer from Brandon Gallas:** The case mix may impact a pathologist binary decision-making behavior, and consequently, impact sensitivity and specificity. That is why ROC analysis has been popular in the radiology world, as it shows the sensitivity and specificity operating point across the full range of decision thresholds. ROC analysis and AUC are not susceptible to class balance issues (enrichment). ROC analysis and AUC are susceptible to differences in the inclusion/exclusion criteria and the balance of patient subgroups in a study population. Clarity on these issues is very important.

**Question from Sean McClure: is it all DICOM compliant?**

**Answer from Brandon Gallas:** No. I have not seen DICOM adoption in the devices we have reviewed within the regulatory space. I think adoption of a standard would have a very positive impact on adoption and innovation in digital pathology.

**Question from Ricardo Gonzalez, M.D. (he/him): Does the FDA has any recommendations about how to perform an External validation (i.e. test them with independent datasets) of a ML model in medical images?**

**Answer from Brandon Gallas:** While probably not exactly what you were hoping for, here are some related guidance documents:

* FDA CDRH, “Guidance for industry and FDA staff - clinical performance assessment: considerations for computer-assisted detection devices applied to radiology images and radiology device data in premarket notification [510(k)] submissions.” FDA, 2020. Accessed: Aug. 05, 2020. [Online]. Available: <https://www.fda.gov/media/77642/download>
* FDA CDRH, “Guidance for industry and FDA staff - Computer-Assisted Detection Devices Applied to Radiology Images and Radiology Device Data - Premarket Notification [510(k)] Submissions.” FDA, 2012. Accessed: Apr. 21, 2020. [Online]. Available: <https://www.fda.gov/media/77635/download>
* FDA CDRH, “General Principles of Software Validation.” FDA, 2021. Accessed: Feb. 05, 2022. [Online]. Available: <https://www.fda.gov/regulatory-information/search-fda-guidance-documents/content-premarket-submissions-device-software-functions>
* {U.S. Food and Drug Administration (FDA)}, {Health Canada}, and {United Kindom’s Medicines and Healthcare products Regulatory Agency (MHRA)}, “Good Machine Learning Practice for Medical Device Development: Guiding Principles.” U.S. FDA, 2021. Accessed: Jul. 11, 2022. [Online]. Available: <https://www.fda.gov/medical-devices/software-medical-device-samd/good-machine-learning-practice-medical-device-development-guiding-principles?utm_medium=email&utm_source=govdelivery>
* {Artificial Intelligence Medical Devices -AIMD- Working Group}, “Machine Learning-enabled Medical Devices: Key Terms and Definitions.” International Medical Device Regulators Forum, 2022. [Online]. Available: <https://www.imdrf.org/documents/machine-learning-enabled-medical-devices-key-terms-and-definitions>

**Answer from Peter Yang:** Your best bet is to look at some of our most recent regulatory decisions – Paige Prostate and other devices that use AI/ML – and understand what was done there. Then, come and discuss your external validation plans with the Agency and understand any specific concerns they have.

**Question from James Harrison: What if the lab changes stain or fixative vendors?**

**Answer from Peter Yang:** Generally speaking, we would seek to understand how robust your system is, i.e. whether it’s designed for processing images created with a specific scanner or slides stained by a certain lab or with certain stains, etc. If you state that your software device can process any and every image, you would generally be asked to provide data to show that is the case. You may also want to read about FDA’s quality systems requirements (see 21 CFR 820 and the link below), which discuss “good manufacturing practices” for developing and making changes to medical devices, including software.

* Quality System (QS) Regulation/Medical Device Good Manufacturing Practices. U.S. FDA, 2022. Accessed: Jul. 11, 2022. [Online]. Available: https://www.fda.gov/medical-devices/postmarket-requirements-devices/quality-system-qs-regulationmedical-device-good-manufacturing-practices

**Question from Dr Kim Blenman: This is a wonderful session and discussion. Thank you to all of the hosts. Should we consider a follow-up discussion?**

**Answer from Joe Lennerz:** Yes, we are happy to announce that Dr. Jansen Seeheult has accepted an invitation to speak on the topic of verification/validation of AI tools in the laboratory.