Dear Dr. ElZarrad -

Thank you for your time today. This is a longer version of a submitted question I have posted to your upcoming session at the Stanford AIMI symposium. Sahiner Berkman and Nick Palatkow and I have spoken in the past, and I'm preparing a manuscript as well as doing additional investigations on this area of great interest to me. I communicate this to you as an advocate and proponent of "AI done right for the benefit of the patient." At this time, I have no commercial conflicts of interest.

The FDA mandate is to "protect the public health by ensuring the safety, efficacy, and security of human and veterinary drugs, biological products, and medical devices"[1] As such, the approval process uses the tried and true metrics of population health: sensitivity, specificity, positive predictive value, the ROC curve and its AUC. However, SaMD, particularly machine learning or AI algorithms, will have impact at the individual, not population level, and are likely to progress from provider-supervised to partially-autonomous operation with high-level supervision at some point in the not-too-distant future.

Machine learning algorithms are impacted by a wide spectrum of inputs and effects, including but not limited to class imbalance, spectrum bias, poor reproducibility in algorithmic training, potential end user decision opacity, and biases inherent in algorithm choice. There may be value in going beyond currently provided metrics.[2]

My question to you is:

1. Is the FDA open to utilizing new metrics beyond those described above?
   - examples: Informedness[3], Markedness[4], Pearson's ρ[4], *F*rechet inception distance[5], Resampling Uncertainty Estimation[6]
2. Has the FDA investigated the utility of newer metrics? Which have been useful?
3. Are regulatory or legislative changes necessary for the FDA to implement and require new metrics to fulfill its mandate if supported by hard scientific evidence?

Thank you.

Stephen M. Borstelmann MD
Boca Raton, FL

1. FDA. What We Do. https://www.fda.gov/about-fda/what-we-do retrieved Aug 3 2020
2. Borstelmann, SM . Beyond the ROC Curve. Presented at the 2019 Radiologic Society of North America Annual Meeting, Chicago IL.
3. Powers DMW. Evaluation: From Precision, Recall and F-measure to ROC, informedness, markedness, and correlation. Journal Machine Learning Technologies 2011 (2) 1.
4. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., & Hochreiter, S. (2017). Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems* (pp. 6626–6637)
5. Schulam, P., Saria, S. Can You Trust This Prediction? Auditing Pointwise Reliability After Learning. Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha,Okinawa, Japan. PMLR: Volume 89.