Q&A from Pathology Innovation Collaborative Community Webinar
https://pathologyinnovationcc.org/presentations/aug-2022-reader-study-designs-and-mrmc-analysis

Evaluating Medical Imaging Devices and Image Based Algorithms with the Clinician in the Loop:

Tutorial on Reader Study Designs and MRMC Analysis

From Sailesh Conjeti: Is there a python version of iMRMC available?

My iMRMC software is available as a java-based graphical user interface (https://github.com/DIDSR/iMRMC) and as R code (https://cran.r-project.org/web/packages/iMRMC/index.html). You can run the iMRMC R code using the python "subprocess" module, and there is a package that allows python to utilize R code (https://appsilon.com/use-r-and-python-together/).

From Sailesh Conjeti: @Brandon: In this study design work, should one not also consider the truther reads, esp. if we are using expert panels with multiple experts?

BDG: When sizing an MRMC study there are three populations to sample: diseased cases, non-diseased cases, and readers. The readers that I focused on in my work are the end users of the imaging device. You are introducing another population of readers: the experts that establish the reference standard. I use "experts" here, because we want to compare the end-user scores to the best possible reference standard. The experts should also have any information that would allow them to make the best reference standard possible, including patient history, demographics, other diagnostic test results, and other information that might only be available after follow-up with additional diagnostic studies. The costs associated with these experts providing the reference standard should be part of the practical considerations when sizing an MRMC study. I would also say that there is a need for more research on how best to utilize experts and ***account for the variability*** in the reference standard from the experts. This is currently part of the current research effort of the Office of Science and Engineering Laboratories in the Center for Devices and Radiological Health at the FDA and my High-Throughput Truthing (HTT) project: https://ncihub.org/groups/eedapstudies/wiki/HTTWhatisHTT.

From Joe Lennerz: Did the readers in the VIPER study know about the prevalence (i.e., the fraction of cancer to non-cancer cases)? It would be of interest to model the cost savings when optimizing MRMC studies.

BDG: We provided the readers in the VIPER study with descriptions of each data set: case mix and approximate prevalence. They weren't told exactly how many cancer and non-cancer cases there were, but we didn't want to fool the readers. We wanted them to make the best recall / don't recall decisions

possible using their clinically developed threshold. We wanted them to then provide an ROC score that was as granular as they could make it and calibrated to "all the cases you have ever seen." We made a special effort to provide clear and explicit instructions to the readers. You can find our scoring instructions in the link below. Regarding the optimization of an MRMC study, I'll point you to the supplemental materials for the VIPER study. You will see the steps I took to estimate the variance components and explore the study size to achieve a certain level of precision in the performance results. The document was created from an R markdown file. The R markdown file includes the R code to create all the results and figures. You can find the document and the corresponding R markdown file here:

- ROC scoring instructions: https://didsr.github.io/viperData/inst/docs/viperSupplementaryMaterials/viperInstructions-scoring-v2.pdf
- VIPER sizing analysis: https://didsr.github.io/viperData/inst/docs/viperSupplementaryMaterials/viperSupplementaryMaterials.pdf,
- R markdown of the VIPER sizing analysis: https://github.com/DIDSR/viperData/raw/master/inst/docs/viperSupplementaryMaterials/viperSupplementaryMaterials.Rmd

From Markus Herrmann: @Brandon: How should one balance the number of "easy" versus "challenging" cases in a reader study to optimally assess a technology?

BDG: Let me start by saying that in a regulatory setting, you should prepare a proposal providing context and justification for your case-sampling plan. Then request a pre-submission meeting to discuss your proposal with the appropriate FDA reviewers. These pre-submission meetings are also referred to as Q-subs, and I recommend them for all impactful questions related to the intended use of your medical device and generating evidence to support marketing of the medical device.

In the research setting, I believe the key is transparency. The goal is to convince your audience that you have considered the device's intended clinical population when creating your sampling plan, how performance results will be affected, and how your results generalize. You probably should discuss the practical challenges in recruiting and curating your samples, as well as acknowledge the ultimate limitations of your sampling process. I feel it is often appropriate to stratify your sampling to adequately represent different subsets of the population, including rare cases. You may find it necessary or worthwhile to analyze the data on the different subsets to provide the full performance picture, whether or not you size the subsets for a rigorous statistical analysis. If the purpose of your study is to compare two similar modalities (devices or interpretation strategies), the concerns about the case mix are mitigated when the two modalities are assessed on the same cases. In this situation, cases that are challenging are often challenging on both systems; one system doesn't gain a structural advantage from the sampling.