I'd like to mention two research projects for which we are seeking to hire a post-doc for at least a year or a doctoral candidate for the summer:

1. The Office of Chief Scientist (OCS) has funded our research project "Performance metrics for multi-class decision support systems" (Mar 1 2022 to Feb 28 2024). Kenny Cha is the principle investigator and I am one of the co-investigators. The Quad Chart and full Research proposal are attached. Kenny is looking would like to hire a post-doc with a statistical background to help with the project as an an ORISE staff fellow. If you have any recommendations, please let us know. I think Kenny will provide me a formal announcement soon.

1. I have some funding left over from my 2021 Critical Path (CP) research project "Statistical Evaluation of Diagnostic Devices for Low Prevalence Conditions" (attachment 3), for which I was seeking a doctoral candidate for the summer to help with developing the statistical methods and with writing RStudio or RShiny software tools for the methods we have already developed. My division is supporting this project through next year, so a post-doc hired for at least a year as an ORISE staff fellow should be doable.

I think the hired candidate(s) could also be involved with papers on evaluating the accuracy of quantitative measurements such as pulse oximeter oxygenation saturation, which has been very controversial lately.

Thanks so much,

Gene

Gene Pennello, PhD, *Mathematical Statistician*
U.S. Food and Drug Administration
Center for Devices and Radiological Health
Office of Science & Engineering Laboratories
Division of Imaging, Diagnostics, & Software Reliability
10903 New Hampshire Avenue, White Oak 64, Room 3020
Silver Spring MD 20993-0002

Tel: 301-796-6038, Fax 301-847-8123, gene.pennello@fda.hhs.gov

Evaluation of classification and prediction performance of non-binary decision support systems

- Kenny Cha, PhD, CDRH/OSEL/DIDSR (Biomedical engineer; image analysis, AI/ML, HPC)

CDRH collaborators

- Berkman Sahiner, PhD; CDRH/OSEL/DIDSR (Electrical engineer; AI, deep learning, imaging)
- Gene Pennello, PhD, CDRH/OSEL/DIDSR (Statistician, statistics, study design)
- Manasi Sheth, PhD, CDRH/OPEQ/OCEA/DCEA2 (Statistician; statistics, study design)
- Daniel Erchul, CDRH/OPEQ/OHT4/DHT4A (Biomedical engineer; reviewer for light-based energy devices)

CDER collaborators

- Abbas Bandukwala, CDER/OND/ODES/DBIRBD (Biomedical engineer, biomarker qualification team)

NCTR collaborators

- Zhichao Liu, PhD, NCTR/OR/DBB (Visiting Scientist; AI, deep learning)

Length of Performance: 2 years

Total Budget: $220K for two years (1 ORISE postdoc fellow).

Specific Aims:

Aim 1: Survey analytical performance metrics in multi-class classification including their meaning and computation methods. Review the metrics being used in non-medical disciplines, such as neural network-based classification in computer vision and multi-spectral analysis and metrics cited in statistical literature for evaluating nominal and ordinal classification variables.

Aim 2: Assess the feasibility of using existing metrics in medical device evaluation, by understanding each metric's strengths, weaknesses, and key properties relevant to various clinical tasks. Study new and underutilized metrics and methods to estimate confidence intervals to quantify sampling variability for tasks where currently existing metrics or methodology are inadequate. Explore how different study designs affects the metrics.

Aim 3: Provide recommendations for metrics that should be utilized in demonstrating effectiveness in evaluation of a multi-class classification task. If no single metric is found to be acceptable, provide a list of characteristics that a set of metrics should have to be acceptable for a medical application. The metrics recommended may depend on study design.

Objective:

With the rise of artificial intelligence (AI) and machine learning (ML), AI/ML-based decision support systems are increasingly being developed for binary classification tasks such as discriminating between presence/absence of disease, e.g., cancer/non-cancer. The methodology used to evaluate the classification and prediction performance of these binary classifiers is well-established. Recently, AI/ML-based systems have been applied to non-binary classification tasks in which an object is classified into one of many classes. For example, a device may classify a skin lesion as a mole, rash, melanoma, or basal cell carcinoma. While the evaluation of these multi-class classification tasks is common in the computer vision field, a consensus is lacking on which performance metrics are appropriate for medical applications because device outputs may influence patient diagnosis and treatment. This research aims to develop principles and methodology for evaluating devices that perform multi-class classification. Our research will help select clinically meaningful metrics based on the task being performed, while addressing common issues such as study design, prevalence and operating point.

Regulatory Impact:

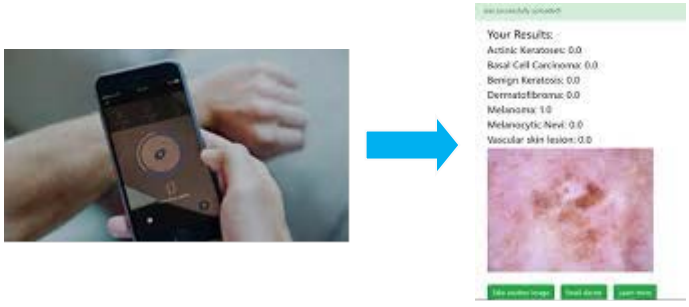
The successful outcomes of this research will grant the agency an initial understanding of different evaluation metrics that can be used for evaluating multi-class classification devices. The acquired knowledge will enable a streamlined (1) CDRH review of AI/ML classification devices, and (2) CDER and NCTR review of biomarkers that classify these indicators into one of many possible types.

Expected Outcomes:

Peer-reviewed papers, a decision tree style flowchart for choosing performance metrics based on the nature of the clinical tasks, and a blueprint for guidance document on multi-class classification.

Background/Project Motivation:

In recent years, there have been several premarket submissions, including breakthrough devices, which expand the current state of decision support systems for binary tasks into multi-class classification problems, where the device is meant to place a case into one of several different categories. Examples include skin lesion classification, cancer subtyping, and computer-aided triage across multiple conditions. The evaluation of these devices is challenging, as there is no consensus on metrics that provide an assurance of effectiveness for these devices. A reason for this is that the effects of prevalence and the trade-offs among the choice of multiple thresholds for multi-class classification are not well-understood. This research aims to fill the knowledge gap that exists in the evaluation of systems that selects a category from a list of possible categories in medical applications, including medical devices and biomarker qualification.



Example of a multi-class classification problem for medical imaging: identifying skin lesion from skin lesion images taken via smartphones.

Project Title: Performance metrics for multi-class decision support systems
Total Amount Requested: $238,469 (see Funds Worksheet for details)
Funding Period: [March 1, 2022]-[February 28, 2024]
Principal Investigator, Center & co-PIs, affiliation:

- Kenny Cha, PhD; CDRH/OSEL/DIDSR (Assistant director; image analysis AI/ML, HPC)
- CDER:
  - Abbas Bandukwala; CDER/OND/ODES/DBIRBD (Biomedical engineer, biomarker qualification team)
- NCTR:
  - Zhichao Liu, PhD; NCTR/OR/DBB (Visiting Scientist; AI, deep learning)
- CDRH:
  - Nicholas Petrick, PhD; CDRH/OSEL/DIDSR (Deputy director; AI, deep learning, imaging)
  - Jana Delfino, PhD; CDRH/OSEL/DIDSR (Assistant director; AI, imaging)
  - Berkman Sahiner, PhD; CDRH/OSEL/DIDSR (Electrical engineer; AI, deep learning, imaging)
  - Gene Pennello, PhD; CDRH/OSEL/DIDSR (Statistician; statistics, study design)
  - Manasi Sheth, PhD; CDRH/OPEQ/OCEA/DCEA2(Statistician; statistics, study design)
  - Daniel Erchul; CDRH/OPEQ/OHT4/DHT4A (Biomedical engineer, reviewer for light-based energy devices)

1. **Public Abstract**: Provide plain language summary of the proposal suitable for public dissemination.
   Up to 300 words (approximately a half page)

   This project will develop methods for evaluation of devices that performs complex classification tasks such as multi-outcome diagnoses. The growth in artificial intelligence (AI) and machine learning (ML) has led to increased development of decision support systems for medical classification and prediction. Binary classification tasks include discriminating between presence/absence of disease, e.g., cancer/non-cancer or malignant/benign. Methods for evaluating the performance of binary classification and prediction are well-established. However, AI/ML-based systems are increasingly being developed for non-binary, polychotomous classification tasks, for which performance evaluation is less understood. For example, a device may classify a skin lesion as a mole, rash, melanoma, or basal cell carcinoma. While the evaluation of these multi-class classification tasks is common in the pattern recognition field – the field for enabling computers to identify and process patterns in data such as images, a consensus is lacking on the study design to determine the safety and effectiveness of devices that perform such tasks. In addition, agreement in the field is lacking regarding which performance metrics are appropriate for medical applications attempting to solve these types of problems. This is important as the device outputs may influence patient diagnosis and treatment, and improper evaluation of these devices can lead to harm to the patients who are affected by these devices, either directly or indirectly.

This research aims to develop principles and methodology for evaluating devices that perform multi-class classification. A team of FDA scientists will study the state of the field, and provide information and data to help industry, as well as the FDA reviewers, in designing a performance evaluation study to show the safety and the effectiveness of a device performing non-binary classification. Our research will provide guidance on how to select clinically meaningful metrics based on the task being performed, while addressing common issues such as study design, prevalence, and operating point. This project will assist in developing assessment metrics for multi-class decision support systems, helping to bring these devices to the US market in a safe and effective manner, giving access to cutting-edge technologies for patients, and providing least-burdensome approaches for device manufacturers.

2. **Technical Abstract**: Provide scientific summary of the proposal suitable for public dissemination.
   Up to 300 words (approximately a half page)

   This project will develop methods for evaluation of devices that performs complex classification tasks such as multi-outcome diagnoses. Medical applications of artificial intelligence (AI) and machine learning (ML) have grown significantly in recent years. An underutilized application area for AI/ML with the potential to benefit patients is multi-class (non-binary or polychotomous) decision support. The output is more granular for these decision support systems than for binary systems, providing additional information that may help patients in receiving better care. However, current knowledge for medical applications is limited in the evaluation of algorithms with multiple classification outputs, making it difficult to come up with consensus approaches for assessing AI/ML algorithms. This project has 3 specific aims: (1) Review analytical performance metrics in medical and non-medical multi-class classification, including their clinical interpretation and methods of computation. Examples for non-medical discipline include neural network-based classification in pattern recognition and multi-spectral analysis and metrics cited in statistical literature for evaluating nominal and ordinal classification variables; (2) Assess the feasibility of the identified metrics in medical device evaluation, by understanding each metric's strengths, weaknesses, and key properties relevant to various clinical tasks. Explore how different study designs affect the metrics. Review methods for estimating confidence intervals to quantify sampling variability. Conduct research on new and underutilized metrics and for tasks where currently existing metrics or methodology are inadequate; (3) Provide recommendations for metrics that should be utilized in demonstrating effectiveness in evaluation of a multi-class classification task. If no single metric is found to be acceptable, provide a list of characteristics that a set of metrics should have to be acceptable for a medical application. The metrics recommended may depend on study design. This research will greatly benefit industry, academic researchers, and regulatory reviewers in developing criteria for

the evaluation of non-binary decision support systems and facilitating the translation of such innovative techniques to patients.

3. **Introduction, and preliminary results:**
For new applications, provide a concise up-to-date status of the field, discuss the PD/PI's preliminary studies, data, and or experience pertinent to this application.
For renewal/revision applications, provide a summary of the project to date with the beginning and ending dates for the period covered. Summarize the specific aims of the previous project period and the importance of the findings, and emphasize the progress made toward their achievement. Explain any significant changes to the specific aims and any new directions including changes to the specific aims.
Up to 600 words (approximately one page)

Machine learning algorithms have shown great promise in performing both non-medical and medical classification tasks, i.e., tasks in which a choice is made between a finite set of possibilities.

In recent years, much literature has appeared on multi-class classification techniques. Examples include prostate Gleason grade grouping, skin lesion classification, and cancer subtyping [7-10]. Studies of these classification systems have used different metrics to evaluate the performance of the algorithms, such as the top-N-accuracy (a common metric used in the computer vision field for natural scene images), agreement with a reference standard using various concordance measures, and other measures derived from the confusion matrix such as the Matthew's Correlation Coefficient. Table 1 shows a preliminary table of metrics used for multi-class classification. While many of these metrics have been used extensively in the past, it is important to understand the implications of using them to evaluate performance in medical applications.

Table 1: a preliminary list of for multi-class classification metrics found in literature

Weighted accuracy	Likelihood ratios
Top-N-accuracy	Confusion matrix
Cross-entropy	Agreement
Kappa	k-class Youden's index
Average precision	Concordance
Matthew's correlation coefficient	Area under the ROC curve (AUC)
F1 score	

To date, most decision support systems have been developed for binary tasks. However, medical devices intended for adjunctive use to support detection and

diagnosis are expanding into multi-class classification problems, where the device is meant to place a case into one of several different categories. The FDA is already seeing premarket submissions of devices with a multi-class diagnosis and will likely see more in the future. In dermatology, a dermatologists' availability limits a patient's access to a health care professional who could help with their skin condition. Computer-aided diagnosis systems that help principal healthcare providers make accurate referrals to dermatologists could aid patients in more timely access to care. Such a system could provide information to the user by providing a result from a multi-class classification algorithm. For example, a device may classify a skin lesion as a mole, rash melanoma, or basal cell carcinoma. Another device area in which multi-class classification is performed is computer-aided triage devices that flag and prioritize cases. These devices can quickly process the relevant data (e.g., images), which can lead to a notification of a serious patient condition being sent to a pertinent medical specialist, who could then provide more timely access to treatment. Multiple triaging systems designed to identify a single condition have received FDA clearance. The next step would be to extend these devices to identify multiple conditions from a single patient, or even into multi-category disease conditions or severity.

Evaluation of multi-class classification devices is challenging. There is no consensus on metrics that provide an assurance of safety and effectiveness for these devices. Reasons include that the effects of prevalence and the trade-offs among the choice of multiple thresholds for multi-class classification are not well-understood. This research aims to fill the knowledge gap that exists in the evaluation of systems that selects a category from a list of possible categories in medical applications, including medical devices and biomarker qualification.

Our research team has extensive experience in performance evaluation and study designs for characterizing binary classification algorithms [1-6]. The information learned from the research that was performed by the members of research team has set the standard for evaluation of binary decision support systems. In addition, many members of our team are part of the quantitative image biomarkers alliance (QIBA), which aim to improve the value and practicality of quantitative image biomarkers by reducing variability across devices, sites, patients, and time. This works hopes to build upon the efforts of the QIBA the Metrology Committee (https://qibawiki.rsna.org/index.php/Metrology_Committee), which is currently exploring gaps and methodology for four multi-parametric use cases (multi-dimensional description, phenotype classification, risk-prediction, and data-driven biomarkers). This research will collaborate with the QIBA multi-parametric metrology task force through the members that are part of both efforts to study how to evaluate multi-class classification outputs and the benefits and the associated drawbacks for specific metrics.

Although we have limited preliminary work on the main goals of this project on multi-

class classification, we believe that our previous and current experience includes all the necessary expertise for success in the proposed project.

4. **Regulatory Impact, Relevance and Significance Statement**: Explain the importance of the problem or critical barrier for the public health and regulatory needs that the proposed project addresses. It should also describe how the project relates and is relevant to the program you are applying (e.g. OWH, MCMi, CORES etc).
   Up to 300 words (approximately a half page)

   Appropriate evaluation of multi-class classification systems directly impacts regulatory decision making for a broad spectrum of device studies and drug trials. For Drug trials, a classification system may be used for enrollment eligibility. Several device submissions to the FDA have involved multi-class classification. Due to the lack of knowledge in evaluation of the device, there were delays in communicating our thinking to the companies. Many such devices are for medium-to-high-risk applications, such as direct to patient-use skin lesion analyzer devices via smartphone app. Correspondingly, the impacted FDA offices include CDRH/OPEQ/OHT7, OHT4, CDRH's Medical Device Qualification Tool program, CDER's Biomarker Qualification program, CDER's Drug Development Tool program, and NCTR's Division of Bioinformatics and Biostatistics. Despite these new developments, there is currently no consensus on methods for multi-class classification assessment nor guidance by the agency. This, according to our intra/inter-center consulting and regulatory review experiences, is a critical barrier for translating these new technologies to the clinic. The regulatory process can be significantly delayed due to lack of constructive guidance, and innovative features/uses can be removed from a device due to lack of appropriate assessment methods. Our research will provide methods to facilitate a streamlined review of FDA submissions related to multi-class classification algorithms. The guidance document and published papers can be used by industry to facilitate the translation of cutting-edge AI/ML technologies from bench to bedside. This will enable a streamlined CDRH review process for AI/ML classification devices and CDER and NCTR review of biomarkers used to classify subjects into one of many possible states. As such, this project is closely aligned with the following strategic priorities in the Advancing Regulatory Science strategic plan:

   - Stimulate Innovation in Clinical Evaluations and Personalized Medicine to Improve Product Development and Patient Outcomes
   - Ensure FDA Readiness to Evaluate Innovative Emerging Technologies

5. **Expected outcomes**: What is the Projected Outcome? Please provide a description of projected outcome(s) such as the development of regulatory science tools like assays, assessment methodologies, and test protocols.
   Up to 300 words (approximately a half page)

The expected outcomes of this project are as follows: 1) Determine the robustness of study designs and metrics to analyze the performance of algorithms that can predict among multiple given conditions. We will start with a survey of currently used metrics in the literature in a variety of application areas, including non-medical computer-vision areas, to further understand the state of the multi-class classification outside of the medical applications. We will publish peer-reviewed papers analyzing the commonly used study designs and metrics, performing statistical analysis to understand the design and metrics, and determining their potential benefits and the risks, and compare the different study designs and metrics. 2) Criteria for selecting study design and metrics with desired properties for a specific clinical task. Using the information learned from the investigation, we will prepare a decision tree-style flowchart to help determine the metric with favorable characteristics for the study design that best fits the clinical task being performed. If no single metric is found to be acceptable, we will provide a list of characteristics that a set of metrics should have to be acceptable for a medical application. We hope that the results of this research will be able to be the blueprint for a potential guidance document on performing evaluation of multi-class classification applications.

6. **Research Strategy**: a) aims, b) innovation & collaboration (for Chief Scientist challenge grant), c) approach & methods.
   Up to 2,400 words (approximately four pages, additional information may be requested during the review)

   a) Aims: Succinctly list the specific objectives of the proposed research, e.g., to test a stated hypothesis, create a novel design, solve a specific problem, challenge an existing paradigm or clinical practice, address a critical barrier to progress in the field, or develop new technology.

   b) Innovation: Describe any novel theoretical concepts, approaches or methodologies, instrumentation or interventions to be developed or used, and any advantage over existing methodologies, instrumentation, or interventions. Collaboration: Describe ongoing or new collaborations outside of the center and explain why these collaborations are essential for the successful completion of the project (compared to alternatives such as commercial services).

   c) Study Design: Describe the study design including the overall strategy, methodology, and analyses to be used to accomplish the specific aims of the project. Discuss how data will be collected, analyzed, and interpreted as well as any resource sharing plans as appropriate. Discuss potential problems, alternative strategies, and benchmarks for success anticipated to achieve the aims. If the project is in the early stages of development, describe any strategy to establish

feasibility, and address the management of any high-risk aspects of the proposed work.

a) Aims

Aim 1: Survey analytical performance metrics in multi-class classification including their meaning, underlying assumptions, and computation methods. Review the metrics being used in non-medical disciplines, such as neural network-based classification in computer vision and multi-spectral analysis, and metrics cited in statistical literature for evaluating nominal and ordinal classification variables.

Aim 2: Assess the feasibility of using existing metrics in medical device evaluation, by understanding each metric's strengths, weaknesses, and key properties relevant to various clinical tasks. Study new and underutilized metrics and methods to estimate confidence intervals to quantify sampling variability for tasks where currently existing metrics or methodology are inadequate. Explore how different study designs affects the metrics.

Aim 3: Study the fit the metrics identified from Aim 2 to different diagnostic areas with different study design condensations. Apply the metrics to simulated data and real-world data to evaluate the metric's fit to the given clinical task. Provide recommendations for metrics that should be utilized in demonstrating effectiveness in evaluation of a multi-class classification task. If no single metric is found to be acceptable, provide a list of characteristics that a set of metrics should have to be acceptable for a medical application. The metrics recommended may depend on study design.

b) Innovation & Collaboration

Innovation: While the methods for performance evaluation of multi-class classification tasks are common in the field of pattern recognition, a consensus is lacking in on the study design to determine the safety and effectiveness of algorithms that perform multi-class classification. In addition, there is no agreement in the field regarding which performance metrics are appropriate for medical applications attempting to solve this type of problems. While there have been attempts in the literature to study such problems [11-13], a conclusion could not be drawn. From the results of our research, we will prepare a decision tree-style flowchart to help determine the metric with favorable characteristics for the study design that best fits the clinical task being performed. If no single metric is found to be acceptable, we will provide a list of characteristics that a set of

metrics should have to be acceptable for a medical application. We hope that the results of this research will be able to be the blueprint for a potential guidance document on performing evaluation of multi-class classification applications.

Collaboration: The determination of appropriate study design and performance metrics for multi-class classification problems has multiple applications that is spread across different FDA centers. To understand the implications across the different application areas, we have planned this project as a coordinated multi-center effort: CDRH – medical devices; CDER – biomarkers; NCTR – precision medicine.

c) Approach & Methods

Aim 1:

We will survey analytical performance metrics in multi-class classification currently being in the field of artificial intelligence and machine learning and pattern recognition, encompassing non-medical applications, as well as medical applications. Metrics cited in statistical literature for evaluating nominal (where order doesn't matter), and ordinal (where order does matter) classification will be explored. We will study the metric's interpretation, underlying assumptions, and computation methods.

In addition to literature, we will survey "challenges" where competitors attempt to solve a common problem or task for competition and rewards, to understand the reasoning behind the choice of metrics used to determine the winners for these challenges. Examples of such challenge include the ImageNet Large Scale Visual Recognition Challenge[14], which was one of the drivers for the current advancement in the field of AI-based computer recognition. Other challenges include Kaggle (https://www.kaggle.com/), which is a large data science community, and the Grand Challenge (https://grand-challenge.org/), which is a platform for end-to-end development of machine learning solutions in biomedical imaging. We will compile a list of various metrics that have been used or proposed, and their associated study design that enabled the usage of the metric. The results of this aim will be a chart of metrics and study design that will be further studied in the later aims.

Aim 2:

We will assess the feasibility of using the existing metrics found in Aim 1 in medical device evaluation. Common metrics for this application currently include top-N-

accuracy, kappa metrics, concordance, and accuracy. Each metric's strengths, weaknesses, and key properties relevant to various clinical tasks will be studied. The metric's tolerance to common issues with performance evaluation, such as study design, prevalence, and operating point, will be assessed.

In order to perform this assessment, algorithm outputs and the associated reference standards are needed for a specific task. We will start by gathering data from the previously mentioned public challenges. The ImageNet challenge is a popular challenge in the computer vision field. We will take common network architectures for ImageNet classification that is pre-trained on the ImageNet training data set, such as the AlexNet[15] VGG[16], and ResNet[17], and apply them to the ImageNet validation/tuning data set to obtain algorithm outputs with known reference standard and varying performances. Published results tend to show high-performing algorithms, leading to information regarding low-performance algorithms being limited. To study how the metrics identified behave with both high-performing and low-performing outputs, we will vary the neural network parameters trained on the same data to get a range of outputs with varying performance. Several members of the research team were also organizers of multiple medical imaging-based classification challenges[7, 18]. With access to the outputs to the algorithms submitted for these challenges, we have access to a rich data set that we can modify for classification purposes and use to study how changing the study design and evaluation metric may affect the conclusions of the challenges. Using the data set, algorithm outputs, and their associated reference standards, we will study existing metrics for multi-class classification.

In addition to using algorithm outputs designed with real data, we will also perform simulations studies. We will simulate the output of classifier outputs for set number of classes allow us to design the performance characteristics of the problem being solved. This will allow us to identify if a metric accurately represents the performance characteristics for the simulated outputs.

If we find that currently existing metrics or methodologies are inadequate, we will study new and underutilized metrics and methods to estimate the confidence intervals to quantify sampling variability for tasks. The results of this aim will be an understanding of the advantages, disadvantages, and important characteristics of each metrics as applied to various clinical tasks.

Aim 3:
Using the information learned from aim 2, we will provide recommendations for metrics that should be utilized in demonstrating effectiveness in evaluation of a multi-class classification task in a variety of specific medical applications. We will perform experiments using the identified metrics on both simulated data and real-world data to evaluate the metric's fit to a given clinical task.

The recommendations will come in the form of a decision tree-style flowchart to help determine the metric with favorable characteristics for the study design that best fits the clinical task being performed to demonstrate the effectiveness in the evaluation of a multi-class classification task. If no single metric is found to be acceptable, we will provide a list of characteristics that a set of metrics should have to be acceptable for a medical application, taking into account that the study design will affect the recommendation. The goal for this aim is to produce a blueprint for a potential guidance document on performing evaluation of multi-class classification applications. Our results are expected to assist in overcoming assessment of non-binary decision support systems, helping to bring these types of devices and treatment methods to the U.S. market in a safe and effective manner, giving access to cutting-edge technologies for patients.

7. **Projected Time Lines:** Provide dates for projected milestones and expected deliverable at least in three month increments in a table or Gantt chart format.

Milestones and Deliverables	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8
Milestone: Survey of multi-class classification metrics commonly used outside of medical applications								
Deliverable: Manuscript of the survey results in medical journal								
Milestone: Acquire algorithm output and truths for specific medical tasks/applications								
Milestone: Study metric characteristics of known metrics								
Deliverable: Annual written report to OCS, oral presentations to Senior Science Council								
Milestone: Survey of statistical literature metrics								
Milestone: Study metric characteristics of under-utilized and new metrics in statistical literature								
Milestone: Performance metrics decision-tree flowchart								
Deliverable: Manuscript on performance metrics								
Milestone: Determine characteristics to be acceptable for medical application								
Deliverable: Manuscript on recommendations for selected metrics in specific clinical tasks								
Deliverable: Final report to OCS								

8. **Facilities:** Describe the shared equipment (i.e., cores) and other laboratory resources necessary to carry out the project: no limit
   This information is used to assess the capability of the resources available to perform the effort proposed. Identify the facilities to be used (Laboratory, Animal, Computer, Office, Clinical and Other). If appropriate, indicate their capacities, pertinent capabilities, relative proximity and extent of availability to the project. Describe only those resources that are directly applicable to the proposed work.

   The Division of Imaging, Diagnostics, and Software Reliability (DIDSR) comprises approximately 15,000 square feet for office and laboratory use within the White Oak Federal Campus in Silver Spring, MD. All investigators have adequate office space with networked personal computers and access to telephones, copiers and fax machines. Investigators are located in close proximity to each other and to the associated laboratories.

   Computing Resources
   The Division computer resources include an approximately 3000-core computing cluster for high-speed computing with 464GB RAM, 0.5 PB of disk storage, with network access via Gigabit Ethernet on a private network dedicated to research, over 30 PC computers running Linux for access to the computational cluster, over 30 PC

computers running Windows for office automation as well as hardware control and data acquisition systems in our laboratories. In addition, DIDSR has assembled a 20-GPU cluster for intensive GPU computing with state-of-the-art graphics cards and CUDA libraries. The division also has 3 artificial intelligence (AI) systems containing 4 or 8 GPUs each, large RAM capacity and software for developing traditional and deep learning machine learning algorithms. Compilers available on these systems include C, FORTRAN, and Pascal for Linux, as well as Visual C, Visual C++ and Visual FORTRAN for Windows. To facilitate image simulation, analysis, and visualization there are 9 simultaneous-user IDL licenses (5 Windows, 4 Linux) and 20 licenses for MATLAB. The division also has access to another larger high-performance computing (HPC) cluster made of up over 5200 processing core, 32 systems with GPUs and access to over 2 PB of additional disc storage. This general purpose HPCs can support a wide array of projects including web-based bioinformatics analysis, artificial intelligence/machine learning, genomics, next-generation sequence analysis and alignment, modeling and simulation, statistical analysis and more.

9. **Bibliography/References:** limit to 2 pages

Applicants should follow scholarly practices in providing citations for source materials relied upon when preparing any section of the application.

1. Sahiner, B., et al., *Stepwise linear discriminant analysis in computer-aided diagnosis: the effect of finite sample size.* Proceedings of the SPIE - Medical Imaging, 1999. **3661**: p. 499-510.
2. Metz, C.E., B.A. Herman, and J.H. Shen, *Maximum-likelihood estimation of receiver operating characteristic (ROC) curves from continuously-distributed data.* Statistics in Medicine, 1998. **17**: p. 1033-1053.
3. Sahiner, B., et al., *Finite sample effects on the area under the ROC curve for linear classifier design.* IEEE Transactions on pattern analysis and machine intelligence, 1997: p. (To be submitted).
4. Pennello, G., *Classification accuracy goals for diagnostic tests based on risk stratification.* Biostatistics & Epidemiology, 2021: p. 1-20.
5. Evans, S.R., et al., *Benefit-risk evaluation for diagnostics: a framework (BED-FRAME).* Clinical Infectious Diseases, 2016. **63**(6): p. 812-817.
6. Raunig, D.L., et al., *Quantitative imaging biomarkers: a review of statistical methods for technical performance assessment.* Statistical methods in medical research, 2015. **24**(1): p. 27-67.
7. Armato, S.G., et al., *PROSTATEx Challenges for computerized classification of prostate lesions from multiparametric magnetic resonance images.* Journal of Medical Imaging, 2018. **5**(4): p. 044501.
8. Liu, Y., et al., *A deep learning system for differential diagnosis of skin diseases.* Nature Medicine, 2020. **26**(6): p. 900-+.
9. Jain, A., et al., *Development and Assessment of an Artificial Intelligence–Based Tool for Skin Condition Diagnosis by Primary Care Physicians and Nurse Practitioners in Teledermatology Practices.* JAMA network open, 2021. **4**(4): p. e217249-e217249.
10. Cascianelli, S., et al., *Machine learning for RNA sequencing-based intrinsic subtyping of breast cancer.* Scientific reports, 2020. **10**(1): p. 1-13.

11. Ferri, C., J. Hernández-Orallo, and M.A. Salido, *Volume under the ROC Surface for Multi-class Problems*, in *Machine Learning: ECML 2003: 14th European Conference on Machine Learning, Cavtat-Dubrovnik, Croatia, September 22-26, 2003. Proceedings*, N. Lavrač, et al., Editors. 2003, Springer Berlin Heidelberg: Berlin, Heidelberg. p. 108-120.
12. Nakas, C.T. and C.T. Yiannoutsos, *Ordered multiple-class ROC analysis with continuous measurements.* Statistics in Medicine, 2004. **23**(22): p. 3437-49.
13. Grandini, M., E. Bagli, and G. Visani, *Metrics for multi-class classification: an overview.* arXiv preprint arXiv:2008.05756, 2020.
14. Russakovsky, O., et al., *ImageNet Large Scale Visual Recognition Challenge.* International Journal of Computer Vision, 2015. **115**(3): p. 211-252.
15. Krizhevsky, A., I. Sutskever, and G.E. Hinton, *ImageNet Classification with Deep Convolutional Neural Networks*, in *Advances in Neural Information Processing Systems*, F. Pereira, et al., Editors. 2012, Curran Associates, Inc. p. 1097--1105.
16. Simonyan, K. and A. Zisserman, *Very deep convolutional networks for large-scale image recognition.* arXiv preprint arXiv:1409.1556, 2014.
17. He, K., et al., *Deep residual learning for image recognition.* arXiv preprint arXiv:1512.03385, 2015.
18. Petrick, N., et al., *SPIE-AAPM-NCI BreastPathQ challenge: an image analysis challenge for quantitative tumor cellularity assessment in breast cancer histology images following neoadjuvant treatment.* Journal of Medical Imaging, 2021. **8**(3): p. 034501.

CDRH Critical Path Application Form - FY 2021

Submissions will be accepted no later than October 16, 2020

For content and technical questions regarding the application, please contact: CDRHRSProposals@fda.hhs.gov

I. Basic Information

Proposal Title *(500 characters including spaces)*	Statistical Evaluation of Diagnostic Devices for Low Prevalence Conditions
Lead Investigator Information	
Full Name (Last, First)	Pennello, Gene
Position/Title	Mathematical Statistician
Office	OSEL - Office of Science and Engineering Laboratories
Division	Division of Imaging, Diagnostics, and Software Reliability
Email	Gene.pennello@fda.hhs.gov
Expected Time Commitment *(Expressed as % of total workload)*	15%
Supervisor	Kyle Myers
Supervisor Email	Kyle.myers@fda.hhs.gov
Application Information	
Critical Path Status	Ongoing CP Project
Was a similar full proposal submitted to the FY21 Office of the Chief Scientist Proposal Call?	No
Research Program Supported *(Applicable if lead investigator is from OSEL: must select the program area supported by this proposal)*	Medical Imaging and Diagnostics

Table of Contents

II. Overview

II.a. Plain Language Summary

Provide a summary of the project in lay language and explain how this proposal helps CDRH achieve its vision. *(3000 characters including spaces)*

The scope of this research proposal is statistical evaluation of diagnostic devices (e.g., tests) intended to detect a low prevalence condition. Many tests are intended to detect a low prevalence condition. For example, screening tests are intended to detect conditions with typically very low prevalence in asymptomatic populations.

Statistical methods are less developed and less well understood for evaluating diagnostic devices (i.e., tests) than therapies. Study design and analysis issues are even less well understood for diagnostic tests intended to detect a low prevalence condition. For example, human papilloma virus (HPV) tests are intended to detect HPV genotypes that can cause cervical cancer. In the U.S., cervical cancer incidence is estimated to be just 8.1 cases per 100,000 women per year and is decreasing thanks to the availability of HPV vaccines. Thus, all-comers studies of HPV tests must have an exceedingly large sample size to enroll enough subjects with cervical cancer to estimate the sensitivity of the HPV test with sufficient precision. Moreover, cervical cancer status is verified based on histology of biopsies taken during colposcopy. In comparative HPV test accuracy studies, cervical cancer status will not be known definitively in women who are test negative on both the index test and the comparator test because referring these women to colposcopy is considered unethical (March 2019 FDA Microbiology Advisory Panel recommendation).

The aims of our project align with the CDRH vision. One aim is to develop valid and efficient study designs that enrich for a low prevalence condition yet can be used to obtain projected estimates of test accuracy in the entire intended use population (Aim 1). Other aims are to develop innovative statistical adjustments for verification bias (Aim 2) and confounding (Aims 4 and 6), use decision analysis to quantify the net benefit of a test to rule-in or rule-out a low prevalence condition based on agreed-upon risk thresholds for these decisions (Aim 3), develop alternative measures of test performance when test accuracy is too difficult to evaluate or interpret (Aims 5, 6), and consider Bayesian approaches for evaluating updates to diagnostic tests based machine learning (i.e., neural networks). If implemented, the innovative study designs and statistical methods to be developed are likely to (1) improve the interpretation of diagnostic device studies for low prevalence conditions, therefore improving the transparency and predictability of regulatory decision making for these tests, (2) provide a less burdensome pathway to reach the same pre-market approval decision, thus increasing U.S. access to safe and effective diagnostic tests first in the world, and (3) improve health care via personalized medicine.

II.b. Regulatory Science Priorities Addressed

Please identify two CDRH Regulatory Science Priorities that apply to your project.

1. Develop methods and tools to improve and streamline clinical trial design
2. Leverage precision medicine and biomarkers for predicting medical device performance, disease diagnosis, and progression

II.c. Evidence of Regulatory Science Gaps/Challenges

Please describe the regulatory science gaps or challenges your proposal is addressing regarding your chosen priorities in Section II.b. Please identify the CDRH stakeholders impacted by these gaps or challenges and provide evidence reflecting the significance of the identified gaps or challenges. *For proposals with lead investigators from OSEL, this section should align with the programmatic gaps or challenges identified in program charters. Contact OSEL program leads/managers for information on program charters. (3000 characters including spaces)*

Diagnostic tests are typically evaluated on real-world (i.e., observational) data and are the key to precision medicine, both CDRH priorities. The aims of this project address gaps identified in the Statistics Subprogram of the OSEL Medical Imaging and Diagnostics Program Charter, namely, identify innovative and efficient study designs to reduce the burden and uncertainty of diagnostic studies (Aim 1), develop methods of estimating diagnostic performance when some study subjects are not verified for presence or absence of condition (Aim 2), and develop methods of estimating diagnostic performance in the presence of confounding variables (Aim 4). Also, Aims 3 and 7 address gaps in the OSEL artificial intelligence/machine learning (AI/ML) charter, namely, to develop study design and analysis methods for rule-out AI/ML devices and develop methods for evaluating updates to AI/ML software as a medical device (SaMD). If implemented, the innovative methods to be developed are likely to (1) impact the review of diagnostic devices for low prevalence conditions, (2) lead to a less burdensome pathway to pre-market approval of these devices, and thus (3) lead to improved health care and personalized medicine.

II.d. Value Proposition

This section should include: (1) impact of this proposal on advancing the regulatory science priorities identified in Section II.b., (2) impact on CDRH mission if this proposal is not funded, and (3) a description of limitations associated with the proposal (including impact of COVID-19 related measures) and approaches used to minimize impact or mitigate risks. *(2000 characters including spaces)*

Statistical methods and tools are desperately needed to improve and streamline clinical trial design for diagnostic tests such as screening tests that detect low prevalence conditions. All-comers studies have become all but infeasible to evaluate such tests. Indeed, at the 2019 Microbiology Advisory panel meeting, the panel all but insisted that FDA stop requiring all-comers studies to evaluate HPV tests because of the tremendous burden to sponsors and time required to conduct such studies and the limited value of study results. All-comers studies are simply too inefficient to meet CDRH's mission that patients and providers have timely and continued access to safe, effective, and high-quality medical devices. If the proposal is not funded, then large all-comers studies with equivocal study results are likely to continue to be conducted in attempts to validate diagnostic tests for low prevalence conditions, leading to more uncertainty, less predictability, less consistency, and less transparency in regulatory decisions making compared with more efficient study designs, if they can be developed and accepted by FDA review staff.

The need for diagnostic tests in personalized medicine is increasing, including tests for disease diagnosis, earlier detection, and disease progression monitoring. For example, in the 2018 FDA guidance "Noncirrhotic Nonalcoholic Steatohepatitis [NASH] with Liver Fibrosis: Developing Drugs for Treatment", FDA states that "noninvasive biomarkers are needed (including imaging biomarkers) to supplant liver biopsy and provide a comparable or superior ability to accurately diagnose and assess various grades of NASH and stages of liver fibrosis. Identification and validation of such biomarkers could significantly accelerate drug development in [nonalcoholic fatty liver disease] NAFLD." Traditional measures for evaluating diagnostic test performance, e.g., sensitivity and specificity, are increasingly being recognized as insufficient for evaluating if a diagnostic test would be clinically useful in practice (Cook, 2007; Pepe, Janes, Li, Bossuyt, Feng, Hilden, 2016). If this research proposal is not funded, the adoption by CDRH of innovative statistical methods designed to evaluate if a test is clinically useful, such as the net benefits of a diagnostic test to rule-in or rule-out a disease condition (Aim 3), or the interchangeability of an index test with a comparator test (Aim 5), would likely be postponed, which could potentially lead to less predictability, transparency, and efficiency in regulatory decision making and ultimately less patient and provider access to safe and effective tests for precision medicine.

II.e. Public Health Impact

Discuss the public health need this proposal addresses (disease area, device type, etc.) and how it relates to CDRH vision. *(2000 characters including spaces)*

The project has great potential for large public health impact. Novel study designs and statistical methods for analysis are desperately needed to ensure that screening tests and other tests for low prevalence conditions can be evaluated in moderately sized studies to decrease burden on sponsors of diagnostic tests and increase the likelihood that safe and effective tests are marketed first in the US.

A main challenge is in designing a moderately sized study enriched with enough subjects with the low prevalence condition such that diagnostic test accuracy can be estimated with sufficient precision and be meaningfully compared with other tests having the same intended use. Similarly, a main focus of clinical epidemiology is efficient study designs and methods of analysis for evaluating potential associations of exposures with rare events. A premise of this research project is that well-understood epidemiologic study design and analysis methods can be exploited for evaluating diagnostic tests for a low prevalence condition and developing innovations in the regulatory review of such tests.

In our Aims, listed in Section III.a, we identify specific challenges. Our researchers already have much experience with these challenges. We therefore think our project has high potential to develop innovative solutions to these challenges faced in pre- and post-market validation of diagnostic tests for low prevalence conditions. Because of our experience, we have confidence that the innovative methods that we will develop will have the potential to greatly impact the review of diagnostic tests for low prevalence conditions such that these critical tests will have a much less burdensome pathway to pre-market approval, leading to improved health care and, in some cases, more personalized medicine. Our researchers also have much experience in presentation and peer-reviewed publication and therefore can effectively disseminate the research to stakeholders.

II.f. Previous Sources of Project Funding *(applicable to ongoing CP projects)*

Please list up to three (3) most recent funding sources for this project.

FY Funded and Project Title	Award Amount	Program Funding
FY20 Statistical Design and Evaluation of Diagnostic Devices for Rare Diseases	$110,402.92	Critical Path
		Choose an item.
		Choose an item.

II.g. Progress to Date *(applicable to ongoing CP projects)*

Describe the progress of your project to date. This may include milestones and deliverables met, input from stakeholders, impact on CDRH mission, and changes to research strategy or aims. *(3000 characters including spaces)*

Our FY20 research project was impeded because we weren't awarded FY20 Critical Path funding until very late in the award cycle: 2020 August 11 to be precise. Only recently, we selected an ORISE IAA research fellow, who will begin on 2020 November 30th. Nonetheless, progress has been made on some of the FY20 aims:

Aims 3-4 have to do with how to evaluate accuracy of a diagnostic test when not everyone is verified for disease status. In extreme settings, an entire subset of subjects are unverified because, for example, for ethical reasons they were not referred to the invasive procedure used to determine disease status (e.g., biopsy). Qin Li (CDRH/OPEQ/OCEA) and I developed a Bayesian model for comparing the diagnostic accuracy of two tests while addressing extreme verification bias. Our model gives good estimates of the accuracy of the diagnostic tests if the disease being diagnosed is rare (i.e., prevalence is low). We achieved this by assuming conditional dependence between the two tests, which is often plausible and is easily implemented in the Bayesian model. We presented our findings in the talk "Comparing Diagnostic Tests in Studies with Extreme Verification Bias--application to HPV testing" given on Aug 3rd at the Joint Statistical

Meetings, the signature meeting of the American Statistical Association. We also presented our findings in the poster "Using Gibbs Sampling and Data Augmentation to Compare Diagnostic Tests in RWE Studies with Extreme Verification Bias" given at FDA Scientific Computing Days (FDASCD) on Sep 30th. The poster emphasized the Gibbs sampler we develeped to do the required Bayesian computations. We intend to develop the model further and then write up the work for publication. I believe this work could have impact on how to analyze comparative diagnosfstic accuracy studies with verification bias.

Aim 6 is on evaluating the net benefit of a diagnostic test for its intended use. I made progress on this aim at the study design stage. I developed a method for determining classification accuracy performance goals that would confer that a diagnostic test would have positive net benefit when used in practice. The method is based on prespecifying acceptable levels of risk stratification by the test. Acceptable levels are often available from a patient management guideline, because the bedrock of such guidelines is risk thresholding. I have submitted a manuscript for publication in the journal Biostatistics & Epidemiology. I believe this work could have impact on how diagnostic accuracy studies are designed to support regulatory approval. I have discussed the work with some review staff and they feel it could have practical impact in the review of submissions.

Aim 9 has do with computing the Mid-P confidence interval for a binomial proportion. Diagnostic performance is typically summarized by binomial proportions such as sensitivity, specificity, positive and negative predictive value. The mid-P confidence interval (CI) has some of the best operating characteristics, but requires solving non-closed form root functions for the lower and upper limits of the CI. The Newton-Raphson method can quickly find the root if it is given a good starting guess, but will fail otherwise. The bisection method will always find the root but is slow. We discovered a fast bracketing method that will always find the root. The fast bracketing method would be useful when many calculations are needed, e.g., for a simulation. Jessie Moon (CDRH/OPEQ/OCEA) and I are writing a manuscript on our fast bracketing method, comparing its rate and order of convergence with classical methods.

III. Research Plan

III.a. Specific Aims

Please list the specific objectives/major tasks of the proposed research, such that it provides an outline of the project. *(2500 characters including spaces)*

Aim 1. Develop valid and efficient study designs that enrich for a low prevalence condition yet can be used to obtain projected estimates of test accuracy in the full intended use population.

Aim 2. Develop innovative statistical adjustments for verification bias (also called referral bias), including adjustments based on Bayesian analysis. In particular, develop valid estimators of test accuracy (e.g., sensitivity, specificity) even when the accuracy parameters are not completely identifiable in a statistical model, e.g., in a study with extreme verification bias.

Aim 3. Develop decision analysis measures of the net benefit of a diagnostic device to rule in or rule out a low prevalence condition based on agreed-upon risk thresholds for making these decisions (e.g., recommendations in guidelines).

Aim 4. Develop statistical methods that adjust diagnostic performance for confounding by numerous variables correlated with both test result and condition status.

Aim 5. Develop statistical measures that quantify the interchangeability of an index diagnostic test with a routinely used comparator test when accuracy of the test for diagnosing true condition status is infeasible to evaluate.

Aim 6. Consider causal inference measures for evaluating test performance. Compare causal inference measures with corresponding association measures commonly used for diagnostic test evaluation but susceptible to confounding.

Aim 7. Develop Bayesian evaluations of updates to machine-learning-based diagnostic tests based on predictive distributions.

III.b. Research Strategy

Please structure this section to provide sufficient information to gauge your likelihood for achieving study objectives. For each specific objectives/major task, please include: (1) preliminary data (if available), (2) approach and methods, (3) risk analysis, and (4) timeline, milestones, benchmarks, and deliverables. *(7500 characters including spaces)*

(1) Preliminary results, and (2) approach and methods.

Aim 1. After the March 2019 Microbiology Advisory Panel Meeting on HPV test evaluation, the National Cancer Institute (NCI) and FDA investigators met on several occasions to discuss efficient alternatives to all-comers study designs for HPV test evaluation. In these discussions, we made preliminary progress on efficient study designs that could be utilized for not just HPV tests but on any test for a low prevalence condition.

Aim 2. We have found recently that Bayesian methods of analysis for comparative diagnostic studies with extreme verification bias can nonetheless yield good estimators of test accuracy when the index and comparator tests are assumed to have a conditional dependence structure in the prior distribution. We presented these preliminary findings at the Joint Statistical Meetings, Aug 2020 (oral presentation) and at FDA Scientific Computing Days (poster).

Aim 3. We discovered that standardized net benefits of a test to rule-in and rule-out a condition have a geometric interpretation on the diagnostic likelihood ratio graph (c) and presented this interpretation at the Regulatory Industry Statistics Workshop (Sep 2018) with application to companion diagnostics. We also submitted a paper to *Biostatistics & Epidemiology* on determining classification accuracy goals (e.g., for sensitivity, specificity, likelihood ratio) based on risk stratification.

Aim 4. The Cochran-Mantel-Haenzel (CMH) odds ratio estimator adjusts for any number of categorical confounders. Agresti and Hartzel (Statist. Med, 2000) review CMH-like estimators for risk difference and relative risk. We anticipate these estimators would be useful for diagnostic test evaluation when as is often the case numerous covariates potentially confound the association of test result with condition status. We presented CMH estimators for diagnostic test evaluation at the Regulatory Industry Statistics Workshop, Sep 2019.

Aim 5. Assessing the interchangeability of an index test with a comparator test is receiving increased attention in the literature (e.g., Obuchowski et al, *Acad Radiol*, 2014) when accuracy of the index test is infeasible to evaluate. We anticipate developing statistical tools for evaluating interchangeability in lieu of accuracy.

Aim 6. Causal inference measures are probability statements about the potential outcomes for subjects who, e.g., receive treatment Rx or C or experience or not an exposure. In a discussion paper, we discovered that the common definitions of causal odds ratio and causal hazard ratio seem to be incorrect because they are based on the marginal distributions of the potential outcomes instead of their joint distribution. We anticipate research in this area can be fruitful for diagnostic test evaluation.

Aim 7. Diagnostic tests are increasingly being based on machine-learning (ML). ML algorithms are updated as data accumulate. An outstanding regulatory science question is how evaluate updates to ML-based diagnostic software as a medical device (SaMD). In a discussion paper (FDA, 2019), FDA sought input on how to regulate an artificial intelligence (AI) or ML-based SaMD having the power to continuously or periodically learn as data accumulate or as the needs the algorithm addresses change. Bayesian assessments of machine learning algorithms as data accumulate are quite natural, being based on the predictive distribution of new data given observed data (Spiegelhalter et al 1996; Gelfand and Dey 1992; Gelfand, Dey, Chang, 1994). We are intrigued that Bayesian assessments of updates to ML-based diagnostic tests could be much more efficient than frequentist attempts to control at 5% the false approval rate across innumerable updates.

(4) Timeline, milestones, benchmarks, and deliverables.

Once sufficient progress has been made on a research aim, we anticipate convening a meeting with CDRH review offices (OHTs), CDRH/OCEA, CDER, and NIH to present findings and seek feedback. These meetings will help us communicate the progress of the project, identify additional goals or adjust existing ones to assure that the project outcomes align with regulatory needs of the centers.

We expect the deliverables to include presentations at scientific conferences and publications in peer-reviewed journals for as many of the Aims as possible in the given time frame. We also expect some of the research to influence FDA pre-market review and perhaps generate FDA guidance development.

III.c. Figures

You may add 1 figure or graphic to your submission by clicking the picture control button below.

III.d. Timeline and Budget

For each specific objectives/major task, provide a brief description of the objective/task along with any milestone or deliverable under 'Description'; indicate if it is a milestone (M), deliverable (D) or not applicable (NA) under 'Type'; highlight the cells under 'Timeline' to show duration for each objective/task; and provide the cost for labor and supplies/equipment under 'Cost'.

Task	Description	Type	Timeline				Cost ($)	
			Q1	Q2	Q3	Q4	Labor	Supplies
Research	The existing personnel on this project have most of their time committed to other projects and device review consults. Considering the mathematical and scientific sophistication required for implementing existing statistical methods and developing new methods, we will need a researcher with a PhD degree in statistics fully committed to this project. We suggest an ORISE fellow. Deliverables include presentations, publications, and freely available statistical software tools for implementing the methods.	D			D	D	$100,000	$10,000
Total Project Cost							$110,000	

III.e. Budget Justification

Please include a detailed justification for requested labor (e.g. research fellow stipend), supplies, and equipment. This section should also include anticipated expenses for continuing the project after YEAR ONE of CP funding. *(2000 characters including spaces)*

Click or tap here to enter text.

III.f. Additional Information

Please check all that apply to this project.

☐ Human Subjects, Sample, or Data Use ☐ Animal Specimen/Data Use ☒ Medical Device Development Tool

☒ Computational Modeling and Simulation ☐ External Collaborations ☐ Technology Transfer

III.g. References

Please use this space to list any references relevant to your proposal. *(3000 characters including spaces)*

1. Agresti A, Hartzel J. Strategies for comparing treatments on a binary response with multi-centre data. Statist Med 2000; 19(8).
2. Biggerstaff BJ. Comparing diagnostic tests: a simple graphic using likelihood ratios. Statist Med 2000; 19(5).
3. Cook NR. Use and Misuse of the Receiver Operating Characteristic Curve in Risk Prediction, Circulation 2007; 115(7): 928–935.
4. FDA. Noncirrhotic Nonalcoholic Steatohepatitis with Liver Fibrosis: Developing Drugs for Treatment: Guidance for Industry, 2018.
5. FDA. Proposed regulatory framework for modifications to artificial intelligence/machine learning (AI/ML)-Based software as a medical device (SaMD)—discussion paper and request for feedback. 2019.
6. Gelfand AE, Dey DK. Bayesian Model Choice: Asymptotics and Exact Calculations. J Royal Statistical Society, Series B (Methodological) 1994; 56(3): 501-514.
7. Gelfand AE, Dey DK, Chang H. Model Determination Using Predictive Distributions with Implementation via Sampling-Based Methods, Technical Report No. 462, 1992 Dec 4; Department of Statistics, Stanford University, Stanford, CA.
8. Obuchowski NA, Subhas N, Schoenhagen P. Testing for interchangeability of imaging tests. Acad Radiol 2014 Nov; 21(11): 1483-9.
9. Pepe MS, Janes H,Li CI, Bossuyt PM, Feng Z, Hilden J, Early-phase studies of biomarkers: what target sensitivity and specificity values might confer clinical utility? Clin Chem. 2016; 62: 737-742.
10. Spiegelhalter D., Thomas A., Best N., and Gilks W. BUGS 0.5: Bayesian inference Using Gibbs Sampling–Manual (version ii). Medical Research Council Biostatistics Unit, Cambridge, 1996.

IV. Metrics

IV.a. Regulatory Science Metrics

Select the metrics that can be used to measure the success and impact of this proposal. For information on how to complete this section, please visit the CDRH Metrics SharePoint.

Level 4: Research Imperatives (max 3)	Level 3: Internal Outputs (max 3)	Level 2: External Impacts (max 2)
Public Health Need	Emerging Methods Research	CDRH Tools in Use by Industry
Leading Edge Science	Updated and Streamlined Review Process	Accelerated Review Timeline
Effective Dissemination	Communications Impact	

IV.b. Outcomes for Evaluating Impact

Using the selected metrics, please describe how this project will change CDRH processes and impact the public health. Provide details that demonstrate how the research imperative metrics will be met, specifically addressing each metric chosen. Then provide plans to produce internal outputs and accomplish external impacts. *(2500 characters including spaces)*

Level 4: ◘ Public Health Need: Novel study designs and statistical methods for analysis are desperately needed to ensure that screening and other rare disease tests can be evaluated in moderately sized studies to decrease burden on sponsors of diagnostic tests and increase the likelihood that safe and effective tests are marketed first in the US. ◘ Leading Edge Science: Develop new methodologies that align with FDA priorities. ◘ Dissemination/Collaboration: Scientific publications in peer-reviewed journals and presentations at scientific conferences. Training of CDRH review staff on new methodologies for evaluating diagnostic tests for low prevalence conditions.

Level 3: ◘ Emerging Methods Research: Impact of methodologies on evaluating diagnostic tests for low prevalence conditions. ◘ Updated and Streamlined Review Process: Guidance for FDA staff and industry that cites our research. Number of reviews or submissions in which the new methodologies are recommended and implemented. ◘ Communications Impact: Number of citations to our publications and presentations. Adoption of similar ideas and approaches in parallel efforts in industry and FDA workgroups.

Level 2: ◘ CDRH Tools Use by Industry: External use of the developed publicly-available software tools. ◘ Accelerated Review Timeline. The newly developed study designs that enrich for disease should reduce study duration and burden. The clarification of study interpretation by utilization of appropriate methods for analysis may reduce FDA review time.

V. Collaborator Information

Full Name (Last, First)	**Role**	**Affiliation**
Luna Zaritsky	Collaborator/Working Group	CDRH
Wei Wang	Collaborator/Working Group	CDRH
Qin Li	Collaborator/Working Group	CDRH
Dandan Xu	Collaborator/Working Group	CDRH
Lei Nie	Collaborator/Working Group	FDA, other Center
Robert Abugov	Collaborator/Working Group	FDA, other Center
Li Cheung	Collaborator/Working Group	Other Government
	Choose an item.	Choose an item.
	Choose an item.	Choose an item.
	Choose an item.	Choose an item.
	Choose an item.	Choose an item.
	Choose an item.	Choose an item.

For each collaborator listed in the table above, include a letter of support. For projects with multiple collaborators, a single letter with signatures from all collaborators may be submitted. You may also include emails from collaborators confirming their intent to participate in the project. The letter of support should include the following information:

- **CDRH Collaborators:** Name, Office/Division, Supervisor, FTE % Time Commitment, and Contribution (describe the expertise of this collaborator and how this contributes to the proposal. *No more than 500 characters including spaces per collaborator*)
- **Non-CDRH Collaborators:** Name, Organization, Collaborative Mechanism*, and Contribution (describe the expertise of this collaborator and how this contributes to the proposal. *No more than 500 characters including spaces per collaborator*)

**Collaborative mechanisms include: (1) Material Transfer Agreement, (2) Research Collaboration Agreement, (3) CRADA, (4) Grant, (5) Inter-Agency Agreement, (6) Memorandum of Understanding, (7) Informal Collaboration, and (8) To be determined.*

VI. Technical Reviewers

You must provide three (3) FDA employees who could be technical reviewers for this project proposal. These reviewers must be familiar with the regulatory science need/area but not involved with this specific proposal.

Full Name (Last, First) **and Title**	**Center/Office**	**E-mail**
Berkman Sahiner	CDRH/OSEL/DIDSR	Berkman.sahiner@fda.hhs.gov
Weijie Chen	CDRH/OSEL/DIDSR	Weijie.chen@fda.hhs.gov
Bipasa Biswas	CDRH/OCEA	Bipasa.Biswas@fda.hhs.govfs

VII. Appendix

Please include any additional information to be shared with the reviewers and leadership. A maximum of 5 documents may be uploaded to this proposal, but it is not guaranteed that all files will be reviewed by reviewers. This can include, but is not limited to CVs/Resumes, and recent publications.

File Name	**Description**

VIII. Submission Instructions

Submission checklist:

☒ All the sections in this proposal are within the permitted character limit.

☒ The table of contents has been updated.

Please email the following documents to CDRHRSProposals@fda.hhs.gov.

- CDRH Critical Path Application Form (MS Word format only)
  - File Naming Convention: LAST_FIRST_Office_FY21CP_Proposal_Title
- Letter of Support from collaborators
  - File Naming Convention: LAST_FIRST_Office_FY21CP_LOS_From
- Additional documents listed under Section VII
  - File Naming Convention: LAST_FIRST_Office_FY21CP_Other_Description from Appendix