Predicting AI model behavior on unrepresented subgroups: A test-time approach to increase variability in a finite test set

Alexis Burgon, Nicholas Petrick, Berkman Sahiner, Gene Pennello, *Ravi K. Samala

CDRH Office of Science and Engineering Laboratories
Artificial Intelligence and Machine Learning Program

Abstract

Background
Evaluation of medical devices in a premarket setting includes performance assessment on a finite test data set that is expected to be representative of the real-world intended use population. However, limited availability of large, diverse, datasets often results in test data that lacks representation of some subgroups limiting the reliability of the estimated generalizability/uncertainty of the artificial intelligence (AI) enabled medical devices.

Project goals
Develop methods that provide an enhanced estimation of model generalizability in circumstances where additional data are not readily available.

Conclusion
Decision region composition analysis using virtual samples from vicinal distributions can detect issues in model generalizability that are not apparent from traditional generalizability estimation based on only the original finite test data set.

Introduction

Represented subgroups have sufficient samples to provide a reasonable estimate of generalizability

Unrepresented subgroups do not have a sufficient number of samples to estimate generalizability or are not present in the finite test set

Vicinal distribution consists of virtual samples generated from the represented subgroups by introducing variabilities that simulate generalizability in the unrepresented subgroups

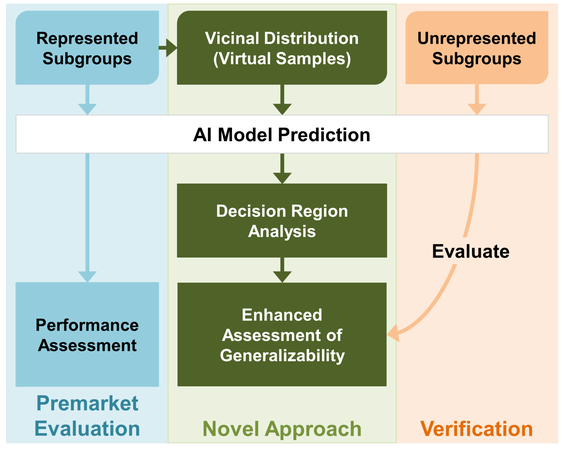

Figure 1. Experimental overview. We present a novel approach for assessing model generalizability by using a vicinal distribution of virtual samples to examine the composition of the regions of the decision space surrounding the available finite test data.

Results

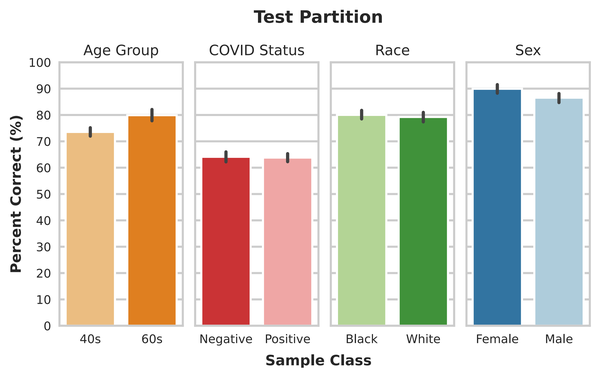

Figure 2. AI model performance on the finite test data set. "Percent correct" refers to the percent of the samples classified as belonging to the correct patient group. For example, the rightmost sub-plot shows the classification of patient sex based on chest x-ray images. The model displays comparable performance between classes for each binary classification task.

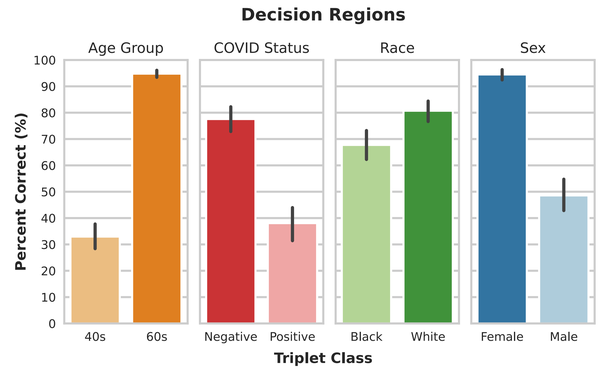

Figure 3. Model classification results for vicinal distributions. "Percent correct" refers to the percent of virtual samples classified as the class of the sample triplet. For each task, the model shows a tendency to overpredict one "*preferred*" class.

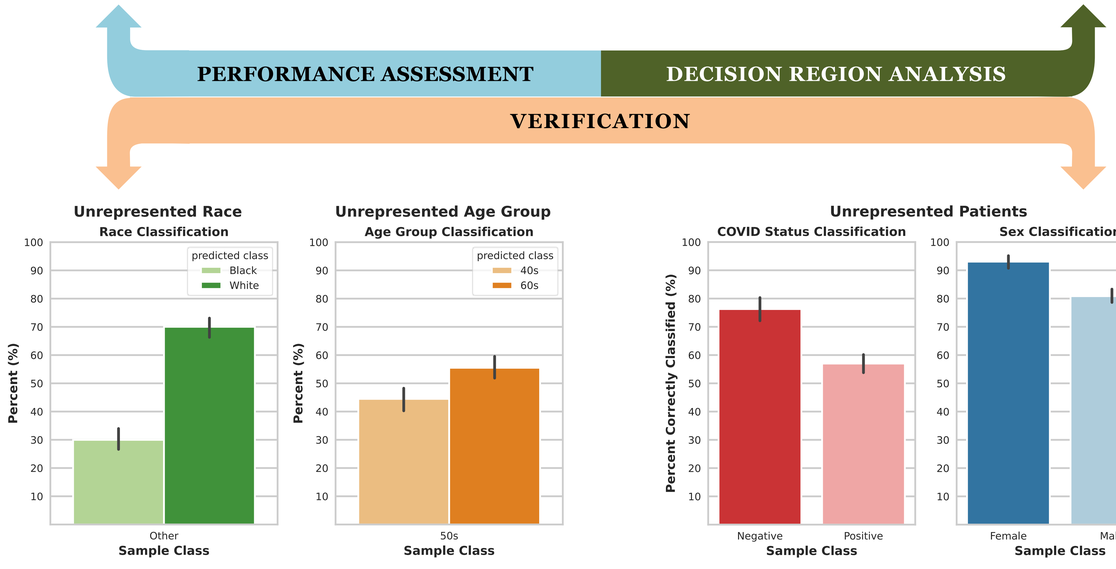

Figure 4. Cross reactivity verification. The left plot shows the AI race classification of patients belonging to 3 unrepresented "Other" races. The right plot shows the AI age classification of patients in the unrepresented age group of 50-59 years. When the true class is not an option, the sample is more likely to be classified as the AI preferred class. White and 60s are the AI preferred classes, as Figure 3 indicates.

Figure 5. Population shift verification. Patient sex and COVID AI classification performance for patients whose race and/or age groups were not represented in the training data. For each task, model shows better generalizability on the AI preferred class than on the non-preferred class. COVID negative and Female are the AI preferred classes, as Figure 3 indicates.

Materials and Methods

Hypothesis
Vicinal distribution analysis of a limited test set allows for the characterization of the regions of the decision space surrounding the available data, which provides insight into how the model will generalize to samples beyond those in the limited test set.

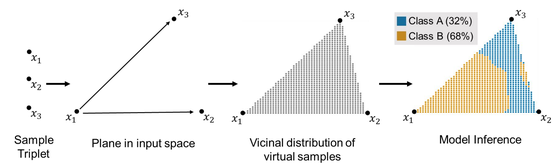

Figure 6. Vicinal distribution generation for decision region composition analysis. Using a "triplet" of three samples from the test partition that belong to the same subgroup, a vicinal distribution of virtual samples is generated by linearly interpolating the triplet along the plane of the input space spanned by the triplet samples.

Experimental Setup
- Patients divided into subgroups based on (1) Sex, (2) Race, (3) COVID status and (4) Age group
- Represented subgroups used for model development and evaluation, unrepresented subgroups used for verification.

Verification Methods: evaluation of unrepresented subgroups
- **Cross-reactivity:** sample is unrepresented with respect to the current task. *Example: Race classification of a patient whose race was unrepresented.*
- **Population shift:** sample is represented with respect to the current task. *Example: Sex classification of a patient whose sex was represented, but whose race was unrepresented.*

Discussion and Conclusion

Decision region composition analysis provides additional information about model generalizability, allowing for the use of a limited dataset to determine how the model is likely to perform when presented with data that is not represented in its training and test datasets.

Demonstration of this approach on the non-clinical task of classification of patient subgroup from chest x-rays reveals that even in cases of comparable performance on the finite test partition:
- A tendency for the model to overpredict one "preferred" class in the decision regions
- The model is likely to overpredict samples from unrepresented subgroups as belonging to the preferred class

Limitations
- Demonstrated on a non-clinical example task and plan to extend it to clinical tasks

* Ravi.Samala@fda.hhs.gov