Generalized Statistical Testing of Interchangeability in Performance Between an AI Segmentation Device and a Multi-Expert Human Panel Without Requiring a Reference Standard

Tingting Hu*[a], Berkman Sahiner[a], Shuyue Guan[a], Mike Mikailov[a], Kenny Cha[a], Frank Samuelson[a], Nicholas Petrick[a]

[a]U.S. Food and Drug Administration, Silver Spring, Maryland, United States
*Corresponding Author, E-mail: Tingting.Hu@fda.hhs.gov

ABSTRACT

AI-based medical imaging devices increasingly include segmentation capabilities for lesions or organs. Conventional performance evaluations rely on comparing device-generated segmentations to an aggregated reference standard annotation using similarity metrics such as the Dice Similarity Coefficient (DSC) or Hausdorff Distance (HD). However, these approaches are limited by the lack of a definitive gold standard annotation and difficulty in defining meaningful success criteria. To address these limitations, we propose a generalizable statistical testing framework that assesses agreement between an AI segmentation device and multiple expert human readers without requiring annotation aggregation. The method compares dissimilarities between the device and each human reader to those observed among the human panel itself. It is compatible with various segmentation similarity metrics such as Dice coefficient (DSC) and the Hausdorff Distance (HD) as demonstrated in our work. Performance was validated through simulation studies involving 2D image-based contours, where a set of ground truth segmentations were transformed using the Medical Image Segmentation Synthesis (MISS) tool, under scenarios where transformation patterns were either shared (transformation-agreeable) or unshared (transformation-disagreeable) between the device and human experts. High-performance computing strategies were used to efficiently scale simulations across a broad range of conditions. Our results show the method effectively controlled Type I error (~0.05) in agreement cases and achieved low Type II error in most disagreement scenarios using either DSC or HD similarity metrics. This method may provide a flexible and practical solution for evaluating segmentation agreement between an image segmentation model and a multi-expert panel without requiring a reference standard.
Keywords: imaging segmentation, interchangeability assessment, AI/ML algorithm, multi-expert human panel, paired testing.

1. INTRODUCTION

AI-based medical imaging devices increasingly include lesion or organ segmentation functionalities. Conventional evaluation approaches typically compare device-generated segmentations to an aggregated reference standard using overlap-based accuracy metrics such as the Dice Similarity Coefficient (DSC [1]) or Hausdorff Distance (HD [2]). However, these methods face challenges due to the absence of a definitive gold standard annotation and ambiguity in defining clinically meaningful success criteria.

To address this, we developed a statistical testing framework [3] that evaluates agreement in overlap-based segmentation performance between an AI device and multiple expert human readers. In the present work, we extend this framework to accommodate distance-based segmentation similarity measures. The key innovation of the proposed approach is that it does not rely on a reference standard and is readily adaptable to a broad class of segmentation similarity metrics.

2. MATERIALS AND METHODS

We present a *generalizable*, paired statistical testing method that assesses whether the segmentation performance of an AI device is significantly different from that of individual human experts. Rather than relying on a reference standard, the method compares the dissimilarity between the device and each human reader to the dissimilarity observed within the human panel. The framework is generalizable to other segmentation similarity metrics.

Our proposed method generalizes a metric proposed by Obuchowski et al. [4], namely the individual equivalence index, that measures the individual equivalence of imaging tests when the health outcome of interest is a numeric variable. We define a segmentation interchangeability metric by modifying the individual equivalence index and derive the point estimate for this new metric. Mathematically, the proposed metric is defined as follows:

$$\delta = E\{SDM(device, reader\ panel)\} - E\{SDM(within\ reader\ panel)\} \qquad (1)$$

where $E\{\cdot\}$ denotes the expected value and SDM refers to a user-specified segmentation *dissimilarity* metric. A smaller value of δ indicates that the device's segmentation performance is closer to that of the human reader panel. In this study, SDM is instantiated using the Dice Similarity Coefficient (DSC) and the Hausdorff Distance (HD), which are representative overlap-based and distance-based segmentation metrics, respectively. Because DSC is a similarity measure, we define SDM=1−DSC when applying the framework to DSC. In contrast, HD is inherently a dissimilarity metric and therefore requires no transformation.

The point estimator for the proposed interchangeability metric (1) can be easily derived as below.

$$\hat{\delta} = \frac{1}{nk}\sum_{j=1}^{n}\sum_{i=1}^{k}\hat{\delta}_i(j) = \frac{1}{nk}\sum_{j=1}^{n}\sum_{i=1}^{k}\{\widehat{SDM}_{Di}(j) - \frac{1}{k-1}\sum_{\substack{i'=1 \\ i'\neq i}}^{k}\widehat{SDM}_{ii'}(j)\} \qquad (2)$$

where i indexes the i^{th} individual reader, j indexes the j^{th} image, the subscript D denotes the segmentation software device, n is the total number of images in the dataset, k is the total number of readers in the panel, and $\hat{\delta}_i(j)$ is the mean difference between device-reader SDM and the reader i-specific reader-reader SDM on image j.

When user-selected SDM is a symmetric measurement (such as DSC, and HD defined by [2]), the point estimate can also alternatively be expressed as:

$$\hat{\delta} = \frac{1}{n}\sum_{j=1}^{n}\hat{\delta}(j) = \frac{1}{n}\sum_{j=1}^{n}\{\frac{1}{k}\sum_{i=1}^{k}[1-\widehat{SDM}_{Di}(j)] - \frac{2}{k(k-1)}\sum_{i=1}^{k}\sum_{i'=i+1}^{k}[1-\widehat{SDM}_{ii'}(j)]\} \qquad (3)$$

where $\hat{\delta}(j)$ is the mean of $\hat{\delta}_i(j)$ across all readers for image j.

The confidence interval of the proposed metric can be obtained via the bootstrap method, following the work by [4].

The performance of our method is validated through simulation studies involving 2D image-based contours. For contour simulation, we used the Medical Image Segmentation Synthesis (MISS) Tool for Guan et al. [5] that has a set of adjustable parameters including affine transformations, Fourier transformations, and spike transformations. Using the MISS Tool, we generate Dice scores and Hausdorff distances through the following process:

(1) Simulate a set of truth contours. For simplicity, we consider images with each comprising a single Region of Interest (ROI) only in this work.
(2) Generate reader and device annotation variations using segmentation transformations based on the MISS tool starting with the true contours.
(3) Calculate the Dice scores and Hausdorff distances from the simulated contours.

The segmentation transformations are either shared *(transformation-agreeable*) or not shared (*transformation-disagreeable*) between the device and the human experts. We evaluated the type I error of our method (the probability of incorrectly rejecting the null hypothesis that the device and human readers are similar) under the transformation-agreeable scenario and the type II error (the probability of failing to reject the null hypothesis that the device and human readers are similar) under transformation-disagreeable scenario.

In the contour simulation step, we used four affine transformation parameters as tunable parameters, and seven parameters (including all Fourier and spike transformation parameters) as "default" parameters. The tunable affine parameters were resizing ratio of width/height (R_x, R_y) and location shift in x and y coordinates (S_x, S_y). Advanced High-Performance Computing scaling techniques were developed to expedite these simulations. Figure 1 illustrates

the effects of the four affine transformation parameters, Rx, Ry, Sx, and Sy, implemented in the MISS tool. The left panel demonstrates resize transformations, where Rx controls width scaling and Ry controls height scaling relative to the original contour. The right panel shows shift transformations, where Sx represents horizontal displacement and Sy represents vertical displacement from the original position.

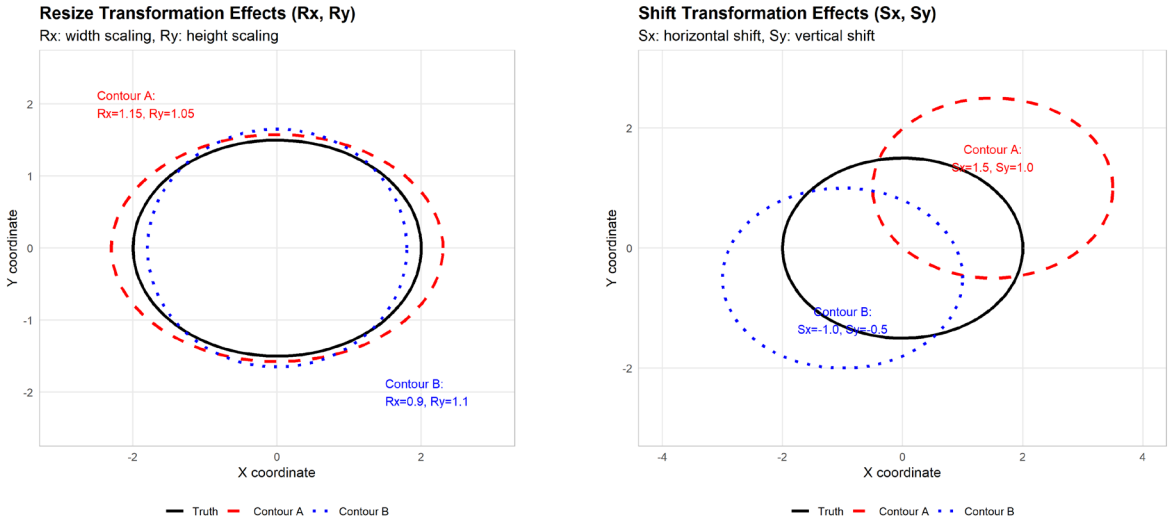

Figure 1. Effects of affine transformation parameters in the Medical Image Segmentation Synthesis (MISS) tool. (Left) Resize transformation effects showing width scaling (Rx) and height scaling (Ry). (Right) Shift transformation effects showing horizontal (Sx) and vertical (Sy) displacement. The black solid line represents the original truth contour, while the red dashed line (Contour A) and blue dotted line (Contour B) show transformed variations used to simulate reader and device segmentation differences.

3. RESULTS

The image-based simulations show that Type I errors are close to 0.05 for transformation agreeable scenarios and Type II errors are close to 0 in terms of DSC and/or HD, most of the time, for transformation disagreeable scenarios across a range of simulation settings with the number of human readers varying between 2 and 9 and the number of images varying between 100 and 500. The results summarized in Figure 2 show that the empirical Type I error rate remains close to the nominal level of 0.05 for the majority of configurations examined in simulated transformation-agreeable scenarios.

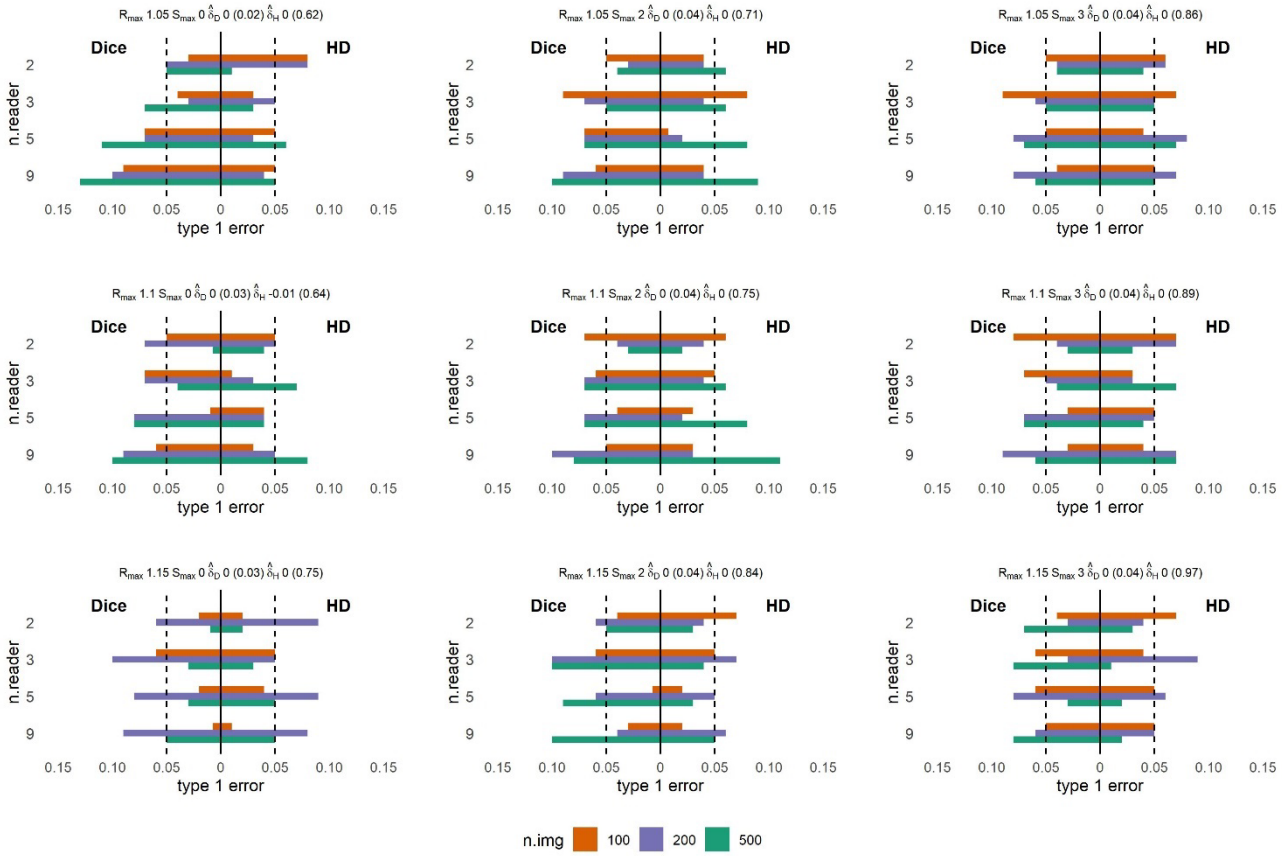

Figure 2. Type I error rates for our proposed method under various experimental conditions, including transformation pattern, number of readers, sample size, and similarity measure (Dice Similarity Coefficient [DSC] vs. Hausdorff Distance [HD]). The black dashed line marks the 0.05 Type I error threshold. In each subplot, the **left bars** show results for DSC (overlap-based), and the **right bars** show results for HD (distance-based). The **orange**, **purple**, and **green** bars represent sample sizes of 100, 200, and 500, respectively. R_{max} and S_{max} are related to the width of the range of the image transformation parameters Rx, Ry, Sx, and Sy; the larger R_{max} (or S_{max}), the larger the range for Rx, Ry (or Sx, Sy). $\hat{\delta}_D$ is the mean (standard deviation) of the difference between the mean DSC (device vs. panel) and the mean DSC (within panel). Similarly, $\hat{\delta}_H$ is the mean (standard deviation) of the difference between the mean HD (device vs. panel) and the mean HD (within panel). From the figure, Type I errors are reasonably close to 0.05 in most scenarios.

Figure 3 illustrates that, under transformation-disagreeable conditions, when holding other parameters fixed, the Type II error decreases as either the sample size or the number of experts in the panel increases. This pattern suggests that enlarging the dataset and expanding the expert panel—assuming comparable segmentation accuracy across experts—leads to improved statistical power for detecting differences in segmentation performance.

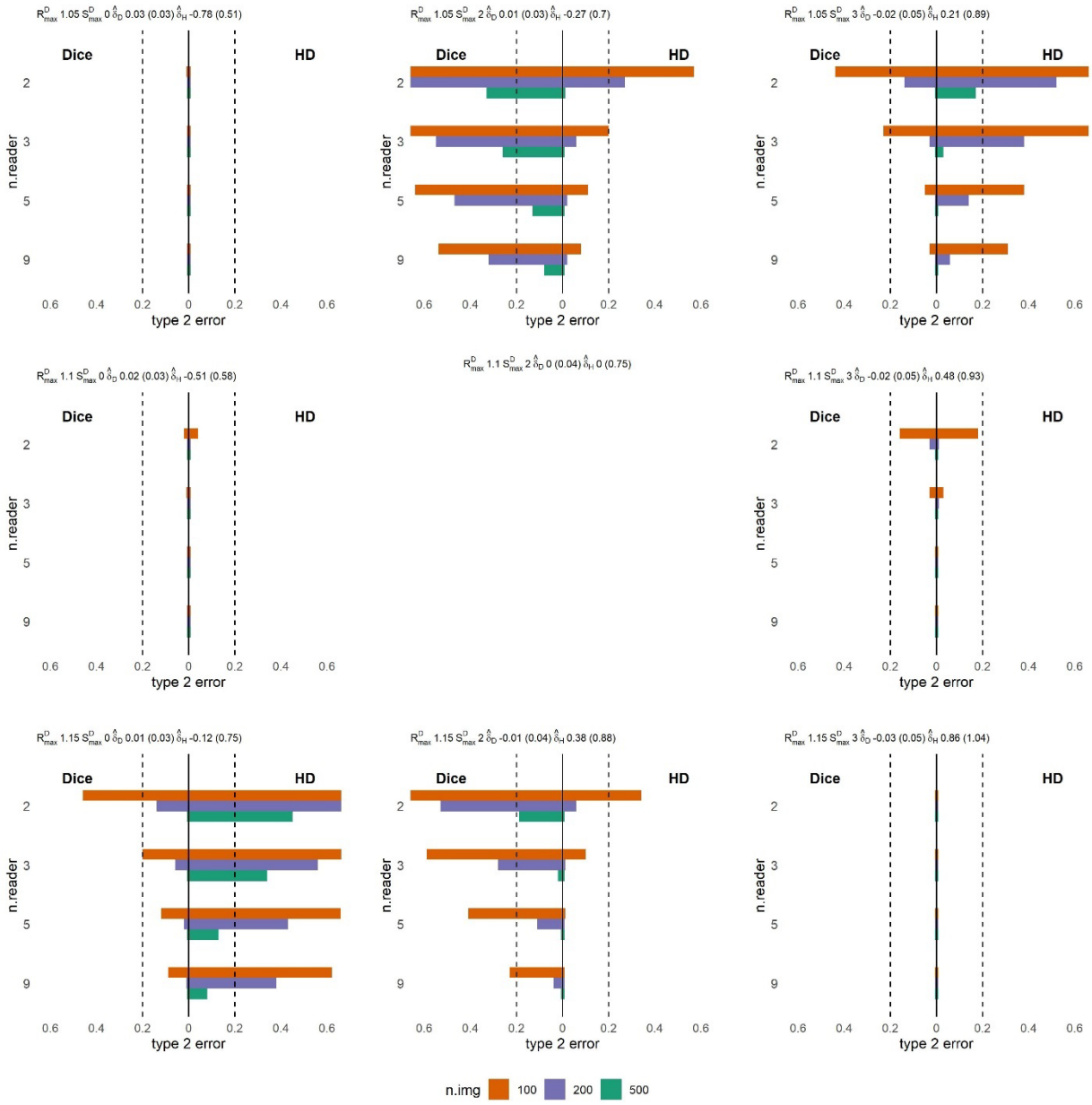

Figure 3. Type II error rates **of** the proposed method under various experimental conditions when reader panel's maximum resize and shift range are fixed at (R_{max}=1.1, S_{max}=2), including transformation pattern, number of readers, sample size, and similarity measure (Dice Similarity Coefficient [DSC] vs. Hausdorff Distance [HD]). The black dashed line marks the 0.20 Type II error threshold. In each subplot, the **left bars** show results for DSC (overlap-based), and the **right bars** show results for HD (distance-based). The **orange**, **purple**, and **green** bars represent sample sizes of 100, 200, and 500, respectively. $R_{max}$and $S_{max}$are related to the width of the range of the image transformation parameters Rx, Ry, Sx, and Sy; the larger R_{max} (or S_{max}), the larger is the range for Rx, Ry (or Sx, Sy). The superscript D in R^D_{max} $(or\ S^D_{max})$ stands for the device, as opposed to the reader panel. $\hat{\delta}_D$ is the mean (standard deviation) of the difference between the mean DSC (device vs. panel) and the mean DSC (within panel). Similarly, $\hat{\delta}_H$ is the mean (standard deviation) of the difference between the mean HD (device vs. panel) and the mean HD (within panel). As expected**,** Type II error generally decreases with increasing reader count or sample size. The scenario in the middle ($R^D_{max} = 1.1\ and\ S^D_{max} = 2$) corresponds to a transformation-agreeable scenario and therefore type-II error analysis is not applicable.

Overall, the proposed method exhibits consistently low Type II error when the discrepancy in segmentation performance, as quantified by the user-specified metric (DSC or HD), is larger. Higher Type II error is observed primarily in transformation-disagreeable settings where the underlying performance differences are smaller. However, the instances where the Type II error exceeds 0.2 correspond to scenarios with negligible differences in the SDM (e.g., DSC=0.01) that are unlikely to be of practical significance. Even in these more challenging scenarios, increases in sample size or the expert panel size reduce the probability of failing to detect disagreement.

An additional noteworthy finding is that, in some transformation-disagreeable scenarios where one metric yielded relatively high Type II error, the alternative metric demonstrated improved error rate (for example, the second and third subplots in the top row of Figure 3). This observation highlights the complementary nature of DSC and HD in characterizing segmentation similarity. When disagreement is insufficiently captured by one metric, the other metric often provided a clearer signal. Therefore, evaluating both classes of metrics may provide a more comprehensive assessment of segmentation performance.

4. DISCUSSION AND CONCLUSIONS

The proposed paired-testing framework offers a flexible and principled approach for evaluating agreement between an AI-based segmentation system and a human expert panel. Using image simulation experiments, we showed that our proposed statistical framework maintains appropriate control of the Type I error rate and exhibits favorable Type II error properties across a range of simulated study settings. A key contribution of this framework is its capacity to evaluate the practical interchangeability of a segmentation AI system with multiple human readers without relying on an aggregated reference standard. In addition, the method is expected to be extendable to a broad class of segmentation similarity measures.

One limitation of our proposed approach is that reader effects are modeled as fixed rather than random. In the context of multi-reader multi-case (MRMC) studies, the approach of modeling a small group of readers as fixed have been employed by Bandos et al. [6] and also discussed by Hillis and Schartz[7]. It is also important to distinguish the present setting from conventional MRMC studies conducted for computer-aided detection (CAD) system evaluation. In typical CAD studies, readers represent device end users and therefore are expected to span a wide range of experience levels. By contrast, in our application the readers are acting as expert annotators who establish reference segmentations to provide a high-performance benchmark against which the standalone performance of the AI device is evaluated; accordingly, the reader panel in segmentation evaluation are generally assumed to consist of highly experienced experts.

ACKNOWLEDGMENTS

This article reflects the views of the authors and does not represent the views or policy of the U.S. Food and Drug Administration, the Department of Health and Human Services, or the U.S. Government. The mention of commercial products, their sources, or their use in connection with material reported herein is not to be construed as either an actual or implied endorsement of such products by the Department of Health and Human Services.

REFERENCES

[1] Dice, L. R. (1945). *Measures of the amount of ecologic association between species*. Ecology, 26(3), 297-302.
[2] Taha, A. A., & Hanbury, A. (2015). *Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool*. *BMC medical imaging*, *15*(1), 29.

[3] Hu, T., Sahiner, B., Guan, S., Mikailov, M., Cha, K., Samuelson, F., & Petrick, N. (2025). *Statistical testing of agreement in overlap-based performance between an AI segmentation device and a multi-expert human panel without requiring a reference standard*. Journal of Medical Imaging, 12(5), 055003-055003.
[4] Obuchowski, N. A., Subhas, N., & Schoenhagen, P. (2014). *Testing for interchangeability of imaging tests*. Academic Radiology, 21(11), 1483-1489
[5] Guan, S., Samala, R. K., Arab, A., & Chen, W. (2023, April). *MISS-tool: medical image segmentation synthesis tool to emulate segmentation errors*. In Medical Imaging 2023: Computer-Aided Diagnosis (Vol. 12465, pp. 273-281). SPIE
[6] Bandos, A. I., Rockette, H. E., & Gur, D. (2006). *A permutation test for comparing ROC curves in multireader studies: a multi-reader ROC, permutation test*. Academic radiology, 13(4), 414-420.
[7] Hillis, S. L., & Schartz, K. M. (2018). *Multireader sample size program for diagnostic studies: demonstration and methodology*. Journal of Medical Imaging, 5(4), 045503-045503.