Statistical Testing of Overlap-Based 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

Abstract

Purpose: AI-based medical imaging devices often include lesion or organ segmentation capabilities. Existing methods for segmentation performance evaluation compare AI results to an aggregated reference standard using accuracy metrics like the Dice coefficient or Hausdorff Distance. However, these approaches are limited for lacking a gold standard and challenges in defining meaningful success criteria. To address this, we developed a statistical method to assess agreement between an AI device and multiple human experts without requiring a reference standard.

Approach: We propose a paired-testing method evaluating whether an AI device's segmentation performance significantly differs from multiple human experts'. The method compares device-to-expert dissimilarity with expert-to-expert dissimilarity, avoiding the need for a reference standard. We validated the method through: (1) statistical simulations where Dice coefficient performance is either shared ("overlap agreeable") or not shared ("overlap disagreeable") between the device and experts; (2) image-based simulations using 2D contours with shared or non-shared transformation parameters ("transformation agreeable or disagreeable"). We also applied the method to compare an AI segmentation algorithm to four radiologists using data from the Lung Image Database Consortium.

Results: Statistical simulations show the method controls type I error (~0.05) for overlap-agreeable and type II error (~0) for overlap-disagreeable scenarios. Image-based simulations show acceptable performance with mean type I error 0.07 (SD 0.03) for transformation-agreeable and mean type II error 0.07 (SD 0.18) for transformation-disagreeable cases.

Conclusions: The paired-testing method offers a new tool for assessing the agreement between an AI segmentation device and multiple human expert panelists without requiring a reference standard.

Keywords: segmentation assessment, multi-expert human panel, paired testing

***First Author**, E-mail: Tingting.Hu@fda.hhs.gov

1 Introduction

The segmentation of structures and lesions within medical images is an increasing focus of artificial intelligence (AI) based medical imaging devices. Examples include devices used for delineating lesions in disease detection and diagnosis, or for outlining organs in surgical planning or radiation therapy. The evaluation of segmentation algorithm performance is critical in the medical device domain to ensure patient safety and benefit, however, there is a relatively small amount of literature addressing the evaluation of segmentation algorithm performance even though there is a significant amount of research focused on AI segmentation development. Furthermore, existing literature on segmentation performance assessment primarily focuses on reviewing and comparing performance evaluation metrics (e.g., [1, 2]) or proposing new evaluation measures (e.g., [3, 4]). However, these proposed metrics and measures cannot be directly used for evaluating segmentation performance in the absence of a reference standard establishing the ground truth for segmentation task assessment. In many medical imaging tasks in the literature, reference contours for each object to be segmented are often defined by multiple human experts ('expert' may be referred to in the literature or submissions by various terms, such as observer, reader, reviewer, or truther). While using only one expert's reference contour to evaluate device performance simplifies the analysis, this approach fails to reflect the truth variability that exists even among high-level experts. Therefore, including multiple experts' reference contours better reflects the true nature of the problem, though this introduces complexity in evaluation and analysis. This complexity necessitates robust statistical methods to appropriately handle comparison of device segmentation against multi-expert references.

A commonly used assessment approach in practice is, with a reference standard contour defined, the AI segmentation output is then compared, through an overlap metric, e.g., Dice [5], Jaccard [6], or a distance-based metric, e.g., Hausdorff distance [1] with the reference. A limitation of this approach is it requires prespecifying a meaningful cutoff for each metric. However, selecting and justifying a clinically meaningful performance criterion can be difficult.

These two challenges with current segmentation assessment methods — the absence of a definitive reference standard contour and the difficulty of establishing a clinically meaningful performance goal — led us to explore a new statistical method that assesses interchangeability, without requiring a reference standard, between an AI segmentation and segmentations from a human expert panel. With the growing number of AI/ML segmentation devices, seeking for a more generalizable segmentation assessment approach has become a pressing need in the medical technology field.

The only relevant work we found is by Zou et al.[7], which addresses a similar problem of evaluating a segmentation algorithm against multiple truthers (Example 2 in their paper). However, their method aggregates the three human annotations into a single STAPLE-based reference, then compares the algorithm to this composite. This would still require defining an arbitrary success cutoff and leave the fundamental challenges unresolved. To address this gap, we examined methodologies for assessing continuous estimation tasks (i.e., estimation tasks for which the target quantity has a continuous value, such as area or volume measurements for organs or lesions), where agreement measures and methods have been more widely discussed. These include the Bland-Altman method [8], individual bioequivalence [9], the individual equivalence index [10], [11], the agreement index [12], and the individual equivalence coefficient and coefficient of individual agreement [13].

To address segmentation agreement, we adapted the interchangeability method proposed by Obuchowski et al. [11] for numerical estimation to the segmentation context. The basic idea is to condense reader-reader pairwise segmentation comparisons into a within-panel agreement score for each image and, likewise, generate a corresponding device-panel agreement score. This reformulates the segmentation comparison as a paired numerical score comparison. By tailoring Obuchowski et al.'s method, we constructed a paired test statistic and apply the resulting confidence interval to determine whether device-panel agreement significantly differs from within-panel agreement.

In the remainder of the paper, we define the problem mathematically and present the proposed methodology. We next present two simulation studies: one statistic-based and the other image-based, and report the Type I and Type II errors for our proposed method. Finally, we conclude with a discussion of the findings and potential directions for future research.

2 Methodology

This section outlines the problem definition and proposed methodology. For illustration simplicity, we focus on a single object to segment throughout this article. But the proposed method can be applied to multi-object scenarios in a similar way as defined here.

2.1 Problem Formulation

Consider a testing dataset containing n images, where each image is obtained from an independent patient and contains a single object to segment. On this dataset, an AI device, denoted as D, segments the object of interest within each image. Concurrently, a human expert panel denoted as P, comprised of k experts, each independently performs manual annotation of the object. Each segmentation, whether by the device or an expert, is represented as a binary image.

The problem of interest is testing whether the segmentation performance of device *D* significantly diverges from that of the human expert panel. Symbolically, the aim is to test whether μ_D - μ_P significantly differs from zero, where μ_D represents the mean (dis)similarity score between the device and panel across images, and μ_P denotes the mean (dis)similarity score between experts within the panel across images, i.e., testing the null hypothesis

$$H_0: \mu_D - \mu_P = 0,$$

against the alternative hypothesis

$$Ha: \mu_D - \mu_P \neq 0.$$

Here, we adopt an equality-based null hypothesis formulation for simplicity, following the fashion of hypothesis formulation in the FDA guidance [14] where the null hypothesis is set as no treatment effect on the selected endpoint. This choice ensures alignment with the conventional definitions of Type I and Type II errors, which will be utilized for method validation later in Section 3.

2.2 Related Work for Numeric Output

Obuchowski et al. [11] proposed a metric called individual equivalence index to measure the individual equivalence of imaging tests when the health outcome of interest is a numeric variable. This metric is defined as below:

$$\gamma = E\left(Y_{jTik} - Y_{jRik'}\right)^2 - E\left(Y_{jRik} - Y_{jRik'}\right)^2 \qquad (1)$$

where Y_{jTik} denotes the result or measurement by the new test modality (T) by reader i for subject j on occasion k, and Y_{jRik} denotes the result or measurement by the existing reference modality (R) by reader i for subject j on occasion k.

While the Obuchowski et al. [11] method was developed for an estimation task with a numeric output, we adapt it to quantify segmentation agreement through suitable modifications.

2.3 Proposed Method

2.3.1 Segmentation Interchangeability Metric

From Eq. (1), it is evident that the central concept behind the interchangeability metric is to compare the dissimilarity between the new test and reference test with the dissimilarity within the reference test. Building on this idea, a natural extension of this approach to segmentation outputs is replacing the dissimilarity metric for numeric outputs adopted by Obuchowski et al. [11] (i.e., the mean squared difference) with an appropriate segmentation dissimilarity metric. One of the most widely used similarity measures for segmentation is the Dice Similarity Coefficient (DSC) [5]. We therefore use 1-DSC as the dissimilarity surrogate to tailor the original individual equivalence index to segmentation. Based on this modification, we propose a segmentation interchangeability metric denoted by δ, to evaluate segmentation agreement between an AI device and a panel of human readers. The proposed metric is defined as follows:

$$\delta = E\{1 - DSC(device, reader\ panel)\} - E\{1 - DSC(within\ reader\ panel)\} \qquad (2)$$

where E denotes the expected value. Clearly, the closer δ is to zero, the more similar the device's segmentation performance is to the human reader panel. The Dice coefficient for paired segmentations on a single image is always positive and ranges between zero and one. These properties make this segmentation interchangeability metric (δ) well-suited for evaluating segmentation performance at the individual image level.

2.3.2 Point Estimate

The point estimator for the proposed interchangeability metric (2) can be easily derived as below.

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

It can also alternatively be expressed as:

$$\hat{\delta} = \frac{1}{\mathrm{nk}}\sum_{j=1}^{n}\sum_{i=1}^{k}\hat{\delta}_{\mathrm{i}}(\mathrm{j}) \;= \frac{1}{\mathrm{nk}}\sum_{j=1}^{n}\sum_{i=1}^{k}\{[1 - DSC_{Di}(j)] - \frac{1}{\mathrm{k-1}}\sum_{\substack{i'=1\\ i\prime\neq i}}^{k}[1 - DSC_{ii'}(j)]\} \qquad (4)$$

From the formulae (3) and (4), we can see $\hat{\delta}_{\mathrm{i}}(\mathrm{j})$ is defined as the mean difference between device and the i^th^ individual reader on j^th^ image, $\hat{\delta}(\mathrm{j})$ is the mean $\hat{\delta}_{\mathrm{i}}(\mathrm{j})$ across all readers for j^th^ image, and $\hat{\delta}$ is the mean of $\hat{\delta}(\mathrm{j})$ across all images.

2.3.3 Confidence Interval

Various approaches can be used to construct confidence intervals (CIs) for $\hat{\delta}$. In this study, we used both a parametric and non-parametric method to construct CIs. The parametric method estimates the z-interval, given by $\hat{\delta} \pm Z_{\alpha/2}\,S$, where $S = \sqrt{\frac{1}{n-1}\sum_{j}^{n}\left(\hat{\delta}(\mathrm{j}) - \hat{\delta}\right)^{2}}$ is the sample standard deviation of $\hat{\delta}(\mathrm{j})$, and the non-parametric method is a bootstrap approach that follows the procedure outlined in [11].

A CI covering zero indicates that no significant difference in overlap-based segmentation performance between device and panel. Note, this should not be interpreted as the two are the same but only that a difference could not be established statistically [14]. A CI entirely below zero indicates device-panel segmentation agreement is statistically higher than the within-panel agreement, while a CI above zero indicates the device-panel agreement is significantly lower than the within-panel agreement.

3 Simulation Studies

This section presents the design and results from our simulation studies conducted as part of the method validation process.

3.1 Overall Study Design

3.1.1 Study Overview

In this article, we undertake two primary types of simulation studies to validate our proposed method. Simulation Study 1 is a statistics-based simulation study, which simulates Dice scores from a predefined statistical distribution. This approach defines the 'agreement/interchangeability' between readers and devices by controlling the Dice distribution characteristics for each. This simulation allows for the assessment of our segmentation interchangeability metric across a range of Dice distributions. A limitation of the Simulation Study 1 design is it may not perfectly capture relationships inherent in image-generated Dice scores. The relationship here refers to the correlations between pairwise Dice scores coming from segmentations of the same objects from different annotators. Simulating Dice scores simply by sampling from a statistical distribution may not fully capture these relationships.

To address this limitation, we conducted Simulation Study 2 where we directly compute Dice scores from contour masks rather than simulating Dice scores directly. To achieve this, we employed the Medical Image Segmentation Synthesis (MISS) Tool, developed by Guan et al. [15] to synthesize multiple segmentation masks. The MISS Tool has a set of adjustable parameters simulating six types of segmentation errors. Using the MISS Tool, we generate Dice scores by applying the following process: (1) simulate a set of truth contours, (2) generate reader and device annotation variations using the MISS tool based on the true contours, and (3) calculate the

Dice scores from the simulated contours. Study 2 establishes 'agreement/interchangeability' between readers and devices by either using the same or different transformation parameters within the MISS Tool. To distinguish between the two simulation studies, we refer to (dis)agreement in Study 1 as “overlap (dis)agreeable” and in Study 2 as “transformation (dis)agreeable”.

3.1.2 Performance Metrics

For Study 1, we evaluate the performance of the proposed method using three metrics:

1. **Type I Error**: The probability of incorrectly rejecting the null hypothesis (H_0) when it is actually true.

2. **Type II Error**: The probability of failing to reject the null hypothesis (H_0) when it is actually false.

3. **Coverage Probability (CP)**: The proportion of times a confidence interval contains the true value of the parameter being estimated (in our case, the true value of δ).

A desirable method is expected to have type I error near 0.05 for agreeable cases, type II error below 0.2 for disagreeable cases assuming a statistical power of 0.80, and CP near the confidence level (set as 0.95 in our study).

For Study 2 (image-based simulation), we assess performance using Type I and Type II errors. Coverage Probability (CP) is not applicable as the true value of δ is unknown because the mean Dice performance cannot be directly controlled or defined using the MISS tool.

3.2 Study 1: Statistics-based Simulation

3.2.1 Parameter Configuration

In Simulation Study 1, the number of readers in the panel, denoted as k, are either 2 or 3 in our experiments. We define the following global parameters for each simulation conduction:

- ($\boldsymbol{\mu_0}$, $\boldsymbol{\sigma_0}$): mean and standard deviation for any reader-reader pair DSC.
- ($\boldsymbol{d_\mu}, \boldsymbol{d_\sigma}$), where $d_\mu = \mu_D - \mu_0$, $d_\sigma = \sigma_D - \sigma_0$, with $\mu_D\ and\ \sigma_D$ representing the mean and standard deviation of DSC for any device-reader pair. d_μ and d_σ correspond to the differences in the means and standard deviations for the device-reader DSC and the reader-reader DSC.
- ($\boldsymbol{\rho_0}$, $\boldsymbol{\rho_D}$, $\boldsymbol{\rho_{D0}}$): pairwise reader-reader DSC correlation $\boldsymbol{\rho_0}$, pairwise device-reader DSC correlation $\boldsymbol{\rho_D}$, and the between device-reader and reader-reader DSC correlation $\boldsymbol{\rho_{D0}}$, respectively.

To mirror real-world data variations, where within-panel and device-panel dissimilarities may have different means and standard deviations, we explore four scenarios:

I. equal mean, equal variance (Overlap Agreeable): $\mu_D = \mu_0$, $\sigma_D = \sigma_0$;

II. unequal mean, equal variance (Overlap Disagreeable): $\mu_D \neq \mu_0$, $\sigma_D = \sigma_0$;

III. equal mean, unequal variance (Overlap Agreeable): $\mu_D = \mu_0$, $\sigma_D \neq \sigma_0$;

IV. unequal mean, unequal variance (Overlap Disagreeable): $\mu_D \neq \mu_0$, $\sigma_D \neq \sigma_0$.

Parameter configurations for these scenarios are presented in Table 1.

Table 1: Parameter configuration scenarios for Study 1: statistic-based simulations.

Scenario	Parameter	configuration
I. equal μ, equal σ (overlap agreeable): 32	(μ_0, σ_0)	μ_0 in {0.75, 0.8, 0.85, 0.9} σ_0 in {0.025, 0.05, 0.1, 0.15}
	(d_μ, d_σ)	$d_\mu = 0$, $d_\sigma = 0$

settings total	$(\rho_0, \rho_D, \rho_{D0})$	$\rho_0 = \rho_D = \rho_{D0}$ in {*moderate, strong/very strong*} (correlation criteria is described above in Section 3.2.1)
II. unequal μ, equal σ (overlap disagreeable): 512 settings total	(μ_0, σ_0)	μ_0 in {0.75, 0.8, 0.85, 0.9} σ_0 in {0.025, 0.05, 0.1, 0.15}
	(d_μ, d_σ)	d_μ in {-0.05, -0.1, -0.25, -0.5}, $d_\sigma = 0$
	$(\rho_0, \rho_D, \rho_{D0})$	ρ_0 in {*moderate, strong/very strong*} ρ_D in {*moderate, strong/very strong*} ρ_{D0} in {*weak, very weak*}
III. equal μ, unequal σ (overlap Dice agreeable): 96 settings total	(μ_0, σ_0)	μ_0 in {0.75, 0.8, 0.85, 0.9} σ_0 in {0.025, 0.05, 0.1}
	(d_μ, d_σ)	$d_\mu = 0$, d_σ in {0.02, 0.05, 0.1, 0.15}
	$(\rho_0, \rho_D, \rho_{D0})$	$\rho_0 = \rho_D = \rho_{D0}$ in {*moderate, strong/very strong*}
IV. unequal μ, unequal σ (overlap disagreeable): 648 settings total	(μ_0, σ_0)	μ_0 in {0.8, 0.85, 0.9} σ_0 in {0.025, 0.05, 0.1}
	(d_μ, d_σ)	d_μ in {-0.05, -0.1, -0.25}, d_σ in {0.02, 0.1, 0.2}
	$(\rho_0, \rho_D, \rho_{D0})$	ρ_0 in {*moderate, strong/very strong*} ρ_D in {*moderate, strong/very strong*} ρ_{D0} in {*weak, very weak*}

(μ_0, σ_0) values in Table 1 were determined based on the mean and standard deviation values reported in the existing literature for human-observer-pair Dice scores in tasks such as CT-scan lung nodule/tumor segmentation and liver tumor/brain hematomas segmentation (e.g., [16, 17]). **($d\mu$, $d\sigma$)** values in Table 1 were defined based on our exploration of LIDC-based synthesized data. In practice device performance is typically lower than human experts, thus **$d\mu$** is set to be non-positive in our studies, and **$d\sigma$** is set to be non-negative across all settings. **$(\rho_0, \rho_D, \rho_{D0})$** configurations in Table 1 are based on the following assumptions: 1) Dice scores drawn from distributions with the same mean exhibit *moderate or higher* correlation, whereas Dice scores drawn from distributions with different means demonstrate *weak or lower* correlation (criteria for defining different correlation strength categories are described in Section 3.2.2); and 2) an equi-correlation-category assumption is used to define correlation structure for between annotator-pair-Dice (details see next section 3.2.2).

3.2.2 Dataset Simulation

For each setting configured in Table 1, we simulated a total of $N_{sim} = 1000$ datasets. Each dataset contained k readers. Thus there is a total of k*(k-1)/2 reader-pair DSCs and k device-reader-pair DSCs for each of n images. Dice scores were generated using a multivariate Beta distribution as follows:

Step 1: Compute the Beta shape parameters (a_0, b_0) and (a_D, b_D) corresponding to the pre-specified mean and standard deviation of (μ_0, σ_0) and (μ_D, σ_D) for the reader-reader Dice distribution and the device-reader Dice distribution, respectively.

Step 2: Generate the correlation matrix for pairwise Dices based on the pre-specified correlation strength categories ($\rho_{0,}$ $\rho_{D,}$ ρ_{D0}) using the following: Set the diagonal elements in the correlation matrix to 1. The upper triangle in the matrix was defined by randomly generating correlation coefficients between pairwise reader-reader DSCs ($DSC_{ii'}$ vs. $DSC_{jj'}$, 1≤i<i'≤k, 1≤j<j'≤k, i<j) based on

$$Corr(DSCii', DSCjj') \sim \begin{cases} U(0, 0.2) & if\ \ \rho_0\ = \text{“}very\ weak\text{”} \\ U[0.2, 0.4) & if\ \ \rho_0\ = \text{“}weak\text{”} \\ U[0.4, 0.6) & if\ \ \rho_0\ = \text{“}moderate\text{”} \\ U[0.6, 0.8) & if\ \ \rho_0\ = \text{“}strong\text{”} \\ U[0.8, 1) & if \rho_0 = \text{“}very\ strong\text{”} \end{cases}$$

where U stands for the uniform distribution. The correlation categorization criteria here is taken from the guidelines established by Evans [18], which are consistent with those used by LaMorte [19] and Swinscow and Campbell [20]. Correlation between pairwise device-reader DSCs (ρ_D), and between reader-reader DSCs and device-reader DSCs (ρ_{D0}) were determined in a similar fashion. The correlation matrix, denoted as Σ and having dimensions k(k+1)/2 by k(k+1)/2, is symmetric; therefore, the lower triangle can be filled by mirroring the upper triangle.

Step 3: Generate all annotator-pair DSCs via multivariate beta distribution as follows.

$$\left(DSC_{12}, \dots, DSC_{(k-1)k}; DSC_{D1}, \dots, DSC_{Dk}\right) \sim MultiBeta\{\vec{a}\,, \vec{b}, \Sigma\}$$

where $\vec{a} = \left(\mathrm{a}_0\vec{1}^{(k-1)k/2}, \mathrm{a}_\mathrm{D}\vec{1}^{k}\right)$, $\vec{b} = \left(\mathrm{b}_0\vec{1}^{(k-1)k/2}, \mathrm{b}_\mathrm{D}\vec{1}^{k}\right)$ and $\vec{1}$ refers to an all-ones vector. The beta shape parameters reflect the desired means and standard deviations from Step 1 and Σ is the correlation matrix generated in Step 2. This step is implemented using modified code from [21].

The statistical simulation was run using a non-synchronized parallelization techniques to efficiently scale the large-scale simulations on a High-Performance Computing (HPC) clusters, and the Pierre L'Ecuyer's [22] 'RngStreams' function was used to generate multiple independent pseudo-random number streams ensuring no overlap across array job tasks.

3.2.3 Results

Below are selected results from the statistics-based simulation study. Note that in the rest of the article, n_{reader} and n_{image} are used interchangeably with k and n, respectively, for ease of reference. Performance metrics are computed for each setting based on N_{Sim}=1000, n_{image}=400, and n_{reader}=2 and 3.

Figure 1 presents boxplots of the coverage probabilities (CP) and Type I/II error results for both 2 and 3 readers across all conditions for the four scenarios. Tables 2–5 provide detailed results for the Type I/II error and CP for the individual settings. The tables include results for the 3-reader scenario and include bootstrap CIs for all 32 settings in Scenario I and selected settings from Scenarios II–IV. The 2-reader scenario and z-interval CIs are similar. The selected settings from scenarios II and IV only include the $d_\mu = -0.05$ setting as the type II error would only be smaller for larger absolute differences $d_\mu < -0.05$.

From Figure 1 and Tables 2–5, it is evident that the proposed method performs well. Specifically, Type I error remains close to 0.05 across all settings for overlap-agreeable cases (Scenarios I and III), Type II error approaches 0 in all settings for overlap-disagreeable cases (Scenarios II and IV), and CP consistently hovers around 0.95 across all the scenarios.

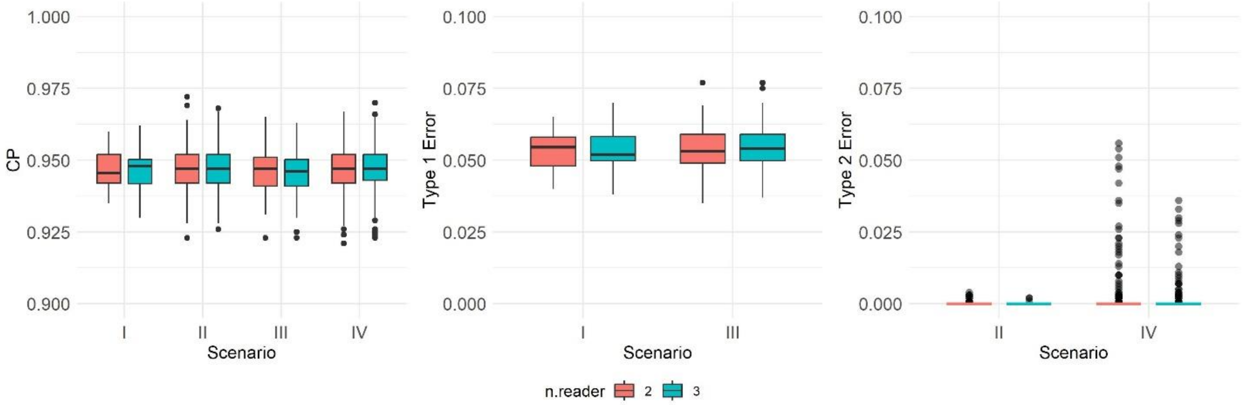


Figure 1. CP, Type I error and Type II error results for the statistics-based Simulation Study 1. The figure includes boxplot results aggregated across all the settings in Scenarios I – IV, and for both the 2-reader and 3-reader scenarios.

Table 2. Results for all 32 settings in Scenario I: Equal Mean μ and Equal Standard Deviation σ fromstatistics-based simulation Study 1.

μ_0	σ_0	$\rho_0 = \rho_D = \rho_{D0}$	CP	Type I Error
0.75	0.025	moderate	0.937	0.063
0.8	0.025	moderate	0.955	0.045
0.85	0.025	moderate	0.947	0.053
0.9	0.025	moderate	0.943	0.057
0.75	0.05	moderate	0.949	0.051
0.8	0.05	moderate	0.93	0.07
0.85	0.05	moderate	0.951	0.049
0.9	0.05	moderate	0.939	0.061
0.75	0.1	moderate	0.942	0.058
0.8	0.1	moderate	0.948	0.052
0.85	0.1	moderate	0.949	0.051
0.9	0.1	moderate	0.953	0.047
0.75	0.15	moderate	0.95	0.05
0.8	0.15	moderate	0.946	0.054

0.85	0.15	moderate	0.962	0.038
0.9	0.15	moderate	0.948	0.052
0.75	0.025	strong/very strong	0.957	0.043
0.8	0.025	strong/very strong	0.961	0.039
0.85	0.025	strong/very strong	0.95	0.05
0.9	0.025	strong/very strong	0.936	0.064
0.75	0.05	strong/very strong	0.957	0.043
0.8	0.05	strong/very strong	0.942	0.058
0.85	0.05	strong/very strong	0.941	0.059
0.9	0.05	strong/very strong	0.935	0.065
0.75	0.1	strong/very strong	0.948	0.052
0.8	0.1	strong/very strong	0.942	0.058
0.85	0.1	strong/very strong	0.932	0.068
0.9	0.1	strong/very strong	0.948	0.052
0.75	0.15	strong/very strong	0.945	0.055
0.8	0.15	strong/very strong	0.941	0.059
0.85	0.15	strong/very strong	0.961	0.039
0.9	0.15	strong/very strong	0.95	0.05

Table 3. Selected results for 8 Settings with $\left(\mu_0, d_\mu, \sigma_0\right) = (0.85, -0.05, 0.15)$ in Scenario II: Unequal Mean μ and Equal Standard Deviation σ from statistics-based Simulation Study 1.

μ_0	d_μ	σ_0	ρ_0	ρ_D	ρ_{D0}	Type II error	CP
0.85	-0.05	0.15	moderate	moderate	very weak	0	0.955
0.85	-0.05	0.15	moderate	moderate	weak	0	0.939
0.85	-0.05	0.15	moderate	strong/very strong	very weak	0.002	0.94
0.85	-0.05	0.15	moderate	strong/very strong	weak	0	0.944
0.85	-0.05	0.15	strong/very strong	moderate	very weak	0	0.946
0.85	-0.05	0.15	strong/very strong	moderate	weak	0	0.94
0.85	-0.05	0.15	strong/very strong	strong/very strong	very weak	0.001	0.942
0.85	-0.05	0.15	strong/very strong	strong/very strong	weak	0	0.941

Table 4. Selected results for 24 of 96 Settings with $\mu_0 = 0.75$ in Scenario III: Equal Mean μ and Unequal Standard Deviation σ from statistics-based Simulation Study 1.

μ_0	σ_0	d_σ	ρ_0= ρ_D= ρ_{D0}	Type I error	CP
0.75	0.025	0.15	strong or very strong	0.054	0.946
0.75	0.1	0.02	moderate	0.048	0.952
0.75	0.05	0.02	moderate	0.058	0.942
0.75	0.025	0.02	moderate	0.05	0.95
0.75	0.1	0.05	moderate	0.049	0.951
0.75	0.05	0.05	moderate	0.054	0.946
0.75	0.025	0.05	moderate	0.048	0.952
0.75	0.1	0.1	moderate	0.067	0.933
0.75	0.05	0.1	moderate	0.056	0.944
0.75	0.025	0.1	moderate	0.061	0.939
0.75	0.1	0.15	moderate	0.055	0.945
0.75	0.05	0.15	moderate	0.058	0.942
0.75	0.025	0.15	moderate	0.054	0.946
0.75	0.1	0.02	strong or very strong	0.066	0.934
0.75	0.05	0.02	strong or very strong	0.048	0.952
0.75	0.025	0.02	strong or very strong	0.052	0.948
0.75	0.1	0.05	strong or very strong	0.051	0.949
0.75	0.05	0.05	strong or very strong	0.05	0.95
0.75	0.025	0.05	strong or very strong	0.055	0.945
0.75	0.1	0.1	strong or very strong	0.042	0.958
0.75	0.05	0.1	strong or very strong	0.055	0.945
0.75	0.025	0.1	strong or very strong	0.059	0.941
0.75	0.1	0.15	strong or very strong	0.046	0.954
0.75	0.05	0.15	strong or very strong	0.061	0.939

Table 5. Selected results for 24 of 648 Settings with $(\mu_0, d_\mu, \sigma_0) = (0.8, -0.05, 0.1)$ in Scenario IV: Unequal Mean μ and Unequal Standard Deviation σ from statistics-based Simulation Study 1.

μ_0	d_μ	σ_0	d_σ	ρ_0	ρ_D	ρ_{D0}	Type II error	CP
0.8	-0.05	0.1	0.2	moderate	moderate	very weak	0.007	0.956
0.8	-0.05	0.1	0.2	moderate	moderate	weak	0.002	0.956
0.8	-0.05	0.1	0.2	moderate	strong/very strong	very weak	0.033	0.952
0.8	-0.05	0.1	0.2	moderate	strong/very strong	weak	0.02	0.932
0.8	-0.05	0.1	0.2	strong/very strong	moderate	very weak	0.009	0.952
0.8	-0.05	0.1	0.2	strong/very strong	moderate	weak	0.003	0.947
0.8	-0.05	0.1	0.2	strong/very strong	strong/very strong	very weak	0.036	0.945
0.8	-0.05	0.1	0.2	strong/very strong	strong/very strong	weak	0.018	0.95
0.8	-0.05	0.1	0.1	moderate	moderate	very weak	0	0.959
0.8	-0.05	0.1	0.1	moderate	moderate	weak	0	0.947
0.8	-0.05	0.1	0.1	moderate	strong/very strong	very weak	0	0.948
0.8	-0.05	0.1	0.1	moderate	strong/very strong	weak	0	0.948
0.8	-0.05	0.1	0.1	strong/very strong	moderate	very weak	0	0.937

0.8	-0.05	0.1	0.1	strong/very strong	moderate	weak	0	0.945
0.8	-0.05	0.1	0.1	strong/very strong	strong/very strong	very weak	0.001	0.946
0.8	-0.05	0.1	0.1	strong/very strong	strong/very strong	weak	0	0.935
0.8	-0.05	0.1	0.02	moderate	moderate	very weak	0	0.951
0.8	-0.05	0.1	0.02	moderate	moderate	weak	0	0.943
0.8	-0.05	0.1	0.02	moderate	strong/very strong	very weak	0	0.954
0.8	-0.05	0.1	0.02	moderate	strong/very strong	weak	0	0.941
0.8	-0.05	0.1	0.02	strong/very strong	moderate	very weak	0	0.938
0.8	-0.05	0.1	0.02	strong/very strong	moderate	weak	0	0.951
0.8	-0.05	0.1	0.02	strong/very strong	strong/very strong	very weak	0	0.952
0.8	-0.05	0.1	0.02	strong/very strong	strong/very strong	weak	0	0.946

3.3 Simulation Study 2: Image-based Simulation Study

In Study 2, we directly compute Dice scores from contour masks rather than simulating Dice scores from statistical distributions. To achieve this, we employed the Medical Image Segmentation Synthesis (MISS) Tool, developed by Guan et al.[15] to synthesize segmentation masks with multiple controlled types of segmentation errors, such as spiculations and shape/alignment changes.

MISS-Tool Overview

The MISS-tool emulates segmentation errors by modifying truth masks of anatomical objects through a set of adjustable parameters that simulate six typical segmentation errors: boundary spiculation, under/over-sizing, centroid location errors, overlap variations, shape/alignment details, and the introduction of satellite structures. These segmentation error types are implemented through four primary image processing methods:

1. **Affine Transformation**: Modifies segmentation contours through resizing (changing height and width ratios), shifting (location changes in x, y coordinates), and rotation

(angle parameter).

2. **Spiculation**: Adjusts contours in polar coordinates by adding Gaussian functions to create spike-like protrusions or indentations on boundaries, representing boundary irregularities.

3. **Fourier Descriptor (FD) Modification**: Modifies contours in the spatial frequency domain by keeping low-frequency components (basic shape) while allowing changes to middle-frequency components and removing high-frequency components (fine details of the shape).

4. **Satellite Structure Synthesis**: Adds separate small objects in nearby regions of the true object, simulating disconnected components that are incorrectly included in segmentation results.

For our simulation study, we employed three of the four image processing methods (affine transformation, spiculation, and Fourier descriptor modification) to generate synthetic segmentation variations that approximate potential segmentation difference between different annotators.

3.3.1 Parameter Configuration

In Study 2, we consider cases containing synthetic contours with 3, 5, 7 and 9 readers. Contours made by all annotators on the same image are provided as 2D binary images of identical dimensions. We define the following parameters to control the synthetic contours used in Study 2 based on the MISS-tool methodology:

- **Affine transformation parameters: Resizing ratios (Rx, Ry):** control the width and height scaling of the segmentation boundaries, where ratios equal to 1 mean no change. **Location shift (Sx, Sy):** controls spatial displacement in x and y coordinates measured in pixels. **Rotation angle (φ):** controls the rotational variation of the segmented contours.
- **Fourier transformation parameters:** (as shown in the Supplementary Figure S1) **Detail:** number of non-zero Fourier descriptors from low to high frequences that will be kept or modified. This controls the level of boundary detail preservation, where higher values retain more fine-grained boundary features. **Range**: number of Fourier descriptors with middle frequency components to be modified. This determines the frequency bandwidth of modifications applied to the contour. **Magnitude**: strength value controlling how the middle-ranged Fourier descriptors are modified. This determines the extent of shape changes applied through frequency domain manipulation.
- **Spiculation parameters:** for adding Gaussian functions to create spike-like protrusions in polar coordinates (as shown in the Supplementary Figure S2), **the following parameters are used: Center (c)**: angular position (in degrees, 0-360°) where spiculation is added to the contour boundary. **Height (h):** magnitude of spike protrusion or indentation, where positive values create convex spiculations and negative values create concave indentations. **Width (w):** angular spread of the spiculation feature, determining how broad the spike modification appears on the boundary.

We then explored two scenarios in Study 2:

I. device segmentation is **Transformation Agreeable** with the reader panel, and

II. device segmentation is **Transformation Disagreeable** with the reader panel.

Each setting is characterized by the 11 parameters defined above. Four affine parameters (Rx, Ry, Sx, Sy) are used as "*tunable*" parameters, which introduce variability across settings. The remaining seven parameters (including all Fourier and spike transformation parameters) are designated as "*default*" parameters, maintaining uniformity across all settings. Table 6 and Appendix A.1 present the configurations for the *tunable* and *default* parameters for the two scenarios being studied. This design allows us to approximate the range of DSC distributions reported in the literature, as confirmed in results later.

Table 6. Configuration for *tunable* parameters in image-based Simulation Study 2.

scenario	parameters	configuration
I. reader panel vs. device: Transformation Agreeable	(n_{image}, n_{reader})	n_{image} in (100, 250, 750) n_{reader} in (3, 5, 7, 9)
	(R_x, R_y) for both panel and device	R_{max} = 1.15, 1.1, or 1.05 $R_x \sim U[2 - R_{max}, R_{max}]$ $R_y \sim U[2 - R_{max}, R_{max}]$
	(S_x, S_y) for both panel and device	S_{max} = 0, 2, or 3 pixels $S_x \sim U[-S_{max}, S_{max}]$ $S_y \sim U[-S_{max}, S_{max}]$
II. reader panel vs. device: Transformation Disagreeable	(n_{image}, n_{reader})	n_{image}in (100, 250, 750) n_{reader} in (3, 5, 7, 9)
	(R_x, R_y, S_x, S_y) for panel	Same as in Scenario I above
	(R_x, R_y, S_x, S_y) for device	For each of the 9 reader panel settings, the device (R_x, R_y, S_x, S_y) parameter set is selected from the remaining 8 settings.

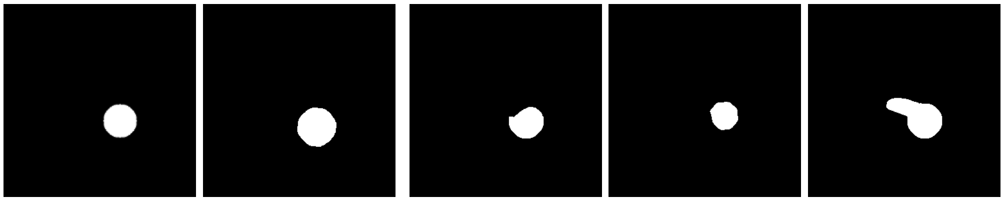
3.3.2 Dataset Simulation

We simulate each Study 2 image dataset and corresponding annotations using the following steps.

Step 1: Generate n true contours, each comprising a single randomly located and pixelized circle. These true contours serve as the foundation for generating both reader and device contours.

Step 2: Simulate $k + 1$ contours for each true contour representing annotations by k readers and the device, respectively. This is achieved by applying the MISS tool transformations using the parameter setting described in Table 6 to the true contours.

Two sets of contours generated using the same set of MISS parameters are defined as Transformation Agreeable. Otherwise, they are deemed Transformation Disagreeable. Figure 2 provides an illustrative example of a true contour and various contours synthesized using the MISS tool transformations for three experts and a Transformation Disagreeable device.



a) truth b) expert 1 c) expert 2 d) expert 3 e) device

Figure 2. Example of synthetic expert and device contours based on an initial true contour. Note, this is a Transformation Disagreeable example where the device contour is based on a different set of MISS parameters compared to that of the experts. a) true contour, b)-d) synthetic reader contours, e) transformation disagreeable device contour.

For each setting in Table 6, we simulate a total of N_{Sim} imaging datasets. For each image dataset, we compute image-level Dice scores for each pair of annotators. Subsequently, based on the Dice scores computed for each image, we calculate the performance metrics as defined in Section 3.1. The results are summarized in next subsection.

3.3.3 Results

Below are results for image-based simulation study. Performance metrics are computed for each setting based on N_{Sim}=100, n_{image}=100, 250, 750, and n_{reader}=3, 5, 7, 9. All results presented here are based on the z-interval CI approach. The results for the bootstrapping approach are omitted as they are similar.

3.3.3.1 Scenario I: Expert panel vs. device transformation-agreeable

Table 7 summarizes the mean DSC performance of the expert panel, the mean DSC performance difference between the device and panel, and the Type I error of the proposed method for each Transformation Agreeable setting, as configured in Table 6. The results in Table 7 indicate that the proposed method performs well across all Transformation Agreeable settings simulated, with Type I error close to 0.05 on average and small standard deviations (ranging from 0.02 to 0.04) across all of the settings evaluated. These findings suggest that when the device and expert panel share the same transformation pattern, their segmentation performance is similar (reflected by a mean DSC difference close to 0 in Table 7), and the proposed method achieves reasonable Type I error (ranging between 0.03 – 0.10 in Table 7, with a mean of 0.065 across all settings). Figure 3 illustrates the Type I error as a function of n_{image} (i.e., sample size) and n_{reader} (i.e., panel size) for each of the nine individual settings under Scenario I: Transformation Agreeable.

Table 7. Mean (standard deviation) for Within-Panel DSC, Device-Panel DSC, and Type 1 Error by panel transformation pattern from Image-Simulation Study 2, Scenario I: Transformation Agreeable.

	Expert Panel Transformation parameters			
		S_{max}=0	S_{max}=2	S_{max}=3
Reader-pair DSC $\hat{\mu}_0$	R_{max}=1.05	0.95 (0.04)	0.88 (0.09)	0.84 (0.13)
	R_{max}=1.10	0.93 (0.04)	0.87 (0.09)	0.83 (0.12)
	R_{max}=1.15	0.90 (0.05)	0.85 (0.09)	0.81 (0.12)
Device-Reader DSC	R_{max}=1.05	0.95 (0.04)	0.88 (0.09)	0.84 (0.13)

$\hat{\mu}_D$	R_{max}=1.10	0.93 (0.04)	0.87 (0.09)	0.83 (0.12)
	R_{max}=1.15	0.90 (0.05)	0.85 (0.09)	0.81 (0.12)
Type 1 error	R_{max}=1.05	0.10 (0.04)	0.05 (0.02)	0.03 (0.03)
α	R_{max}=1.10	0.09 (0.02)	0.06 (0.02)	0.06 (0.02)
	R_{max}=1.15	0.07 (0.02)	0.07 (0.03)	0.05 (0.02)

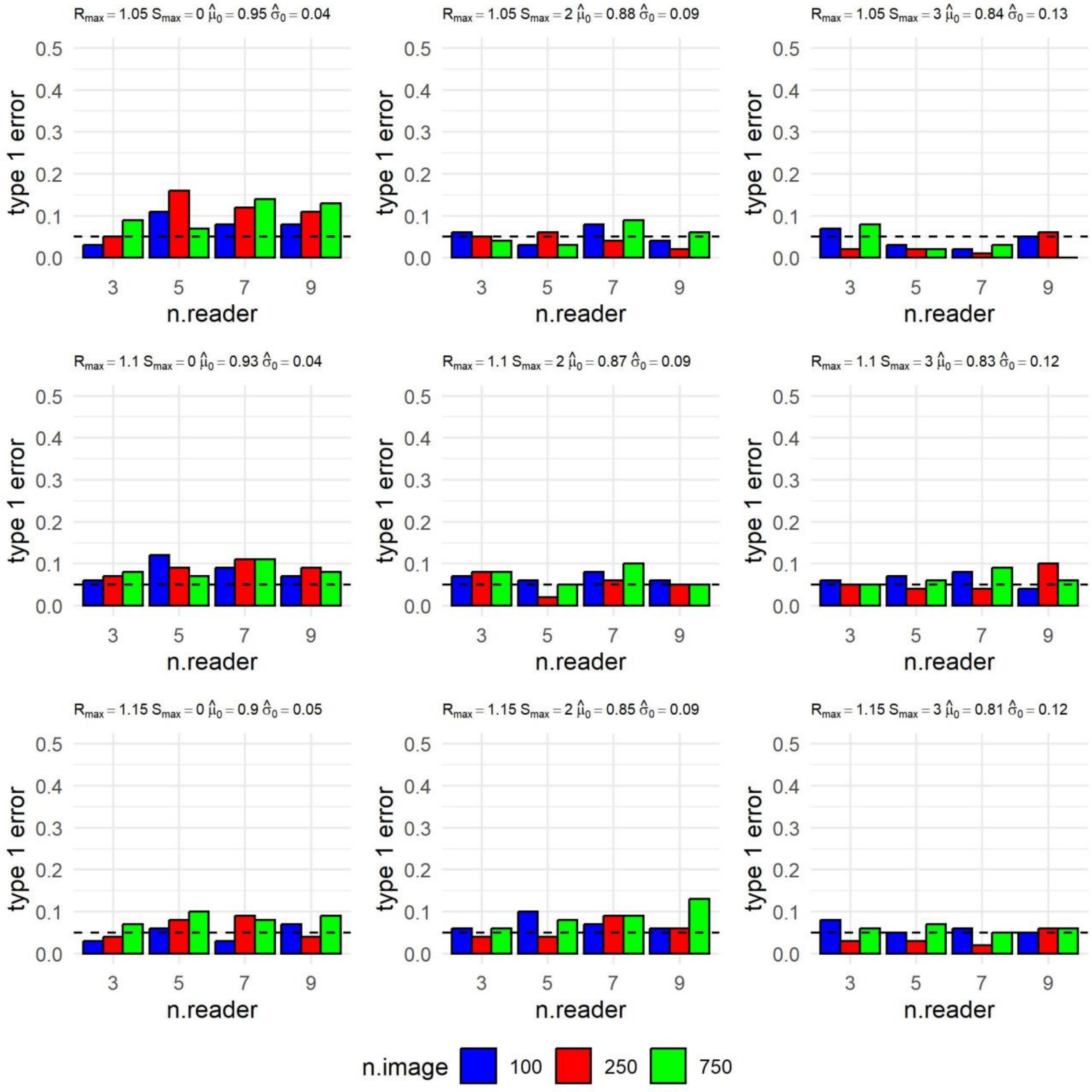


Figure 3. Plots of the Type 1 error results by transformation pattern (R_{max}, S_{max}) for Image Simulation Study 2, Scenario I: Transformation Agreeable. The black dashed horizontal line marks the expected type 1 error level of 0.05.

3.3.3.2 Scenario II: expert panel vs. device transformation-disagreeable

Scenario II (Transformation Disagreeable) includes 72 settings. Table 8 summarizes the mean within-panel DSC, the mean device-panel DSC, and the Type II error of the proposed method by expert panel transformation pattern defined by defined by (Rmax, Smax). Each cell's summary statistics are aggregated over the eight Transformation Disagreeable device settings corresponding to that panel pattern.

Table 8. Mean (standard deviation) of Within-Panel DSC, Device-Panel DSC, and Type 2 Error by Panel Transformation Pattern, Image-Simulation Study 2, Scenario II: Transformation Disagreeable.

		Expert Panel Transformation Pattern		
		S_{max}=0	S_{max}=2	S_{max}=3
Reader-pair DSC $\hat{\mu}_0$ ($\hat{\sigma}_0$)	R_{max}=1.05	0.95 (0.04)	0.88 (0.09)	0.84 (0.13)
	R_{max}=1.10	0.93 (0.04)	0.87 (0.09)	0.83 (0.12)
	R_{max}=1.15	0.90 (0.05)	0.85 (0.09)	0.81 (0.12)
Device-Reader DSC: $\hat{\mu}_D$ ($\hat{\sigma}_D$)	R_{max}=1.05	0.89 (0.07)	0.87 (0.09)	0.85 (0.1)
	R_{max}=1.10	0.89 (0.07)	0.87 (0.09)	0.85 (0.1)
	R_{max}=1.15	0.88 (0.07)	0.86 (0.08)	0.84 (0.1)
Type 2 error β	R_{max}=1.05	0.00 (0.00)	0.08 (0.18)	0.16 (0.28)
	R_{max}=1.10	0.01 (0.03)	0.07 (0.16)	0.14 (0.26)
	R_{max}=1.15	0.01 (0.04)	0.09 (0.18)	0.07 (0.17)

Table 8 reveals that the average Type II error for both interval approaches is below the threshold of 0.20, demonstrating the proposed method performs well in transformation-disagreeable settings overall. Across all Transformation Disagreeable settings, the mean (standard deviation) Type II error is 0.072 (0.182) for the z-interval approach. While not explicitly detailed, the mean (standard deviation) of the Type II error based on bootstrap confidence intervals is similar, at 0.070 (0.177).

The results highlight an interesting contrast in error fluctuations. Unlike Type I error, which shows relatively small variability for both Overlap Agreeable (Figure 1) and Transformation

Agreeable scenarios (std = 0.03, Table 7), the Type II error exhibits low variability in Overlap Disagreeable scenarios (Figure 1) but high variability in Transformation-Disagreeable scenarios (std = 0.18, Table 8). This disparity arises because Transformation Disagreeable scenarios do not necessarily equate to Overlap Disagreeable. Certain Transformation Disagreeable patterns can still yield similar overlap-based segmentation performance. For example, the transformation patterns (R_{max}, S_{max}) = (1.05, 2) and (1.10, 2) produce similar mean (standard deviation) DSC values of 0.88 (0.09) and 0.87 (0.09)**,** respectively. The inclusion of both Overlap Agreeable and Overlap Disagreeable cases within the transformation-disagreeable scenario contributes to higher variability in Type II error.

Table 8 also demonstrates that as the absolute mean DSC difference $\hat{d}_{\mu} = \hat{\mu}_D - \hat{\mu}_0$ increases, the mean Type II error approaches 0. This is as expected, as the method is expected to perform better when the DSC performance difference between the device and the panel becomes more pronounced.

An analysis of subgroup trends reveals that as the maximum reader shift (R_{max}) increases from 0 to 2 to 3, the mean Type II error rises from 0.01 to 0.08 to 0.13, respectively. This increase corresponds to a greater uncertainty in DSC performance as well because higher pixel shift counts are allowed. In Table 8, the DSC standard deviation increases from 0.04 to 0.09 to 0.13 as R_{max} increases from 0 to 2 to 3. This trend suggests that statistical power increases when expert transformation patterns exhibit less variability. This would be akin to a situation where all panel members have substantial experience and expertise, resulting is less variation among their segmentation contours. However, it is important to note that deliberately selecting experts to ensure near-identical patterns (e.g., all from one institution with identical training) is inappropriate, as this level of agreement is unrealistic in the broader clinical context.

To further visualize the method's performance for individual settings, Figure 4 illustrates the Type II error for each setting with fixed panel transformation parameters (R_{max}, S_{max}) = (1.10, 2). Results for other (Rmax, Smax) settings are omitted here and provided instead in Supplementary Figures S3–S10, as their patterns are similar to the trends shown in Figure 4. Each subplot in Figure 4 displays the Type II error of the proposed method when the device adopts a transformation pattern disagreeable to the panel across different numbers of images and experts.

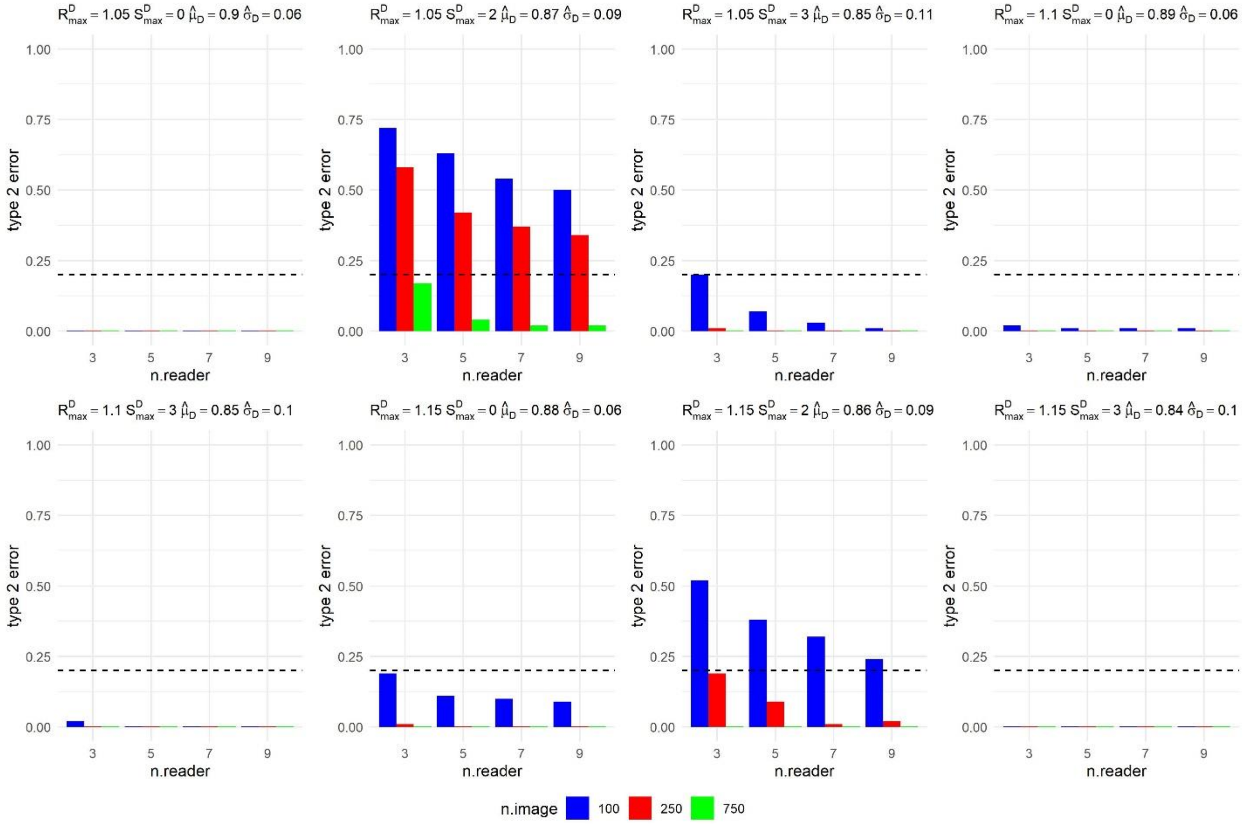


Figure 4. Type 2 error by device transformation pattern (R^D_{max}, S^D_{max}), given panel transformation $(R_{max}, S_{max}) = (1.10, 2)$, with mean(std) within-panel Dice 0.87(0.09) for Image Simulation Study 2, Scenario II: Transformation Disagreeable. The black dashed horizontal line marks type 2 error level at 0.2.

Figure 4 shows a decreasing Type II error as sample size or panel size increases. These trends indicate that increasing the sample size and the panel size (assuming all experts have similar DSC performance) enhances the power of the proposed method. The results presented in Figure 4 further corroborate the findings in Table 8, demonstrating that the proposed method achieves consistently low Type II error in settings with larger difference in Dice performance, except in transformation-disagreeable but with smaller difference in the Dice performance. But even in these challenging cases, increasing the sample size or the number of experts in the panel effectively reduces the Type II error. Notably, the instances of Type II error exceeding 0.2 in Figure 4 are unlikely to pose practical concerns, as the DSC performance differences in these settings are minimal (e.g., as small as 0.01) and are shown to be mitigated to below 0.2 by increasing the sample size.

4 Application

This section illustrates a real-world example to demonstrate the practical application of the segmentation interchangeability metric for comparison of an AI segmentation and multiple human experts.

4.1 Database and Data Preparation

The data are taken from the LIDC-IDRI database [23] developed by the Lung Image Database Consortium. This is one of the largest publicly available datasets for lung nodule detection and segmentation. It contains data from 1010 patients (1018 studies) and 2660 nodules, with slice thickness varying from 0.45 mm to 5.0 mm. Each study includes clinical thoracic CT images accompanied by an XML file documenting the annotations made by four experienced thoracic radiologists. This dataset is anonymized, and all protected health information (PHI) is removed [23].

A U-Net segmentation model with a ResNeXt encoder and ImageNet transfer learning was trained using a subset of LIDC-IDRI data. This U-Net model was then compared to the segmentation performance of 4 radiologist annotators using a testing dataset based on 124 independent LIDC-IDRI images. Details of training and testing data preparation as well as model training are provided in the Appendix A2.

4.2 Results

Figure 5 shows the distribution of DSC values for pairwise annotators. The results suggest the AI model may not agree with the panel, as evidenced by a lower DSC between the AI model and the readers compared to the DSC between pairwise readers. However, this visual comparison does not substantiate a statistical assessment of the difference.

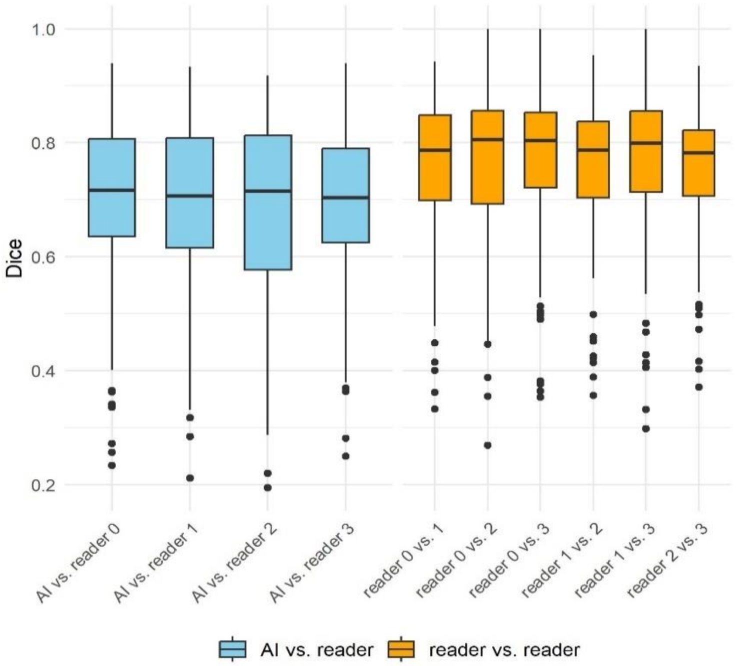


Figure 5. Distribution of DSC values for each pair of annotators.

We now apply our proposed statistical testing method to assess the interchangeability of the AI model and four radiologists in the panel. Table 10 reports the mean (standard deviation) of within-panel DSC $\hat{\mu}_P(\hat{\sigma}_P)$, the mean (standard deviation) for the AI-panel comparison $\hat{\mu}_A(\hat{\sigma}_A)$, the difference between the two $\hat{d} = \hat{\mu}_P - \hat{\mu}_A$ and 95% CI derived from bootstrapping and z-interval approaches derived using the proposed method.

Table 9. Agreement assessment results for the U-Net model (device) and the expert panel on the LIDC-IDRI test dataset.

Annotator vs. Reference Panel	$\hat{\mu}_P$ $(\hat{\sigma}_P)$	$\hat{\mu}_A$ $(\hat{\sigma}_A)$	$\hat{d} = \hat{\mu}_P - \hat{\mu}_A$ (95% CI) bootstrapping approach	$\hat{d} = \hat{\mu}_P - \hat{\mu}_A$ (95% CI) z-interval approach
U-Net vs. 4 readers	0.7619 (0.0937)	0.6896 (0.1172)	0.0723 (0.0584, 0.087)	0.0723 (0.0583, 0.0863)

The results in Table 9 indicate that the trained U-Net model vs. 4-reader panel comparison yields a 95% z-interval CI of (0.0583, 0.0856) (with a similar result from the bootstrapping CI, as shown in Table 9). This indicates that the U-Net model has significantly lower agreement with the panel than the within-panel agreement, and thus it is not interchangeable with the reader panel. We note that the trained AI model presented here is used exclusively for illustrative purposes, and the results of this specific use case should ***not*** be interpreted as evidence that U-Net segmentation models more generally underperform relative to human experts.

5 Discussion

In this work, we developed a segmentation interchangeability metric and statistical method for evaluating agreement between an AI device and a panel of human experts. Through a statistical and an image-based simulation studies, we demonstrated that the proposed method exhibits well-controlled Type I error and good Type II error behavior. The novelty of this method lies in its ability to directly assess the interchangeability of a segmentation AI device with multiple human experts without requiring a reference standard. This distinguishes our method from the traditional approach that do require defining reference standard contours. Additionally, setting an acceptable performance goal with conventional methods is challenging due to the lack of a widely accepted clinical cutoff for the Dice score. In contrast, our method compares device performance directly with an expert panel as a control, eliminating the need for a surrogate ground truth. Setting a performance goal based on the mean DSC difference is also simpler, as the target value is typically close to 0.

A limitation of the proposed method is it treats reader effect as fixed. This may not be a major concern, as treating a small reader panel as fixed is not uncommon. In the context of multi-reader studies, this approach has been used by Bandos et al. [25] and discussed by Hillis and Schartz [26]. It would be an interesting future direction to incorporate methods accounting for truther panel variability to refine our proposed method as an extension. Another limitation of the method is it is designed to assess differences in overlap-based segmentation performance not distance-based performance. One can easily substitute other metric into the proposed method, such as a distance-based metric, but we have not determined how well the method's assumptions hold.

The simulation study results indicate that increasing the sample size or the expert panel size can enhance the power of a study utilizing the proposed method. In practice, it may be challenging to expand the panel size while maintaining a high within-panel agreement level. As such, increasing the sample size may be a most practical and feasible approach compared to enlarging the panel.

Appendix A

A1 Table. Configuration for Default parameters in Image-based Simulation

transform	parameter	configuration
fourier	(detail, range, magnitude)	detail = # of all pixels in original contour
		range = round(detail * 0.2) + 1
		magnitude = 2

spike	(center, height, width)	center ~ U[0°, 360°)
		height ~ estimate lesion diameter * U[0.01, 0.2] (randomly assign ± sign)
		width ~ estimate lesion diameter * U[0.01-0.2] / 2
affine - rotation	φ	φ ~ U[0°, 360°)

A2 Training/Testing Data Preparation and Model Training

Based on LIDC-IDRI database, we prepare the final data to be used for method demonstration using following steps:

1. For each nodule, we draw the central slice from each scan (if there is even number of slices (say, 2m slices) in a scan, we take (m+1)th slice as the central slice for this scan).
2. From the imaging dataset created in Step 1, we randomly draw a 70% sample (at the patient level) and use these cases to train an AI segmentation model (a **U-net** architecture with ResNeXt Encoder that is pretrained on ImageNet Database) using an IoU metric. The reference standard mask for the training data, which is required for AI model training purposes, is created by applying Majority Vote (MV) rule to the 4 radiologists' manual annotations. For illustration, the figure below presents an example of the original slice, alongside annotations from four radiologists, the aggregated consensus derived using the Majority Vote (MV) criterion, and the annotation extracted by the U-Net algorithm.

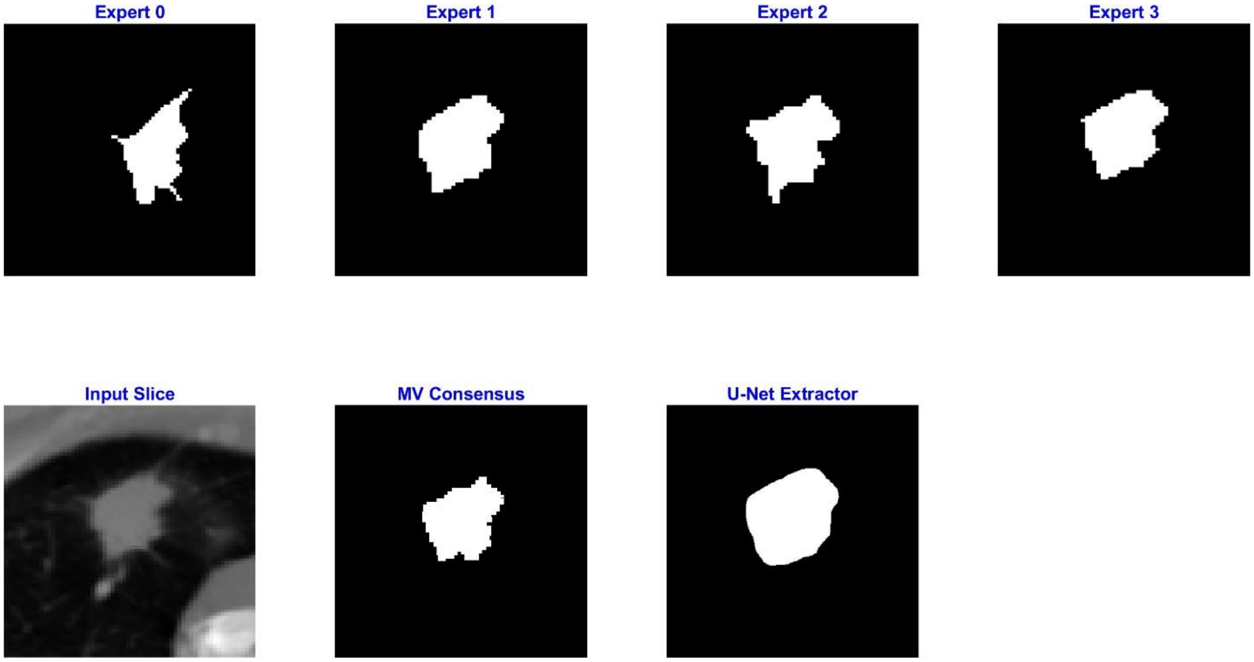


Figure A1: Radiologists annotations of an LIDC-IDRI lesion along with the majority vote (MV) consensus, and the contour produced by our U-Net segemantation algorithm.

3. From the remaining 30% of patients not used in training, we collected cases where the 4 radiologists agreed on the lesion (defined here as having at least a 1 pixel overlap between annotations from any pair of radiologists in panel). One image per patient is then randomly chosen to form the final testing dataset, resulting in 124 independent images.
4. Using contours generated by the algorithm and those annotated by the four radiologists on the testing data, compute paired annotator Dice scores. And then the computed Dice scores will be fed to the proposed method for annotator vs. reference panel interchangeability assessment.

Declaration of interest statement

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Disclaimer

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. This is a contribution of the U.S. Food and Drug Administration and is not subject to copyright.

Acknowledgments

The authors would like to thank our colleagues for their generous help with reviewing and revising.

Biographies

Tingting Hu, PhD, is a visiting scientist in the Division of Imaging, Diagnostics, and Software Reliability, Center for Devices and Radiological Health, U.S. Food and Drug Administration. She received her PhD in statistics from Florida State University. Her research interests include statistics.

Berkman Sahiner is a senior biomedical research scientist with the Division of Imaging, Diagnostics and Software Reliability, Center for Devices and Radiological Health, U.S. Food and Drug Administration. He has a PhD in electrical engineering and computer science from the University of Michigan, Ann Arbor. His research is focused on the evaluation of medical imaging and computer-assisted diagnosis devices, including devices that incorporate machine learning and artificial intelligence. He is a fellow of SPIE, AAPM and AIMBE.

Shuyue Guan, PhD, a current staff fellow in the Division of Imaging, Diagnostics, and Software Reliability, Center for Devices and Radiological Health, U.S. Food and Drug Administration. PhD received in Biomedical Engineering from the George Washington University. Primary research interests are image processing and machine learning.

Nicholas Petrick, Ph.D. has expertise in medical artificial intelligence development and assessment. He is currently the Deputy Director for the Division of Imaging, Diagnostics, and Software Reliability in the Center for Devices and Radiological Health, U.S. Food and Drug Administration and is a member of FDA’s Senior Biomedical Research and Biomedical Product Assessment Service.

Code, Data, and Materials Availability

The data used in the application section—LIDC-IDRI images—are publicly available at: https://www.cancerimagingarchive.net/collection/lidc-idri/. The code developed for the proposed method is available at ######link to be inserted after manuscript clearance######.

Figure Caption List

Figure 1.CP, Type I error and Type II error results for the statistics-based Simulation Study 1. The figure includes boxplot results aggregated across all the settings in Scenarios I – IV, and for both the 2-reader and 3-reader scenarios.

Figure 2. Example of synthetic expert and device contours based on an initial true contour. Note, this is a Transformation Disagreeable example where the device contour is based on a different set of MISS parameters compared to that of the experts. a) true contour, b)-d) synthetic reader contours, e) Transformation Disagreeable device contour.

Figure 3. Plots of the Type 1 error results by transformation pattern (Rmax, Smax) for Image Simulation Study 2, Scenario I: Transformation Agreeable. The black dashed horizontal line marks the expected type 1 error level of 0.05.

Figure 4.Type 2 error by device transformation pattern (R^D_{max}, S^D_{max}), given panel transformation $(R_{max}, S_{max}) = (1.10, 2)$, with mean(std) within-panel Dice 0.87(0.09) for image simulation study Scenario II: Transformation Disagreeable. The black dashed horizontal line marks desirable type 2 error level at 0.2.

Figure 5. Distribution of DSC values for each pair of annotators.

Figure A1: Radiologists annotations of an LIDC-IDRI lesion along with the majority vote (MV) consensus, and the contour produced by our U-Net segmentation algorithm.

Table Caption List

Table 1: Parameter configuration scenarios for Study 1: statistic-based simulations.

Table 2. Results for all 32 settings in Scenario I: Equal Mean μ and Equal Standard Deviation σ from statistics-based simulation Study 1.

Table 3. Selected results for 8 Settings with $(\mu_0, d_\mu, \sigma_0) = (0.85, -0.05, 0.15)$ in Scenario II: Unequal Mean μ and Equal Standard Deviation σ from statistics-based Simulation Study 1.

Table 4. Selected results for 24 of 96 Settings with $\mu_0 = 0.75$ in Scenario III: Equal Mean μ and Unequal Standard Deviation σ from statistics-based Simulation Study 1.

Table 5. Selected results for 24 of 648 Settings with $(\mu_0, d_\mu, \sigma_0) = (0.8, -0.05, 0.1)$ in Scenario IV: Unequal Mean μ and Unequal Standard Deviation σ from statistics-based Simulation Study 1.

Table 6. Configuration for tunable parameters in image-based Simulation Study 2.

Table 7. Mean (standard deviation) for Within-Panel DSC, Device-Panel DSC, and Type 1 Error by panel transformation pattern from Image-Simulation Study 2, Scenario I: Transformation Agreeable.

Table 8. Mean (standard deviation) of Within-Panel DSC, Device-Panel DSC, and Type 2 Error by Panel Transformation Pattern, Image-Simulation Study 2, Scenario II: Transformation Disagreeable.

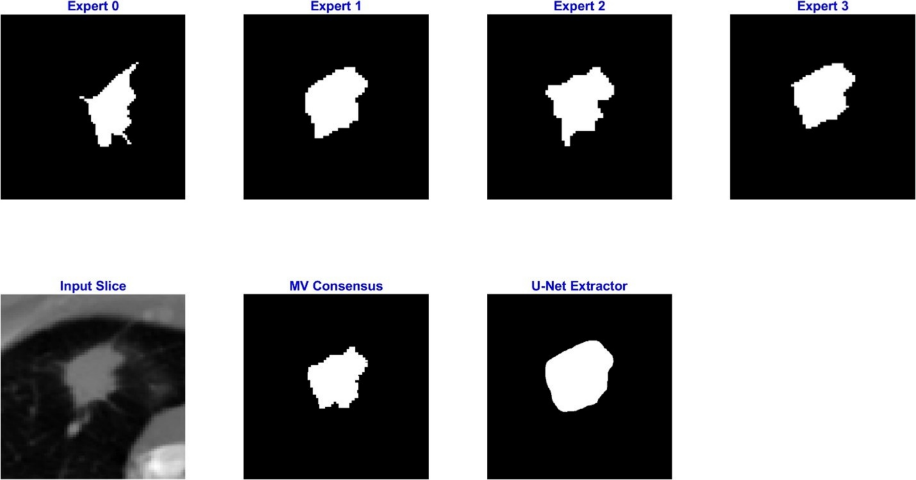
Table 9. Agreement assessment results for the U-Net model (device) and the expert panel using the LIDC-IDRI test dataset.

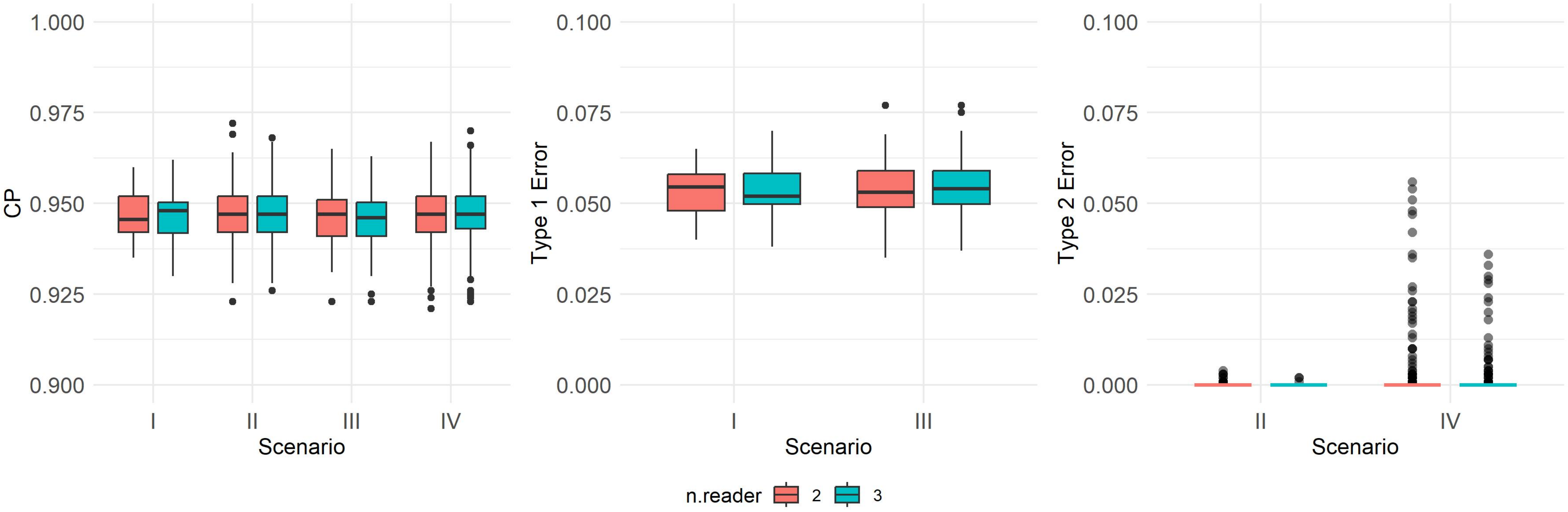
References

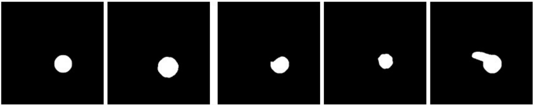
1. Taha, A. A., & Hanbury, A. (2015). Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. BMC medical imaging, 15(1), 1-28.
2. Wang, Z., Wang, E., & Zhu, Y. (2020). Image segmentation evaluation: a survey of methods. Artificial Intelligence Review, 53(8), 5637-5674.
3. Udupa, J. K., LeBlanc, V. R., Zhuge, Y., Imielinska, C., Schmidt, H., Currie, L. M., ... & Woodburn, J. (2006). A framework for evaluating image segmentation algorithms. Computerized medical imaging and graphics, 30(2), 75-87.
4. Cárdenes, R., de Luis-Garcia, R., & Bach-Cuadra, M. (2009). A multidimensional segmentation evaluation for medical image data. Computer methods and programs in biomedicine, 96(2), 108-124.
5. Dice, L. R. (1945). Measures of the amount of ecologic association between species. Ecology, 26(3), 297-302.
6. Jaccard, P. (1912). The distribution of the flora in the alpine zone. 1. New phytologist, 11(2), 37-50.
7. Zou, K. H., Warfield, S. K., Bharatha, A., Tempany, C. M., Kaus, M. R., Haker, S. J., ... & Kikinis, R. (2004). Statistical validation of image segmentation quality based on a spatial overlap index1: scientific reports. Academic radiology, 11(2), 178-189.
8. Bland, J. M., & Altman, D. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. The lancet, 327(8476), 307-310.
9. FDA (2001). Guidance for industry: Statistical approaches to establishing bioequivalence, Food and Drug Administration, Center for Drug Evaluation and Research (CDER).

10. Obuchowski, N. A. (2001). Can electronic medical images replace hard-copy film? Defining and testing the equivalence of diagnostic tests. Statistics in medicine, 20(19), 2845-2863.
11. Obuchowski, N. A., Subhas, N., & Schoenhagen, P. (2014). Testing for interchangeability of imaging tests. Academic Radiology, 21(11), 1483-1489.
12. Shao, J., & Zhong, B. (2004). Assessing the agreement between two quantitative assays with repeated measurements. Journal of Biopharmaceutical Statistics, 14(1), 201-212.
13. Barnhart, H. X., Kosinski, A. S., & Haber, M. J. (2007). Assessing individual agreement. Journal of biopharmaceutical statistics, 17(4), 697-719.
14. FDA CDER (2022). Guidance for industry: Multiple Endpoints in Clinical Trials. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/multiple-endpoints-clinical-trials
15. 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.
16. Wang, S., Zhou, M., Liu, Z., Liu, Z., Gu, D., Zang, Y., ... & Tian, J. (2017). Central focused convolutional neural networks: Developing a data-driven model for lung nodule segmentation. Medical image analysis, 40, 172-183.
17. Joskowicz, L., Cohen, D., Caplan, N., & Sosna, J. (2019). Inter-observer variability of manual contour delineation of structures in CT. European radiology, 29, 1391-1399.
18. Evans, J. D. (1996). Straightforward statistics for the behavioral sciences. Thomson Brooks/Cole Publishing Co.
19. LaMorte W. W. The Correlation Coefficient (r) [Internet]. [Accessed January 2024]. Boston University School of Public Health. Available at: https://sphweb.bumc.bu.edu/otlt/MPH-Modules/PH717-QuantCore/PH717-Module9-Correlation-Regression/PH717-Module9-Correlation-Regression4.html.

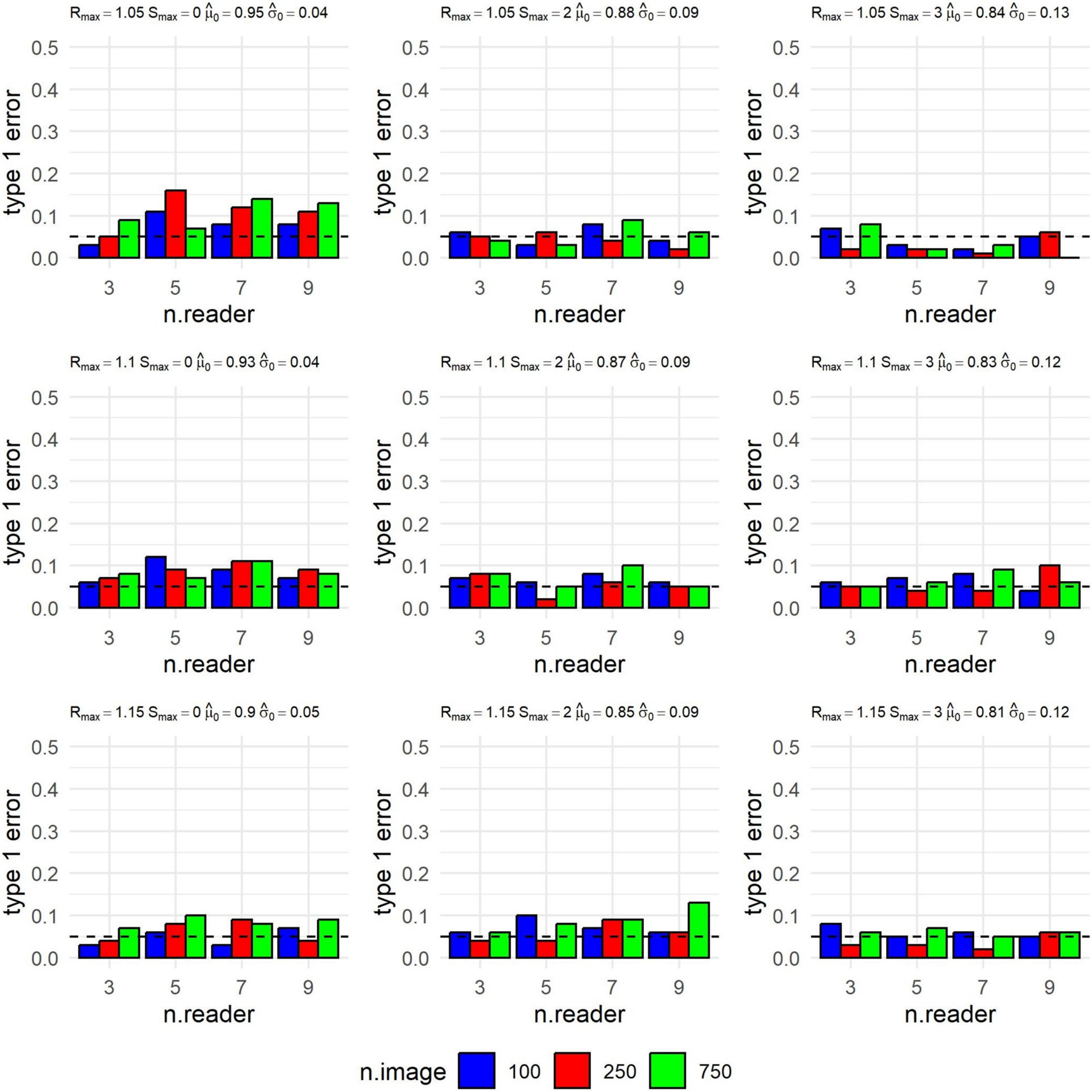
20. Swinscow, T. D. V., & Campbell, M. J. (2002). Statistics at square one (pp. 111-25). London: Bmj. Available at: https://www.bmj.com/about-bmj/resources-readers/publications/statistics-square-one/11-correlation-and-regression.
21. Whuber (https://stats.stackexchange.com/users/919/whuber), How to construct a multivariate Beta distribution?, URL (version: 2021-10-21): https://stats.stackexchange.com/q/549262
22. L'Ecuyer, P. (2015, December). Random number generation with multiple streams for sequential and parallel computing. In 2015 Winter Simulation Conference (WSC) (pp. 31-44). IEEE.
23. Armato III, S. G., McLennan, G., Bidaut, L., McNitt-Gray, M. F., Meyer, C. R., Reeves, A. P., ... & Clarke, L. P. (2011). The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans. Medical physics, 38(2), 915-931.
24. FDA CDRH (2022). Guidance for Industry and Food and Drug Administration Staff Computer-Assisted Detection Devices Applied to Radiology Images and Radiology Device Data - Premarket Notification [510(k)] Submissions. https://www.fda.gov/media/77635/download
25. 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.
26. 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.

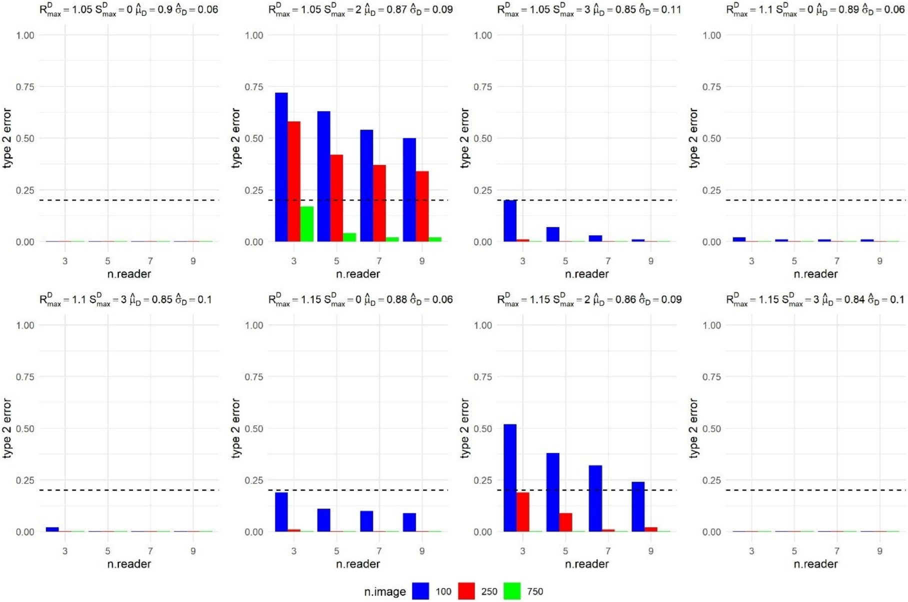

Expert 0
Expert 1
Expert 2
Expert 3
Input Slice
MV Consensus
U-Net Extractor

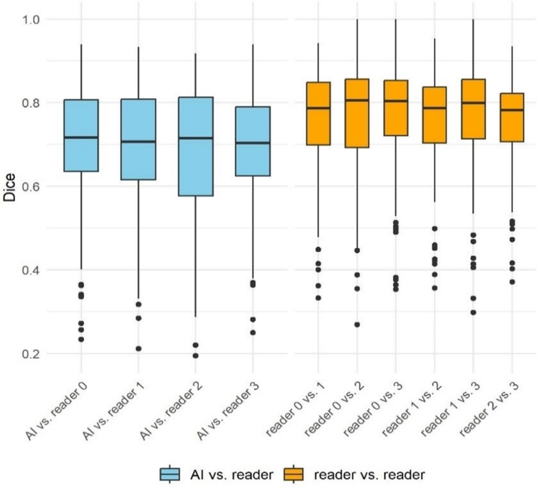

1.000
0.975
0.950
0.925
0.900
CP
I
II
III
IV
Scenario
0.100
0.075
0.050
0.025
0.000
Type 1 Error
I
III
Scenario
0.100
0.075
0.050
0.025
0.000
Type 2 Error
II
IV
Scenario
n.reader
2
3



a) truth b) expert 1 c) expert 2 d) expert 3 e) device


$R_{max} = 1.05$ $S_{max} = 0$ $\hat{\mu}_0 = 0.95$ $\hat{\sigma}_0 = 0.04$
$R_{max} = 1.05$ $S_{max} = 2$ $\hat{\mu}_0 = 0.88$ $\hat{\sigma}_0 = 0.09$
$R_{max} = 1.05$ $S_{max} = 3$ $\hat{\mu}_0 = 0.84$ $\hat{\sigma}_0 = 0.13$
$R_{max} = 1.1$ $S_{max} = 0$ $\hat{\mu}_0 = 0.93$ $\hat{\sigma}_0 = 0.04$
$R_{max} = 1.1$ $S_{max} = 2$ $\hat{\mu}_0 = 0.87$ $\hat{\sigma}_0 = 0.09$
$R_{max} = 1.1$ $S_{max} = 3$ $\hat{\mu}_0 = 0.83$ $\hat{\sigma}_0 = 0.12$
$R_{max} = 1.15$ $S_{max} = 0$ $\hat{\mu}_0 = 0.9$ $\hat{\sigma}_0 = 0.05$
$R_{max} = 1.15$ $S_{max} = 2$ $\hat{\mu}_0 = 0.85$ $\hat{\sigma}_0 = 0.09$
$R_{max} = 1.15$ $S_{max} = 3$ $\hat{\mu}_0 = 0.81$ $\hat{\sigma}_0 = 0.12$
type 1 error
0.0
0.1
0.2
0.3
0.4
0.5
n.reader
3
5
7
9
n.image
100
250
750


$R^D_{max} = 1.05\ S^D_{max} = 0\ \hat{\mu}_D = 0.9\ \hat{\sigma}_D = 0.06$
$R^D_{max} = 1.05\ S^D_{max} = 2\ \hat{\mu}_D = 0.87\ \hat{\sigma}_D = 0.09$
$R^D_{max} = 1.05\ S^D_{max} = 3\ \hat{\mu}_D = 0.85\ \hat{\sigma}_D = 0.11$
$R^D_{max} = 1.1\ S^D_{max} = 0\ \hat{\mu}_D = 0.89\ \hat{\sigma}_D = 0.06$
$R^D_{max} = 1.1\ S^D_{max} = 3\ \hat{\mu}_D = 0.85\ \hat{\sigma}_D = 0.1$
$R^D_{max} = 1.15\ S^D_{max} = 0\ \hat{\mu}_D = 0.88\ \hat{\sigma}_D = 0.06$
$R^D_{max} = 1.15\ S^D_{max} = 2\ \hat{\mu}_D = 0.86\ \hat{\sigma}_D = 0.09$
$R^D_{max} = 1.15\ S^D_{max} = 3\ \hat{\mu}_D = 0.84\ \hat{\sigma}_D = 0.1$
type 2 error
1.00
0.75
0.50
0.25
0.00
n.reader
3
5
7
9
n.image
100
250
750


Dice
1.0
0.8
0.6
0.4
0.2
AI vs. reader 0
AI vs. reader 1
AI vs. reader 2
AI vs. reader 3
reader 0 vs. 1
reader 0 vs. 2
reader 0 vs. 3
reader 1 vs. 2
reader 1 vs. 3
reader 2 vs. 3
AI vs. reader
reader vs. reader