Journal of *Personalized Medicine*

MDPI

Article

Design and Analysis Methods for Trials with AI-Based Diagnostic Devices for Breast Cancer

Lu Liu [1], Kevin J. Parker [2] and Sin-Ho Jung [1,*]

[1] Department of Biostatistics and Bioinformatics, Duke University, Durham, NC 27710, USA; lu.liu@duke.edu
[2] Department of Electrical and Computer Engineering, University of Rochester, Rochester, NY 14627, USA; kevin.parker@rochester.edu
* Correspondence: sinho.jung@duke.edu

Abstract: Imaging is important in cancer diagnostics. It takes a long period of medical training and clinical experience for radiologists to be able to accurately interpret diagnostic images. With the advance of big data analysis, machine learning and AI-based devices are currently under development and taking a role in imaging diagnostics. If an AI-based imaging device can read the image as accurately as experienced radiologists, it may be able to help radiologists increase the accuracy of their reading and manage their workloads. In this paper, we consider two potential study objectives of a clinical trial to evaluate an AI-based device for breast cancer diagnosis by comparing its concordance with human radiologists. We propose statistical design and analysis methods for each study objective. Extensive numerical studies are conducted to show that the proposed statistical testing methods control the type I error rate accurately and the design methods provide required sample sizes with statistical powers close to pre-specified nominal levels. The proposed methods were successfully used to design and analyze a real device trial.

Keywords: artificial intelligence (AI); breast cancer; clinical device trial; concordance rate; generalized estimating equation; sample size calculation; statistical test

check for updates

Citation: Liu, L.; Parker, K.J.; Jung, S.-H. Design and Analysis Methods for Trials with AI-Based Diagnostic Devices. *J. Pers. Med.* **2021**, *11*, 1150. https://doi.org/10.3390/jpm11111150

Academic Editors: Gianluca Franceschini, Alba Di Leone and Alejandro Martin Sanchez

Received: 8 September 2021
Accepted: 2 November 2021
Published: 4 November 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Introduction

There are different types of device trials depending on the use of device and the study objectives. In this paper, we introduce statistical design and analysis methods for a trial on an artificial intelligence (AI)-based device for the diagnosis of breast cancer.

Imaging technologies play a major role in the diagnosis of breast cancer. The reading and interpretation of imaging requires intensive medical training and significant clinical experience. With the advance of big data analysis methods, machine learning and AI-based imaging systems are currently under active development [1]. If an AI-based imaging device can read the image as accurately as experienced radiologists, it may be able to help radiologists increase the accuracy of their reading, manage their workloads, or possibly replace radiologists in remote clinics that would not have an experienced radiologist available for consultation.

In the assessment of breast lesions, the BI-RADS reporting system and classification are widely used [2]. This system includes categories between 1 and 5 (benign to malignant) with a key diagnostic transition subdivided into categories 4a (low suspicion of malignancy), 4b (moderate suspicion) and 4c (high suspicion, greater than 50% likelihood but less than 95% likelihood of malignancy). Furthermore, the BI-RADS lexicon covers radiological descriptive features that are important in diagnostic assessments, and these vary by modality. Examples of ultrasound lexicon used in AI-based classifications is given in Table A1. The earlier approaches to breast ultrasound technology concentrated on the extraction of features of lesions such as size, shape, texture, and boundaries within a clustering or classification or rule-based decision making algorithms [3–6]. More recent developments in AI, machine learning, and deep learning systems have utilized layers of convolution neural network models, a variety of approaches and extensive training sets to produce differentiated output classifications [7,8].

In this paper, we consider the requirements for a clinical device trial to evaluate the performance of an AI-based imaging device using BI-RADS reporting system for the diagnosis of breast cancer. Since BI-RADS reporting system does not have a gold standard, we evaluate the performance of the device by how well its reading aligns with those of radiologists. We propose design and analysis methods for two different types of study objectives that can be used for such a trial. The first objective is to test if the reading of the AI-based device concurs with those of radiologists as much as the readings concord among radiologists. The second objective is to test if the reading of the AI-based device is more concordant with those of experienced radiologists than with those of junior radiologists. For each objective, we propose a statistical testing method and its sample size calculation formula. The proposed testing methods will be used to analyze the data for each of the five BI-RADS lexicon classification category listed in Table A1, but the sample size calculation for a trial may be conducted only for the most important one. The performance of these methods are evaluated using simulations.

2. Materials and Methods

We consider two types of study objectives to evaluate the performance of an AI-based device for the diagnosis of breast cancer. For each study objective, we propose a testing method and its sample size formula. Suppose that we have images from n subjects.

2.1. Objective 1: Is the Concordance Rate between the AI-Based Device and Radiologists as High as That among Radiologists?

The image of each subject is read by m radiologists and the AI-based device. BI-RADS lexicon does not have a gold standard. So, in order to validate a device with an AI-based algorithm, we should show that the reading of the device concurs with those with radiologists. For example, in Table A1, for BI-RADS lexicon classification, Shape, two radiologists will be declared to be concurrent for an image if they both read oval, round or irregular. The question is how high the concordance rate should be between the device and the radiologists. The concordance rate among radiologists is used as a reference.

For each category of BI-RADS lexicon classifications, let p_r and p_s denote the concordance rate among radiologists and that between radiologists and the device, respectively. Since the latter can not be higher than the former, we specify a similarity margin $\delta_1 (> 0)$. That is, we will not be interested in the AI-based device if $p_s \leq p_r - \delta_1$ and will be highly interested in it if $p_s = p_r$. So, we want to test a null hypothesis $H_1 : p_s = p_r - \delta_1$ against the alternative hypothesis $\bar{H}_1 : p_s > p_r - \delta_1$.

2.1.1. Statistical testing method

Suppose that there are n patients, and the image of each patient is read by the AI-based device and m radiologists. For subject $i(= 1, ..., n)$ and radiologist $j(= 1, ..., m)$, let $r_{ijj'} = 1$ if radiologists j and j' concur and $= 0$ otherwise, and let $s_{ij} = 1$ if radiologist j and the device concur and $= 0$ otherwise. Note that we have $p_r = E(r_{ijj'})$ and $p_s = E(s_{ij})$. Since $r_{ijj'} = r_{ij'j}$ and $r_{ijj} = 1$ for $j, j' = 1, ..., m$, the number of informative concordance scores among m radiologists is $m(m-1)/2$ for each image, the concordance rate among radiologists for subject i is estimated by

$$r_i = \frac{\sum_{j=1}^{m-1} \sum_{j'=j+1}^{m} r_{ijj'}}{m(m-1)/2}.$$

On the other hand, for subject i, the concordance rate between the device and m radiologists is estimated by

$$s_i = \frac{\sum_{j=1}^{m} s_{ij}}{m}$$

Using the images from n subjects, concordance rate among radiologists is estimated by

$$\hat{p}_r = \frac{2}{nm(m-1)} \sum_{i=1}^{n} \sum_{j=1}^{m-1} \sum_{j'=j+1}^{m} r_{ijj'} = \frac{1}{n} \sum_{i=1}^{n} r_i$$

and that between the device and radiologists is estimated by

$$\hat{p}_s = \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} s_{ij} = \frac{1}{n} \sum_{i=1}^{n} s_i$$

Those estimates are unbiased because

$$E(\hat{p}_r) = E(\frac{2}{nm(m-1)} \sum_{i=1}^{n} \sum_{j=1}^{m-1} \sum_{j'=j+1}^{m} r_{ijj'}) = \frac{2}{nm(m-1)} \frac{nm(m-1)}{2} p_r = p_r$$

and

$$E(\hat{p}_s) = E(\frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} s_{ij}) = \frac{1}{nm} nmp_s = p_s.$$

Since $r_1, ..., r_n$ are independent random variables with mean p_r, for large n by the central limit theorem,

$$\sqrt{n}(\hat{p}_r - p_r) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (r_i - p_r)$$

is asymptotically normal with mean 0 and variance σ_r^2 that can be consistently estimated by

$$\hat{\sigma}_r^2 = \frac{1}{n} \sum_{i=1}^{n} (r_i - \hat{p}_r)^2$$

Similarly, $s_1, ..., s_n$ are independent random variables with mean p_s, so that for large n,

$$\sqrt{n}(\hat{p}_s - p_s) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (s_i - p_s)$$

is asymptotically normal with mean 0 and variance σ_s^2 that can be consistently estimated by

$$\hat{\sigma}_s^2 = \frac{1}{n} \sum_{i=1}^{n} (s_i - \hat{p}_s)^2$$

Since each subject's image is read by the device and radiologists, r_i and s_i are correlated. However, $(s_i - r_i + \delta_1, i = 1, ..., n)$ are independent, with mean 0 under the null hypothesis H_1. Hence, by the central limit theorem under H_1,

$$\sqrt{n}(\hat{p}_s - \hat{p}_r + \delta_1) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (s_i - r_i + \delta_1)$$

is asymptotically normal with mean 0 and variance σ_1^2 that can be consistently estimated by

$$\hat{\sigma}_1^2 = \frac{1}{n} \sum_{i=1}^{n} (s_i - r_i + \delta_1)^2$$

Hence, we reject the null hypothesis $H_1 : p_s \leq p_r - \delta_1$ if $Z_1 > z_{1-\alpha}$, where

$$Z_1 = \frac{\sqrt{n}(\hat{p}_s - \hat{p}_r + \delta_1)}{\hat{\sigma}_1}$$

and $z_{1-\alpha}$ is the $100(1-\alpha)$ percentile of the standard normal distribution. Note that we use a 1-sided test because the hypotheses are 1-sided and to avoid too large a sample size with a small δ_1.

Note that $\hat{p}_r$ and $\hat{p}_s$ are the generalized estimating Equation [9] (GEE) estimators of p_r and p_s, respectively, using the working independent correlation. Furthermore, the robust estimator of σ_1^2 by the GEE method is given as

$$\tilde{\sigma}_1^2 = \frac{1}{n}\sum_{i=1}^{n}(s_i - r_i + \hat{p}_s - \hat{p}_r)^2$$

Since $\hat{p}_s - \hat{p}_r$ is a consistent estimator of $p_s - p_r$, $\tilde{\sigma}_1^2$ is asymptotically identical to $\hat{\sigma}_1^2$ under H_1. Hence, Z_1 can be counted as a test statistic based on the GEE method with the working independent correlation.

2.1.2. Power and Sample Size Calculation

We calculate the sample size for the test statistic Z_1 under a specific alternative hypothesis $\bar{H}_1 : p_r = p_s$. An accurate sample size calculation for the statistical test requires specification of correlation coefficients between $r_{ij_1j_2}$ and $r_{ij_1'j_2'}$, between $r_{ij_1j_2}$ and s_{ij}, and between s_{ij} and $s_{ij'}$. The dependency between $r_{ij_1j_2}$ and $r_{ij_1j_2'}$ is expected to be higher than that between $r_{ij_1j_2}$ and $r_{ij_1'j_2'}$ for $j_1 \neq j_1' \neq j_2 \neq j_2'$ because the former pair includes the same reader j_1 while the latter pair contains four different readers. Similarly, we expect that the dependency between $r_{ij_1j_2}$ and s_{ij_1} is expected to be higher than that between $r_{ij_1j_2}$ and $s_{ij_1'}$ for $j_1, j_2 \neq j_1'$.

For a simplified sample size formula, we just specify $\rho_1 = \text{corr}(r_i, s_i)$. We define the correlation coefficients among the concordance scores $\rho_{r1} = \text{corr}(r_{i12}, r_{i13})$, $\rho_{r2} = \text{corr}(r_{i12}, r_{i34})$, $\rho_{s1} = \text{corr}(r_{i12}, s_{i1}) = \text{corr}(r_{i12}, s_{i2})$, and $\rho_{s2} = \text{corr}(r_{i12}, s_{i3})$, and $\rho_{ss} = \text{corr}(s_{i1}, s_{i2})$. Appendix A.1 shows that we have $\rho_1 = \text{corr}(r_i, s_i)$ is expressed as

$$\rho_1 = \frac{\frac{2}{m}\rho_{s1} + \frac{m-2}{m}\rho_{s2}}{\sqrt{(\frac{2}{m(m-1)} + \frac{4(m-2)}{m(m-1)}\rho_{r1} + \frac{(m-2)(m-3)}{m(m-1)}\rho_{r2})(\frac{1}{m} + \frac{m-1}{m}\rho_{ss})}}$$

Under $\bar{H}_1 : p_r = p_s$, $\{(s_i - r_i), i = 1, ..., n\}$ are independent random variables with mean 0, so that $\sqrt{n}(\hat{p}_s - \hat{p}_r)$ is asymptotically normal with mean 0 and variance σ_1^2 that can be consistently estimated by $s_1^2 = \sqrt{n}^{-1}\sum_{i=1}^{n}(s_i - r_i)^2$. Since $\hat{\sigma}_1^2$ is asymptotically identical to $s_1^2 + \delta_1^2$ under $\bar{H}_1$, it converges to $\sigma_1^2 + \delta_1^2$. Hence, the power for a given sample size n is

$$\begin{aligned} 1 - \beta &= P\Big(\frac{\sqrt{n}(\hat{p}_s - \hat{p}_r + \delta_1)}{\hat{\sigma}_1} > z_{1-\alpha} | p_r = p_s\Big) \\ &= P\Big(\frac{\sqrt{n}(\hat{p}_s - \hat{p}_r) + \sqrt{n}\delta_1)}{\sqrt{\sigma_1^2 + \delta_1^2}} > z_{1-\alpha} | p_r = p_s\Big) \\ &= P\Big(\frac{\sqrt{n}(\hat{p}_s - \hat{p}_r)}{\sigma_1} > \frac{z_{1-\alpha}\sqrt{\sigma_1^2 + \delta_1^2} - \sqrt{n}\delta_1}{\sigma_1} | p_r = p_s\Big) \\ &= \Phi\Big(\frac{z_{1-\alpha}\sqrt{\sigma_1^2 + \delta_1^2} - \sqrt{n}\delta_1}{\sigma_1}\Big) \end{aligned} \quad (1)$$

where $\Phi(.)$ is the survivor function of the standard normal distribution and σ_1^2 is the limit of $n^{-1}\sum_{i=1}^{n}(s_i - r_i)^2$.

By solving the power Equation (1) with respect to n, we obtain the required sample size for power $1-\beta$

$$n = \frac{(z_{1-\alpha}\sqrt{\sigma_1^2+\delta_1^2}+z_{1-\beta}\sigma_1)^2}{\delta_1^2} \tag{2}$$

where, as shown in the Appendix A.2,

$$\sigma_1^2 = \mathrm{var}(s_i)+\mathrm{var}(r_i)-2\rho_1\sqrt{\mathrm{var}(s_i)\mathrm{var}(r_i)} \tag{3}$$

$$\mathrm{var}(s_i) = p_r(1-p_r)\{1/m+\rho_{ss}(m-1)/m\}$$

and

$$\mathrm{var}(r_i) = p_r(1-p_r)\Big(\frac{2}{m(m-1)}+\frac{4(m-2)}{m(m-1)}\rho_{r1}+\frac{(m-2)(m-3)}{m(m-1)}\rho_{r2}\Big).$$

The process of calculating the required sample size is summarized as follows:

1. Specify $(\alpha, 1-\beta)$, expected concordance rate among radiologists p_r, similarity margin δ_1 and hypothetical correlation coefficients $\rho_{r1}, \rho_{r2}, \rho_{ss}, \rho_{s1}$ and ρ_{s2}.
2. Calculate σ_1^2 using (3).
3. Obtain sample size using (2).

It may be difficult to specify the correlation coefficients $\rho_{r1}, \rho_{r2}, \rho_{ss}, \rho_{s1}$ and ρ_{s2}. If pilot data are available, we may estimate them from the pilot data. Otherwise, we may conduct a two-stage trial to estimate these correlation coefficients from the first stage data and recalculate the sample size for the whole trial based on the estimated correlation coefficients.

2.2. *Objective 2: Is the AI-Based Device More Concordant with Experienced Radiologists Than with Junior Radiologists?*

As another study objective, we may want to test if the reading of the AI-based device agrees more with those of experienced radiologists than with those of junior radiologists for each BI-RADS lexicon classification category.

Let p_x and p_y denote the concordance rate between the AI-based device and highly experienced radiologists and that between the AI-based device and less experienced radiologists, respectively. We want to test the null hypothesis $H_2 : p_x = p_y$ against the alternative hypothesis $\bar{H}_2 : p_x > p_y$.

2.2.1. Statistical Testing Method

Let m (= 5, say) denote the number of radiologists in each group (highly experienced group and less experienced group). For subject i $(= 1, ..., n)$ and senior radiologist j $(= 1, ..., m)$, let $x_{ij} = 1$ if the reading by senior radiologist j and that by the AI-based device agree and =0 otherwise, and let $y_{ij} = 1$ if the reading by less experienced radiologist j $(= 1, ..., m)$ and that by the AI-based device agree and =0 otherwise. Then, we have $p_x = E(x_{ij})$ and $p_y = E(y_{ij})$. Using the data from subject i,, p_x is estimated by

$$x_i = \frac{\sum_{j=1}^m x_{ij}}{m}$$

and p_y is estimated by

$$y_i = \frac{\sum_{j=1}^m y_{ij}}{m}$$

Using the whole data, we estimate p_x and p_y by

$$\hat{p}_x = \frac{1}{nm}\sum_{i=1}^n\sum_{j=1}^m x_{ij} = \frac{1}{n}\sum_{i=1}^n x_i$$

and

$$\hat{p}_y = \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} y_{ij} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

respectively. Note that those estimates are unbiased.

For large n under H_2, $\sqrt{n}(\hat{p}_x - \hat{p}_y)$ is approximately normal with mean 0 and variance σ_2^2 that can be estimated by

$$\hat{\sigma_2}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2$$

Hence, we reject $H_2 : p_x = p_y$ if $|Z_2| > z_{1-\alpha/2}$, where

$$Z_2 = \frac{\sqrt{n}(\hat{p}_x - \hat{p}_y)}{\hat{\sigma}_2}.$$

Note that we use a standard 2-sided test since usually there is no small effect size issue in this case.

2.2.2. Power and Sample Size Calculation

We calculate the sample size under a specific alternative hypothesis $\bar{H}_2 : p_x = p_y + \delta_2$. Since each subject's image is read by all experienced and inexperienced radiologists as well as the device, $\{(x_{ij}, y_{ij}), j = 1, ..., m\}$ are correlated.

Let $\rho_{xx} = corr(x_{i1}, x_{i2})$ denote the correlation coefficient between the concordance score between the AI-based device and a highly experienced radiologist and the concordance score between the device and another highly experienced radiologist, and $\rho_{yy} = corr(y_{i1}, y_{i2})$ denote the correlation coefficient between the AI-based device and a less experienced radiologist and the concordance score between the device and another less experienced radiologist. Furthermore, let $\rho_{xy} = \mathrm{corr}(x_{ij}, y_{ij'})$ for $j, j' = 1, ..., m$ denote the correlation coefficient between the concordance score between the device and a highly experienced radiologist and that between the device and a less experienced radiologist. Let $\rho_2 = corr(x_i, y_i)$. As shown in Appendix A.3. ρ_2 is a function of ρ_{xx}, ρ_{yy}, and ρ_{xy}.

Under $\bar{H}_2 : p_x = p_y + \delta_2$, $\{(x_i - y_i - \delta_2), i = 1, ..., n\}$ are independent random variables with mean 0, so that $\sqrt{n}(\hat{p}_x - \hat{p}_y - \delta_2)$ is asymptotically normal with mean 0 and variance σ_2^2 that can be consistently estimated by $s_2^2 = \sqrt{n}^{-1}\sum_{i=1}^{n}(x_i - y_i - \delta_2)^2$. Note that $\hat{\sigma}_2^2$ is asymptotically identical to $s_2^2 + \delta_2^2$ under $\bar{H}_2$, so that it converges to $\sigma_2^2 + \delta_2^2$. Hence, the power for a given sample size n is

$$\begin{aligned}
1 - \beta &= P\Big(\frac{\sqrt{n}(\hat{p}_x - \hat{p}_y)}{\hat{\sigma}_2} > z_{1-\alpha/2}|p_x = p_y + \delta_2\Big) \\
&= P\Big(\frac{\sqrt{n}(\hat{p}_x - \hat{p}_y - \delta_2) + \sqrt{n}\delta_2}{\sqrt{s_2 + \delta_2^2}} > z_{1-\alpha/2}|p_x = p_y + \delta_2\Big) \\
&= P\Big(\frac{\sqrt{n}(\hat{p}_x - \hat{p}_y - \delta_2)}{\sigma_2} > \frac{z_{1-\alpha/2}\sqrt{\sigma_2^2 + \delta_2^2} - \sqrt{n}\delta_2}{\sigma_2}|p_x = p_y + \delta_2\Big) \\
&= \Phi\Big(\frac{z_{1-\alpha/2}\sqrt{\sigma_2^2 + \delta_2^2} - \sqrt{n}\delta_2}{\sigma_2}\Big) \qquad (4)
\end{aligned}$$

since $\sqrt{n}(\hat{p}_x - \hat{p}_y - \delta_2)/\sigma_2$ is $N(0, 1)$ under $\bar{H}_2$, where σ_2^2 is the limit of $n^{-1}\sum_{i=1}^{n}(x_i - y_i - \delta_2)^2$ under $\bar{H}_2$.

By solving (4) with respect to n, we obtain the required sample size for power $1 - \beta$

$$n = \frac{(z_{1-\alpha/2}\sqrt{\sigma_2^2 + \delta_2^2} + z_{1-\beta}\sigma_2)^2}{\delta_2^2} \qquad (5)$$

Appendix A.4 shows that

$$\sigma_2^2 = \text{var}(x_i) + \text{var}(y_i) - 2\rho_2\sqrt{\text{var}(x_i)\text{var}(y_i)} \tag{6}$$

where

$$\text{var}(x_i) = p_x(1-p_x)(\frac{1}{m} + \frac{m-1}{m}\rho_{xx})$$

and

$$\text{var}(y_i) = (p_x - \delta_2)(1 - p_x + \delta_2)(\frac{1}{m} + \frac{m-1}{m}\rho_{yy})$$

under $\bar{H}_2$.

The process of calculating the required sample size is summarized as follows:

1. Specify $(\alpha, 1-\beta)$, expected concordance rate between the AI-based device and a highly experienced radiologist p_x, clinically meaningful difference in concordance rates δ_2 and correlation coefficients ρ_{xx}, ρ_{yy} and ρ_{xy}.
2. Calculate σ_2^2 using (6).
3. Obtain the required sample size using (5).

It will be difficult to specify the correlation coefficients ρ_{xx}, ρ_{yy} and ρ_{xy}. If pilot data are available, we may estimate them from the pilot data. Otherwise, we may use a two-stage design to estimate these correlation coefficients from the first stage data and calculate the sample size for the whole trial based on the estimated correlation coefficients.

3. Numerical Studies and Results

Note that our test statistics and sample size formulas are derived based on large sample approximations. We want to conduct simulation studies to evaluate their finite sample performance.

We consider the first type of study objective to test if the concordance rate between an AI-based device and radiologists is as high as that among radiologists. Suppose that each subject's image is read by the AI-based device and $m = 10$ radiologists. We set $\alpha = 0.05$, $1-\beta = 0.8$ or 0.9, $p_r = 0.3$, 0.5 or 0.7, $\delta_1 = 0.05$ or 0.1 and $\rho_1 = 0.1$, 0.3, 0.5, or 0.7. Assuming $\rho_{s1} = \rho_{s2} + 0.1, \rho_{r1} = \rho_{r2} + 0.1, \rho_{r1} = \rho_{ss} = \rho_{s1} + 0.1$, we calculate the correlation coefficients for a given $\rho = corr(r_i, s_i)$. That is, we obtain $(\rho_{s1}, \rho_{s2}, \rho_{ss}, \rho_{r1}, \rho_{r2}) = (0.101, 0.001, 0.201, 0.201, 0.101)$, $(0.16, 0.06, 0.26, 0.26, 0.16)$, $(0.26, 0.16, 0.36, 0.36, 0.26)$ and $(0.48, 0.38, 0.58, 0.58, 0.48)$ for $\rho_1 = 0.1$, 0.3, 0.5, and 0.7, respectively.

For each design setting, we calculate the required sample size n using our proposed formula (2) and generate 10,000 simulation data sets of size n under the design setting and H_1 or $\bar{H}_1$. Then, we apply the statistical testing using Z_1 to each simulation data set, and compute the empirical type I error $(\hat{\alpha})$ and power $(1-\hat{\beta})$ by the proportion of samples that reject H_1 among the 10,000 samples simulated under $H_1 : p_s = p_r - \delta_1$ and $\bar{H}_1 : p_s = p_r$, respectively. The correlated concordance (binary) data are generated by first generating multivariate normal data and then dichotomizing them with corresponding proportion level [10].

Table 1 reports the sample size n, empirical type I error rate $\hat{\alpha}$, and power $1-\hat{\beta}$ under each design setting. We observe that the required sample size increases in $1-\beta$ and decreases in δ_1 and ρ_1. With other design parameters fixed, we have the same sample sizes for $p_r = 0.3$ and $p_r = 0.7$. We have this result because, from (2), the sample size depends on p_r only through $p_r(1-p_r)$. Since the empirical type I errors are very close to the nominal $\alpha = 0.05$ overall, our test statistic Z_1 controls the type I error rate accurately. On the other hand, the empirical powers are close to the corresponding nominal level $1-\beta = 0.8$ or 0.9 overall, so that we conclude that our sample size formula is accurate too.

Now we conduct simulations for the second type of study objective to test if the AI-based device is more concordant with experienced radiologists than with junior radiologists. We assume that each subject's image is read by $m = 5$ experienced radiologists and $m = 5$ junior radiologists. We set $\alpha = 0.05$, $1-\beta = 0.8$ or 0.9, $p_x = 0.3$, 0.5, or 0.7,

$\delta_2 = 0.05$ or 0.1, and $\rho_2 = 0.1$, 0.3, 0.5, or 0.7. Assuming $\rho_{xx} = \rho_{yy} = \rho_{xy} + 0.1$, we solve the corresponding correlation coefficients for given $\rho_2 = corr(x_i, y_i)$. So, we have $(\rho_{xx}, \rho_{yy}, \rho_{xy}) = (0.13, 0.13, 0.03)$, $(0.21, 0.21, 0.11)$, $(0.33, 0.33, 0.23)$, and $(0.55, 0.55, 0.45)$ for $\rho_2 = 0.1$, 0.3, 0.5, and 0.7, respectively. For each design setting, we calculate sample size n using (5), and generate 10,000 samples of size n under the design setting and H_2 or $\bar{H}_2$. We apply the test statistic Z_2 to each sample, and calculate the empirical type I error rate and power $(\hat{\alpha}, 1 - \hat{\beta})$ under $H_2 : p_x = p_y$ and $\bar{H}_2 : p_x = p_y + \delta_0$, respectively.

Table 2 summarizes the required sample size n, and empirical type I error rate and power, $(\hat{\alpha}, 1 - \hat{\beta})$, under each design setting. We observe that the required sample size increases in $1 - \beta$ and decreases in δ_2 and ρ_2. Since the empirical type I errors are very close to the nominal $\alpha = 0.05$ overall, our test statistic Z_2 controls the type I error accurately. On the other hand, the empirical powers are close to the corresponding nominal level $1 - \beta = 0.8$ or 0.9 overall, so that we conclude that our sample size formula is accurate too.

Table 1. Sample size (empirical type I error rate, empirical power), $n(\hat{\alpha}, 1 - \hat{\beta})$, under various design settings of $(p_r, \delta_1, \rho_1, 1 - \beta)$ for the first type of study objective.

p_r	δ_1	ρ_1	$1 - \beta = 0.8$	$1 - \beta = 0.9$
0.3	0.05	0.1	$210(0.044, 0.808)$	$290(0.051, 0.910)$
		0.3	$206(0.048, 0.812)$	$285(0.047, 0.903)$
		0.5	$200(0.049, 0.805)$	$275(0.049, 0.910)$
		0.7	$186(0.054, 0.811)$	$256(0.056, 0.903)$
	0.1	0.1	$56(0.041, 0.829)$	$76(0.045, 0.915)$
		0.3	$55(0.047, 0.823)$	$75(0.042, 0.914)$
		0.5	$53(0.048, 0.822)$	$73(0.052, 0.921)$
		0.7	$50(0.061, 0.822)$	$68(0.060, 0.913)$
0.5	0.05	0.1	$249(0.047, 0.804)$	$344(0.048, 0.904)$
		0.3	$245(0.051, 0.808)$	$338(0.053, 0.901)$
		0.5	$237(0.045, 0.812)$	$327(0.050, 0.907)$
		0.7	$220(0.053, 0.798)$	$304(0.050, 0.904)$
	0.1	0.1	$66(0.052, 0.815)$	$90(0.048, 0.911)$
		0.3	$65(0.049, 0.824)$	$88(0.046, 0.914)$
		0.5	$63(0.051, 0.831)$	$86(0.048, 0.912)$
		0.7	$58(0.054, 0.813)$	$80(0.054, 0.909)$
0.7	0.05	0.1	$210(0.052, 0.804)$	$290(0.054, 0.902)$
		0.3	$206(0.050, 0.800)$	$285(0.048, 0.906)$
		0.5	$200(0.052, 0.802)$	$275(0.051, 0.906)$
		0.7	$186(0.055, 0.806)$	$256(0.049, 0.899)$
	0.1	0.1	$56(0.055, 0.821)$	$76(0.054, 0.909)$
		0.3	$55(0.055, 0.816)$	$75(0.049, 0.904)$
		0.5	$53(0.052, 0.814)$	$73(0.054, 0.911)$
		0.7	$50(0.060, 0.821)$	$68(0.058, 0.912)$

Table 2. Sample size (empirical type I error rate, empirical power), $n(\hat{\alpha}, 1-\hat{\beta})$, under various design settings of $(p_x, \delta_2, \rho_2, 1-\beta)$ for the second type of study objective.

p_x	δ_2	ρ_2	$\mathbf{1-\beta=0.8}$	$\mathbf{1-\beta=0.9}$
0.3	0.05	0.1	$348(0.052, 0.810)$	$465(0.052, 0.898)$
		0.3	$328(0.050, 0.812)$	$438(0.048, 0.894)$
		0.5	$298(0.049, 0.807)$	$397(0.047, 0.900)$
		0.7	$245(0.053, 0.807)$	$327(0.048, 0.907)$
	0.1	0.1	$86(0.054, 0.816)$	$113(0.050, 0.905)$
		0.3	$81(0.053, 0.820)$	$107(0.047, 0.912)$
		0.5	$74(0.047, 0.825)$	$98(0.047, 0.914)$
		0.7	$63(0.053, 0.844)$	$83(0.048, 0.934)$
0.5	0.05	0.1	$434(0.050, 0.798)$	$580(0.048, 0.902)$
		0.3	$409(0.050, 0.810)$	$546(0.049, 0.904)$
		0.5	$370(0.049, 0.806)$	$495(0.052, 0.905)$
		0.7	$304(0.054, 0.802)$	$406(0.050, 0.905)$
	0.1	0.1	$111(0.053, 0.816)$	$148(0.053, 0.909)$
		0.3	$105(0.047, 0.814)$	$140(0.051, 0.909)$
		0.5	$96(0.050, 0.811)$	$127(0.052, 0.911)$
		0.7	$79(0.052, 0.829)$	$105(0.050, 0.913)$
0.7	0.05	0.1	$382(0.051, 0.803)$	$511(0.052, 0.900)$
		0.3	$360(0.052, 0.797)$	$481(0.048, 0.901)$
		0.5	$327(0.046, 0.804)$	$436(0.055, 0.902)$
		0.7	$269(0.047, 0.811)$	$359(0.053, 0.909)$
	0.1	0.1	$103(0.050, 0.815)$	$136(0.049, 0.914)$
		0.3	$97(0.054, 0.817)$	$129(0.049, 0.912)$
		0.5	$89(0.050, 0.821)$	$117(0.050, 0.917)$
		0.7	$74(0.048, 0.840)$	$98(0.045, 0.925)$

4. Discussion and Conclusions

Existing papers on comparing correlated concordance rates mainly focus on comparing two (or more) competitive diagnosis methods using their concordance rates with a gold standard on multiple sites [11]. In this paper, there is no gold standard and we compare the concordance rate between an AI-based diagnostic device and human radiologists and that among radiologists. We also compare the concordance rate between an AI-based diagnostic device and highly experienced radiologists and that between AI-based device and less experienced radiologists. In our design setting, each study subject has single site but is rated by the AI-based device and multiple human radiologists. We extend existing methods to perform design and analysis in this new setting.

We provide design and analysis plan for two types of study objectives to perform different comparisons of concordance between the AI-based diagnostic device and human radiologists. For each type of study objective, we propose a test statistic using GEE method with independent working correlation to account for the dependency in the observations from the device and the radiologists for each study subject, and derive its sample size formula based on large sample theory. Through extensive simulations, we show that the test statistics control the type I error accurately and the sample size formulas estimate sample sizes with powers close to the specified ones accounting for the dependency of images read by radiologists and device.

Since each subject's image is read by the device and many radiologists, the concordance scores have complicated dependency structure, while the test statistics do not require specification of the multiple correlation coefficients by using the GEE method, the sample size formulas require specification of these correlation coefficients. Since it is difficult to accurately specify the correlation coefficients, we propose to conduct a two-stage device trial to estimate these correlation coefficients from the first stage data and recalculate the required sample size for the whole trial based on the estimated correlation coefficients.

We use concordance rate as a measure of agreement among multiple raters. Cohen's kappa is another measure of agreement that is popularly used, e.g., Qureshi et al. [12].

Unlike concordance rate, however, it is not clear how similar two kappa values should be to conclude similarity of two different groups of raters. The proposed methods were successfully used by O'Connell et al. [8] to design and analyze a device trial.

Author Contributions: Conceptualization, L.L. and S.-H.J.; methodology, L.L.; software, L.L.; validation, L.L. and S.-H.J.; formal analysis, L.L.; investigation, L.L. and S.-H.J.; resources, K.J.P. and S.-H.J.; data curation, L.L.; writing—original draft preparation, L.L.; writing—review and editing, S.-H.J.; visualization, L.L.; supervision, K.J.P. and S.-H.J. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Table A1. Examples of ultrasound lexicon.

Ground Truth	Lesion Type
Shape	Oval
	Round
	Irregular
Margin	Circumscribed
	Indistinct
	Angular
	Microlobulated
	Spiculated
Orientation	Parallel
	Not parallel
Echo pattern	Anechoic
	Hypoechoic
	Complex cystic and solid
	Isoechoic
	Hyperechoic
	Heterogeneous
Posterior features	No features
	Enhancement
	Shadowing
	Combined pattern

Appendix A.1. Derivation of ρ_1

Since $r_i = \{m(m-1)/2\}^{-1} \sum_{j=1}^{m-1} \sum_{j'=j+1}^{m} r_{ijj'}$ and $s_i = m^{-1} \sum_{j=1}^{m} s_{ij}$, we have

$$\rho_1 = \mathrm{corr}(r_i, s_i) = \frac{\mathrm{cov}(r_i, s_i)}{\sqrt{\mathrm{var}(r_i)\mathrm{var}(s_i)}}$$

Here

$$\mathrm{cov}(r_i, s_i) = \mathrm{cov}(\frac{\sum_{j=1}^{m-1} \sum_{j'=j+1}^{m} r_{ijj'}}{m(m-1)/2}, \frac{\sum_{j=1}^{m} s_{ij}}{m}) = \frac{2}{m^2(m-1)} cov(\sum_{j=1}^{m-1} \sum_{j'=j+1}^{m} r_{ijj'}, \sum_{j=1}^{m} s_{ij})$$

$$= \frac{2}{m^2(m-1)} \Big\{ m(m-1)\mathrm{cov}(r_{ij_1j_2}, s_{ij_1}) + \frac{m(m-1)(m-2)}{2} \mathrm{cov}(r_{ij_1j_2}, s_{ij'_1}) \Big\}$$

$$= \frac{2}{m}\mathrm{cov}(r_{i12}, s_{i1}) + \frac{m-2}{m}\mathrm{cov}(r_{i12}, s_{i3})$$

$$\mathrm{var}(r_i) = \frac{4}{m^2(m-1)^2} \mathrm{var}(\sum_{j=1}^{m-1} \sum_{j'=j+1}^{m} r_{ijj'})$$

$$= \frac{4}{m^2(m-1)^2}\Big\{\frac{m(m-1)}{2}\mathrm{var}(r_{ij_1j_2}) + m(m-1)(m-2)\mathrm{cov}(r_{ij_1j_2}, r_{ij_1j_2'})$$

$$+\frac{m(m-1)(m-2)(m-3)}{4}\mathrm{cov}(r_{ij_1j_2}, r_{ij_1'j_2'})\Big\}$$

$$= \frac{2}{m(m-1)}var(r_{i12}) + \frac{4(m-2)}{m(m-1)}cov(r_{i12}, r_{i13}) + \frac{(m-2)(m-3)}{m(m-1)}cov(r_{i12}, r_{i34})$$

and

$$\mathrm{var}(s_i) = \frac{1}{m^2}\{m\mathrm{var}(s_{i1}) + m(m-1)\mathrm{cov}(s_{i1}, s_{i2})\}$$

$$= \frac{1}{m}\mathrm{var}(s_{i1}) + \frac{m-1}{m}\mathrm{cov}(s_{i1}, s_{i2})$$

Hence,

$$\rho_1 = \frac{\frac{2}{m}\mathrm{cov}(r_{i12}, s_{i1}) + \frac{m-2}{m}\mathrm{cov}(r_{i12}, s_{i3})}{\sqrt{\{\frac{2}{m(m-1)}\mathrm{var}(r_{i12}) + \frac{4(m-2)}{m(m-1)}\mathrm{cov}(r_{i12}, r_{i13}) + \frac{(m-2)(m-3)}{m(m-1)}\mathrm{cov}(r_{i12}, r_{i34})\}}}$$

$$* \frac{1}{\sqrt{\{\frac{1}{m}\mathrm{var}(s_{i1}) + \frac{m-1}{m}cov(s_{i1}, s_{i2})\}}}$$

$$= \frac{\frac{2}{m}\rho_{s1} + \frac{m-2}{m}\rho_{s2}}{\sqrt{\{\frac{2}{m(m-1)} + \frac{4(m-2)}{m(m-1)}\rho_{r1} + \frac{(m-2)(m-3)}{m(m-1)}\rho_{r2}\}(\frac{1}{m} + \frac{m-1}{m}\rho_{ss})}}$$

Appendix A.2. The Limit of $\hat{\sigma}_1^2$ under $\bar{H}_1$

The limit of $\hat{\sigma}_1^2 = n^{-1}\sum_{i=1}^n (s_i - r_i)^2$ is its expected value $\sigma_1^2 = E(s_i - r_i)^2$. Since $E(s_i - r_i) = p_s - p_r = 0$ under $\bar{H}_1 : p_r = p_s$, $E(s_i - r_i)^2 = \mathrm{var}(s_i - r_i) = \mathrm{var}(s_i) + \mathrm{var}(r_i) - 2\rho_1\sqrt{\mathrm{var}(s_i)\mathrm{var}(r_i)}$ where $\rho_1 = \mathrm{corr}(r_i, s_i)$,

$$\mathrm{var}(s_i) = \frac{1}{m^2}var(\sum_{j=1}^m s_{ij}) = \frac{1}{m^2}\{m\mathrm{var}(s_{ij}) + m(m-1)\mathrm{cov}(s_{ij}, s_{ij'})\}$$

$$= \frac{1}{m}\mathrm{var}(s_{ij}) + \frac{m-1}{m}\rho_{ss}\mathrm{var}(s_{ij}) = p_s(1-p_s)(\frac{1}{m} + \frac{m-1}{m}\rho_{ss})$$

and

$$\mathrm{var}(r_i) = var(\frac{1}{m(m-1)/2}\sum_{j_1=1}^{m-1}\sum_{j_2=j+1}^{m} r_{ij_1j_2})$$

$$= \Big\{\frac{1}{m(m-1)/2}\Big\}^2\Big\{\frac{m(m-1)}{2}\mathrm{var}(r_{i12}) + m(m-1)(m-2)\mathrm{cov}(r_{i12}, r_{i13})$$

$$+\frac{m(m-1)(m-2)(m-3)}{4}\mathrm{cov}(r_{i12}, r_{i34})\Big\}$$

$$= (\frac{2}{m(m-1)}\mathrm{var}(r_{i12}) + \frac{4(m-2)}{m(m-1)}\rho_{r1}\mathrm{var}(r_{i12}) + \frac{(m-2)(m-3)}{m(m-1)}\rho_{r2}\mathrm{var}(r_{i12}))$$

$$= p_r(1-p_r)\Big\{\frac{2}{m(m-1)} + \frac{4(m-2)}{m(m-1)}\rho_{r1} + \frac{(m-2)(m-3)}{m(m-1)}\rho_{r2}\Big\}$$

since $s_{ij} \sim \mathrm{Bernoulli}(p_s)$ and $r_{ij_1j_2} \sim \mathrm{Bernoulli}(p_r)$.

Appendix A.3. Derivation of ρ_2

Since $x_i = m^{-1}\sum_{j=1}^{m} x_{ij}$ and $y_i = m^{-1}\sum_{j=1}^{m} y_{ij}$,

$$\rho_2 = \mathrm{corr}(x_i, y_i) = \frac{\mathrm{cov}(x_i, y_i)}{\sqrt{\mathrm{var}(x_i)\mathrm{var}(y_i)}}$$

Here,

$$\mathrm{cov}(x_i, y_i) = \mathrm{cov}\Big(\frac{\sum_{j=1}^{m} x_{ij}}{m}, \frac{\sum_{j=1}^{m} y_{ij}}{m}\Big) = \mathrm{cov}(x_{ij}, y_{ij})$$

$$\mathrm{var}(x_i) = \frac{1}{m}\mathrm{var}(x_{i1}) + \frac{m-1}{m}\mathrm{cov}(x_{i1}, x_{i2})$$

and, similarly,

$$\mathrm{var}(y_i) = \frac{1}{m}\mathrm{var}(y_{i1}) + \frac{m-1}{m}\mathrm{cov}(y_{i1}, y_{i2})$$

Hence,

$$\rho_2 = \frac{\mathrm{cov}(x_{i1}, y_{i1})}{\sqrt{\{\frac{1}{m}\mathrm{var}(x_{i1}) + \frac{m-1}{m}\mathrm{cov}(x_{i1}, x_{i2})\}\{\frac{1}{m}\mathrm{var}(y_{i1}) + \frac{m-1}{m}\mathrm{cov}(y_{i1}, y_{i2})\}}}$$

$$= \frac{\rho_{xy}}{\sqrt{(\frac{1}{m} + \frac{m-1}{m}\rho_{xx})(\frac{1}{m} + \frac{m-1}{m}\rho_{yy})}}$$

Appendix A.4. The Limit of $\hat{\sigma}_2^2$ under $\bar{H}_2$

The limit of $\hat{\sigma}_2^2 = n^{-1}\sum_{i=1}^{n}(x_i - y_i - \delta_2)^2$ is $\sigma_2^2 = E(x_i - y_i - \delta_2)^2$. Since $E(x_i - y_i - \delta_2) = p_x - p_y - \delta_2 = 0$ under $\bar{H}_2 : p_x = p_y + \delta_0$, $E(x_i - y_i - \delta_2)^2 = \mathrm{var}(x_i - y_i - \delta_2) = \mathrm{var}(x_i - y_i) = \mathrm{var}(x_i) + \mathrm{var}(y_i) - 2\rho_2\sqrt{\mathrm{var}(x_i)\mathrm{var}(y_i)}$. Here, $\rho_2 = \mathrm{corr}(x_i, y_i)$,

$$\mathrm{var}(x_i) = \frac{1}{m^2}\{m\mathrm{var}(x_{i1}) + m(m-1)\mathrm{cov}(x_{i1}, x_{i2})\}$$

$$= \frac{\mathrm{var}(x_{i1})}{m} + \frac{m-1}{m}\rho_{xx}\mathrm{var}(x_{i1})) = p_x(1 - p_x)\Big(\frac{1}{m} + \frac{m-1}{m}\rho_{xx}\Big)$$

and, similarly,

$$\mathrm{var}(y_i) = p_y(1 - p_y)\Big(\frac{1}{m} + \frac{m-1}{m}\rho_{yy}\Big)$$

$$= (p_x - \delta_0)(1 - p_x + \delta_0)\Big(\frac{1}{m} + \frac{m-1}{m}\rho_{yy}\Big)$$

since $x_{ij} \sim Bernoulli(p_x), y_{ij} \sim Bernoulli(p_y)$.

References

1. Zhang, Z.; Sejdic, E. Radiological images and machine learning: trends, perspectives, and prospects. *Comput. Biol. Med.* **2019**, *108*, 354–370.
2. DSickles, E.A.;D'Orsi, C.J.;Bassett, L.W.; Appleton, C.M.; Berg, W.A.; Burnside, E.S. *ACR BI-RADS®Atlas, Breast Imaging Reporting and Data System*; American College of Radiology: Reston, VA, USA, 2013.
3. Wu, W.J., Lin, S.W.; Moon, W.K. Combining support vector machine with genetic algorithm to classify ultrasound breast tumor images. *Comput. Med. Imaging Graph.* **2012**, *36*, 627–633.
4. Liu, B.; Cheng, H.D.; Huang, J.; Tian, J.; Tang, X.; Liu, J. Fully automatic and segmentation-robust classification of breast tumors based on local texture analysis of ultrasound images. *Pattern Recogn.* **2010**, *43*, 280–298.
5. Shan, J.; Cheng, H.D.; Wang, Y. Completely automated segmentation approach for breast ultrasound images using multiple-domain features. *Ultrasound Med. Biol.* **2012**, *38*, 262–275.
6. Cheng, H.D.; Shan, J.; Ju, W.; Guo, Y.; Zhang, L. Automated breast cancer detection and classification using ultrasound images: A survey. *Pattern Recogn.* **2010**, *43*, 299–317.
7. Wu, G.G.; Zhou, L.Q.; Xu, J.W.; Wang, J.Y.; Wei, Q.; Deng, Y.B.; Cui, X.W.; Dietrich, C.F. Artificial intelligence in breast ultrasound. *World J. Radiol.* **2019**, *11*, 19–26.

8. O'Connell, A.M. Diagnostic Performance of An Artificial Intelligence System in Breast Ultrasound. *J. Ultrasound Med.* **2021**, doi:10.1002/jum.15684.
9. Liang, K.Y.; Zeger, S. Longitudinal data analysis using generalized linear models. *Biometrika* **1986**, *73*, 13–22.
10. Emrich, I.J.; Piedmonte, M.R. A method for generating high dimensional multivariate binary variables. *Am. Stat.* **1991**, *45*, 302–304.
11. Jung, S.H.; Barnhart, H.X.; Sohn, I.; Stinnett, S.S.; Wallace, D.K. Sample Size for Comparing Correlated Concordance Rates. *J. Biopharm. Stat.* **2008**, *18*, 359–369.
12. Qureshi, A.; Lakhtakia, R.; Bahri, M.A.; Al Haddabi, I.; Saparamadu, A.; Shalaby, A.; Al Riyami, M.; Rizvi, G. Gleason's Grading of Prostatic Adenocarcinoma: Inter-Observer Variation Among Seven Pathologists at a Tertiary Care Center in Oman. *Asian Pac. J. Cancer Prev.* **2016**, *17*, 4867–4868.