Chapter 7
Statistical Inference for Non-inferiority of a Diagnostic Procedure Compared to an Alternative Procedure, Based on the Difference in Correlated Proportions from Multiple Raters

Hiroyuki Saeki and Toshiro Tango

Abstract In a clinical trial of diagnostic procedures to indicate non-inferiority, the efficacy is generally evaluated on the basis of the results from multiple raters who interpret and report their findings independently. Although we can handle the multiple results from the multiple raters as if there were a single rater by considering consensus evaluations or majority votes, this handling is not recommended for the primary evaluation. Therefore, all results from the multiple independent raters should be used in the analysis. This chapter addresses a non-inferiority test, confidence interval and sample size formula, for inference of the difference in correlated proportions between the two diagnostic procedures based on the multiple raters. Moreover, we illustrate the methods with data from studies of diagnostic procedures for the diagnosis of oesophageal carcinoma infiltrating the tracheobronchial tree and for the diagnosis of aneurysm in patients with acute subarachnoid hemorrhage.

7.1 Introduction

In situations where an accepted standard diagnostic procedure exists, it is possible to plan a clinical trial to confirm that a new diagnostic procedure is superior to the standard diagnostic procedure. However, if it will be expected that the efficacy of the new diagnostic procedure is not lower than that of the standard diagnostic procedure and the new diagnostic procedure is less or non-invasive, less or non-toxic, inexpensive or easy to operate in comparison with the standard

H. Saeki (✉)
FUJIFILM RI Pharma Co. LTD., Chuo-ku, Tokyo, Japan
e-mail: sahiroyuki@ffri.co.jp

T. Tango
Center for Medical Statistics, Minato-ku, Tokyo, Japan
e-mail: tango@medstat.jp

K. van Montfort et al. (eds.), *Developments in Statistical Evaluation of Clinical Trials*, DOI 10.1007/978-3-642-55345-5_7

procedure, we can plan a non-inferiority study. A non-inferiority study of two diagnostic procedures is designed to indicate that the sensitivity or specificity of the new diagnostic procedure is no more than 100Δ percent inferior compared with the sensitivity or specificity of the standard procedure, respectively, where $\Delta (0 < \Delta \leq 1)$ is a pre-specified acceptable difference between the two proportions. In general, sensitivity is defined as the probability that a result of a diagnostic procedure is positive when the subject has the disease, and specificity is defined as the probability that a result of a diagnostic procedure is negative when the subject does not have the disease. These two measures are very important to evaluate the performance of the diagnostic procedure. However, these measures are calculated on the basis of different populations of subjects. Therefore, we consider the statistical inference for the difference in sensitivities in this chapter. However, the same methods can be applied to examine the difference in the specificities using a different study population.

If two diagnostic procedures are performed on each subject, the difference in proportions for matched-pair data has a correlation between the two diagnostic procedures. Nam [10] and Tango [17] derived the same non-inferiority test for the difference in proportions for matched-pair categorical data based on the efficient score in which the pairs were independent. Tango [17] also derived the confidence interval based on the efficient score. However, these methods are only applicable to the case where the results of the two diagnostic procedures are evaluated by a single rater. Multiple independent raters often evaluate the diagnoses obtained from these diagnostic procedures (see, e.g., [6]). If multiple raters are involved in the evaluation, the differences in proportions for matched-pair data also have correlations between different raters. Although we can apply the aforementioned methods by considering consensus evaluations or majority votes to handle multiple results from the multiple raters as if there were a single rater, these methods are not recommended for the primary evaluation [1, 2, 12]. The consensus evaluations may produce a bias caused by non-independent evaluations. For example, senior or persuasive raters may affect the evaluations of junior or passive raters. Moreover, the majority votes cannot take into account the variability in results of the multiple raters. Therefore, all results from the multiple independent raters should be used in the analysis.

In this chapter, we introduce a non-inferiority test, confidence interval and sample size formula proposed by Saeki and Tango [14], for inference of the difference in correlated proportions between two diagnostic procedures on the basis of the results from the multiple independent raters where the matched pairs are independent. Furthermore, we consider a possible procedure based on majority votes and we conduct Monte Carlo simulation studies to examine the validity of the proposed methods in comparison with the procedure based on majority votes. Finally, we illustrate the methods with data from studies of diagnostic procedures for the diagnosis of oesophageal carcinoma infiltrating the tracheobronchial tree [13] and for the diagnosis of aneurysm in patients with acute subarachnoid hemorrhage [4].

7.2 Design

7.2.1 Data Structure and Model

Consider a clinical experimental design where a new diagnostic procedure (or treatment) and a standard diagnostic procedure (or treatment) that are independently performed on the same subject (or matched pairs of subjects) and independently evaluated by K raters are compared. Each rater's judgment is assumed to take on one of two values: 1 represents that the subject is diagnosed as 'positive', and 0 indicates that the subject is diagnosed as 'negative'. Suppose we have n subjects. If we consider only subjects with a pre-specified disease, we use a positive probability as a measure, that is, sensitivity. On the other hand, if we consider subjects without the disease, we use a negative probability as a measure, that is, specificity. In the following, we consider a situation on the basis of sensitivity.

For ease of explanation, let us consider the case of $K = 2$ first. The resulting types of matched observations and probabilities are naturally classified as a 4×4 contingency table shown in Table 7.1, where $+(1)$ or $-(0)$ denotes a positive or negative judgment on a procedure, respectively. For example, y_{1101} denotes the observed number of matched type $\{+$ on the new procedure by rater 1, $+$ on the new procedure by rater 2, $-$ on the standard procedure by rater 1, $+$ on the standard procedure by rater 2$\}$ and r_{1101} indicates its probability.

Let $\pi_N^{(k)}$ ($\pi_S^{(k)}$) denote the probability that rater k judges as positive on the new (standard) diagnostic procedure of a randomly selected subject. Then, it will be naturally calculated as

$$\pi_N^{(1)} = r_{11\cdot\cdot} + r_{10\cdot\cdot}, \quad \pi_N^{(2)} = r_{11\cdot\cdot} + r_{01\cdot\cdot} \tag{7.1}$$

Table 7.1 A 4×4 contingency table for matched-pair categorical data in the case of two raters

		Standard procedure				
	Judgment of (Rater 1, Rater 2)	(+,+)	(+,−)	(−,+)	(−,−)	Total
New procedure	(+,+)	r_{1111}	r_{1110}	r_{1101}	r_{1100}	$r_{11\cdot\cdot}$
		(y_{1111})	(y_{1110})	(y_{1101})	(y_{1100})	$(y_{11\cdot\cdot})$
	(+, −)	r_{1011}	r_{1010}	r_{1001}	r_{1000}	$r_{10\cdot\cdot}$
		(y_{1011})	(y_{1010})	(y_{1001})	(y_{1000})	$(y_{10\cdot\cdot})$
	(−, +)	r_{0111}	r_{0110}	r_{0101}	r_{0100}	$r_{01\cdot\cdot}$
		(y_{0111})	(y_{0110})	(y_{0101})	(y_{0100})	$(y_{01\cdot\cdot})$
	(−, −)	r_{0011}	r_{0010}	r_{0001}	r_{0000}	$r_{00\cdot\cdot}$
		(y_{0011})	(y_{0010})	(y_{0001})	(y_{0000})	$(y_{00\cdot\cdot})$
	Total	$r_{\cdot\cdot 11}$	$r_{\cdot\cdot 10}$	$r_{\cdot\cdot 01}$	$r_{\cdot\cdot 00}$	1
		$(y_{\cdot\cdot 11})$	$(y_{\cdot\cdot 10})$	$(y_{\cdot\cdot 01})$	$(y_{\cdot\cdot 00})$	(n)

and $\pi_S^{(1)}$ and $\pi_S^{(2)}$ are defined in a similar manner. Let π_N and π_S denote the probability of a positive judgment on the new and standard diagnostic procedures, respectively. Then, these probabilities can, in general, be defined as follows:

$$\pi_N = \omega^{(1)}\pi_N^{(1)} + \omega^{(2)}\pi_N^{(2)} , \tag{7.2}$$

$$\pi_S = \omega^{(1)}\pi_S^{(1)} + \omega^{(2)}\pi_S^{(2)} , \tag{7.3}$$

where $\omega^{(k)}$ ($\omega^{(1)} + \omega^{(2)} = 1$) denotes the weight for rater k, showing the difference in the raters' evaluation skill. However, raters are usually selected among the raters with *at least equivalent skill*, and it is assumed in this paper that

$$\omega^{(k)} = 1/K \quad (k = 1, \ldots, K) . \tag{7.4}$$

Therefore, these probabilities can be defined as follows:

$$\pi_N = \frac{\pi_N^{(1)} + \pi_N^{(2)}}{2} = r_{11\cdot\cdot} + \frac{r_{10\cdot\cdot} + r_{01\cdot\cdot}}{2} , \tag{7.5}$$

$$\pi_S = \frac{\pi_S^{(1)} + \pi_S^{(2)}}{2} = r_{\cdot\cdot 11} + \frac{r_{\cdot\cdot 10} + r_{\cdot\cdot 01}}{2} . \tag{7.6}$$

On the basis of the form of the expressions of (7.5) and (7.6), the 4×4 contingency table is found to be reduced to the 3×3 contingency table shown in Table 7.2, where $p_{\ell m}$ ($x_{\ell m}$) denotes the probability (observed number of observations) that ℓ raters judge as positive on the new procedure and m raters judge as positive on the standard procedure. Then, we have

$$\begin{aligned} \pi_N &= p_{2\cdot} + \frac{1}{2}p_{1\cdot} \\ &= p_{20} + (p_{21} + \frac{1}{2}p_{10}) + (p_{22} + \frac{1}{2}p_{11}) + \frac{1}{2}p_{12} , \end{aligned} \tag{7.7}$$

Table 7.2 A 3×3 contingency table for matched-pair categorical data in the case of two raters

		Standard procedure			
	Judgment of (Rater 1, Rater 2)	(+, +)	(+,−) or (−,+)	(−,−)	Total
New procedure	(+, +)	p_{22}	p_{21}	p_{20}	$p_{2\cdot}$
		(x_{22})	(x_{21})	(x_{20})	($x_{2\cdot}$)
	(+, −) or (−, +)	p_{12}	p_{11}	p_{10}	$p_{1\cdot}$
		(x_{12})	(x_{11})	(x_{10})	($x_{1\cdot}$)
	(−, −)	p_{02}	p_{01}	p_{00}	$p_{0\cdot}$
		(x_{02})	(x_{01})	(x_{00})	($x_{0\cdot}$)
	Total	$p_{\cdot 2}$	$p_{\cdot 1}$	$p_{\cdot 0}$	1
		($x_{\cdot 2}$)	($x_{\cdot 1}$)	($x_{\cdot 0}$)	(n)

$$\begin{aligned}\pi_S &= p_{\cdot 2} + \frac{1}{2} p_{\cdot 1} \\ &= p_{02} + (p_{12} + \frac{1}{2} p_{01}) + (p_{22} + \frac{1}{2} p_{11}) + \frac{1}{2} p_{21} \ . \end{aligned} \tag{7.8}$$

Let λ denote the difference in positive probabilities; that is,

$$\begin{aligned}\lambda &= \pi_N - \pi_S \\ &= p_{20} + \frac{1}{2}(p_{21} + p_{10}) - p_{02} - \frac{1}{2}(p_{12} + p_{01}) \ , \end{aligned} \tag{7.9}$$

and its sample estimate will be

$$\tilde{\lambda} = \frac{1}{n} \left\{ x_{20} + \frac{1}{2}(x_{21} + x_{10}) - x_{02} - \frac{1}{2}(x_{12} + x_{01}) \right\} \ , \tag{7.10}$$

which clearly shows that the inference on λ can be made by the observed vector $\boldsymbol{x} = (x_{20}, x_{21} + x_{10}, x_{02}, x_{12} + x_{01}, x_{22} + x_{11} + x_{00})$ following a multinomial distribution with parameters n and $\boldsymbol{p} = (p_{20}, p_{21} + p_{10}, p_{02}, p_{12} + p_{01}, p_{22} + p_{11} + p_{00})$.

It should be noted that x_{20} is the frequency such that the number of raters judging as positive on the new procedure is larger than the number of raters judging as positive on the standard procedure by 2 and that $(x_{21} + x_{10})$ is the frequency such that the number of raters judging as positive on the new procedure is larger than the number of raters judging as positive on the standard procedure by 1. Similarly, x_{02} is the frequency such that the number of raters judging as positive on the standard procedure is larger than the number of raters judging as positive on the new procedure by 2 and $(x_{12} + x_{01})$ is the frequency such that the number of raters judging as positive on the standard procedure is larger than the number of raters judging as positive on the new procedure by 1. These observations lead to a generalization to K raters. The resulting types of matched observations and probabilities are classified as a $(K + 1) \times (K + 1)$ contingency table similar to Table 7.2. However, the method is reduced to the following. Let n_{Nk} denote the frequency such that the number of raters who judge as positive on the new procedure is larger than the number of raters who judge as positive on the standard procedure by k and let q_{Nk} indicate such probability. Namely, we have

$$\begin{aligned} n_{Nk} &= \sum_{\ell - m = k} x_{\ell m} \ , \\ q_{Nk} &= \sum_{\ell - m = k} p_{\ell m} \ , \end{aligned}$$

where ℓ is the number of raters who judge as positive on the new procedure, and m is the number of raters who judge as positive on the standard procedure. Similarly,

let n_{Sk} denote the frequency such that the number of raters who judge as positive on the standard procedure is larger than the number of raters who judge as positive on the new procedure by k and let q_{Sk} indicate such probability. Then, we have

$$n_{Sk} = \sum_{\ell - m = -k} x_{\ell m} \ ,$$

$$q_{Sk} = \sum_{\ell - m = -k} p_{\ell m} \ ,$$

and $q_{N0} = q_{S0}$ and $n_{N0} = n_{S0}$. Namely, for K raters, the inference on λ can be made by the vector of random variables $\boldsymbol{n} = (n_{N0}, n_{N1}, \ldots, n_{NK}, n_{S1}, \ldots, n_{SK})$ following a multinomial distribution with parameters n and $\boldsymbol{q} = (q_{N0}, q_{N1}, \ldots, q_{NK}, q_{S1}, \ldots, q_{SK})$. Then, we have

$$\begin{aligned}
\pi_N &= \sum_{k=1}^{K} \omega^{(k)} \pi_N^{(k)} = \frac{1}{K} \sum_{k=1}^{K} k \sum_{m=0}^{K} p_{km} = \frac{1}{K} \sum_{k=1}^{K} k p_{k\cdot} \\
&= \frac{1}{K} \sum_{k=1}^{K} k q_{Nk} + \frac{1}{K} \sum_{k=1}^{K} k p_{kk} + \frac{1}{K} \sum_{\substack{\ell, m \in K \\ \ell < m}} \ell p_{\ell m} + \frac{1}{K} \sum_{\substack{\ell, m \in K \\ m < \ell}} m p_{\ell m} \ , \\
\pi_S &= \sum_{k=1}^{K} \omega^{(k)} \pi_S^{(k)} = \frac{1}{K} \sum_{k=1}^{K} k \sum_{\ell=0}^{K} p_{\ell k} = \frac{1}{K} \sum_{k=1}^{K} k p_{\cdot k} \\
&= \frac{1}{K} \sum_{k=1}^{K} k q_{Sk} + \frac{1}{K} \sum_{k=1}^{K} k p_{kk} + \frac{1}{K} \sum_{\substack{\ell, m \in K \\ \ell < m}} \ell p_{\ell m} + \frac{1}{K} \sum_{\substack{\ell, m \in K \\ m < \ell}} m p_{\ell m} \ . \qquad (7.11)
\end{aligned}$$

Therefore, the difference in positive probabilities (7.9) is generalized to

$$\begin{aligned}
\lambda = \pi_N - \pi_S &= \Big(\frac{1}{K} \sum_{k=1}^{K} k p_{k\cdot} \Big) - \Big(\frac{1}{K} \sum_{k=1}^{K} k p_{\cdot k} \Big) \\
&= \frac{1}{K} \sum_{k=1}^{K} k (q_{Nk} - q_{Sk}) \ . \qquad (7.12)
\end{aligned}$$

Then, the estimate $\tilde{\lambda}$ given in (7.10) is generalized to

$$\tilde{\lambda} = \frac{1}{nK} \sum_{k=1}^{K} k (n_{Nk} - n_{Sk}) \ . \qquad (7.13)$$

7.2.2 Problems in Consensus Evaluations or Majority Votes

Although we can handle multiple results from the multiple raters as if there were a single rater by considering consensus evaluations or majority votes, these handlings are not recommended for the primary evaluation [1,2,12]. The consensus evaluations may produce a bias caused by non-independent evaluation, even if the consensus evaluations are performed after individual evaluations by the multiple raters are completed. For example, senior or persuasive raters may affect the evaluations of junior or passive raters. Moreover, the majority votes cannot take into account the variability in results of the multiple raters. For ease of explanation, let us consider the case of $K = 3$. The resulting types of matched observations are classified as a 4×4 contingency table in Table 7.3. In this case, $\tilde{\lambda}_{K=3}$ can be addressed from (7.13) as

$$\tilde{\lambda}_{K=3} = \frac{1}{n}\left\{(n_{N3} - n_{S3}) + \frac{2}{3}(n_{N2} - n_{S2}) + \frac{1}{3}(n_{N1} - n_{S1})\right\} ,$$

where $(n_{N3} - n_{S3}) = (x_{30} - x_{03})$, $(n_{N2} - n_{S2}) = \{(x_{31} + x_{20}) - (x_{13} + x_{02})\}$ and $(n_{N1} - n_{S1}) = \{(x_{32} + x_{21} + x_{10}) - (x_{23} + x_{12} + x_{01})\}$. If we adopt the majority votes, the 4×4 contingency table shown in Table 7.3 is transformed to the 2×2 contingency table shown in Table 7.4, and the estimate of the difference between π_N and π_S on the basis of the results from the majority votes will be

$$\tilde{\lambda}_{MV} = \frac{(b-c)}{n} = \frac{1}{n}\{(n_{N3} - n_{S3}) + (n_{N2} - n_{S2}) + (x_{21} - x_{12})\} .$$

We should focus on two problems in $\tilde{\lambda}_{MV}$.

Table 7.3 A 4×4 contingency table for matched-pair categorical data in the case of three raters

		Standard procedure			
	Judgment of (Rater 1, Rater 2, Rater 3)	(+, +, +)	(+, +, −) or (+, −, +) or (−, +, +)	(+, −, −) or (−, +, −) or (−, −, +)	(−, −, −)
New procedure	(+, +, +)	x_{33}	x_{32}	x_{31}	x_{30}
	(+, +, −) or (+, −, +) or (−, +, +)	x_{23}	x_{22}	x_{21}	x_{20}
	(+, −, −) or (−, +, −) or (−, −, +)	x_{13}	x_{12}	x_{11}	x_{10}
	(−, −, −)	x_{03}	x_{02}	x_{01}	x_{00}

Table 7.4 A 2×2 contingency table transformed from Table 7.3 by majority votes

		Standard procedure	
	Judgment	(+)	(−)
New procedure	(+)	a	b
		$(= x_{33} + x_{32} + x_{23} + x_{22})$	$(= x_{30} + x_{31} + x_{20} + x_{21})$ $(= n_{N3} + n_{N2} + x_{21})$
	(−)	c	d
		$(= x_{03} + x_{13} + x_{02} + x_{12})$ $(= n_{S3} + n_{S2} + x_{12})$	$(= x_{11} + x_{10} + x_{01} + x_{00})$

1. $\tilde{\lambda}_{MV}$ involves $(n_{N2} - n_{S2})$ and $(x_{21} - x_{12})$ without the weights of the contribution for π_N and π_S from $\pi_N^{(1)}$, $\pi_N^{(2)}$, $\pi_N^{(3)}$ and $\pi_S^{(1)}$, $\pi_S^{(2)}$, $\pi_S^{(3)}$.
2. x_{32}, x_{10} and x_{23}, x_{01} do not take part in $\tilde{\lambda}_{MV}$, because these values are involved in the cells 'a' and 'd' in Table 7.4.

Therefore, it is important that all results from the multiple independent raters are used in the analysis appropriately.

7.3 Methods for Statistical Inference

In this section, we shall introduce methods for statistical inference of the difference λ, that is, a non-inferiority test, confidence interval and formula for determination of sample size.

7.3.1 Non-inferiority Test

The non-inferiority hypothesis will be formulated as

$$H_0 : \pi_N = \pi_S - \Delta, \; H_1 : \pi_N > \pi_S - \Delta \, ,$$

where Δ $(0 < \Delta \leq 1)$ is a pre-specified acceptable difference in two probabilities. Let

$$\delta = \lambda + \Delta = \pi_N - (\pi_S - \Delta) = \frac{1}{K} \sum_{k=1}^{K} k q_{Nk} - \left(\frac{1}{K} \sum_{k=1}^{K} k q_{Sk} - \Delta \right) . \qquad (7.14)$$

Then, under the null hypothesis, the log-likelihood function without constant terms is expressed as

$$L = L(\boldsymbol{\theta}) = n_{N0}\log(q_{N0}) + n_{NK}\log(q_{NK}) + \sum_{k=1}^{K-1} n_{Nk}\log(q_{Nk}) + \sum_{k=1}^{K} n_{Sk}\log(q_{Sk})$$
$$= n_{N0}\log(1-\delta+\Delta-A-B-C) + n_{NK}\log(\delta-\Delta+A)$$
$$+ \sum_{k=1}^{K-1} n_{Nk}\log(q_{Nk}) + \sum_{k=1}^{K} n_{Sk}\log(q_{Sk}) \ , \tag{7.15}$$

where $\boldsymbol{\theta} = (\delta, q_{N1}, \ldots, q_{N(K-1)}, q_{S1}, \ldots, q_{SK})^T$ is the parameter vector of dimension $2K$ and

$$A = \frac{1}{K}\Big(\sum_{k=1}^{K} k q_{Sk} - \sum_{k=1}^{K-1} k q_{Nk}\Big), \quad B = \sum_{k=1}^{K-1} q_{Nk} \ , \quad C = \sum_{k=1}^{K} q_{Sk} \ .$$

Then, the score test for testing the null hypothesis $H_0 : \delta = 0$ against $H_1 : \delta > 0$ is expressed as

$$Z_S = \left[\frac{\partial L}{\partial \delta}\Big|_{\delta=0,\, q_{Nk}=\hat{q}_{Nk},\, q_{Sk}=\hat{q}_{Sk}}\right] \sqrt{\left(\hat{I}^{-1}\right)_{11}\Big|_{\delta=0,\, q_{Nk}=\hat{q}_{Nk},\, q_{Sk}=\hat{q}_{Sk}}} \sim_{H_0} N(0,1) \ , \tag{7.16}$$

where $(\hat{q}_{N1}, \ldots, \hat{q}_{N(K-1)}, \hat{q}_{S1}, \ldots, \hat{q}_{SK})$ is the vector of the maximum likelihood estimators under the null hypothesis, which is the unique solution for the following equations:

$$\frac{\partial L}{\partial q_{Nk}}\Big|_{\delta=0} = 0, \quad (k = 1, \ldots, K-1) \ , \tag{7.17}$$

$$\frac{\partial L}{\partial q_{Sk}}\Big|_{\delta=0} = 0, \quad (k = 1, \ldots, K) \ . \tag{7.18}$$

These equations can be obtained iteratively using the quasi-Newton method with constraints. The R function 'constrOptim' is useful for the quasi-Newton method with constraints. Further, $(\hat{I}^{-1})_{11}$ indicates the $(1,1)$th element of the $(2K \times 2K)$ inverse Fisher information matrix evaluated at the maximum likelihood estimators. On the other hand, we can consider a test based on the sample estimate T for the difference δ

$$T = \tilde{\lambda} + \Delta = \frac{1}{nK}\sum_{k=1}^{K} k(n_{Nk} - n_{Sk}) + \Delta \ . \tag{7.19}$$

The variance of T evaluated at the null hypothesis $\delta = 0$ is

$$\mathrm{Var}_{H_0}(T) = \frac{1}{n}\left[\frac{1}{K^2}\sum_{k=1}^{K} k^2(q_{Nk} + q_{Sk}) - \Delta^2\right] \ .$$

Therefore, the normal deviate for testing $H_0 : \delta = 0$ against $H_1 : \delta > 0$ is expressed as

$$Z_{ND} = \frac{\frac{1}{nK}\sum_{k=1}^{K} k(n_{Nk} - n_{Sk}) + \Delta}{\sqrt{\frac{1}{n}\left[\frac{1}{K^2}\sum_{k=1}^{K} k^2(\hat{q}_{Nk} + \hat{q}_{Sk}) - \Delta^2\right]}} \sim_{H_0} N(0,1) . \tag{7.20}$$

It can be shown that when $K = 1$, the normal deviate test statistic, Z_{ND}, is equivalent to the score test statistic Z_S [10, 17]. When $K = 2$ or 3, we confirmed that Z_S and Z_{ND} were approximately equal using the example data (see Sect. 7.5). However, we have not been able to show the equivalence between Z_S and Z_{ND} analytically. On the other hand, by using the observed proportions $\tilde{q}_{Nk} = n_{Nk}/n$, $\tilde{q}_{Sk} = n_{Sk}/n$ instead of the maximum likelihood estimators, we can construct a Wald-type test statistic for testing $H_0 : \delta = 0$ against $H_1 : \delta > 0$:

$$Z_W = \frac{\frac{1}{nK}\sum_{k=1}^{K} k(n_{Nk} - n_{Sk}) + \Delta}{\sqrt{\frac{1}{n}\left[\frac{1}{nK^2}\sum_{k=1}^{K} k^2(n_{Nk} + n_{Sk}) - \Delta^2\right]}} \sim_{H_0} N(0,1) . \tag{7.21}$$

When $\Delta = 0$, the Wald-type test Z_W is identical to Schouten's [15] generalized McNemar test although Schouten's test statistic is presented in a different form. When $K = 1$, the Wald-type test Z_W is identical to the unconditional test for non-inferiority of Lu and Bean [7]. When $\Delta = 0$ and $K = 1$, both the normal deviate test Z_{ND} and the Wald-type test Z_W are identical to the McNemar test [9].

7.3.2 *Confidence Interval*

Testing non-inferiority with an acceptable difference Δ at a one-sided significance level $\alpha/2$ is equivalent to judging whether the lower limit of the $1 - \alpha$ level confidence interval is greater than $-\Delta$. The score-type approximate confidence limits for the difference in two proportions, λ, are the two solutions to the equation

$$\frac{\frac{1}{nK}\sum_{k=1}^{K} k(n_{Nk} - n_{Sk}) - \lambda}{\sqrt{\frac{1}{n}\left[\frac{1}{K^2}\sum_{k=1}^{K} k^2(\hat{q}_{Nk} + \hat{q}_{Sk}) - \lambda^2\right]}} = \pm Z_{\alpha/2} , \tag{7.22}$$

where the plus and minus signs indicate the lower limit λ_{low} and the upper limit λ_{up}, respectively, and $Z_{\alpha/2}$ is the upper $\alpha/2$ percentile of the standard normal distribution. These two limits can be found using an iterative numerical method such as the secant method (see, e.g., [17]). On the other hand, we can easily derive the Wald-type confidence interval:

$$\mathrm{CI}_W : \frac{1}{nK}\left(\sum_{k=1}^{K} k(n_{Nk} - n_{Sk}) \pm Z_{\alpha/2}\sqrt{\sum_{k=1}^{K} k^2(n_{Nk} + n_{Sk})}\right) . \tag{7.23}$$

Equation (7.23) utilizes the variance evaluated under the null hypothesis and is identical to Schouten's [15] Wald-type confidence interval.

7.3.3 Sample Size

To calculate the sample size required for testing the null hypothesis $H_0 : \delta = 0$ against the alternative hypothesis $H_1 : \delta > 0$, we only have to consider the following properties of the statistic T:

$$\begin{aligned} E_{H_0}(T) &= 0 , \\ E_{H_1}(T) &= \lambda + \Delta , \\ S = \lim_{n\to\infty} n\mathrm{Var}_{H_1}(T) &= \left[\frac{1}{K^2}\sum_{k=1}^{K} k^2(q_{Nk} + q_{Sk}) - \lambda^2\right] . \end{aligned}$$

On the other hand, we have

$$R = \lim_{n\to\infty} n\mathrm{Var}_{H_0}(T) = \left[\frac{1}{K^2}\sum_{k=1}^{K} k^2(\bar{q}_{Nk} + \bar{q}_{Sk}) - \Delta^2\right] ,$$

where $(\bar{q}_{Nk}, \bar{q}_{Sk})$, $k = 0, \ldots, K$, are the asymptotic values of the maximum likelihood estimators $(\hat{q}_{Nk}, \hat{q}_{Sk})$, $k = 0, \ldots, K$. These asymptotic values are solutions to (7.17) and (7.18). From the aforementioned equations, the approximate sample size n required for $100(1-\beta)$ power of a one-sided normal deviate test at $\alpha/2$ level is given by

$$n = \left(\frac{Z_{\alpha/2}\sqrt{R} + Z_\beta\sqrt{S}}{\lambda + \Delta}\right)^2 . \tag{7.24}$$

When $K = 1$, the derived formula for determining the sample size agrees with that proposed by Nam [10]. The sample sizes required for 80 % power of a one-sided non-inferiority test at $\alpha/2 = 2.5\,\%$ for $K = 2, 3$, $\Delta = 0.1, 0.05$, and various values of $(q_{N3}, q_{N2}, q_{N1}, q_{S3}, q_{S2}, q_{S1})$ with $\pi_N - \pi_S = \lambda = 0$ are shown in Table 7.5.

Table 7.5 Sample sizes calculated by formula (7.24) for nominal power = 80 % of a non-inferiority test at $\alpha/2 = 2.5\,\%$ for $K = 2, 3$, $\Delta = 0.1, 0.05$, $\pi_N - \pi_S = \lambda = 0$, $q_{N3} = q_{S3}$, $q_{N2} = q_{S2}$, $q_{N1} = q_{S1}$

K	Δ	$q_{N3} = q_{S3}$	$q_{N2} = q_{S2}$	$q_{N1} = q_{S1}$	Sample size	
2	0.1	–	0.05	0.05	117	(81.7)
		–	0.05	0.1	132	(81.9)
		–	0.1	0.05	187	(80.7)
		–	0.1	0.1	204	(80.7)
	0.05	–	0.05	0.05	417	(80.6)
		–	0.05	0.1	487	(81.0)
		–	0.1	0.05	718	(80.8)
		–	0.1	0.1	793	(80.2)
3	0.1	0.05	0.02	0.05	120	(81.5)
		0.05	0.02	0.1	126	(81.5)
		0.05	0.05	0.1	142	(80.8)
		0.1	0.02	0.05	190	(80.4)
		0.1	0.02	0.1	197	(80.3)
		0.1	0.05	0.1	215	(79.8)
	0.05	0.05	0.02	0.05	428	(80.2)
		0.05	0.02	0.1	459	(80.0)
		0.05	0.05	0.1	536	(80.2)
		0.1	0.02	0.05	730	(81.1)
		0.1	0.02	0.1	763	(80.7)
		0.1	0.05	0.1	844	(80.5)

The parenthetical values are empirical power (%) based on 10,000 replicates

7.4 Simulation

We have indicated here the results of simulation studies for the methods at a one-sided 2.5 % level for the case of $K = 3$ and sample size $n = 25$, 50 or 100 with 10,000 replicates. Simulation data were generated on the basis of a multinomial distribution by considering typical situations for parameter values $(q_{N3}, q_{N2}, q_{N1}, q_{S3}, q_{S2}, q_{S1})$ and non-inferiority margin $\Delta = 0.1$. In assessing the performance of the methods based on the majority votes, we transformed the simulation data based on the following definitions: $q_N = (q_{N3} + q_{N2} + \frac{1}{3} \times q_{N1})$, $q_S = (q_{S3} + q_{S2} + \frac{1}{3} \times q_{S1})$.

7.4.1 Non-inferiority Test

We performed Monte Carlo simulation studies to assess the empirical size and power of the normal deviate test statistic Z_{ND}, the Wald-type test statistic Z_W and the test

Table 7.6 Empirical sizes of the normal deviate test Z_{ND}, the Wald-type test Z_W and the test based on majority votes Z_{MV} at $\alpha/2 = 2.5\,\%$ for $K = 3$, $\pi_N - \pi_S = \lambda = -0.1$, $\Delta = 0.1$ based on 10,000 replicates

							Size (%)		
n	q_{N3}	q_{N2}	q_{N1}	q_{S3}	q_{S2}	q_{S1}	Z_{ND}	Z_W	Z_{MV}
100	0.01	0.02	0.05	0.11	0.02	0.05	2.2	4.6	1.7
	0.01	0.02	0.1	0.11	0.02	0.1	2.3	4.3	1.3
	0.01	0.05	0.1	0.11	0.05	0.1	2.2	3.7	1.6
50	0.01	0.02	0.05	0.11	0.02	0.05	2.0	5.9	1.5
	0.01	0.02	0.1	0.11	0.02	0.1	2.2	5.5	1.3
	0.01	0.05	0.1	0.11	0.05	0.1	2.2	4.6	1.4
25	0.01	0.02	0.05	0.11	0.02	0.05	1.6	8.0	1.1
	0.01	0.02	0.1	0.11	0.02	0.1	1.9	7.3	0.9
	0.01	0.05	0.1	0.11	0.05	0.1	2.4	5.9	1.2

Table 7.7 Empirical powers of the normal deviate test Z_{ND}, the Wald-type test Z_W and the test based on majority votes Z_{MV} at $\alpha/2 = 2.5\,\%$ for $K = 3$, $\pi_N - \pi_S = \lambda = 0$, $\Delta = 0.1$ based on 10,000 replicates

							Power (%)		
n	q_{N3}	q_{N2}	q_{N1}	q_{S3}	q_{S2}	q_{S1}	Z_{ND}	Z_W	Z_{MV}
100	0.01	0.02	0.05	0.01	0.02	0.05	97.2	99.3	85.8
	0.01	0.02	0.1	0.01	0.02	0.1	95.7	98.4	78.6
	0.01	0.05	0.1	0.01	0.05	0.1	89.5	93.2	62.2
50	0.01	0.02	0.05	0.01	0.02	0.05	70.6	89.6	45.8
	0.01	0.02	0.1	0.01	0.02	0.1	68.4	85.2	38.1
	0.01	0.05	0.1	0.01	0.05	0.1	60.1	72.7	29.7
25	0.01	0.02	0.05	0.01	0.02	0.05	22.7	69.6	15.2
	0.01	0.02	0.1	0.01	0.02	0.1	22.9	65.5	10.0
	0.01	0.05	0.1	0.01	0.05	0.1	25.0	50.6	10.8

statistic based on the majority votes Z_{MV}. Z_{MV} was calculated using the method of Nam [10] and Tango [17]. Table 7.6 presents the empirical sizes. For the set of parameter values $(q_{N3}, q_{N2}, q_{N1}, q_{S3}, q_{S2}, q_{S1})$ considered here, the empirical sizes for the normal deviate test Z_{ND} are generally closer to the nominal $\alpha/2$-level of 2.5 % than those for the Wald-type test Z_W or the test based on the majority votes Z_{MV}. The empirical sizes of Z_W tend to be quite inflated. The empirical sizes of Z_{MV}, on the other hand, tend to be quite reduced. Table 7.7 presents the empirical powers for the alternative hypothesis $H_1 : \pi_N = \pi_S$ for the case of $\Delta = 0.1$. The differences in powers between Z_{ND} and Z_W are generally small. When the sample size is small, however, the empirical powers of Z_W are far greater than those of Z_{ND}. On the other hand, the empirical powers of Z_{MV} are far smaller than those of Z_{ND} under all situations.

Table 7.8 Coverage probabilities of the score-type 95 % confidence interval, the Wald-type 95 % confidence interval and the 95 % confidence interval based on the majority votes for $K = 3$ based on 10,000 replicates generated under the null hypothesis $\pi_N - \pi_S = \lambda = -0.1$

							Coverage prob. (%)		
n	q_{N3}	q_{N2}	q_{N1}	q_{S3}	q_{S2}	q_{S1}	score-type	CI_W	CI_{MV}
100	0.01	0.02	0.05	0.11	0.02	0.05	95.0	94.8	96.4
	0.01	0.02	0.1	0.11	0.02	0.1	94.9	94.9	97.3
	0.01	0.05	0.1	0.11	0.05	0.1	94.7	95.2	96.7
50	0.01	0.02	0.05	0.11	0.02	0.05	94.7	94.2	96.7
	0.01	0.02	0.1	0.11	0.02	0.1	94.7	94.4	97.7
	0.01	0.05	0.1	0.11	0.05	0.1	95.0	95.1	97.1
25	0.01	0.02	0.05	0.11	0.02	0.05	95.3	93.7	97.7
	0.01	0.02	0.1	0.11	0.02	0.1	95.4	93.9	98.4
	0.01	0.05	0.1	0.11	0.05	0.1	95.9	94.6	97.9

7.4.2 Confidence Interval

We performed Monte Carlo simulation studies to evaluate the coverage probability of the score-type confidence interval, the Wald-type confidence interval CI_W and the confidence interval based on the majority votes CI_{MV}. CI_{MV} was calculated using the method of Tango [17]. Table 7.8 shows the empirical coverage probabilities of the score-type 95 % confidence interval, the Wald-type 95 % confidence interval and the 95 % confidence interval based on the majority votes under the hypothesis $\pi_N - \pi_S = \lambda = -0.1$. It shows that the score-type confidence interval and the Wald-type confidence interval both generally perform very well. However, when $n = 25$, the score-type confidence interval outperforms the Wald-type confidence interval. On the other hand, the confidence interval based on the majority votes shows a conservative property.

7.5 Example

7.5.1 Study of Diagnostic Procedures for the Diagnosis of Oesophageal Carcinoma Infiltrating the Tracheobronchial Tree

Here, we shall consider the data presented by Rapp-Bernhardt et al. [13]. They compared the sensitivities between axial computed tomography (CT) slices and minimal intensity projection (MIP) in 21 patients with oesophageal carcinoma infiltrating the tracheobronchial tree. The bronchoscopic findings were determined as the gold standard. Three radiologists, working independently of each other and without knowledge of the findings on the gold standard, assessed separately the

Table 7.9 A 4 × 4 contingency table ($K = 3$) of the assessments of MIP and axial CT slices by three radiologists (True positive (TP: +) and false negative (FN: −) by three radiologists (1, 2, 3): I (+, +, +), II (+, +, − or +, −, + or −, +, +), III (+, −, − or −, +, − or −, −, +), IV (−, −, −)) (Rapp-Bernhardt et al. [13])

		Axial CT slices				
	TP and FN by three radiologists	I	II	III	IV	Total
MIP	I	14	2	1	0	17
	II	0	0	0	0	0
	III	0	0	2	0	2
	IV	0	0	2	0	2
	Total	14	2	5	0	21

CT, computed tomography; FN, false negative;
MIP, minimal intensity projection; TP, true positive

axial CT slices and MIP. In these diagnostic procedures, stenoses were localized, and the degree of stenosis was assessed as in real bronchoscopy. The resulting type of matched observations was classified as a 4 × 4 contingency table for MIP versus axial CT slices and is shown in Table 7.9 (similar to Table 7.3), where '+' indicates a true positive and '−' indicates a false negative based on binary assessment where 0–50 % of total occlusion was considered as negative and 50–100 % of total occlusion was considered as positive. MIP is one of the reconstruction techniques of making three-dimensional images. MIP images make it easier to appreciate the condition of the whole tracheobronchial tree than axial CT slices. Therefore, we are interested in the non-inferiority of MIP to axial CT slices where the non-inferiority margin is set as $\Delta = 0.1$. From Table 7.9, we have $\tilde{p}_{3\cdot} = 17/21$, $\tilde{p}_{2\cdot} = 0/21$, $\tilde{p}_{1\cdot} = 2/21$, $\tilde{p}_{\cdot 3} = 14/21$, $\tilde{p}_{\cdot 2} = 2/21$ and $\tilde{p}_{\cdot 1} = 5/21$. Then, the sensitivities of MIP and axial CT slices are estimated as $\tilde{\pi}_{MIP} = (17 + 2/3 \times 0 + 1/3 \times 2)/21 = 0.841$ and $\tilde{\pi}_{CT} = (14 + 2/3 \times 2 + 1/3 \times 5)/21 = 0.810$, respectively. Moreover, we have $\tilde{q}_{N3} = 0/21$, $\tilde{q}_{N2} = (1 + 0)/21$, $\tilde{q}_{N1} = (2 + 0 + 0)/21$, $\tilde{q}_{S3} = 0/21$, $\tilde{q}_{S2} = (0+0)/21$ and $\tilde{q}_{S1} = (0+0+2)/21$. Then, the difference in the sensitivities between MIP and axial CT slices based on the three raters is $\tilde{\lambda}_{K=3} = 0.032$, and the normal deviate test has $Z_{ND} = 1.753 \approx Z_S$ (one-sided p-value $= 0.040$). The score-type 95 % confidence interval is −0.141 to 0.181 where the lower limit is not greater than $-\Delta = -0.1$. These results suggest that the non-inferiority of MIP to axial CT slices cannot be claimed at the one-sided 2.5 % significance level. The Wald-type test statistic, on the other hand, suggests non-inferiority because $Z_W = 3.358$ with one-sided p-value <0.001 and because the Wald-type 95 % confidence interval under the null hypothesis is −0.056 to 0.120. However, the simulation study suggests that the Wald-type test result here is not reliable because of its inflated empirical sizes for a quite small sample size such as $n = 21$. The result of the normal-deviate test, on the other hand, may or may not be reliable because its empirical sizes for $\Delta = 0.1$ and $n = 25$ are shown to be around $1.6 \sim 2.4$.

7.5.2 Study of Diagnostic Procedures for the Diagnosis of Aneurysm in Patients with Acute Subarachnoid Hemorrhage

Jäger et al. [4] performed a blinded multi-rater study comparing magnetic resonance angiography (MRA) and digital subtraction angiography (DSA) in 34 prospectively enrolled patients who presented with acute subarachnoid hemorrhage (SAH). Two raters independently evaluated the MRA and DSA images. The presence of an aneurysm was evaluated on a 4-point ordinal scale (1, absent; 2, probably absent; 3, probably present; 4, definitely present). Additionally, all aneurysms for which the two raters had given different evaluations on the 4-point scale were subsequently reviewed by consensus evaluations. Because the authors intended to study the inter-rater and inter-procedure agreement, neither method was a priori taken as the gold standard. However, they showed the data of evaluation of the MRA and DSA images by the two raters with details of the clinical follow-up of all patients. Therefore, we considered comparing the difference in sensitivities between MRA and DSA on the basis of the data of 27 patients with aneurysms among the patients with SAH. Data were analyzed on a patient-basis, taking into account only the aneurysm with the highest ranking on the 4-point scale in each patient. We assigned the rating of true positive ('+') for scores of 3 and 4 or false negative ('−') for scores of 1 and 2. The resulting types of matched observations based on the two independent raters and the consensus evaluations were classified as a 3×3 and 2×2 contingency tables, respectively (Tables 7.10 and 7.11). DSA is a procedure in which radiographic images of blood vessels filled with a contrast agent are digitized and then subtracted from images obtained before administration of the contrast agent. This method increases the contrast between the vessels and the background. However, as a catheter (a long, thin, flexible tube) is inserted into an artery, DSA is considered to be invasive. MRA is a procedure to image blood vessels based on MRI. Unlike DSA that involves placing a catheter into the body, MRA is considered noninvasive. Therefore, we are interested in the non-inferiority of MRA to DSA where the non-inferiority margin is set as $\Delta = 0.1$. From Table 7.10 based on the multiple raters, we have $\tilde{p}_{2\cdot} = 20/27$, $\tilde{p}_{1\cdot} = 5/27$, $\tilde{p}_{\cdot 2} = 22/27$ and $\tilde{p}_{\cdot 1} = 2/27$. Then, the sensitivities of MRA and DSA are estimated as $\tilde{\pi}_{MRA} = (20 + 1/2 \times 5)/27 = 0.833$ and $\tilde{\pi}_{DSA} = (22 + 1/2 \times 2)/27 = 0.852$, respectively. Moreover, we have $\tilde{q}_{N2} = 1/27$, $\tilde{q}_{N1} = (0 + 2)/27$, $\tilde{q}_{S2} = 0/27$ and $\tilde{q}_{S1} = (3 + 2)/27$. Then, the difference in the sensitivities between MRA and DSA based on the two raters is $\tilde{\lambda}_{K=2} = -0.019$, and the normal deviate test has $Z_{ND} = 1.393 \approx Z_S$ (one-sided p-value $= 0.082$). The score-type 95 % confidence interval is -0.141 to 0.144 where the lower limit is not greater than $-\Delta = -0.1$. Furthermore, the Wald-type test has $Z_w = 1.397$ (one-sided p-value $= 0.081$) and the Wald-type 95 % confidence interval under the null hypothesis is -0.139 to 0.102. From Table 7.11 based on the consensus evaluations, on the other hand, the sensitivities of MRA and DSA are estimated as $\tilde{\pi}_{MRA_{CE}} = 0.926$ and $\tilde{\pi}_{DSA_{CE}} = 0.889$, respectively. Then, the difference in the sensitivities between MRA and DSA based on the

Table 7.10 A 3 × 3 contingency table ($K = 2$) of the assessments of MRA and DSA by two neuroradiologists (True positive (TP: +) and false negative (FN: −) by two neuroradiologists (1, 2): I (+, +), II (+, − or −, +), III (−, −)) (Jäger et al. [4])

	TP and FN by two radiologists	DSA			Total
		I	II	III	
MRA	I	19	0	1	20
	II	3	0	2	5
	III	0	2	0	2
	Total	22	2	3	27

DSA, digital subtraction angiography; FN, false negative; MRA, magnetic resonance angiography; TP, true positive

Table 7.11 A 2 × 2 contingency table of the assessments of MRA and DSA by consensus evaluations (True positive (TP: +) and false negative (FN: −)) (Jäger et al. [4])

	TP and FN by consensus evaluations	DSA		Total
		+	−	
MRA	+	22	3	25
	−	2	0	2
	Total	24	3	27

DSA, digital subtraction angiography; FN, false negative; MRA, magnetic resonance angiography; TP, true positive

consensus evaluations is $\tilde{\lambda}_{CE} = 0.037$, and the score test derived from Nam [10] and Tango [17] has $Z_S = 1.510$ (one-sided p-value $= 0.066$). Moreover, the score-based 95 % confidence interval derived from Tango [17] is −0.150 to 0.227. These results suggest that the non-inferiority of MRA to DSA cannot be claimed at the one-sided significance level. However, although the difference in the sensitivities based on the two raters $\tilde{\lambda}_{K=2}$ is a negative value, the difference in the sensitivities based on the consensus evaluations $\tilde{\lambda}_{CE}$ is a positive value. We consider that bias from the consensus evaluations caused this phenomenon.

7.6 Conclusion

A non-inferiority trial of diagnostic procedures is generally evaluated on the basis of the results from multiple independent raters who are independent of the study centers. However, consensus evaluations or majority votes to handle multiple results from the multiple raters are not recommended in terms of bias or loss of information [1, 2, 12]. Therefore, it is important that all of the results from the multiple raters are utilized appropriately in the statistical analysis. The methods addressed in this chapter are available for inference of the difference in correlated proportions between the two diagnostic procedures based on the multiple raters. In this chapter, we introduced methods on the basis of sensitivity. However, the methods can be applied to inference of the difference in specificity. Furthermore, if we need to consider the simultaneous non-inferiority of a new diagnostic procedure to the

standard diagnostic procedure in sensitivity and specificity, we can extend the methods using an approach proposed by Lu et al. [8]. Lu et al. extended the score test proposed by Nam [10] and Tango [17] for a single proportion to a simultaneous test for both sensitivity and specificity based on the principle of intersection-union test.

We carried out Monte Carlo simulation studies to evaluate the performance of these methods. The normal deviate test for non-inferiority was shown to have an empirical size closer to a nominal significance level of one-sided 2.5 % than the Wald-type test or the test based on the majority votes. Moreover, the score-type confidence interval had better performance than the Wald-type confidence interval under the null-hypothesis in terms of coverage probability, when the sample size was small. On the other hand, the confidence interval based on the majority votes shows a conservative property.

When we plan a clinical trial to compare the efficacies between two diagnostic procedures, it is very important to take into account the study design. The methods addressed in this chapter are only useful for a study design in which two diagnostic procedures are applied to each subject and all raters evaluate all subjects, that is, paired-patient, paired-rater design. Zhou et al. [18] provided information on study designs for diagnostic procedures in detail. Moreover, it is noted that these methods may not be appropriate for clustered matched-pair data. Schwenke and Busse [16] proposed a Wald-type test for clustered matched-pair data based on multiple raters. However, the test of Schwenke and Busse is a so-called *test for superiority* and cannot be used as a test for non-inferiority. If the results of the two diagnostic procedures are evaluated by a single rater, we can apply several non-inferiority tests for clustered matched-pair data [3, 5, 11]. Therefore, we expect that a non-inferiority test for clustered matched-pair data on the basis of the results from multiple raters will be developed. If there are missing data among the results from the multiple raters in some subject, we would have to apply some kind of imputation method, which would require future research. Furthermore, if the presence of a qualitative interaction between the two diagnostic procedures and the multiple raters is demonstrated, we would not be able to apply these methods for those data. However, this problem could probably be solved by a non-statistical study, for example, by training all of the raters on the criteria of judgment about diagnostic procedures before the start of evaluation.

7.7 Program

The R programs for the methods of this chapter can be downloaded at http://www.medstat.jp/downloadsaeki.html.

References

1. Guidance for industry. Developing medical imaging drugs and biological products. Part 3: design, analysis, and interpretation of clinical studies (2004). URL http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/ucm071604.pdf. Cited 21 May 2012
2. Appendix 1 to the guideline on clinical evaluation of diagnostic agents (CPMP/EWP/1119/98 REV. 1) on imaging agents (Doc. Ref. EMEA/CHMP/EWP/321180/2008) (2009). URL http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500003581.pdf. Cited 21 May 2012
3. Durkalski, V., Palesch, Y., Lipsitz, S., Rust, P.: Analysis of clustered matched-pair data for a non-inferiority study design. *Statistics in Medicine* **22**, 279–290 (2003). DOI 10.1002/sim.1385
4. Jäger, H., Mansmann, U., Hausmann, O., Partzsch, U., Moseley, I., Taylor, W.: MRA versus digital subtraction angiography in acute subarachnoid haemorrhage: a blinded multireader study of prospectively recruited patients. *Neuroradiology* **42**, 313–326 (2000)
5. Jin, H., Lu, Y.: Comparison of correlated proportions based on paired binary data from clustered samples. *Journal of Statistical Planning and Inference* **139**, 4206–4212 (2009). DOI 10.1016/j.jspi.2009.06.005
6. Lehr, R., Kashanian, F.: Three persistent issues in analysis of clinical trials involving diagnostic contrast agents. *Drug Information Journal* **43**, 525–532 (2009). DOI 10.1177/009286150904300501
7. Lu, Y., Bean, J.: On the sample size for one-sided equivalence of sensitivities based upon McNemar's test. *Statistics in Medicine* **14**, 1831–1839 (1995). DOI 10.1002/sim.4780141611
8. Lu, Y., Jin, H., Genant, H.: On the non-inferiority of a diagnostic test based on paired observations. *Statistics in Medicine* **22**, 3029–3044 (2003). DOI 10.1002/sim.1569
9. McNemar, Q.: Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* **12**, 153–157 (1947). DOI 10.1007/BF02295996
10. Nam, J.: Establishing equivalence of two treatments and sample size requirements in matched-pairs design. *Biometrics* **53**, 1422–1430 (1997)
11. Nam, J., Kwon, D.: Non-inferiority tests for clustered matched-pair data. *Statistics in Medicine* **28**, 1668–1679 (2009). DOI 10.1002/sim.3580
12. Obuchowski, N., Lieber, M.: Statistics and methodology. *Skeletal Radiology* **37**, 393–396 (2008). DOI 10.1007/s00256-008-0448-1
13. Rapp-Bernhardt, U., Welte, T., Budinger, M., Bernhardt, T.: Comparison of three-dimensional virtual endoscopy with bronchoscopy in patients with oesophageal carcinoma infiltrating the tracheobronchial tree. *The British Journal of Radiology* **71**, 1271–1278 (1998)
14. Saeki, H., Tango, T.: Non-inferiority test and confidence interval for the difference in correlated proportions in diagnostic procedures based on multiple raters. *Statistics in Medicine* **30**, 3313–3327 (2011). DOI 10.1002/sim.4364
15. Schouten, H.: Estimating kappa from binocular data and comparing marginal probabilities. *Statistics in Medicine* **12**, 2207–2217 (1993). DOI 10.1002/sim.4780122306
16. Schwenke, C., Busse, R.: Analysis of differences in proportions from clustered data with multiple measurements in diagnostic studies. *Methods of Information in Medicine* **46**, 548–552 (2007). DOI 10.1160/ME0433
17. Tango, T.: Equivalence test and confidence interval for the difference in proportions for the paired-sample design. *Statistics in Medicine* **17**, 891–908 (1998). DOI 10.1002/(SICI)1097-0258(19980430)17:8⟨891::AID-SIM780⟩3.0.CO;2-B
18. Zhou, X., Obuchowski, N., McClish, D.: *Statistical Methods in Diagnostic Medicine*, 2nd edn. Wiley & Sons, New York (2011)