Improvement of Sample Size Calculations for Binary Diagnostic Test Assessment

Sébastien Bailly, MSc,†‡§ Cyrielle Dupont, MSc,*†‡§ Jean Iwaz, PhD,*†‡§ Nadine Bossard, MD, PhD,*†‡§ and Muriel Rabilloud, MD, PhD*†‡§*

Objective: This study aimed to formulate a new R function to improve sample size calculation for more accurate estimations of sensitivity (Se) and specificity (Sp).

Methods: The developed function is based on the binDesign function of the binGroup R package. This allowed the use of an "exact" method based on the binomial distribution. In addition, the function takes into account a joint testing of Se and Sp and a nonmonotonous behavior of the power function.

Results: Four tables were generated to display the number of cases (or controls) in joint or separate assessments for an expected combination of Se (or Sp) and a determined difference between the expected Se (or Sp) and the minimum acceptable Se (or Sp). Using the formula for a joint testing of Se and Sp, it resulted in a higher increase of the sample sizes than simply allowing for the sawtooth shape of the power curve.

Conclusion: Whenever equal Se and Sp values are important, a joint testing should be favored and used for sample size determination.

Key Words: diagnostic tests, sample sizes, binomial distribution, sensitivity and specificity

(*J Investig Med* 2014;62: 687–689)

Assessing the accuracy of a new diagnostic test with binary outcome (yes/no or diseased/healthy) requires precise estimates of sensitivity (Se) and specificity (Sp). Sensitivity is the probability of a positive test in a diseased subject. Specificity is the probability of a negative test in a nondiseased subject. Determining the sample size that allows a given precision level in estimating these 2 parameters is an important step in a research study protocol and should always be reported.[1] In fact, although the importance of sample size calculation is generally well recognized, a literature survey has shown that few diagnostic studies have reported details on sample size calculations and that these studies are often underdimensioned, which leads to inaccurate estimates of Se and Sp.[1]

In 2005, Flahault et al.[2] provided sample size tables for binary diagnostic tests. In these tables, the sample size is calculated so as to obtain Se and Sp values significantly greater than the minimal acceptable values specified by the experimenter given a specified power and expected Se and Sp values. Because, in the context of diagnostic test assessment, the value of Se and/or Sp is often close to 1, the use of a normal approximation of the binomial distribution to calculate the sample size may lead to an underestimation of this sample size.[3] This is why Flahault et al.[2] used an "exact" method based on a binomial distribution rather than on a normal approximation; they produced tables for case-control studies with separate sample sizes for cases and controls according to various test performance values and various statistical risk levels.

However, in the latter sample size determinations, 2 factors were not taken into account, which are as follows: (1) the nonmonotonous increase of the power with the increase of the sample size in the case of a binomial distribution[3,4]; and (2) the possibility of testing jointly Se and Sp values when these are considered to have equal importance in the diagnosis. To our knowledge, diagnostic studies did not often consider these 2 factors simultaneously.

The objective of the present work was to develop an R function that is able to provide sample sizes for binary diagnostic tests using an exact method, taking into account the nonmonotonous shape of the power function, and testing jointly Se and Sp by using joint probabilities for alpha and beta risks. This work led to the development of tables for case-control studies that give the number of cases and controls for the most usual combinations of Se and Sp.

MATERIAL AND METHODS

The newly developed function uses the binDesign function from the R package binGroup. The binDesign function computes the sample sizes for testing separately the values of Se and Sp with a given power.[5] It gives the minimum sample size (n_1) needed to reach a prespecified power. By default, the method that computes the sample sizes in binDesign is the most frequently used exact method, that is, the Clopper-Pearson interval. The parameters to be specified are the following: (1) the minimum acceptable value for Se (Se_{min}) or the minimum acceptable value for Sp (Sp_{min}); and (2) δSe (or δSp) the distance between the expected Se (or Sp) and Se_{min} (or Sp_{min}) within which the 1-α lower confidence limit of Se (or Sp) is required to fall with probability 1-β. This is equivalent to a unilateral test of the alternative hypothesis that is Se is greater than Se_{min} with a power 1-β against the null hypothesis that is Se is less than or equal to Se_{min} with a type I error α.

The new function developed here offers 2 improvements. The first is that it determines the "improved" minimum sample size (n_2) needed to reach a prespecified power; that is, all greater sample sizes will allow reaching that power (because n_1 might not be that minimum). The second is that it allows testing jointly the Se and the Sp values of the diagnostic test using joint probabilities based on the rectangular method.[6] In such a context, the null hypothesis is $H0:\{Se \leq Se_{min} \text{ or } Sp \leq Sp_{min}\}$, and the alternative hypothesis is $H1:\{Se \leq Se_{min} \text{ and } Sp > Sp_{min}\}$. If we denote, respectively, $1-\beta^*$ and α^* as the joint probabilities for power and type I error, testing jointly Se and Sp with the rectangular method will be equivalent to carrying out 2 separate

From the *Service de Biostatistique, Hospices Civils de Lyon; †Université de Lyon, Lyon; ‡Université Lyon 1; and §Centre National de la Recherche Scientifique Unité Mixte de Recherche Laboratoire de Biométrie et Biologie Evolutive, Equipe Biostatistique-Santé, Villeurbanne, France.

Received November 22, 2013, and in revised form January 14, 2014.
Accepted for publication January 14, 2014.

Reprints: Sébastien Bailly, PhD, Service de Biostatistique, Hospices Civils de Lyon, 162, avenue Lacassagne, F-69003, Lyon, France. E-mail: sbailly@chu-grenoble.fr.

ISSN: 1081-5589
DOI: 10.231/JIM.0000000000000066

tests, each with power 1-β equal to $\sqrt{1-\beta^*}$ and type I error α equal to $1-\sqrt{1-\alpha^*}$.

Considering the most frequent settings in diagnostic test studies, we computed the n_2 sample sizes for 2 joint powers $1-\beta^*$ (namely, 90% and 80%) and a single joint type I error α^* (namely, 5%) for various Se_{min} (Sp_{min}) and various δSe (δSp) values.

To quantify the impact of using joint probabilities, n_2 sample sizes were also computed for separate testing of Se and Sp. The effect of the sawtooth shape of the power curve was assessed by computing the n_1 sample sizes.

RESULTS

Tables 1 and 2 give, respectively, the improved n_2 number of cases (or controls) for joint and separate testings of Se and Sp with 90% power and 5% type I error. For example, the sample size for a Se_{min} (or Sp_{min}) of 0.75 and a δSe (or δSp) of 0.1 is 220 cases (or controls; Table 1). As expected, the sample size required increases progressively as Se_{min} (or Sp_{min}) decreases (gets closer to 0.5) and as δSe (or δSp) decreases too. At same power and type I error, the sample sizes were always higher with a joint than with a separate testing by 30% on average.

The n_1 classically required to reach a 90% power and a 5% type I error in a joint testing of Se and Sp in the same conditions of Se_{min} (or Sp_{min}) and δ_{Se} (or δ_{Sp}) as previously mentioned are shown in Table 3. These sample sizes are lower than with the improved method by 6% on average. Table 4 shows the sample sizes in the same conditions but only with 80% power. The numbers found are lower than those required for a 90% power by 12.5% on average.

DISCUSSION

We developed here a new R function to calculate the sample sizes for studies of binary diagnostic methods and proposed several tables that correspond to the most common settings.

In improving previous methods for sample size calculations for accuracy, we followed the recommendations highlighted in the previous articles that advocated the use of an "exact" method based on the binomial distribution rather than the use of the standard method with a normal approximation of the binomial distribution,[2] took into account the sawtooth shape of the power curve,[4] and computed the sample sizes for a joint determination of Se and Sp.

The effect of using a joint testing on the sample size is greater than that of allowing for the nonmonotonous shape of the power curve. Although the sample size is smaller with a separate than with a joint determination, we believe that a joint testing should be favored whenever both Se and Sp are of equal importance.

The sample sizes were determined here using the exact method of Clopper-Pearson. This method guarantees that the actual coverage probability of the confidence interval is always equal or greater than the nominal confidence level. Consequently, the conservativeness of the method may lead to overestimated sample sizes.[3,7,8] Some authors proposed a correction for continuity to apply to the Clopper-Pearson exact method.[9] This "mid-P method" is less conservative than the original exact method but still achieves a good coverage probability. Anyway, this method cannot be implemented yet with the new function developed here. Besides, as shown by Agresti and Coull,[7] approximation-based methods may have better properties than exact methods. The binDesign function proposes 2 of the latter methods—the score method of Wilson and the Agresti-Coull method. Both are less conservative than the

TABLE 1. Improved Minimum Sample Sizes of Cases (or Controls) for a Joint Determination of Se and Sp With 90% Joint Power and 5% Joint Type I Error According to Whom It May Concern: Various Se_{min} (or Sp_{min}) and Various δ_{Se} (or δ_{Sp}) Values

	Se_{min} or Sp_{min}								
	0.5	**0.55**	**0.6**	**0.65**	**0.7**	**0.75**	**0.8**	**0.85**	**0.9**
δ_{Se} or δ_{Sp}									
0.05	1308	1288	1236	1164	1059	929	775	595	387
0.1	331	320	306	287	256	220	179	127	
0.15	147	143	134	122	105	89	69		
0.2	81	78	75	66	53	44			
0.25	54	50	45	38	31				
0.3	35	31	29	22					
0.35	25	23	19						
0.4	17	14							
0.45	12								

TABLE 2. Improved Minimum Sample Sizes of Cases (or Controls) With Separate Determinations of Se and Sp with 90% Power and 5% Type I Error According to Various Se_{min} (or Sp_{min}) and the δ_{Se} (or δ_{Sp}) Values

	Se_{min} or Sp_{min}								
	0.5	**0.55**	**0.6**	**0.65**	**0.7**	**0.75**	**0.8**	**0.85**	**0.9**
δ_{Se} or δ_{Sp}									
0.05	891	871	835	778	716	634	528	408	263
0.1	224	220	211	196	175	153	124	85	
0.15	102	100	94	85	73	65	44		
0.2	58	55	51	45	37	29			
0.25	35	34	30	28	24				
0.3	26	24	21	16					
0.35	16	15	14						
0.4	13	12							
0.45	8								

TABLE 3. Minimum Sample Sizes of Cases or Controls for 90% Joint Power and 5% Joint Type I Error According to Various Se_{min} or Sp_{min} and δ_{Se} or δ_{Sp} Values

	Se_{min} or Sp_{min}								
	0.5	**0.55**	**0.6**	**0.65**	**0.7**	**0.75**	**0.8**	**0.85**	**0.9**
δ_{Se} or δ_{Sp}									
0.05	1283	1260	1212	1134	1034	903	748	565	351
0.1	320	313	292	274	241	206	167	110	
0.15	143	135	129	116	101	84	62		
0.2	76	73	69	59	53	38			
0.25	49	47	42	34	27				
0.3	32	28	26	22					
0.35	23	20	16						
0.4	17	14							
0.45	12								

TABLE 4. Improved Minimum Sample Sizes of Cases or Controls for 80% Joint Power and 5% Joint Type I Error According to Various Se_{min} or Sp_{min} and δ_{Se} or δ_{Sp} Values

	Se_{min} or Sp_{min}								
	0.5	**0.55**	**0.6**	**0.65**	**0.7**	**0.75**	**0.8**	**0.85**	**0.9**
δ_{Se} or δ_{Sp}									
0.05	1055	1038	1000	931	852	750	630	484	316
0.1	272	261	247	234	212	179	143	101	
0.15	121	116	109	103	90	75	55		
0.2	65	66	60	52	45	38			
0.25	42	39	29	34	27				
0.3	30	26	23	18					
0.35	20	20	16						
0.4	15	14							
0.45	9								

exact method but ensure good coverage probabilities.[3] They may be used instead of the Clopper-Pearson method adopted by default by binDesign.

The sample sizes were calculated here with the assumption of a binomial distribution of the Se (or Sp). However, the binomial distribution may be overdispersed because of the mix of populations with various Se (or Sp) values.[10] The next step will be to calculate the sample sizes, taking into account the inflation of the variance.

In summary, we developed here a function that allows optimal calculations of sample sizes for diagnostic tests with binary result. This function may be used in cohort studies and take into account the prevalence of the disease. Tables representative of the most common clinical contexts are presented. For other hypotheses and other type I or II errors, the function can be obtained from the corresponding author.

REFERENCES

1. Bachmann LM, Puhan MA, ter Riet G, et al. Sample sizes of studies on diagnostic accuracy: literature survey. *BMJ*. 2006;332:1127–1129.
2. Flahault A, Cadilhac M, Thomas G. Sample size calculation should be performed for design accuracy in diagnostic test studies. *J Clin Epidemiol*. 2005;58:859–862.
3. Brown LD, Cai TT, DasGupta A. Interval estimation for a binomial proportion. *Stat Sci*. 2001;16:101–133.
4. Chu H, Cole SR. Sample size calculation using exact methods in diagnostic test studies. *J Clin Epidemiol*. 2007;60:1201–1202.
5. Bilder CR, Zhang B, Schaarschmidt F, et al. binGroup: a package for group testing. *R J*. 2010;2:56–61.
6. Pepe MS. *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford: Oxford University Press; 2004.
7. Agresti A, Coull BA. Approximate is better than "exact" for interval estimation of binomial proportions. *Am Stat*. 1998;52:119–126.
8. Brown LD, Cai TT, DasGupta A. Confidence intervals for a binomial proportion and asymptotic expansions. *Ann Stat*. 2002;30:160–201.
9. Fosgate GT. Modified exact sample size for a binomial proportion with special emphasis on diagnostic test parameter estimation. *Stat Med*. 2005;24:2857–2866.
10. Chen C, Tipping RW. Confidence interval of a proportion with over-dispersion. *Biom J*. 2002;7:877–886.