User’s guide to sample size estimation in diagnostic accuracy studies

Haldun Akoglu*
Department of Emergency Medicine, Marmara University School of Medicine, Istanbul, Turkey
*Corresponding author

Abstract:
Sample size estimation is an overlooked concept and rarely reported in diagnostic accuracy studies, primarily because of the lack of information of clinical researchers on when and how they should estimate sample size. In this review, readers will find sample size estimation procedures for diagnostic tests with dichotomized outcomes, explained by clinically relevant examples in detail. We hope, with the help of practical tables and a free online calculator (https://turkjemergmed.com/calculator), researchers can estimate accurate sample sizes without a need to calculate from equations, and use this review as a practical guide to estimating sample size in diagnostic accuracy studies.

Keywords:
Calculator, diagnostic accuracy, online, sample size, sensitivity, specificity

Introduction

Diagnostic accuracy studies are essential to achieve a better clinical decision-making process. In estimating the diagnostic accuracy of a test and obtaining the desired statistical power, the investigators need to know the minimal sample size required for their experiments. As in all kinds of research, studies with small sample sizes fail to determine an accurate estimate, with wide confidence intervals, and studies with large sample sizes may lead to the wasting of resources. Indeed, sample size estimation is an overlooked concept and rarely reported in diagnostic accuracy studies. Bochman et al. reported in 2005 that only 1 in 40 of the diagnostic accuracy studies published in the top 5 journals of ophthalmology reported a sample size calculation. This is primarily because of the lack of information of clinical researchers on when and how they should estimate sample size.

Therefore, this review aims to help clinical researchers by defining practical sample size estimation techniques for different study designs. We will start with the description of the clinical diagnostic evaluation process. Then, we will define the characteristics and measures of diagnostic accuracy studies. After we summarize the design options, we will define how to estimate the sample size for each of those different designs.

Definitions

In diagnostic accuracy studies, the test in question is called the index test. The comparative and probably the better test is called the reference standard. The diagnostic evaluation process starts with a list of differential diagnoses, where each one of them has a different probability. Those probabilities are generated with the use of the local epidemiological data, the "gestalt" of the experienced physician, and results of the previous tests. The probability of disease before performing a test is called the prior probability. Physicians order consecutive tests to increase or decrease
the probability of those specific diagnoses and narrow down the list. Each diagnosis in this list has its own probability scale (from 0% to 100%) for that patient. There are two important thresholds on that scale: the test threshold marks the disease probability that is high enough to warrant further testing to rule in or out that diagnosis; treatment threshold marks the disease probability that is high enough to accept that diagnosis and start treatment. The prior probability of each disease changes according to the result of each test, which is called the posterior probability. The aim is to move the posterior probabilities above the treatment or below the test threshold with the results of consecutive tests to rule in or out every diagnosis. In the clinical setting, each procedure performed to gather information about the disease probability is a test, such as history taking (age, sex, and presence of comorbidities), measurements (RR, HR, or pSO2), or physical examination (rales, rhonchi, Romberg, etc.). We combine the results of those tests and increase or decrease the probabilities of diagnoses we have in mind, decide to test further, or treat.

For better comprehension, let us assume that a 75-year-old bedridden female patient with Alzheimer’s disease presented to an emergency department with tachypnea of 30/min, peripheral oxygen saturation of 90%, and tachycardia of 110 bpm. As soon as those data were gathered, a few diagnoses could be listed where pulmonary embolism makes it to the top. In this patient, the probability of pulmonary embolism is above the treatment threshold and ordering a treatment with LMWH (Low Molecular Weight Heparin) is warranted. One may still order tests to rule in or rule out pneumonia, pneumothorax, or other diagnoses, or may order antibiotics if pneumonia makes it above the treatment threshold, too. On the contrary, an X-ray may lower the probability of pneumothorax below the test threshold; therefore, pneumothorax could be ruled out. A clinical diagnostician is a detective investigating multiple diagnoses simultaneously, using a bunch of tests to move the probabilities of several diagnoses below or above the test and treatment thresholds.

In classical diagnostic accuracy studies, a categorical or continuous index test variable is compared against a categorical, dichotomized reference standard variable. In this review, we will focus on index tests with a dichotomized outcome (positive or negative). We evaluate the accuracy of the index test by its sensitivity and specificity, which are calculated from the values in the cells of the contingency table comparing those two tests. The sensitivity indicates the proportion of true positives in diseased subjects, and specificity determines the proportion of true negatives in nondiseased subjects. Positive predictive value (PPV) determines the proportion of diseased subjects out of all the positives, and negative predictive value (NPV) determines the proportion of nondiseased subjects out of all negatives.

PPV and NPV are affected by the prior probability (prevalence) of disease in the target population and are rarely used. On the other hand, sensitivity and specificity are not influenced by the prevalence of disease, which is why they are so popular. Their total is a more important metric than the individual values, and they should always be considered together. Tests with the total of sensitivity and specificity closer to 100% are almost perfect. It is no good than tossing a coin if the total of sensitivity and specificity is closer to 100, even one of the values were close to 100. For example, a test with a sensitivity of 90% and specificity of 10 is a test without any clinical diagnostic benefit. Therefore, both metrics were combined in a one-dimensional index called likelihood ratio (LR). The positive LR is the ratio of the probability of a positive test in diseased to nondiseased, and the negative LR is the ratio of the probability of a negative test in diseased to nondiseased [Table 1]. Any test with a positive LR above 10 is considered a good test for ruling in, and tests with a negative LR below 0.1 are considered good for ruling out a diagnosis. LRs are not affected by the prevalence of the disease. They are beneficial in comparing two separate tests. Furthermore, the posterior probability of a diagnosis can be calculated with the help of the positive and negative LRs (see online calculator at https://turkjemrgmed.com/calculator).

In a comparative analysis, a Type 1 error happens if we reject the null hypothesis (no difference) incorrectly and report a difference, whereas a Type II error happens if we accept the null hypothesis incorrectly and report that there is no difference [Table 1]. Sample size estimation is performed to calculate how many patients are required to avoid a Type 1 or a Type 2 Error. The classical design is a cross-sectional cohort study, or single-test design, where all consecutive patients suspected of the target disease or condition are tested with the index test and the reference standard [Figure 1]. This approach may be modified to delayed-type cross-sectional, case-referent, or test result-based sampling designs, or cohort and case-control designs may be used instead. In a comparative design, the index test is compared to a previously evaluated comparator test in a paired or unpaired fashion [Figure 1]. In the comparative unpaired design (between-subjects), study participants are randomly assigned to either the index or comparator test. Participants are tested with one...
of the two tests, not both. Then, the disease status of every participant is confirmed with the reference standard. This design is preferred when researchers aim to evaluate the impact of diagnostic testing on clinical decision-making, patient prognosis, and real-life utility of the index test. These are the “diagnostic randomized controlled trial” and the before-after type studies. In the comparative paired design (within-subjects), index, comparator, and reference standard tests were performed on all subjects.

Since the variability of the study results is decreased, the paired design is preferred if feasible and justifiable. Sample Size Estimation in Diagnostic Accuracy Studies

There are four major designs to compare a dichotomized index test with a dichotomized reference standard. The appropriate equations that should be used for the estimation

| Metric               | Definition                                                                 | Formula                                           |
|----------------------|-----------------------------------------------------------------------------|---------------------------------------------------|
| Sensitivity          | The proportion of true positives in diseased subjects                      | True positives/(true positives + false negatives) |
| Specificity          | The proportion of true negatives in nondiseased subjects                   | True negatives/(true negatives + false positives) |
| PPV                  | The proportion of diseased subjects out of all positives                   | True positives/(true positives + false positives) |
| NPV                  | The proportion of nondiseased subjects out of all negatives                | True negatives/(true negatives + false negatives) |
| Positive likelihood ratio | The ratio of the probability of positive test in diseased to nondiseased | Sensitivity/(1-specificity)                        |
| Negative likelihood ratio | The ratio of the probability of negative test in diseased to nondiseased | (1-sensitivity)/specificity                        |
| Type 1 error         | Finding a difference in fact there is none (false positive)                | None                                              |
| Type 2 error         | Finding no difference in fact there is (false negative)                    | None                                              |
| Power                | Number of patients required to avoid a type II error                       |                                                   |

PPV: Positive predictive value, NPV: Negative predictive value

Figure 1: Major study designs that are used to compare the diagnostic accuracy of tests
of sample size in each of those situations are previously summarized by Obuchowski [Table 2].\[9\] We prepared offline tables [Tables 2-6] and an online calculator (https://turkjemergmed.com/calculator) for the use of researchers to estimate the sample size for their diagnostic accuracy studies.

### Table 2: Sample size estimation formulas

| Equations | Explanations |
|-----------|--------------|
| Equation 1 | Those formulas are defined using normal approximation to construct a confidence interval for the true sensitivity and specificity value with a confidence level of (1-\(\alpha\))% and a maximum marginal error of \(d\). \(Se\) and \(Sp\) are predetermined values ascertained by previously published data or clinician experience/judgment. Estimated sample sizes should be adjusted for disease prevalence (Equations 4a and 4b) |
| \(n_{Se,\text{unadj}} = \frac{Z^2 \times Se(1-Se)}{d^2}\) | |
| \(n_{Sp,\text{unadj}} = \frac{Z^2 \times Sp(1-Sp)}{d^2}\) | |
| \(n_{Se,\text{prev.adj}} = \frac{n_{Se,\text{unadj}}}{\text{Prevalance}}\) | |
| \(n_{Sp,\text{prev.adj}} = \frac{n_{Sp,\text{unadj}}}{(1-\text{Prevalance})}\) | |
| Equation 2, Comparison of a proportion with null | The estimated proportion (sensitivity or specificity) of the index test (\(P_1\)), the proportion that we plan to find a statistically significant difference (\(P_0\)), type 1 error (\(\alpha\)), power (\(\beta\)), and disease prevalence are needed for the calculations. The sample size should be calculated for sensitivity and specificity separately for a power of 90%, so the final power of the study would be 80%. Estimated sample sizes should be adjusted for disease prevalence (Equations 4a and 4b) |
| \(n(\text{unadj}) = \frac{Z_\frac{\alpha}{2} \times P_1 - P_2 + Z_\frac{\beta}{2} \times P_1 + P_2}{(P_1 - P_2)^2}\) | |
| \(n(\text{adj}) = \frac{n(\text{unadj})}{\text{Prevalance}}\) | |
| \(n(\text{adj}) = \frac{n(\text{unadj})}{(1-\text{Prevalance})}\) | |
| \(n(\text{Yates continuity correction}) = \frac{n}{4} \left[1 + \sqrt{\frac{4}{n}} \right] \) | |
| Equation 3a, Comparison of two unpaired proportions | The Formula Set 2 is extended to include both tests. \(P\) denotes the average of the tests' estimated proportions (\(P_1\) and \(P_2\), sensitivity or specificity). One-sided \(P\) is preferred since we want to test if one of the paths is different from the other. Yates’ continuity correction should be applied (Equation 5) |
| \(n = \frac{Z_\frac{\alpha}{2} \times \sqrt{2 \times P_1(1-P_1) + Z_\frac{\beta}{2} \times (P_1(1-P_1) + P_2(1-P_2))}}{(P_1 - P_2)^2}\) | |
| \(n(\text{Yates continuity correction}) = \frac{n}{4} \left[1 + \sqrt{\frac{4}{n}} \right] \) | |
| Equation 3b, Comparison of two paired proportions | \(\Psi\) is the probability of disagreement between the two tests. Bounds on the Probability of Disagreement (\(\Psi\)): The minimum probability of disagreement is \(P_2 - P_1\). The maximum probability of disagreement is when agreement occurs only by chance, equal to \(P_1 \times (1 - P_2) + (1 - P_1) \times P_2\). One-sided \(P\) is preferred since we want to test if one of the tests is different from the other. Yates’ continuity correction should be applied (Equation 5) |
| \(n = \frac{Z_\frac{\alpha}{2} \times \sqrt{\Psi + Z_\frac{\beta}{2} \times (P_2 - P_1)}}{(P_2 - P_1)^2}\) | |
| \(\Psi_{\min} = P_2 - P_1\) | |
| \(\Psi_{\max} = P_1 \times (1 - P_2) + P_2 \times (1 - P_1)\) | |
| \(n(\text{Yates continuity correction}) = \frac{n}{4} \left[1 + \sqrt{\frac{4}{n}} \right] \) | |
| Adjusting for disease prevalence (Equations 4a and 4b) | Estimated sample sizes should be adjusted for disease prevalence with those equations |
| \(n_{Se} = \frac{n}{\text{Prevalance}}\) | |
| \(n_{Sp} = \frac{n}{(1-\text{Prevalance})}\) | |
| Yates’ Continuity Correction (Equation 5) | Yates’ Continuity Correction should be applied to all calculations comparing two proportions, as described by Beam et al.[11] |
| \(n(\text{Yates continuity correction}) = \frac{n}{4} \left[1 + \sqrt{\frac{4}{n}} \right] \) | |

Single-test design (new diagnostic tests)

If a new diagnostic test (new test or new to the study population) is investigated in a prospective cohort that the disease status and prevalence are known, this approach is preferred [Table 2, Equation 1].\[11\] Researchers...
### Table 3: Sample size estimates at predetermined sensitivity and specificity values for various disease prevalence states

Values are at marginal of (A) 3%, (B) 5%, and (C) 7%.

| Disease prevalence (%) | Sensitivity | Specificity | Sample size estimates |
|-------------------------|-------------|-------------|-----------------------|
|                         |             |             | 5%                    |
|                         |             |             | 7%                    |
|                         |             |             | 10%                   |
|                         |             |             | 15%                   |

Type I error is accepted as 5% in all the calculations. If disease prevalence is more than 50%, read the complement of the prevalence to 100%, but switch sensitivity for specificity. For example, in a study with a disease prevalence of 60%, read the line with a prevalence of 40% (100%–60%) but use the sample size under sensitivity columns for specificity.
try to be sure with a confidence level of 95% that their predetermined sensitivity or specificity lies within the marginal error of d (desired width of one-half of the confidence interval [CI]). Sensitivity and specificity values are ascertainment by previously published data or clinician experience/judgment.

For example, let us assume that we are investigating the value of a new test for diagnostic screening. We aim for a sensitivity of 90% in a cohort with a known disease prevalence of 10%. We want maximum marginal error of the estimate not to exceed 5% with a CI of 95%. So, we select Table 3B, find the row for the disease prevalence of 10%, and read the cell for the column of 90% sensitivity, which is 1383. We estimate that 10% of the 1383 subjects will be diseased (n = 138), and 90% will be nondiseased.

**Single-test design, comparing the accuracy of a single test to a null value**

If the true disease status of the patients is unknown at the time of enrollment, those studies are called confirmatory diagnostic accuracy studies.[7] Obuchowski defined this approach as “comparing the sensitivity of a test to a prespecified value” [Table 2, Equation 2].[8] For example, surgery is the reference standard test for the diagnosis of acute appendicitis, but it is invasive. The prevalence of acute appendicitis confirmed by surgery is around 40%, which means that 60% of the patients suspected of acute appendicitis had an unnecessary surgery. Therefore, noninvasive alternatives such as noncontrast-enhanced computed tomography (CT) have emerged, and it has been shown to have a sensitivity of 90%.[10] We hypothesize that contrast-enhanced CT is better, with a sensitivity around 95%. How many patients do we need to recruit if we need to be sure the sensitivity of 95% is statistically significant from 90% with a power of 90% and type I error of 5%?

Table 4 presents precalculated sample size estimates for studies comparing the accuracy of single index test to a null value. Table 4 includes estimates for a type 1 error of 5% and power of 90%. The cell intersecting expected sensitivity around 95%. How many patients do we need to recruit if we need to be sure the sensitivity of 95% is statistically significant from 90% with a power of 90% and type I error of 5%?

Table 5 presents precalculated sample size estimates for studies comparing two independent proportions, unpaired groups for A) Power of 90% and B) 80%. Type 1 error is accepted as 5% for all calculations. Yates’ continuity correction is applied. Estimates here are for each of the independent groups.
Table 6A: Sample size estimation for the comparison of two dependent proportions, paired groups. A) N (Ψ_min), B1) N (Ψ_max) for power of 90%, B2) N (Ψ_max) for power of 80%

| Difference P2−P1 (%) | Power 80% | Power 90% |
|----------------------|-----------|-----------|
|                      |           |           |
| 1                    | 804       | 1043      |
| 3                    | 266       | 345       |
| 5                    | 159       | 206       |
| 10                   | 78        | 101       |
| 15                   | 52        | 66        |
| 20                   | 38        | 48        |
| 25                   | 30        | 38        |
| 30                   | 25        | 31        |
| 35                   | 21        | 26        |
| 40                   | 18        | 22        |
| 45                   | 16        | 19        |
| 50                   | 14        | 17        |
| 55                   | 12        | 15        |
| 60                   | 11        | 13        |
| 65                   | 10        | 12        |
| 70                   | 9         | 11        |
| 75                   | 8         | 9         |
| 80                   | 7         | 8         |
| 85                   | 7         | 8         |
| 90                   | 6         | 7         |
| 95                   | 5         | 6         |
| 99                   | 5         | 5         |
| 100                  | 4         | 4         |

Yates’ continuity correction is applied

Table 6B: Sample size estimation for the comparison of two dependent proportions, paired groups. 1) N (Ψ_max) for power of 90%, 2) N (Ψ_max) for power of 80%

| P1 | P2 (%) |
|----|--------|
| 55 | 60%    | 65% |
| 70% | 75% | 80% | 85% | 90% | 95% | 100% |
| 1749 | 444 | 200 | 113 | 72 | 50 | 37 | 28 | 21 | 17 | 50 |
| 1715 | 431 | 192 | 108 | 68 | 47 | 34 | 25 | 19 | 15 | 55 |
| 1646 | 410 | 181 | 100 | 63 | 42 | 30 | 22 | 17 | 15 | 60 |
| 1543 | 380 | 165 | 90 | 56 | 42 | 30 | 23 | 19 | 15 | 65 |
| 1406 | 341 | 146 | 79 | 48 | 38 | 30 | 24 | 20 | 16 | 70 |
| 1235 | 294 | 123 | 65 | 48 | 30 | 23 | 20 | 16 | 18 | 75 |
| 1029 | 238 | 97 | 48 | 30 | 23 | 20 | 16 | 18 | 20 | 80 |
| 789 | 174 | 66 | 48 | 30 | 23 | 20 | 16 | 18 | 20 | 85 |
| 515 | 101 | 90 | 206 | 95 |

(B1) N (Ψ_max), Power 90%

| 1274 | 327 | 148 | 85 | 55 | 39 | 28 | 22 | 17 | 14 | 50 |
| 1249 | 317 | 143 | 81 | 52 | 36 | 26 | 20 | 16 | 14 | 55 |
| 1200 | 302 | 135 | 76 | 48 | 33 | 24 | 18 | 16 | 18 | 60 |
| 1126 | 280 | 124 | 69 | 43 | 29 | 21 | 18 | 16 | 18 | 65 |
| 1027 | 252 | 110 | 60 | 37 | 25 | 21 | 18 | 16 | 18 | 70 |
| 903 | 218 | 93 | 50 | 30 | 23 | 19 | 17 | 14 | 17 | 75 |
| 755 | 178 | 74 | 38 | 23 | 20 | 18 | 15 | 13 | 13 | 80 |
| 581 | 132 | 52 | 28 | 20 | 18 | 15 | 13 | 13 | 13 | 85 |
| 383 | 78 | 90 | 159 | 95 |

(B2) N (Ψ_max), Power 80%

Yates’ continuity correction is applied
Sometimes researchers aim for sensitivity and specificity simultaneously and want to estimate a sample size that is enough for both. Since sensitivity and specificity are calculated in different groups (diseased vs. nondiseased), two separate sample sizes are calculated for a power of 90%, so the final power of the study would be 80%. Let’s enhance the example above and assume that we also want an adequate sample size for a specificity hypothesis, too. We think that the specificity of contrast-enhanced CT would be 85%, and we want to be sure that it is significantly higher than the specificity of noncontrast-enhanced CT (80%). To calculate the sample size estimate for specificity at a power of 90%, we again use Table 4. The cell intersecting P1 (noncontrast-enhanced CT) of 85% and P0 (null, contrast-enhanced CT) of 80% reveals that we need at least 656 nondiseased subjects (patients without acute appendicitis confirmed with surgery). We use Equations 4a and 4b in Table 2 to adjust specificity for disease prevalence (n/(1 − prevalence) = 656/(1 − 0.4)) and find that we need to recruit 1093 subjects. Since the higher of the two estimates (849 for sensitivity and 1093 for specificity) is 1093, we select this estimate for a power of 80% and type 1 error of 5% for both outcomes.

According to Beam, Yates’ continuity correction should be used to compare proportions. Therefore, we present corrected values in Tables 4-6 and both corrected and uncorrected values on the online calculator.[11] Several authors reported calculations that did not incorporate disease prevalence, and several others did, which we also preferred in this review.[12,13]

**Studies comparing two diagnostic tests**

As mentioned above, comparative design can be unpaired or paired [Figure 1]. Beam described the formulas to estimate sample sizes for both designs [Table 2, Equation 3a and b].[11] Since we want to be sure if one of the tests is significantly different than the other, calculations for one-sided significance levels are sufficient.

**Unpaired design (between-subjects)**

Proportions will be compared between different groups (unpaired) with a Chi-squared test. Therefore, the sample size for each group would be estimated for the Chi-squared test with Yates’ continuity correction, using the method given by Casagrande and Pike [Table 2, Equation 5].[14]

Let us assume we want to compare the sensitivity of two alternative diagnostic pathways, where the contender has 70% sensitivity. We want to design our study so that there is an 80% chance of detecting a difference when our index test has at least a sensitivity of 80% (or a difference of 10%). We accept the significance level as 5%, with a one-sided hypothesis. In Table 5 (for the power of 80%), we check the cell intersecting 70% and 80%, and find that at least 250 subjects are needed for each pathway, making the total estimate 500 subjects.

**Paired design (within-subjects)**

In this design, proportions will be compared between paired samples. Therefore, the sample size for the entire study would be estimated for McNemar’s test, using the method defined by Connor et al.[15] Those two diagnostics tests agree with each other with variable degrees (probability of disagreement [Ψ]), which affects the estimated sample size. On one end, tests disagree with each other just with the degree of the difference in proportions (sensitivity or specificity [Ψ\(_{\text{min}}\) = P\(_2\) − P\(_1\)]). Conversely, they agree with each other just by chance, where the probability of disagreement is maximum (Ψ\(_{\text{max}}\) = (1−P\(_1\)+P\(_2\))(1−P\(_1\))). Those are the two boundaries of the estimated sample size range for the paired design, and the mean of those two ends may be enough in most situations.

Let us work the same example above for a paired design: first, we check Table 6 (lower boundary) for a 10% difference in proportions and 80% power. If the disagreement probability of the tests is minimum, a sample size of 78 subjects would be enough. Second, we check Table 6 (higher boundary) for a power of 80% and read the cell intersecting 70% and 80%. If both tests agree with each other just by chance (maximum disagreement), we would need at least 252 subjects. The mean value of this range (78 to 252, n = 165) or the higher boundary (n = 252) can be selected as the sample size. Please note that, even at the highest probability of disagreement, almost half of the sample size would be enough with paired design compared to the unpaired design.

**Discussion**

We reviewed methods for estimating the minimum required sample size for different study designs in diagnostic accuracy research. This review is performed by a clinical researcher with ease of use for clinical researchers in mind. There are alternative and better methods to estimate the sample size for the procedures described above. Researchers should consult a statistician whenever they need a more accurate or sophisticated approach.

The accuracy of sample size estimates heavily depends on how closely the required assumptions are met.[11] Study results may fall far from the researchers’ assumptions, and post hoc (or interim) power and sample size analyses may be needed in those extreme conditions.
Debates are ongoing if Yates’ continuity correction should be used, if correcting for the disease prevalence is needed when it is unknown before the enrollment phase, or if Connor et al.’s (Equation 3b) formula is too optimistic by underestimating the sample size.\cite{11,15} Researchers should include a safe limit to control for those debatable points and aim for an optimal sample size.

**Conclusion**

Sample size estimation is an overlooked concept and rarely reported in diagnostic accuracy studies, primarily because of the lack of information of clinical researchers on when and how they should estimate sample size. We hope the tables and the online calculator supplemented to this review may be used as a guide to estimate sample size in diagnostic accuracy studies.

**Supplement**

Online Calculator: https://turkjemergmed.com/calculator

**Author contributions (CRediT)**

HA completed this review on his own.

**Acknowledgments**

I thank Ozan Konrot for being one of the best in problem-solving with his charming smile. The online calculator could not have been prepared without his devoted input and relentless efforts. I also would like to thank Gokhan Aksel and Seref Kerem Corbacıoglu. They have been and hopefully will be my closest peers during my journey through being an enthusiast, a student, a mentor, and a teacher of biostatistics, journalology, and clinical research methodology.

**Conflicts of interest**

None declared.

**Funding**

None.

**References**

1. Hajian-Tilaki K. Sample size estimation in diagnostic test studies of biomedical informatics. J Biomed Inform 2014;48:193-204.
2. Bachmann LM, Puhan MA, ter Riet G, Bossuyt PM. Sample sizes of studies on diagnostic accuracy: Literature survey. BMJ 2006;332:1127-9.
3. Bochmann F, Johnson Z, Azuara-Blanco A. Sample size in studies on diagnostic accuracy in ophthalmology: A literature survey. Br J Ophthalmol 2007;91:898-900.
4. Jones SR, Carley S, Harrison M. An introduction to power and sample size estimation. Emerg Med J 2003;20:453-8.
5. Holtman GA, Berger MY, Burger H, Deeks JJ, Donner-Banzhoff N, Fanshawe TR, et al. Development of practical recommendations for diagnostic accuracy studies in low-prevalence situations. J Clin Epidemiol 2019;114:38-48.
6. Knottnerus JA, Buntinx F, eds. The Evidence Base of Clinical Diagnosis: Theory and Methods of Diagnostic Research. 2nd edition. Blackwell Publishing Ltd; 2011.
7. Stark M, Hesse M, Brannath W, Zapf A. Blinded sample size re-estimation in a comparative diagnostic accuracy study. BMC Med Res Methodol 2022;22:115.
8. Sitch AJ, Dekkers OM, Scholefield BR, Takwoingi Y. Introduction to diagnostic test accuracy studies. Eur J Endocrinol 2006;154:501-9.
9. Obuchowski NA. Sample size calculations in studies of test accuracy. Stat Methods Med Res 1998;7:371-92.
10. Rud B, Vejborg TS, Rappeport ED, Reitsma JB, Wille-Jørgensen P. Computed tomography for diagnosis of acute appendicitis in adults. Cochrane Database Syst Rev 2019;2019:CD009977.
11. Beam CA. Strategies for improving power in diagnostic radiology research. AJR Am J Roentgenol 1992;159:631-7.
12. Buderer NM. Statistical methodology: I. Incorporating the prevalence of disease into the sample size calculation for sensitivity and specificity. Acad Emerg Med 1996;3:895-900.
13. European Medicines Agency Committee for Medicinal Products for Human Use. Guideline on Clinical Evaluation of Diagnostic Agents. Published Online; July 23, 2009. Available from: https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-clinical-evaluation-diagnostic-agents_en.pdf. [Last accessed on 2022 Jul 15].
14. Casagrande JT, Pike MC. An improved approximate formula for calculating sample sizes for comparing two binomial distributions. Biometrics 1978;34:483-6.
15. Connor RJ. Sample size for testing differences in proportions for the paired-sample design. Biometrics 1987;43:207-11.