TURKISH JOURNAL OF
EMERGENCY MEDICINE

Invited Review Article

Access this article online
Quick Response Code:

Website:
www.turkjemergmed.com

DOI:
10.4103/2452-2473.357348

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.[1] Indeed, sample size estimation is an overlooked concept and rarely reported in diagnostic accuracy studies.[2,3] 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.[3] 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

Submitted: 25-07-2022
Revised: 30-07-2022
Accepted: 05-08-2022
Published: 30-09-2022

ORCID:
HA: 0000-0002-1316-0308

Address for correspondence:
Prof. Haldun Akoglu,
Department of Emergency
Medicine, Marmara
University School of
Medicine, Istanbul, Turkey.
E-mail: drhaldun@gmail.com

How to cite this article: Akoglu H. User's guide to sample size estimation in diagnostic accuracy studies. Turk J Emerg Med 2022;22:177-85.

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.[1] 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 200% 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://turkjemergmed.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.[4]

Design Options of the Diagnostic Accuracy Studies

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].[6] 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.[5] 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

Table 1: Definition of major diagnostic utility metrics

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

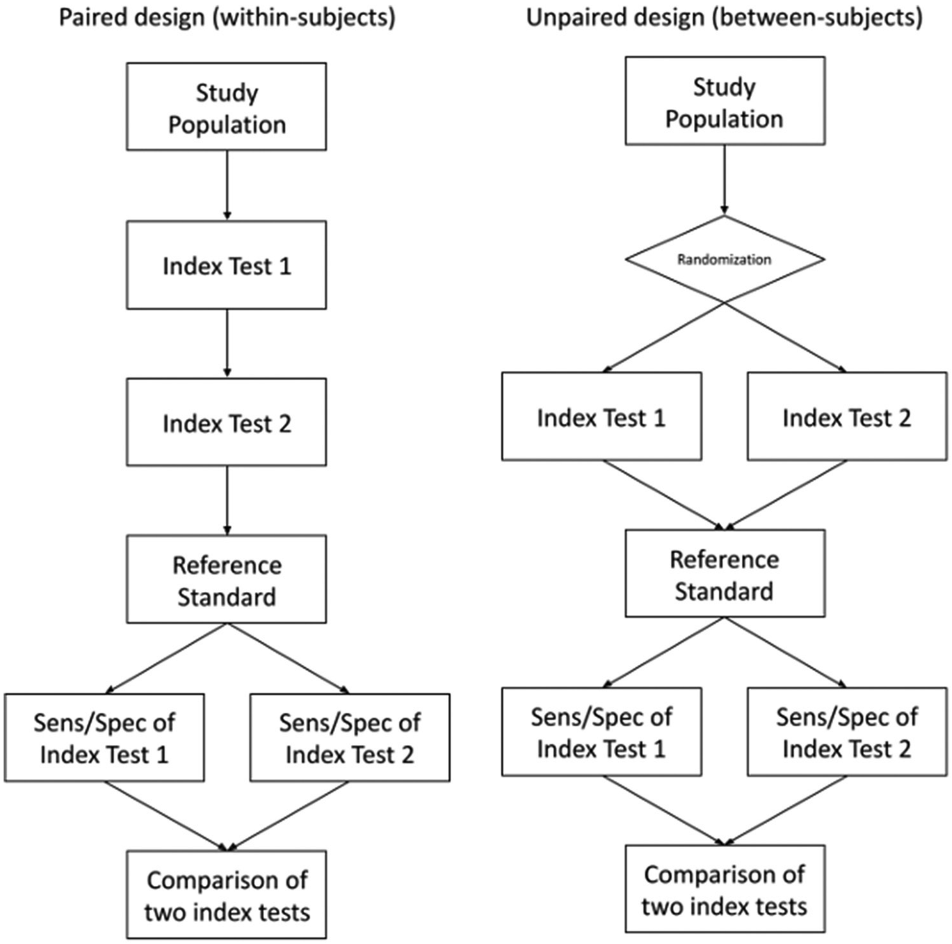
Paired design (within-subjects)
Study Population
Index Test 1
Index Test 2
Reference Standard
Sens/Spec of Index Test 1
Sens/Spec of Index Test 2
Comparison of two index tests
Unpaired design (between-subjects)
Study Population
Randomization
Index Test 1
Index Test 2
Reference Standard
Sens/Spec of Index Test 1
Sens/Spec of Index Test 2
Comparison of two index tests

Figure 1: Major study designs that are used to compare the diagnostic accuracy of tests

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.[5] 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.[7,8]

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 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.

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].[1] Researchers

Table 2: Sample size estimation formulas

Equations	Explanations
Equation 1	
$n_{Se(unadj)} = \dfrac{Z^2_{\frac{\alpha}{2}} \times Se(1-Se)}{d^2}$ $n_{Sp(unadj)} = \dfrac{Z^2_{\frac{\alpha}{2}} \times Sp(1-Sp)}{d^2}$ $n_{Se\ (prev.adj)} = \dfrac{n_{Se\ (unadj)}}{Prevalance}$ $n_{Sp\ (prev.adj)} = \dfrac{n_{Sp\ (unadj)}}{(1-Prevalance)}$	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-α)% 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)
Equation 2, Comparison of a proportion with null	
$n\ (unadj) = \dfrac{\left[Z_{\frac{\alpha}{2}} \sqrt{P_0(1-P_0)} + Z_\beta \sqrt{P_1(1-P_1)}\right]^2}{(P_1-P_0)^2}$ $n_{Se\ (adj)} = \dfrac{n\ (unadj)}{Prevalance}$ $n_{Sp\ (adj)} = \dfrac{n\ (unadj)}{(1-Prevalance)}$ $n(Yates\ continuity\ correction) = \dfrac{n}{4}\left(1+\sqrt{1+4/(n\lvert P_1-P_2\rvert)}\right)^2$	The estimated proportion (sensitivity or specificity) of the index test (P1), the proportion that we plan to find a statistically significant difference (P0), type 1 error (α), power (β), 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) Yates' continuity correction should be applied (Equation 5)
Equation 3a, Comparison of two unpaired proportions	
$n = \dfrac{\left[Z_\alpha \sqrt{2 \times \bar{P}(1-\bar{P})} + Z_\beta \sqrt{P_1(1-P_1)+P_2(1-P_2)}\right]^2}{(P_1-P_2)^2}$ $n\ (Yates\ continuity\ correction) = \dfrac{n}{4}\left(1+\sqrt{1+4/(n\lvert P_1-P_2\rvert)}\right)^2$	The Formula Set 2 is extended to include both tests $\bar{P}$ denotes the average of the tests' estimated proportions (P1 and P2, 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)
Equation 3b, Comparison of two paired proportions	
$n = \dfrac{\left[Z_\alpha \sqrt{\Psi} + Z_\beta \sqrt{\Psi - (P_2-P_1)^2}\right]^2}{(P_2-P_1)^2}$ $\Psi_{min} = P_2 - P_1$ $\Psi_{max} = P_1 \times (1-P_2) + P_2 \times (1-P_1)$ $n(Yates\ continuity\ correction) = \dfrac{n}{4}\left(1+\sqrt{1+4/(n\lvert P_1-P_2\rvert)}\right)^2$	Ψ is the probability of disagreement between the two tests. Bounds on the Probability of Disagreement (Ψ): The minimum probability of disagreement is P2 - P1. The maximum probability of disagreement is when agreement occurs only by chance, equal to P1 x (1 - P2) + (1 - P1) x P2 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)
Adjusting for disease prevalence (Equations 4a and 4b)	
$n_{Se} = \dfrac{n}{Prevalance}$ $n_{Sp} = \dfrac{n}{(1-Prevalance)}$	Estimated sample sizes should be adjusted for disease prevalence with those equations
Yates' Continuity Correction (Equation 5)	
$n\ (Yates\ continuity\ correction) = \dfrac{n}{4}\left(1+\sqrt{1+4/(n\lvert P_1-P_2\rvert)}\right)^2$	Yates' Continuity Correction should be applied to all calculations comparing two proportions, as described by Beam *et al.*[11]

Table 3: Sample size estimates at predetermined sensitivity and specificity values for various disease prevalence states Values are calculated at marginal errors of (A) 3%, (B) 5%, and (C) 7%.

Sensitivity											Disease prevalence (%)	Specificity										
0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95	0.99		0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95	0.99
											(A) Marginal error of 3%											
106711	105644	102443	97107	89637	80033	68295	54423	38416	20275	4226	1	1078	1067	1035	981	905	808	690	550	388	205	43
21342	21129	20489	19421	17927	16007	13659	10885	7683	4055	845	5	1123	1112	1078	1022	944	842	719	573	404	213	44
10671	10564	10244	9711	8964	8003	6830	5442	3842	2028	423	10	1186	1174	1138	1079	996	889	759	605	427	225	47
7114	7043	6830	6474	5976	5336	4553	3628	2561	1352	282	15	1255	1243	1205	1142	1055	942	803	640	452	239	50
5336	5282	5122	4855	4482	4002	3415	2721	1921	1014	211	20	1334	1321	1281	1214	1120	1000	854	680	480	253	53
4268	4226	4098	3884	3585	3201	2732	2177	1537	811	169	25	1423	1409	1366	1295	1195	1067	911	726	512	270	56
3557	3521	3415	3237	2988	2668	2277	1814	1281	676	141	30	1524	1509	1463	1387	1281	1143	976	777	549	290	60
3049	3018	2927	2774	2561	2287	1951	1555	1098	579	121	35	1642	1625	1576	1494	1379	1231	1051	837	591	312	65
2668	2641	2561	2428	2241	2001	1707	1361	960	507	106	40	1779	1761	1707	1618	1494	1334	1138	907	640	338	70
2371	2348	2277	2158	1992	1779	1518	1209	854	451	94	45	1940	1921	1863	1766	1630	1455	1242	990	698	369	77
2134	2113	2049	1942	1793	1601	1366	1088	768	406	85	50	2134	2113	2049	1942	1793	1601	1366	1088	768	406	85
											(B) Marginal error of 5%											
38416	38032	36879	34959	32269	28812	24586	19592	13830	7299	1521	1	388	384	373	353	326	291	248	198	140	74	15
7683	7606	7376	6992	6454	5762	4917	3918	2766	1460	304	5	404	400	388	368	340	303	259	206	146	77	16
3842	3803	3688	3496	3227	2881	2459	1959	1383	730	152	10	427	423	410	388	359	320	273	218	154	81	17
2561	2535	2459	2331	2151	1921	1639	1306	922	487	101	15	452	447	434	411	380	339	289	230	163	86	18
1921	1902	1844	1748	1613	1441	1229	980	691	365	76	20	480	475	461	437	403	360	307	245	173	91	19
1537	1521	1475	1398	1291	1152	983	784	553	292	61	25	512	507	492	466	430	384	328	261	184	97	20
1281	1268	1229	1165	1076	960	820	653	461	243	51	30	549	543	527	499	461	412	351	280	198	104	22
1098	1087	1054	999	922	823	702	560	395	209	43	35	591	585	567	538	496	443	378	301	213	112	23
960	951	922	874	807	720	615	490	346	182	38	40	640	634	615	583	538	480	410	327	230	122	25
854	845	820	777	717	640	546	435	307	162	34	45	698	691	671	636	587	524	447	356	251	133	28
768	761	738	699	645	576	492	392	277	146	30	50	768	761	738	699	645	576	492	392	277	146	30
											(C) Marginal error of 7%											
19600	19404	18816	17836	16464	14700	12544	9996	7056	3724	776	1	198	196	190	180	166	148	127	101	71	38	8
3920	3881	3763	3567	3293	2940	2509	1999	1411	745	155	5	206	204	198	188	173	155	132	105	74	39	8
1960	1940	1882	1784	1646	1470	1254	1000	706	372	78	10	218	216	209	198	183	163	139	111	78	41	9
1307	1294	1254	1189	1098	980	836	666	470	248	52	15	231	228	221	210	194	173	148	118	83	44	9
980	970	941	892	823	735	627	500	353	186	39	20	245	243	235	223	206	184	157	125	88	47	10
784	776	753	713	659	588	502	400	282	149	31	25	261	259	251	238	220	196	167	133	94	50	10
653	647	627	595	549	490	418	333	235	124	26	30	280	277	269	255	235	210	179	143	101	53	11
560	554	538	510	470	420	358	286	202	106	22	35	302	299	289	274	253	226	193	154	109	57	12
490	485	470	446	412	368	314	250	176	93	19	40	327	323	314	297	274	245	209	167	118	62	13
436	431	418	396	366	327	279	222	157	83	17	45	356	353	342	324	299	267	228	182	128	68	14
392	388	376	357	329	294	251	200	141	74	16	50	392	388	376	357	329	294	251	200	141	74	16

Type 1 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

Table 4: Sample size estimates for a difference of at least 5% in co-primary endpoints with a Type 1 error of 5% and A) Power of 90%, and B) 80%

Predetermined Sens/Spec (P0)										Expected Sens/Spec (P1) (%)
50%	55%	60%	65%	70%	75%	80%	85%	90%	95%	
(A) Power 90%										
1086										55
278	1067									60
126	271	1027								65
71	121	259	966							70
45	68	115	242	884						75
31	43	64	106	219	781					80
22	29	40	58	95	190	656				85
16	20	26	35	51	80	156	510			90
12	14	18	23	30	41	63	115	340		95
7	9	10	12	15	19	24	34	53	109	100
(B) Power 80%										
822										55
213	809									60
98	209	781								65
56	95	201	737							70
36	54	91	188	677						75
25	35	52	85	172	601					80
19	24	33	48	77	151	510				85
14	18	23	30	43	67	127	402			90
11	13	16	20	26	36	54	97	277		95
7	9	10	12	15	19	24	34	53	109	100

Type 1 error is accepted as 5% for all calculations, Yates' continuity correction is applied

Table 5: Sample size estimation for comparing two independent proportions, unpaired groups for A) Power of 90% and B) 80%

P1										P2 (%)
55%	60%	65%	70%	75%	80%	85%	90%	95%	100%	
(A) Power 90%										
1746	442	197	111	70	48	34	25	19	15	50
	1712	429	190	105	66	44	31	23	17	55
		1644	408	178	98	60	40	28	20	60
			1541	378	163	88	54	35	24	65
				1404	339	144	76	45	29	70
					1232	292	121	62	36	75
						1027	236	94	46	80
							787	172	64	85
								513	98	90
									203	95
(B) Power 80%										
1272	325	146	83	53	37	26	20	15	12	50
	1247	315	141	79	50	34	24	18	14	55
		1198	300	133	74	46	31	22	16	60
			1124	278	121	67	41	27	19	65
				1025	250	108	58	35	23	70
					901	216	91	48	28	75
						753	176	72	36	80
							579	129	50	85
								381	76	90
									157	95

Type 1 error is accepted as 5% for all calculations. Yates' continuity correction is applied. Estimates here are for each of the independent groups

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 ascertained 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].[9] 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 1 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 probability of 95% (P1, contrast-enhanced CT) and null value of 90% (P0, noncontrast-enhanced CT) reveals that at least 340 diseased subjects are needed (patients with acute appendicitis confirmed with surgery). We use Equations 4a and 4b in Table 2 to adjust for prevalence (acute appendicitis prevalence is 40%, we divide 340 by 0.4 = 849). For this study, at least 849 subjects with a suspected acute appendicitis are needed. Please be reminded that those calculations are corrected with Yates' continuity correction.

Table 6A: Sample size estimation for the comparison of two dependent proportions, paired groups. A) N (▯_min), B1) N (Ψ_max) for power of 90%, B2) N (Ψ_max) for power of 80%

	(A) N (Ψ_min)	
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%**	
				(B1) N (Ψ_max), Power 90%						
1749	444	200	113	72	50	37	28	21	17	50
	1715	431	192	108	68	47	34	25	19	55
		1646	410	181	100	63	42	30	22	60
			1543	380	165	90	56	37	26	65
				1406	341	146	79	48	31	70
					1235	294	123	65	38	75
						1029	238	97	48	80
							789	174	66	85
								515	101	90
									206	95
				(B2) N (Ψ_max), Power 80%						
1274	327	148	85	55	39	28	22	17	14	50
	1249	317	143	81	52	36	26	20	16	55
		1200	302	135	76	48	33	24	18	60
			1126	280	124	69	43	29	21	65
				1027	252	110	60	37	25	70
					903	218	93	50	30	75
						755	178	74	38	80
							581	132	52	85
								383	78	90
									159	95

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 ($n/(1 - \text{prevalence}) = 656/(1 - 0.4)$) 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 [$\Psi_{min}=P_2\text{-}P_1$]). Conversely, they agree with each other just by chance, where the probability of disagreement is maximum ($\Psi_{max}=P_1\times(1\text{-}P_1)+P_2(1\text{-}P_1)$). 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.[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, journalogy, 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 2021;184:E5-9.
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.