mathematics

MDPI

Article

Common Medical and Statistical Problems: The Dilemma of the Sample Size Calculation for Sensitivity and Specificity Estimation

M. Rosário Oliveira [1,*,†], Ana Subtil [1,†] and Luzia Gonçalves [2]

[1] Department of Mathematics and CEMAT, Instituto Superior Técnico, Universidade de Lisboa, Av. Rovisco Pais, 1049-001 Lisbon, Portugal; anasubtil@tecnico.ulisboa.pt

[2] Unidade de Saúde Pública Internacional e Bioestatística, Global Health and Tropical Medicine, Instituto de Higiene e Medicina Tropical, Universidade Nova de Lisboa and Centro de Estatística e Aplicações da Universidade de Lisboa, Rua da Junqueira, 100, 1349-008 Lisbon, Portugal; luziag@ihmt.unl.pt

* Correspondence: rosario.oliveira@tecnico.ulisboa.pt

† These authors contributed equally to this work.

Received: 29 June 2020; Accepted: 29 July 2020; Published: 1 August 2020

Abstract: Sample size calculation in biomedical practice is typically based on the problematic Wald method for a binomial proportion, with potentially dangerous consequences. This work highlights the need of incorporating the concept of conditional probability in sample size determination to avoid reduced sample sizes that lead to inadequate confidence intervals. Therefore, new definitions are proposed for coverage probability and expected length of confidence intervals for conditional probabilities, like sensitivity and specificity. The new definitions were used to assess seven confidence interval estimation methods. In order to determine the sample size, two procedures—an optimal one, based on the new definitions, and an approximation—were developed for each estimation method. Our findings confirm the similarity of the approximated sample sizes to the optimal ones. R code is provided to disseminate these methodological advances and translate them into biomedical practice.

Keywords: sample size; sensitivity; specificity; conditional probability; coverage probability

1. Introduction

Diagnostic tests are helpful tools in medical decision-making, since they give indication of the disease or infection status. Assessing the relevance and utility of such tests through the estimation of sensitivity (Se) and specificity (Sp), among other measures, is crucial to make informed choices on their use. During the COVID-19 pandemic, the concepts of diagnostic tests and their performance have been a hot topic of debate, even outside the scientific community.

Sensitivity and specificity are conventional measures of diagnostic test accuracy. Sensitivity is the probability of a positive test result given that the individual has the condition (infection or disease), i.e., the probability of correctly identifying an individual with the condition. Specificity is the probability of a negative test result given that the individual does not have the condition, i.e., the probability of correctly identifying an individual without the condition.

Interval estimation has been strongly recommended by [1] and increasingly adopted in the biomedical literature. To improve the quality of reporting of diagnostic accuracy studies, the Standards for Reporting Diagnostic Accuracy (STARD) statement was originally published in 2003 and updated in 2015—STARD 2015 [2]. Since the first version of STARD, 95% confidence intervals (CI) for the accuracy estimates of diagnostic tests have been recommended. However, several works have failed to comply with these guidelines [3] and obsolete methods continue to be used and not well-described to obtain 95% confidence intervals.

For a binomial proportion, the Wald method has been the most popular and widespread method over the years, because it is simple to teach, understand, and use. However, in the last two decades, theoretical statistical research (e.g., [4–8]) has reported its drawbacks, which are particularly severe for small sample sizes and proportions near 0 or 1. Some researchers have addressed the subject of interval estimation for a binomial proportion [7–9], exploring and comparing alternative methods, and providing recommendations concerning their performance. Regarding criteria to evaluate the performance of interval estimation methods, coverage probability (CP) and expected length (EL) are two of the most frequently used.

Reporting how the sample size determination was conducted is also an important step in scientific research. In the biomedical literature, there is an under-reporting of the required sample size and the achieved sample size at the end of the study. The updated list of STARD 2015 added the need for methodological details on sample size [2]. Once again, sample size calculation is dominated by the Wald method to estimate a binomial proportion. [10] emphasizes that confidence intervals for proportions are influenced, not only by the actual estimate of the proportion, but also by the sample size, and argues that there is no totally reliable sample size choice to achieve a desired goal.

In diagnostic test accuracy, an additional problem derived from the wrong replacement of the conditional probabilities (sensitivity and specificity) by a binomial proportion, at least from a theoretical point of view. The maximum likelihood estimator of sensitivity (specificity) is a ratio that depends on the number of individuals in the sample with (without) the condition, which is a random variable with a binomial distribution that depends on the prevalence. This situation requires a new approach to modify the expressions of coverage probability and expected length, in order to use these two criteria to compare the performance for different interval estimation methods and for the sample size calculation, avoiding the traditional Wald method.

In this work, new expressions for coverage probability and expected length of a conditional probability interval, like sensitivity and specificity, were developed. These expressions are used to compare alternative interval estimation methods and to determine optimal sample sizes using simple random sampling. Moreover, we compare these sample sizes based on the rewritten formula of the expected length with approximated sample sizes. The approximated procedure has two steps. The first one considers sample sizes based on a binomial proportion (as calculated in [11]) to estimate the number of patients. The second step determines the sample size by the ratio of the previous number of subjects with (without) the condition to the prevalence (1-prevalence), following [12]. Differences between optimal and approximate procedures were explored, in order to evaluate if the simpler procedure based on an approximation has some practical use.

Practical problems arising from real studies are also discussed. Among several studies presented in the literature, we focus on examples related to two diseases, hepatitis B [13] and depression [14], which illustrate the diversity and complexity of problems found in practice. Regarding the diagnostic accuracy of tests to detect hepatitis B surface antigen, [13] presented a systematic review of the literature, including a meta-analysis. Sample sizes of the 40 reported studies ranged from 50 to 3956 (median size: 284). A survey of published papers on the diagnostic accuracy of depression screening tools [14] included 89 studies with sample sizes varying from 34 to 42,676 (median size: 224).

In different studies for the same disease, a substantial heterogeneity was found in terms of sample size, sensitivity and specificity estimates, reporting of their uncertainty, and identification of the statistical methods used. Another important issue is the prevalence of the disease in the population under study. For hepatitis B, [13] referred values range from 1.9% to 84.0%. Samples are frequently drawn from sub-populations with a higher probability of having the disease than the overall population. Despite efforts to improve reporting of diagnostic accuracy (e.g., STARD 2015 guidelines [2]), individual studies based on small sample sizes are often published. This usually leads to confidence intervals with an undesirably large width. In the survey of published studies on depression, even in journals with a high impact factor, authors reported that only "...34% of the studies provided reasonably accurate intervals for estimates of sensitivity and specificity ..." (p. 147

in [14]). This fact can mislead researchers without solid statistical background on the trustfulness of statistical tools in the biomedical practice. A major challenge in this context lies in the absence of a single statistical response to the high diversity of practical situations. Thus, statistically sound methods and associated software need to be developed to help researchers in the biomedical areas.

2. Materials and Methods

2.1. Interval Estimation Using Different Methods

The methods for constructing confidence or credibility intervals under analysis in the present study were selected among the ones that seem most promising according to previous research [6–8,11]: Clopper–Pearson, Anscombe, Agresti–Coull, Bayesian with Uniform prior and with Jeffreys prior, and Wilson, for a binomial proportion. For comparison reasons, the Wald method was also included in the study, given its common application in biomedical research. These methods provide two-sided confidence intervals for an unknown binomial proportion in the population, p, based on a sample of size n. A nominal confidence level, $100 \times (1-\alpha)\%$, is pre-specified for the intervals, which means that the probability of including p, the so-called coverage probability, is intended to be $(1-\alpha)$.

Some notation and expressions for coverage probability and expected length are summarized in Appendix A. The expressions for the lower and upper bounds of these methods can be found in Appendix B and the code for their implementation is available in Supplement S1. For more details about these methods see [8].

In terms of coverage probability, [8] distinguish three classes of interval estimation methods. The strictly conservative methods have minimum coverage probability greater or equal to $(1-\alpha)$ for all the values of p and n. A second class includes the methods which are correct on average, i.e., for each n fixed, the mean coverage probability over all possible values of p is greater than or equal to $(1-\alpha)$. There is a third class comprising the remaining methods, which are neither strictly conservative, nor correct on average. Accordingly, Clopper–Pearson and Anscombe methods can be stated as strictly conservative, while Bayesian-U, Jeffreys, Wilson, and Agresti–Coull can be classified as correct on average. The Wald method belongs to the third group and stands out for displaying quite low coverage probabilities for low or high values of p, which makes it an unreliable method.

2.2. New Expressions for Coverage Probability and Expected Length of a Conditional Probability Interval

In order to become adequate for the sensitivity and specificity, as well as other conditional probabilities, the conventional formulas for coverage probability and expected length of a confidence interval (presented in Appendix A) should be updated. If the proportion p, for which we aim to find an interval estimate, is the sensitivity or specificity, then p is a conditional probability. Accordingly, the usual estimator of the sensitivity (specificity) of a given diagnostic test is a ratio that depends on the state of the patient. Although the sample size, n, is known in advance, the number of individuals with and without the condition among the n individuals actually observed, N_D and $N_{\overline{D}}$, respectively, are random variables. As a consequence, for example, N_D has a binomial distribution with the parameters n and η, in which η is the prevalence, i.e., the proportion of individuals with the condition in the population under study. In fact, if $p = \text{Se}$, then $(X|N_D = m_D)$ is the number of individuals that tested positive among the m_D that have the condition in the sample of size n. It follows that $(X|N_D = m_D) \sim binomial(m_D, \text{Se})$. Similarly, being $(Y|N_{\overline{D}} = m_{\overline{D}})$ the number of individuals that tested negative among the $m_{\overline{D}}$ that do not have the condition in the sample of size n, we have $(Y|N_{\overline{D}} = m_{\overline{D}}) \sim binomial(m_{\overline{D}}, \text{Sp})$.

Instead of the expressions for the conventional coverage probability and expected length (see Appendix A, respectively), the proper definitions for sensitivity are the following:

$$\text{CP}(\text{Se}, n, \eta) = \sum_{m=1}^{n} \text{CP}(m, \text{Se}) Pr\{N_D = m | N_D > 0\}, \quad (1)$$

$$\mathrm{EL}(\mathrm{Se}, n, \eta) = \sum_{m=1}^{n} \mathrm{EL}(m, \mathrm{Se}) Pr\{N_D = m | N_D > 0\}. \quad (2)$$

Since it makes no sense to build an interval estimate if the number of individuals with the condition in the sample is null ($N_D = 0$), the restriction $N_D > 0$ in Equations (1) and (2) was added. In a simple random sampling scheme like the one we adopt, N_D has a binomial distribution. This approach may be extended to other sampling schemes by using appropriate distributions for N_D.

According to these new definitions, the coverage probability (expected length) associated to a sensitivity interval emerges as an expected value of coverage probability (expected length) taken for all possible non-null values of N_D. Formulas for the coverage probability and expected length of a specificity interval or other conditional probability derive from the same rationale and have analogous interpretations.

2.3. Optimal and Approximate Sample Size Determination

Regarding sample size determination, the previous definitions have some implications to calculate the sample size, n, to obtain a $100 \times (1 - \alpha)\%$ interval estimate for p, with a desired expected width, ω, i.e., we seek for n such that

$$\mathrm{EL}(n, p) = \omega \, . \quad (3)$$

According to [11], for different confidence interval methods, these equations may not have neither a closed-form, nor an integer solution, but it is always possible to find an integer, n, that minimizes $|\mathrm{EL}(n, p) - \omega|$ within a certain tolerance, ξ, which in most situations will be such that $\mathrm{EL}(n, p) \simeq \omega$. Consequently, we can use a procedure to determine all the values n verifying:

$$|\mathrm{EL}(n, p) - \omega| \leq \xi \, . \quad (4)$$

If multiple solutions are found, we select the one that satisfies a chosen criteria, e.g., the one that maximizes coverage probability. Besides this option, the provided code enables other alternatives. If no solution is found, it is possible to increase ξ.

To calculate adequate sample sizes for the estimation of sensitivity and specificity with the desired confidence, we followed this procedure using the expected length new definition (2) stated above (an analogous expression was applied in the case of specificity). This procedure is designated *optimal*.

In parallel, following a rationale similar to the one described in [12] for the Wald method, an approximated procedure is also used and compared with the optimal one, now considering other methods besides the Wald. This approximated procedure took previously reported values for interval estimates of binomial proportions obtained by [11] as the number of subjects with (without) the disease required for sensitivity (specificity) estimation, $n_{(\mathrm{Se})}$ (or $n_{(\mathrm{Sp})}$). Theoretically, $\mathrm{E}(N_D) = \eta \cdot n$. In practice, $\mathrm{E}(N_D)$ is estimated by $n_{(\mathrm{Se})}$, hence $n = n_{(\mathrm{Se})}/\eta$ is an approximation of the optimal sample size for sensitivity estimation. Following the same reasoning, in the case of specificity, the corresponding sample size is approximated by $n = n_{(\mathrm{Sp})}/(1 - \eta)$.

All calculations were performed using the statistical software R [15] and the code can be found in Supplementary Materials.

3. Results

3.1. New Expressions for Coverage Probability of Interval Estimation Methods

Comparisons of new and conventional expressions of coverage probability and expected length were performed, using the seven methods, for some practical examples. The choice of the best methods is essentially based on their coverage probability. Nevertheless, this indicator varies with the sensitivity (specificity) of the diagnostic test under study, the prevalence of the condition, and the sample size. Given the diversity of practical examples, we adopted specific values for prevalence, sensitivity,

and specificity motivated by a study presented in [16] concerning the diagnosis of dengue fever, an important vector-borne disease common in tropical areas.

Coverage probabilities were calculated considering different range values of sensitivity, prevalence, and sample size.

Figure 1 shows the coverage probability of sensitivity intervals as a function of sensitivity, for $\eta = 0.25$ and $n = 500$, obtained with the new formula and the conventional formula, using Clopper–Pearson (strictly conservative method) and Wilson (correct on average). The new formula originates much smoother coverage probability curves, in contrast with the sharper indentations obtained with the conventional expressions.

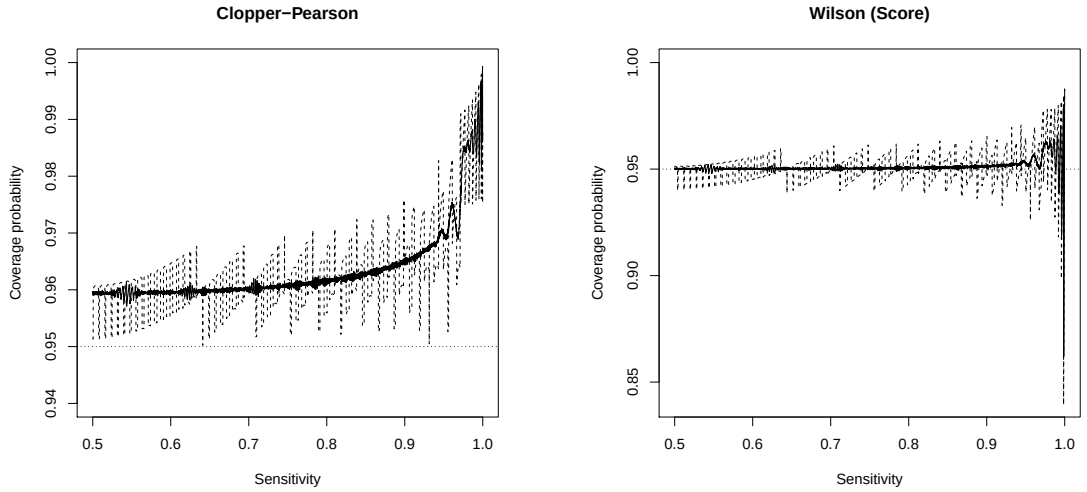

Figure 1. Coverage probability of sensitivity intervals varying with the sensitivity, for $\eta = 0.25$ and $n = 500$, obtained with the new formula (solid line) and the conventional formula (dashed line), for Clopper–Pearson (left) and Wilson (right). The horizontal spotted line marks the nominal confidence level, 95%.

Figure 2 illustrates the coverage probability of a sensitivity interval, admitting two prevalence values (0.10 and 0.25) and a sample size of 500 individuals, according to four methods (Clopper–Pearson, Agresti–Coull, Wilson, and Wald).

Anscombe, Jeffreys, and Bayesian-U are not shown in the plots given their similarity with other methods. For most sensitivity values, the coverage probability corresponding to the highest prevalence, $\eta = 0.25$, tends to be closer to the nominal confidence level than $\eta = 0.10$. This tendency is not so clear for sensitivity closer to 1, where the coverage probability is quite unstable and erratic.

For the studied methods, regarding coverage probability values, we can recognise a section closer to Se $= 1$ with erratic and unstable values, a section closer to Se $= 0.5$ with stable values, and an intermediate section in between. Strictly conservative methods (Clopper–Pearson and Anscombe), present higher coverage probability than the nominal confidence level, with curves that seem detached from this target. The Wilson, Jeffreys, and Bayesian-U methods are similar, exhibiting stability around the nominal confidence level, except for erratic values close to 1. The Agresti–Coull method seems in between the Wilson and the higher detached values calculated for the strictly conservative methods.

It is also important to study the coverage probability associated with the sensitivity interval as a function of the prevalence. The coverage probability tends to be nearer the target as the prevalence increases towards $\eta = 0.5$. Figure 3 presents two sensitivities, 0.96 and 0.73, for a sample size of 500 individuals. The curve tends to be nearer the nominal confidence level in the case of the lower sensitivity (0.73).

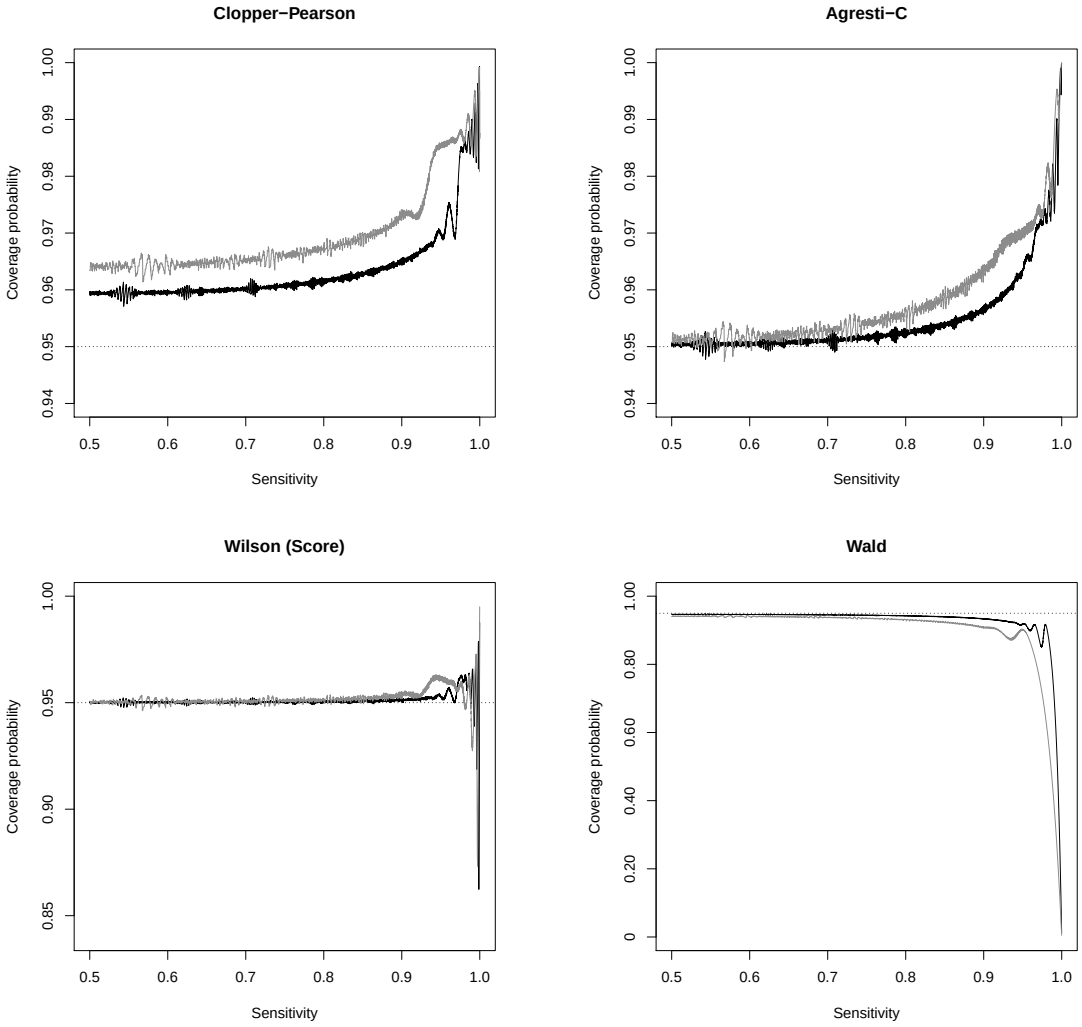

Figure 2. Coverage probability of sensitivity intervals varying with the sensitivity, admitting $n = 500$, obtained for $\eta = 0.10$ (grey line) and $\eta = 0.25$ (black line), for Clopper–Pearson (top left), Agresti–Coull (top right), Wilson (bottom left), and Wald (bottom right). The horizontal spotted line marks the nominal confidence level, 95%.

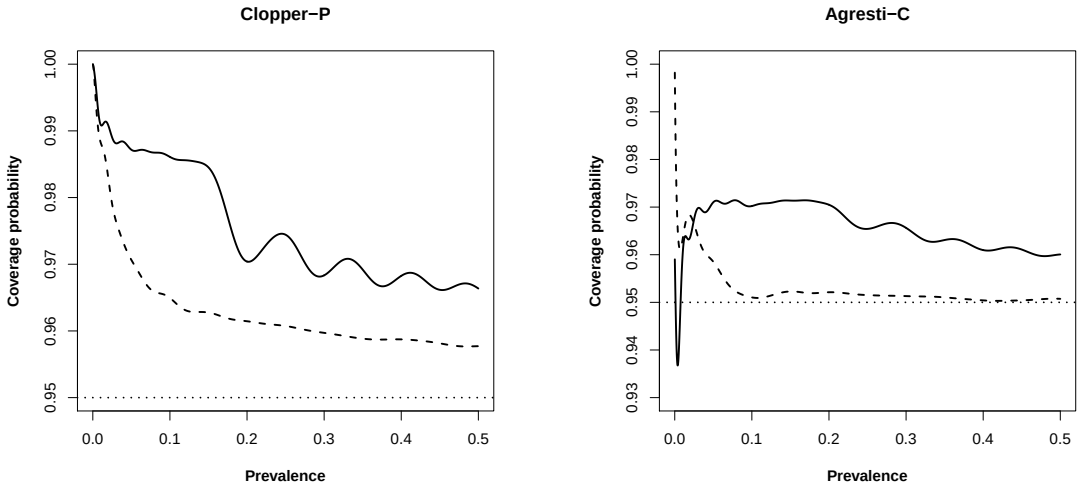

Figure 3. *Cont.*

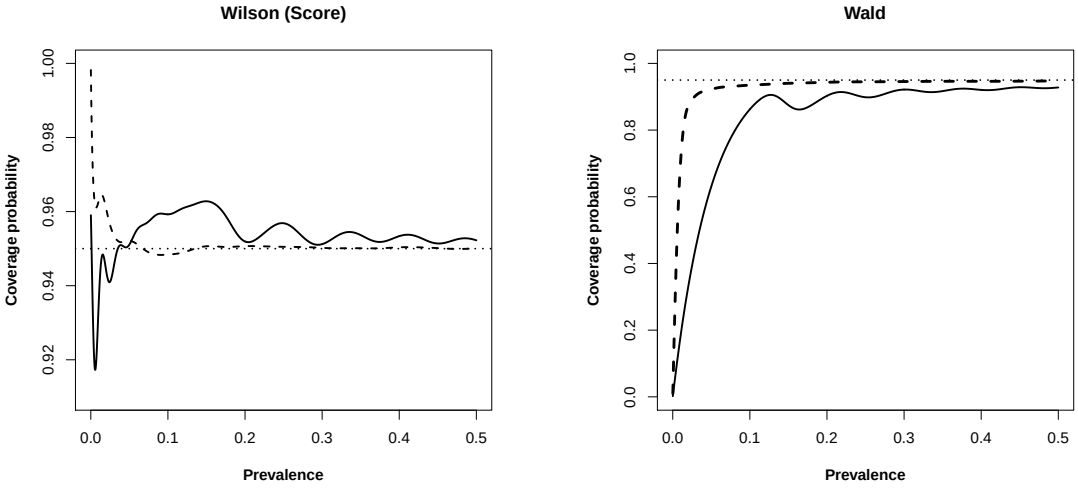

Figure 3. Coverage probability of sensitivity intervals varying with the prevalence, admitting $n = 500$, obtained for Se $= 0.96$ (solid line) and Se $= 0.73$ (dashed line), for Clopper–Pearson (top left), Agresti–Coull (top right), Wilson (bottom left), and Wald (bottom right). The horizontal spotted line marks the nominal confidence level, 95%.

3.2. Impact on Sample Sizes

The sample sizes based on the optimal procedure were calculated for 95% intervals for sensitivity or specificity, with expected width $\omega = 0.05$ and a tolerance of $\xi = 10^{-4}$. Approximated sample sizes for sensitivity intervals were obtained from the values reported by [11], divided by the prevalence, $\eta = 0.10$, and, for specificity, the divisor was $(1 - \eta) = 0.90$.

Tables 1 and 2 show the optimal sample sizes corresponding to sensitivity and specificity, respectively. Both tables also present the differences between optimal and approximate sample sizes, δ.

Only small differences were detected between optimal and approximate sample sizes. As sensitivity (specificity) increases, the sample sizes required to satisfy the established criteria decrease. Moreover, the sample sizes needed in the case of sensitivity are much higher than the sample sizes for the particular case of specificity, as expected if the prevalence is smaller than 0.5. In the majority of practical situations, the prevalence is less than 0.5 and therefore the infected or diseased subjects are less represented in the sample, thus demanding higher sample sizes to guarantee an adequate sensitivity interval.

Table 1. Optimal sample sizes ($n_{optimal}$) corresponding to several sensitivities, and differences between the optimal and approximate (n_{aprox}) sample sizes, $\delta = n_{optimal} - n_{aprox}$, admitting $\eta = 0.10$, $\omega = 0.05$, $\xi = 10^{-4}$, and 95% nominal confidence level.

	Clopper-Pearson		Anscombe		Agresti-Coull		Bayesian Uniform		Jeffreys		Wilson		Wald	
Se	$n_{optimal}$	δ	$n_{optimal}$	δ	$n_{optimal}$	δ	$n_{optimal}$	δ	$n_{optimal}$	δ	$n_{optimal}$	δ	$n_{optimal}$	δ
0.75	11848	−2	11855	5	11459	−1	11449	−11	11452	2	11538	78	11472	2
0.80	10165	−45	10172	−8	9869	39	9771	1	9847	37	9847	77	9864	54
0.85	8176	−34	8184	4	7829	−1	7852	22	7781	−29	7853	23	7855	45
0.90	5880	−10	5890	0	5589	−21	5557	27	5488	−2	5546	26	5532	12
0.91	5385	5	5395	5	5112	2	5024	−26	5032	22	5031	−19	5029	29
0.92	4877	7	4887	−23	4626	6	4559	39	4521	31	4532	2	4514	14
0.93	4357	7	4368	−22	4165	35	4045	35	3998	8	4024	4	3984	14
0.94	3824	−16	3836	6	3638	−2	3495	5	3463	33	3535	25	3441	11
0.95	3280	0	3293	3	3142	2	2991	21	2916	26	3008	28	2882	12
0.96	2723	3	2737	7	2653	−7	2440	0	2341	1	2461	1	2302	12
0.97	2170	20	2185	15	2178	−2	1915	5	1796	16	1944	4	1613	−7
0.98	1584	4	1595	5	1722	2	1406	6	1272	12	1462	12	489	9
0.99	1079	9	1079	9	1284	14	931	11	913	13	1040	10	NA	NA

Table 2. Optimal sample sizes ($n_{optimal}$) corresponding to several specificities, and differences between the optimal and approximate (n_{aprox}) sample sizes, $\delta = n_{optimal} - n_{aprox}$, admitting $\eta = 0.10$, $\omega = 0.05$, $\xi = 10^{-4}$, and 95% nominal confidence level.

	Clopper-Pearson		Anscombe		Agresti-Coull		Bayesian Uniform		Jeffreys		Wilson		Wald	
Sp	$n_{optimal}$	δ	$n_{optimal}$	δ	$n_{optimal}$	δ	$n_{optimal}$	δ	$n_{optimal}$	δ	$n_{optimal}$	δ	$n_{optimal}$	δ
0.50	1741	−2	1742	−1	1696	0	1709	11	1697	−1	1696	0	1700	0
0.55	1724	−1	1738	12	1691	2	1685	−4	1680	0	1685	−4	1695	1
0.60	1673	−1	1674	−2	1628	−1	1640	11	1629	0	1640	11	1632	−2
0.65	1600	12	1601	12	1543	−2	1543	−2	1556	11	1543	−2	1547	0
0.70	1469	−1	1471	1	1428	2	1425	−1	1436	10	1424	−2	1433	1
0.75	1325	8	1317	0	1273	−1	1275	1	1282	9	1281	7	1274	−1
0.80	1130	−5	1138	6	1093	0	1093	7	1093	3	1093	7	1095	5
0.85	908	−5	915	6	870	0	865	−5	865	−3	871	1	870	2
0.90	657	2	654	−1	621	−3	616	1	610	0	617	3	614	0
0.91	602	4	603	4	571	3	561	−1	558	1	559	−3	558	2
0.92	545	3	546	0	514	0	502	−1	498	−1	503	−1	500	0
0.93	484	0	485	−3	462	3	446	0	443	−1	447	0	442	0
0.94	425	−2	426	0	405	0	390	2	384	2	392	2	381	−1
0.95	366	1	366	0	349	0	331	1	323	1	333	1	319	0
0.96	304	1	305	1	294	−2	271	−1	260	0	274	0	254	−1
0.97	240	1	241	−1	242	−1	213	0	198	0	216	0	179	−1
0.98	176	0	177	0	191	−1	156	0	140	0	161	−1	54	0
0.99	119	0	119	0	142	0	103	0	100	0	114	−1	NA	NA

In the hepatitis B meta-analysis performed by [13], the pooled sensitivity and specificity are 0.90 and 0.995, respectively. Considering a prevalence of 0.10 for a particular population of suspected cases, $\omega = 0.05$, and $\xi = 10^{-4}$, the optimal sample sizes for the reported sensitivity range from 5446 (Wilson) to 5890 (Anscombe), according to Table 1. Being so, all the 40 studies fail to provide reliable accurate sensitivity intervals. By contrast, if the target is the Sp = 0.99, optimal sample sizes range from 100 (Jeffreys) to 142 (Agresti–Coull), as presented in Table 2. Therefore, 90% and 80% of the studies fulfil the Jeffreys and Agresti–Coull requirements, respectively.

The conservative methods, Clopper–Pearson and Anscombe, demand higher sample sizes than the remaining methods. The sample sizes corresponding to the Agresti–Coull method are intermediate between the conservative methods and Bayesian-U, Jeffreys, and Wilson.

Tables 1 and 2 show optimal sample sizes associated with expected interval width $\omega = 0.05$. However, taking a closer look at the studies reported in the hepatitis B meta-analysis [13], we can see that many studies provide results with much wider confidence intervals, which, although of limited interest in some scenarios, may be useful in others, by ethical and economical reasons. Therefore, ranging the interval width from 0.05 to 0.10, Tables 3 and 4 present optimal sample sizes for sensitivity and specificity intervals, admitting a prevalence of 0.10, for Clopper–Pearson and Wilson methods.

As expected, wider confidence intervals correspond to smaller optimal sample sizes. For example, if the researcher anticipates a sensitivity of 0.95 (and a prevalence of 0.10), a Clopper–Pearson interval of length $\omega = 0.05$ would require a sample of 3280 observations to fulfil the desired requirements. Increasing the interval width by 0.01 to $\omega = 0.06$, the number of observations ($n_{optimal} = 2326$) needed is reduced by 954. Moreover, if the initial width doubles ($\omega = 0.10$), the optimal sample size decreases to 905, which is only 27.6% of the initial value.

Figures 4 and 5 allow us to explore the expected length for desired confidence intervals as a function of optimal sample size, for specific values of prevalence, sensitivity (Figure 4) and specificity (Figure 5). In general, the expected lengths of the intervals decrease for high sensitivity and specificity, for the same sample size and prevalence.

Table 3. Optimal sample sizes ($n_{optimal}$) corresponding to several sensitivities, for the Clopper–Pearson and Wilson methods, with ω varying between 0.05 and 0.10, admitting $\eta = 0.10$, $\xi = 10^{-4}$, and 95% nominal confidence level.

	Interval Width (ω)											
	0.05		**0.06**		**0.07**		**0.08**		**0.09**		**0.10**	
Se	**Clopper-Pearson**	**Wilson**	**Clopper-Pearson**	**Wilson**	**Clopper-Pearson**	**Wilson**	**Clopper-Pearson**	**Wilson**	**Clopper-Pearson**	**Wilson**	**Clopper-Pearson**	**Wilson**
0.75	11,848	11,538	8280	7997	6119	5831	4711	4459	3742	3532	3046	2854
0.80	10,165	9847	7110	6826	5259	5006	4053	3825	3222	3003	2625	2428
0.85	8176	7853	5728	5445	4243	5445	3275	3995	2607	3054	2126	2399
0.90	5880	5546	4133	3838	3071	2839	2377	2165	1897	1719	1552	1389
0.91	5385	5031	3789	3523	2818	2578	2183	1987	1744	1567	1428	1272
0.92	4877	4532	3436	3158	2559	2340	1984	1797	1587	1419	1301	1158
0.93	4357	4024	3074	2808	2293	2074	1781	1604	1426	1270	1170	1036
0.94	3824	3535	2705	2454	2021	1827	1573	1404	1262	1121	1040	917
0.95	3280	3008	2326	2098	1743	1569	1360	1217	1094	976	905	801
0.96	2723	2461	1951	1742	1460	1307	1148	1028	925	829	768	691
0.97	2170	1944	1555	1403	1174	1064	930	849	763	697	639	586
0.98	1584	1462	1167	1080	903	845	732	686	613	576	525	494
0.99	1079	1040	833	807	679	659	571	554	491	476	431	417

Table 4. Optimal sample sizes ($n_{optimal}$) corresponding to several specificities, for the Clopper–Pearson and Wilson methods, with ω varying between 0.05 and 0.10, admitting $\eta = 0.10$, $\xi = 10^{-4}$, and 95% nominal confidence level.

	Interval Width (ω)											
	0.05		**0.06**		**0.07**		**0.08**		**0.09**		**0.10**	
Sp	**Clopper-Pearson**	**Wilson**	**Clopper-Pearson**	**Wilson**	**Clopper-Pearson**	**Wilson**	**Clopper-Pearson**	**Wilson**	**Clopper-Pearson**	**Wilson**	**Clopper-Pearson**	**Wilson**
0.50	1741	1696	1222	1184	901	864	692	663	547	521	445	421
0.55	1724	1685	1203	1172	888	855	683	656	542	517	440	418
0.60	1673	1640	1171	1130	862	833	665	636	527	500	428	405
0.65	1600	1543	1115	1077	822	786	630	601	501	474	406	384
0.70	1469	1424	1028	988	760	726	583	556	464	437	377	354
0.75	1325	1281	925	887	680	650	523	497	416	391	339	316
0.80	1130	1093	790	753	586	553	450	424	358	334	291	270
0.85	908	871	636	604	471	443	364	338	289	267	236	215
0.90	657	617	459	426	342	314	264	240	211	190	172	154
0.91	602	559	421	390	313	286	242	220	193	174	158	141
0.92	545	503	383	351	284	259	220	199	176	157	144	128
0.93	484	447	341	313	255	231	198	177	158	141	130	115
0.94	425	392	301	273	225	202	174	156	140	124	115	101
0.95	366	333	258	233	193	173	151	134	121	108	100	88
0.96	304	274	216	194	162	145	127	113	102	92	85	76
0.97	240	216	172	155	130	118	103	93	84	77	70	64
0.98	176	161	129	119	100	93	81	76	67	63	57	54
0.99	119	114	92	89	75	72	63	61	54	52	47	45

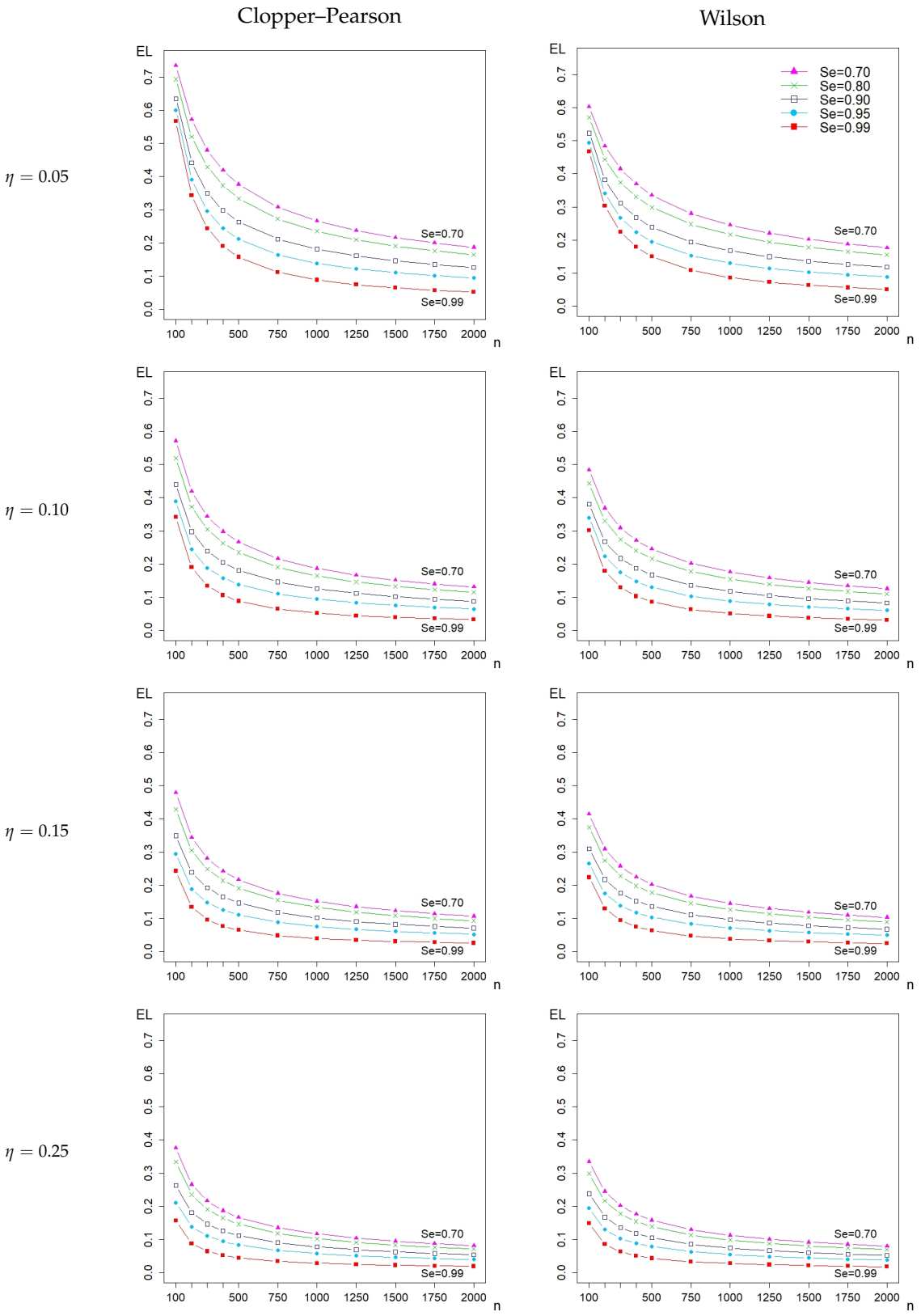

Figure 4. Expected length of sensitivity intervals varying with the sample size (n), for $\eta = 0.05; 0.10; 0.15; 0.25$, and Se $= 0.70; 0.80; 0.90; 0.95; 0.99$, using Clopper–Pearson and Wilson methods with 95% nominal confidence level.

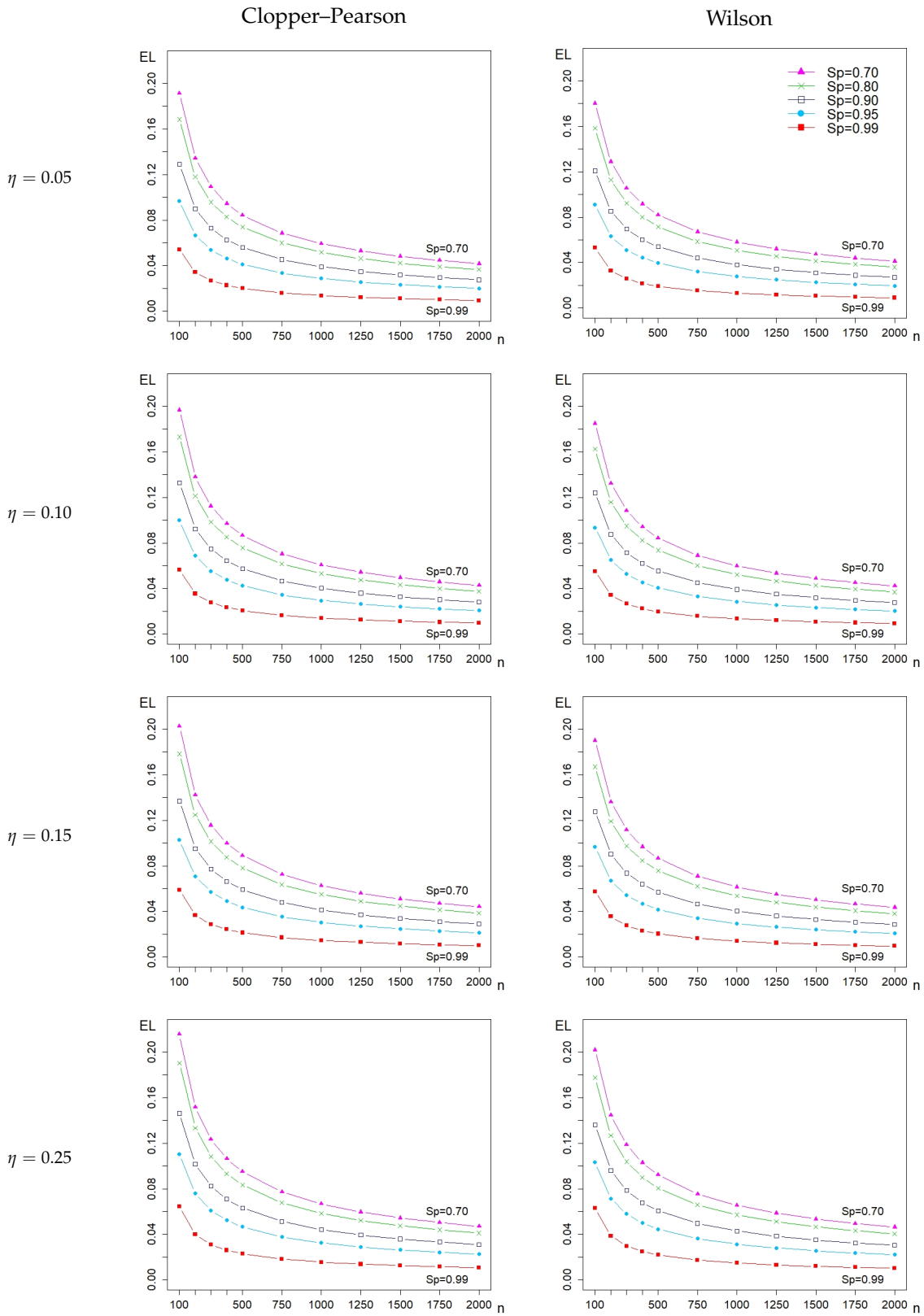

Figure 5. Expected length of specificity intervals varying with the sample size (n), for $\eta = 0.05; 0.10; 0.15; 0.25$, and Sp $= 0.70; 0.80; 0.90; 0.95; 0.99$, using Clopper–Pearson and Wilson methods with 95% nominal confidence level.

Higher sample sizes and higher prevalence lead also to confidence intervals with lower expected length. The same figures also confirmed findings already discussed: (i) Wilson, being correct on average, leads to smaller optimal sample sizes than Clopper–Pearson, a strictly conservative method; (ii) optimal sample sizes designed for specificity intervals require smaller values than the ones demanded for sensitivity intervals, when the prevalence of the condition is smaller that 0.5.

Illustrating again with the work of hepatitis B surface antigen [13], when the experiment was designed, if the authors anticipate a sensitivity of 0.95 for the new test and they want a confidence interval with an expected width of $\omega = 0.10$, then, according to Figure 4, for a prevalence of the population under study equal to 0.05, a sample of size 1750 would be needed. If this prevalence doubles ($\eta = 0.10$) the optimal sample size decreases to 1000, and for $\eta = 0.25$ a sample size between 300 and 400 would be enough to obtain a Clopper–Pearson sensitivity interval with the desired width.

4. Discussion and Conclusions

New formulas for coverage probability and expected length were derived in this work. These expressions are more suitable for comparing alternative confidence interval methods and for determining optimal sample size when a conditional probability is the estimation target and the individuals' condition is not known in advance, under simple random sampling. The new formulas are transposable to other equally useful proportions that are conditional probabilities, besides sensitivity and specificity, such as the positive and negative predictive values. This approach may be extended to other types of random sampling.

The new coverage probability expression originates much smoother curves, in contrast with the sharper indentations obtained with the classical binomial proportion approach.

Due to the diversity and complexity of the problems found in practice, there seems to be no confidence interval method that performs better than all the others in every situation. When sensitivity and specificity are not very close to 1, the Wilson method appears to be a good recommendation. However, in risk situations and near the boundary, strictly conservative methods, such as Anscombe and Cloper–Pearson, are better choices. This work showed, using the new expressions specifically tailored for conditional probabilities like sensitivity and specificity, as others before [4–8] have shown more generally, the performance problems of the Wald method, which may reach very low coverage probability, particularly for sensitivity (specificity) near 1.

Although some reported sample sizes (Tables 1–4) may seem excessive when compared to certain sample sizes in medical practice, works such as [12,17] also discuss sample sizes of the same magnitude. Moreover, large samples can actually be found in medical research [18,19] and this is likely to be the case in the context of the COVID 19 pandemic, given the wide scale application of serological and RT-PCR tests [20].

Some of the accuracy studies included in the systematic review addressing the diagnosis of hepatitis B [13] provide confidence intervals for the tests' sensitivity and specificity which are very wide and therefore useless. For example, a study reports a sensitivity of 0.60 (95% CI [0.15, 0.95]) and another a specificity of 1.00 (95% CI [0.03, 1.00]), corresponding to interval widths of 0.80 and 0.97, respectively. These are stark examples of how pointless confidence intervals can be from the estimation point of view, since they comprise almost the complete range of possible values.

Similarly, in the survey of published studies on depression [14], among 86 studies where 95% confidence intervals were provided or could be calculated, merely 8% had sensitivity intervals' widths not exceeding 0.10% and 62% had widths higher or equal to 0.21.

Determining in advance the sample size required to meet a predefined interval width, based on a sound procedure, is crucial to obtain informative confidence intervals.

Since, even in suspected clinical populations, the prevalence may be smaller than 0.5, sample sizes required for adequate estimation of sensitivities are higher than for specificities, in most of the cases. Choosing sample sizes based on sensitivity values, particularly for lower sensitivities, may lead to quite high sample sizes. Consequently, these sample sizes often guarantee adequate expected length

and coverage probability of the confidence interval also for the remaining performance measures simultaneously estimated. In essence, if confidence intervals for both sensitivity and specificity are desired, the maximum of the calculated sample sizes assures that both cases are attended.

In an experimental design, when the sample size to estimate the sensitivity or specificity is determined without taking into account that a conditional probability is involved, inadequate sizes may be obtained.

The proposed approach, however, inevitably requires a conjecture regarding the unknown prevalence in the target population, because the number of individuals with (without) the condition is a random variable depending on the prevalence. However, this is always the case when the individuals' condition is not known in advance.

The approximate method leads to sample sizes comparable to the optimal method. Given the availability of the R code (see Supplement 1), the new procedure is easily accessible, having the great advantage of leading to the optimal result based on a solid theoretical framework. When an optimal solution is easily available, there is no need for an approximation.

Supplementary Materials: The code is available online at http://www.mdpi.com/2227-7390/8/8/1258/s1, Supplement S1.

Author Contributions: M.R.O., A.S., and L.G. equally contributed for conceptualization; methodology; validation; formal analysis; investigation; writing—original draft preparation; writing—review and editing. M.R.O. and A.S. developed the software and the results' visualization. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded partially by Fundação para a Ciência e a Tecnologia (FCT), Portugal, under the projects UIDB/00006/2020 and UIDP/00006/2020 of CEAUL, UIDB/04621/2020 and UIDP/04621/2020 of CEMAT/IST-ID, and UID/04413/2020 of GHTM.

Acknowledgments: We are thankful for the fruitful suggestions and encouragement from Prof. Polychronis Kostoulas, Chair of CA18208 Action, *Novel tools for test evaluation and disease prevalence estimation*, from COST (European Cooperation in Science and Technology), funded by the Horizon Framework Programme of the European Union, and also from António Pacheco, from Instituto Superior Técnico, Universidade de Lisboa. We would also like to thank Ana Pires and Conceição Amado, from Instituto Superior Técnico, Universidade de Lisboa, for the R code used in the initial steps of this research.

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

Abbreviations

The following abbreviations are used in this manuscript:

CI	Confidence interval
CP	Coverage probability
EL	Expected length
Se	Sensitivity
Sp	Specificity

Appendix A

The notation and definitions presented here closely follow [8,11]. Suppose that a random sample of fixed size n is drawn from a large or infinite population and that we aim to find an interval estimate with a desired confidence level for an unknown proportion of interest in the population, p.

Let X ($0 \leq X \leq n$) be the random variable that counts the number of observations in the sample belonging to the category of interest and $[L(X); U(X)]$ represent the random interval for p. We aim to attach to this interval a nominal confidence level fixed in advance as $100 \times (1-\alpha)\%$. This means that the probability of $[L(X); U(X)]$ containing the unknown p should be $(1-\alpha)$. Under the previously described conditions, X has a $binomial(n, p)$ distribution. Since this is a discrete distribution, the probability of $[L(X); U(X)]$ containing the unknown p, designated coverage probability, may not achieve the target value of $(1-\alpha)$ for all possible values of the parameter.

As we mentioned above, the coverage probability of the random confidence interval $[L(X); U(X)]$ is the probability that the random interval contains the unknown parameter p, i.e,

$$\mathrm{CP}(n,p) = \sum_{j=0}^{n} \binom{n}{j} p^j (1-p)^{n-j} I_{[L(j);U(j)]}(p), \tag{A1}$$

where $I_{[a,b]}(x) = 1$ if $x \in [a,b]$ and $I_{[a,b]}(x) = 0$, otherwise.

The expected length of the random confidence interval $[L(X); U(X)]$ for p is given by:

$$\mathrm{EL}(n,p) = \sum_{j=0}^{n} \binom{n}{j} p^j (1-p)^{n-j} (U(j) - L(j)). \tag{A2}$$

When the coverage probability does not achieve the nominal confidence level $(1-\alpha)$, it should be at least close to $(1-\alpha)$. Between two methods with similar coverage probabilities, the choice of method may be based on additional criteria such as minimizing the expected length, among others. Coverage probability and expected length are the indicators of confidence interval performance covered in the present work. Discussion on additional criteria can be found in [21,22].

Appendix B

The expressions for the lower and upper bounds of the confidence interval methods for a binomial proportion are presented in Table A1 together with the R commands (from the code available in Supplement S1) needed to call the corresponding functions to obtain the confidence intervals. For more details about these methods see [8].

Table A1. Two-sided $100 \times (1-\alpha)\%$ confidence intervals for a binomial proportion, $[L(X), U(X)]$, where X is the number of successes, and R commands.

Method	`R command`	$[L(X), U(X)]$
Clopper–Pearson	`prop.ci`$(X, n, \alpha,$ `"ClopperP"`$)$	
$X = 0$		$\left[0, 1 - (\frac{\alpha}{2})^{\frac{1}{n}}\right]$
$0 < X < n$		$\left[Beta_{\frac{\alpha}{2}}(X, n-X+1), Beta_{1-\frac{\alpha}{2}}(X+1, n-X)\right]$
$X = n$		$\left[(\frac{\alpha}{2})^{\frac{1}{n}}, 1\right]$
Bayesian-U	`prop.ci`$(X, n, \alpha,$ `"BayesianU"`$)$	
$X = 0$		$\left[0, 1 - \alpha^{\frac{1}{n+1}}\right]$
$0 < X < n$		$\left[Beta_{\frac{\alpha}{2}}(X+1, n-X+1), Beta_{1-\frac{\alpha}{2}}(X+1, n-X+1)\right]$
$X = n$		$\left[\alpha^{\frac{1}{n+1}}, 1\right]$
Jeffreys	`prop.ci`$(X, n, \alpha,$ `"Jef"`$)$	
$X = 0$		$\left[0, 1 - (\frac{\alpha}{2})^{\frac{1}{n}}\right]$
$X = 1$		$\left[0, Beta_{1-\frac{\alpha}{2}}(2, n)\right]$
$1 < X < n-1$		$\left[Beta_{\frac{\alpha}{2}}(X+\frac{1}{2}, n-X+\frac{1}{2}), Beta_{1-\frac{\alpha}{2}}(X+\frac{1}{2}, n-X+\frac{1}{2})\right]$
$X = n-1$		$\left[Beta_{\frac{\alpha}{2}}(n, 2), 1\right]$
$X = n$		$\left[(\frac{\alpha}{2})^{\frac{1}{n}}, 1\right]$

Table A1. *Cont.*

Agresti–Coull	`prop.ci`$(X, n, \alpha,$ "`AgrestC`"$)$
	$\left[\max\left\{\frac{X+2}{n+4} - z_{1-\frac{\alpha}{2}}\sqrt{\frac{X+2}{(n+4)^2}\left(1-\frac{X+2}{n+4}\right)}; 0\right\}, \min\left\{\frac{X+2}{n+4} + z_{1-\frac{\alpha}{2}}\sqrt{\frac{X+2}{(n+4)^2}\left(1-\frac{X+2}{n+4}\right)}; 1\right\}\right]$
Wilson	`prop.ci`$(X, n, \alpha,$ "`Wils`"$)$
	$\left[\frac{2X+z^2_{1-\frac{\alpha}{2}} - z_{1-\frac{\alpha}{2}}\sqrt{z^2_{1-\frac{\alpha}{2}}+4X\left(1-\frac{X}{n}\right)}}{2\left(n+z^2_{1-\frac{\alpha}{2}}\right)}, \frac{2X+z^2_{1-\frac{\alpha}{2}} + z_{1-\frac{\alpha}{2}}\sqrt{z^2_{1-\frac{\alpha}{2}}+4X\left(1-\frac{X}{n}\right)}}{2\left(n+z^2_{1-\frac{\alpha}{2}}\right)}\right]$
$X = 0$	$\left[0, \sin^2\left(\min\left\{\arcsin\sqrt{\frac{\frac{3}{8}+\frac{1}{2}}{n+\frac{3}{4}}} + \frac{z_{1-\frac{\alpha}{2}}}{2\sqrt{n+\frac{1}{2}}}; \frac{\pi}{2}\right\}\right)\right]$
$0 < X < n$	$\left[\sin^2\left(\max\left\{\arcsin\sqrt{\frac{\frac{3}{8}+X-\frac{1}{2}}{n+\frac{3}{4}}} - \frac{z_{1-\frac{\alpha}{2}}}{2\sqrt{n+\frac{1}{2}}}; 0\right\}\right), \sin^2\left(\min\left\{\arcsin\sqrt{\frac{\frac{3}{8}+X+\frac{1}{2}}{n+\frac{3}{4}}} + \frac{z_{1-\frac{\alpha}{2}}}{2\sqrt{n+\frac{1}{2}}}; \frac{\pi}{2}\right\}\right)\right]$
$X = n$	$\left[\sin^2\left(\max\left\{\arcsin\sqrt{\frac{\frac{3}{8}+n-\frac{1}{2}}{n+\frac{3}{4}}} - \frac{z_{1-\frac{\alpha}{2}}}{2\sqrt{n+\frac{1}{2}}}; 0\right\}\right), 1\right]$
Wald	`prop.ci`$(X, n, \alpha,$ "`Wald`"$)$
	$\left[\max\left\{\frac{X}{n} - z_{1-\frac{\alpha}{2}}\sqrt{\frac{X}{n^2}\left(1-\frac{X}{n}\right)}; 0\right\}, \min\left\{\frac{X}{n} + z_{1-\frac{\alpha}{2}}\sqrt{\frac{X}{n^2}\left(1-\frac{X}{n}\right)}; 1\right\}\right]$

Note: z_γ and $Beta_\gamma(a, b)$ represent the $\gamma-$quantiles of the $N(0, 1)$ and the $Beta(a, b)$ distributions, respectively.

References

1. Altman, D.; Machin, D.; Bryant, T.; Gardner, M. *Statistics with Confidence. Confidence Intervals and Statistical Guidelines*; BMJ: London, UK, 2000.
2. Bossuyt, P.M.; Reitsma, J.B.; Bruns, D.; Gatsonis, C.A.; Glasziou, P.; Irwig, L.; Lijmer, J. G.; Moher, D.; Rennie, D.; de Vet, H. C.; et al. STARD 2015: An updated list of essential items for reporting diagnostic accuracy studies. *Clin. Chem.* **2015**, h5527. [CrossRef]
3. Korevaar, D.; Wang, J.; van Enst, W.A.; Leeflang, M.M.; Hooft, L.; Smidt, N.; Bossuyt, P.M.M. Reporting diagnostic accuracy studies: Some improvements after 10 years of STARD. *Radiology* **2015**, *274*, 781–789. [CrossRef] [PubMed]
4. Newcombe, R. Two-sided confidence intervals for the single proportion: Comparison of seven methods. *Stat. Med.* **1998**, *17*, 857–872. [CrossRef]
5. Agresti, A.; Coull, B. Approximate is better than "exact" for interval estimation of binomial proportions. *Am. Stat.* **1998**, *52*, 119–126.
6. Brown, L.; Cai, T.; DasGupta, A. Interval estimation for a binomial proportion. *Stat. Sci.* **2001**, *16*, 101–117.
7. Brown, L.; Cai, T.; Dasgupta, A. Confidence intervals for a Binomial proportion and asymptotic expansions. *Ann. Stat.* **2002**, *30*, 160–201.
8. Pires, A.; Amado, C. Interval estimators for a binomial proportion: Comparison of twenty methods. *REVSTAT J.* **2008**, *6*, 165–197.
9. Andrés, A.M.; Hernández, M.Á. Two-tailed asymptotic inferences for a proportion. *J. Appl. Stat.* **2014**, *41*, 1516–1529. [CrossRef]
10. Zelmer, D.A. Estimating prevalence: A confidence game. *J. Parasitol.* **2013**, *99*, 386–389. [CrossRef] [PubMed]
11. Gonçalves, L.; Oliveira, M.; Pascoal, C.; Pires, A. Sample size for estimating a binomial proportion: Comparison of different methods. *Appl. Stat.* **2012**, *39*, 2453–2473. [CrossRef]
12. 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. [CrossRef] [PubMed]

13. Amini, A.; Varsaneux, O.; Kelly, H.; Tang, W.; Chen, W.; Boeras, D.I.; Falconer, J.; Tucker, J.D.; Chou, R.; Ishizaki, A.; et al. Diagnostic accuracy of tests to detect hepatitis B surface antigen: A systematic review of the literature and meta-analysis. *BMC Infect. Dis.* **2017**, *17*, 698. [CrossRef] [PubMed]
14. Thombs, B.D.; Rice, D.B. Sample sizes and precision of estimates of sensitivity and specificity from primary studies on the diagnostic accuracy of depression screening tools: A survey of recently published studies. *Int. J. Methods Psychiat. Res.* **2016**, *25*, 145–152. [CrossRef] [PubMed]
15. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2018.
16. Sathish, N.; Vijayakumar, T.; Abraham, P.; Sridharan, G. Dengue fever: Its laboratory diagnosis, with special emphasis on IgM detection. *Dengue Bull.* **2003**, *27*, 116–125.
17. Dendukuri, N.; Rahme, E.; Bélisle, P.; Joseph, L. Bayesian sample size determination for prevalence and diagnostic test studies in the absence of a gold standard test. *Biometrics* **2004**, *60*, 388–397. [CrossRef] [PubMed]
18. Gonçalves, L.; Subtil, A.; Oliveira, M.; Rosário, V.; Lee, P.; Shaio, M.F. Bayesian latent class models in malaria diagnosis. *PLoS ONE* **2012**, e40633. [CrossRef] [PubMed]
19. Qiu, S.F.; Poon, W.Y.; Tang, M.L. Confidence intervals for proportion difference from two independent partially validated series. *Stat. Methods Med. Res.* **2016**, *25*, 2250–2273. [CrossRef] [PubMed]
20. Gudbjartsson, D.F.; Helgason, A.; Jonsson, H.; Magnusson, O.T.; Melsted, P.; Norddahl, G.L.; Saemundsdottir, J.; Sigurdsson, A.; Sulem, P.; Agustsdottir, A.B.; et al. Spread of SARS-CoV-2 in the Icelandic population. *N. Engl. J. Med.* **2020**. [CrossRef] [PubMed]
21. Vos, P.; Hudson, S. Evaluation criteria for discrete confidence intervals: Beyond coverage and length. *Am. Stat.* **2005**, *59*, 137–142. [CrossRef]
22. Newcombe, R. Measures of location for confidence intervals for proportions. *Commun. Stat.-Theor. Methods* **2011**, *40*, 1743–1767. [CrossRef]

© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).