2015

Estimating Negative Likelihood Ratio Confidence When Test Sensitivity is 100%: A Bootstrapping Approach

Keith A. Merill

Yuchiao Chang

Kim F. Wong

Ari B. Friedman
University of Pennsylvania

Recommended Citation

Merill, K. A., Chang, Y., Wong, K. F., & Friedman, A. B. (2015). Estimating Negative Likelihood Ratio Confidence When Test Sensitivity is 100%: A Bootstrapping Approach. *Statistical Methods in Medical Research,* 1-16. http://dx.doi.org/10.1177/0962280215592907

Estimating Negative Likelihood Ratio Confidence When Test Sensitivity is 100%: A Bootstrapping Approach

Abstract

Objectives: Assessing high-sensitivity tests for mortal illness is crucial in emergency and critical care medicine. Estimating the 95% confidence interval (CI) of the likelihood ratio (LR) can be challenging when sample sensitivity is 100%. We aimed to develop, compare, and automate a bootstrapping method to estimate the negative LR CI when sample sensitivity is 100%.

Methods: The lowest population sensitivity that is most likely to yield sample sensitivity 100% is located using the binomial distribution. Random binomial samples generated using this population sensitivity are then used in the LR bootstrap. A free R program, "bootLR," automates the process. Extensive simulations were performed to determine how often the LR bootstrap and comparator method 95% CIs cover the true population negative LR value. Finally, the 95% CI was compared for theoretical sample sizes and sensitivities approaching and including 100% using: (1) a technique of individual extremes, (2) SAS software based on the technique of Gart and Nam, (3) the Score CI (as implemented in the StatXact, SAS, and R PropCI package), and (4) the bootstrapping technique.

Results: The bootstrapping approach demonstrates appropriate coverage of the nominal 95% CI over a spectrum of populations and sample sizes. Considering a study of sample size 200 with 100 patients with disease, and specificity 60%, the lowest population sensitivity with median sample sensitivity 100% is 99.31%. When all 100 patients with disease test positive, the negative LR 95% CIs are: individual extremes technique (0,0.073), StatXact (0,0.064), SAS Score method (0,0.057), R PropCI (0,0.062), and bootstrap (0,0.048). Similar trends were observed for other sample sizes.

Conclusions: When study samples demonstrate 100% sensitivity, available methods may yield inappropriately wide negative LR CIs. An alternative bootstrapping approach and accompanying free open-source R package were developed to yield realistic estimates easily. This methodology and implementation are applicable to other binomial proportions with homogeneous responses.

Keywords

Sensitivity and specificity, confidence intervals, Monte Carlo method, data interpretation, statistical, bootstrapping, biostatistics

Estimating Negative Likelihood Ratio Confidence when Test Sensitivity is 100%: A Bootstrapping Approach

Journal:	*Statistical Methods in Medical Research*
Manuscript ID:	SMM-14-0124.R2
Manuscript Type:	Original Article
Keywords:	Sensitivity and Specificity, Confidence Intervals, Monte Carlo Method, Data Interpretation, Statistical, Bootstrapping, Biostatistics
Abstract:	Objectives: Assessing high sensitivity tests for mortal illness is crucial in emergency and critical care medicine. Estimating the 95% confidence interval (CI) of the likelihood ratio (LR) can be challenging when sample sensitivity is 100%. We aimed to develop, compare, and automate a bootstrapping method to estimate the negative LR CI when sample sensitivity is 100%. Methods: The lowest population sensitivity that is most likely to yield sample sensitivity 100% is located using the binomial distribution. Random binomial samples generated using this population sensitivity are then used in the LR bootstrap. A free R program, "bootLR," automates the process. Extensive simulations were performed to determine how often the LR bootstrap and comparator method 95% CI's cover the true population negative LR value. Finally, the 95% CI was compared for theoretical sample sizes and sensitivities approaching and including 100% using: (1) a technique of individual extremes, (2) SAS software based on the technique of Gart and Nam, (3) the Score CI (as implemented in the StatXact, SAS, and R PropCI package), and (4) the bootstrapping technique. Results: The bootstrapping approach demonstrates appropriate coverage of the nominal 95% CI over a spectrum of populations and sample sizes. Considering a study of sample size 200 with 100 patients with disease, and specificity 60%, the lowest population sensitivity with median sample sensitivity 100% is 99.31%. When all 100 patients with disease test positive, the negative LR 95% CI's are: individual extremes technique (0,0.073), StatXact (0,0.064), SAS Score method (0,0.057), R PropCI (0,0.062), and bootstrap (0,0.048). Similar trends were observed for other sample sizes. Conclusions: When study samples demonstrate 100% sensitivity, available methods may yield inappropriately wide negative LR CIs. An alternative bootstrapping approach and accompanying free open source R package

	were developed to yield realistic estimates easily. This methodology and implementation are applicable to other binomial proportions with homogeneous responses.

Estimating Negative Likelihood Ratio Confidence when Test Sensitivity is 100%: A Bootstrapping Approach

Authors:

Keith A. Marill, M.D.
University of Pittsburgh
Department of Emergency Medicine

Yuchiao Chang, Ph.D.
Harvard Medical School
Department of Medicine

Kim F. Wong, Ph.D.
University of Pittsburgh
Center for Simulation and Modeling
Department of Chemistry

Ari B. Friedman, B.S., M.A.
Fellow, Leonard Davis Institute of Health Economics
University of Pennsylvania

Author correspondence:
Keith A. Marill, M.D.
3600 Forbes Avenue
Suite 400A Iroquois Building
Pittsburgh, PA 15261
Phone: 617-312-0106
Fax: 412-647-6999
marillka@upmc.edu

No reprints are available.

Running title: Bootstrapping Neg LR Confidence when Test Sensitivity is 100%

Key words: Sensitivity and Specificity; Confidence Intervals; Monte Carlo Method; Data Interpretation, Statistical; Bootstrapping; Biostatistics

Presented at the Society For Academic Emergency Medicine Annual Meeting, May 15, 2013, Atlanta, Georgia.

The project described was supported by Award Number 5K12HL109068 (KAM) from the National Heart, Lung, and Blood Institute. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Heart, Lung, and Blood Institute or the National Institutes of Health.

Abstract:

Objectives: Assessing high sensitivity tests for mortal illness is crucial in emergency and critical care medicine. Estimating the 95% confidence interval (CI) of the likelihood ratio (LR) can be challenging when sample sensitivity is 100%. We aimed to develop, compare, and automate a bootstrapping method to estimate the negative LR CI when sample sensitivity is 100%.

Methods: The lowest population sensitivity that is most likely to yield sample sensitivity 100% is located using the binomial distribution. Random binomial samples generated using this population sensitivity are then used in the LR bootstrap. A free R program, "bootLR," automates the process. Extensive simulations were performed to determine how often the LR bootstrap and comparator method 95% CI's cover the true population negative LR value. Finally, the 95% CI was compared for theoretical sample sizes and sensitivities approaching and including 100% using: (1) a technique of individual extremes, (2) SAS software based on the technique of Gart and Nam, (3) the Score CI (as implemented in the StatXact, SAS, and R PropCI package), and (4) the bootstrapping technique.

Results: The bootstrapping approach demonstrates appropriate coverage of the nominal 95% CI over a spectrum of populations and sample sizes. Considering a study of sample size 200 with 100 patients with disease, and specificity 60%, the lowest population sensitivity with median sample sensitivity 100% is 99.31%. When all 100 patients with disease test positive, the negative LR 95% CI's are: individual extremes technique (0,0.073), StatXact (0,0.064), SAS Score method (0,0.057), R PropCI (0,0.062), and bootstrap (0,0.048). Similar trends were observed for other sample sizes.

Conclusions: When study samples demonstrate 100% sensitivity, available methods may yield inappropriately wide negative LR CIs. An alternative bootstrapping approach and accompanying

free open source R package were developed to yield realistic estimates easily. This methodology and implementation are applicable to other binomial proportions with homogeneous responses.

Introduction:

The primary mission of emergency and critical care medicine is to detect rapidly and manage appropriately life-threatening illness. Consequently, physicians seek highly sensitive diagnostic tests for serious illness. The emphasis on test sensitivity often comes at an acknowledged cost of relatively low test specificity.

Assessing highly sensitive diagnostic tests for life threatening illness is critical in emergency and critical care clinical research. Likelihood ratios (LR) based on sensitivity and specificity (Figure 1) are a popular metric used to assess diagnostic tests.[1-3] The negative likelihood ratio (Neg LR) is commonly used to assess the decrease in the odds of disease after a negative test result. When investigating highly sensitive diagnostic tests, emergency researchers may obtain perfect 100% sensitivity for disease in their study sample. An important example includes Perry, et al.'s observation that head computed tomography (CT) is 100% sensitive and specific for subarachnoid hemorrhage within 6 hours of headache onset, potentially obviating the need for lumbar puncture if head CT is negative.[4] Other critical and emergency care examples include: use of the serum d-dimer to diagnose acute aortic dissection, chest CT angiography for pulmonary embolism, C-reactive protein for bacterial infection, recognition of cardiac syncope in the setting of trauma, emergency imaging of ovarian torsion, and others.[5-11]

Sample sensitivity of 100%, as found in these examples, yields a sample estimate Neg LR of zero, which has limited practical utility. In reality, the absence of a false negative result in the study sample does not preclude the existence of a false negative in the population of interest.

This is true regardless of sample size determinations and the sample size studied. There is always some uncertainty, and the 95% confidence interval (CI) for this Neg LR will range from zero to some upper bound. The upper bound of the 95% CI for the Neg LR becomes the critical metric for evaluating a diagnostic test when the Neg LR sample estimate is zero.

An accurate upper bound of the 95% CI for the Neg LR is important for a highly sensitive test to determine the minimal amount by which a negative test result can be expected to lower the odds of disease. The Neg LR metric can be used to compare the benefit of one highly sensitive test to another, and to weigh its benefit versus its cost and risks. For example, if the upper bound estimate of the 95% CI is too high, it will incorrectly lower the perceived minimum benefit of the test.

Calculating the 95% CI for sensitivity or specificity, which are binomial proportions, is accurately accomplished using techniques based on the binomial distribution.[12-14] Calculating the 95% CI for the LR, a ratio of binomial proportions that incorporates both sensitivity and specificity, can be challenging. A number of modified Taylor series and other methods have been developed to estimate the result.[15-19]

We hypothesized that when study samples yield 100% sensitivity, techniques relying on modified Taylor series methods to estimate the population Neg LR CI may yield inappropriately wide intervals with high upper bound values. The primary objective of this study was to develop and compare an alternative bootstrapping method based more directly on the binomial distribution to estimate the negative LR CI particularly when study sample test sensitivity is 100%. The

secondary objective was to write and distribute a freely available automated software program to perform the calculations for researchers and clinicians.

Methods:

A bootstrapping technique was developed to estimate the upper bound of the Neg LR for samples with 100% sensitivity and compared to a conservative approach and to other conventional methods available from commercial and open source statistical packages.

Bootstrapping approach: problem

Bootstrapping is a method of estimation which repeatedly resamples with replacement and calculates the statistic of interest on each hypothetical sample.[20,21] Bootstrapping can be used "non-parametrically" by resampling the observed data without any distributional assumptions, or "parametrically," by sampling repeatedly from a distribution whose parameters (e.g. mean and standard deviation for the normal distribution) match those of the observed data. Traditionally, Neg LR CIs are calculated by non-parametric bootstrapping. This is the convention we use throughout for bootstrap resampling except for the case when sample sensitivity or specificity equal 100%. Repeated random samples are drawn from a theoretical population based on the observed data, and the sample elements withdrawn are replaced in the population each time. The distribution of the summary statistics of that collection of random samples withdrawn is the item of interest. For instance, characteristics such as the distribution of the means of the samples drawn can be used to estimate CIs. However, if all of the population elements are the same, such as when the sensitivity is 100% (all positive), then the collection of samples withdrawn are identical with all positive elements and no useful distribution can be discerned.

Bootstrapping approach: solution

When a diagnostic test is found to have *X* sensitivity in a given sample, then *X* also serves as the best estimate for the population sensitivity. This may also be true when the sample sensitivity is at the extreme of 100% or zero. However, *X* is not the only population sensitivity that could have yielded a sample sensitivity of *X* result. For example, both a population sensitivity of 70% and 72% are likely to have yielded a sample sensitivity of seven positive out of ten subjects (7/10 or 70%). Similarly, for any sample with 100% sensitivity, there are population sensitivities less than 100% that could have yielded the 100% sample sensitivity result.

Our approach to computing the Neg LR 95% CI for a study sample with 100% sensitivity starts with calculating the lowest population sensitivity that is likely to yield a sample sensitivity of 100% (Figure 2). Assuming the population sensitivity is *p*, the probability of observing a sample sensitivity of 100% is p^n for a sample size of *n*. We define that any population that is likely to yield a sample sensitivity of 100% has the probability of observing a sample sensitivity of 100% at least 0.5. Thus, the lowest population sensitivity that is most likely to yield sample sensitivity 100% is calculated as $p = e^{\frac{1}{n}(\log 0.5)}$ (Appendix, section 1). For a sample size of 20, the lowest population sensitivity *p* that satisfies our definition is 96.6%.

The next step is to draw 10,000 bootstrap samples each from those with and without the condition of interest. For sensitivity, bootstrapping is achieved by sampling repeatedly from a binomial distribution using parameters *n* and *p*, where *n* is the original sample size and *p* is the calculated lowest population sensitivity. For specificity, bootstrapping is achieved by resampling

the observed data. The two sets of bootstrap samples were then combined to calculate the Neg LR and its 95% CI. The CI of the bootstrap samples is determined using the bias corrected and accelerated (BCa) percentile method.[22] The BCa confidence interval method is used to correct for bias or skewness in the bootstrap distribution. Given the approximate 10% or less instability of the 95% CI results due to random variation particularly noted with low subject numbers, the entire procedure was repeated 50 times and the average value of the upper 95% CI was used.

bootLR

All computations were performed with the boot package in R, version 3.1.2. (Appendix, section 2). Given the multiple steps involved in determining the population sensitivity with median of samples equal to 100%, generating the bootstrap samples, and repeatedly performing the bootstrap procedure, a program, "bootLR," was written in R to automate the process. The user simply provides the number of true positives and total subjects with disease and true negatives and total subjects without disease in the study sample. The program performs the entire bootstrapping procedure and provides the positive and negative LR point estimates and 95% CI's (Figure 3). When sensitivity or specificity is 100%, bootLR generates samples from a parametric binomial (n, p) using the procedure described above. In the case that sensitivity or specificity is not 100%, bootLR resamples from the observed, non-parametric data distributions. A protocol repeating 10,000 samples 50 times was used because it provides low variance and high stability on repeat testing while using computing power generally available on a desktop PC. Using a variety of sample sizes, this protocol demonstrated better stability than protocols repeating larger numbers of bootstrap samples fewer times. The bootLR package also permits specifying a

different number of bootstrap samples from the default 10,000, or repetitions to average over from the default 50, if desired.

To evaluate the performance of the bootstrapping approach and bootLR, we compared results to a conservative approach, and methods available from commercial and open source statistical packages in two aspects: (1) the 95% CI coverage and (2) the upper bound of the 95% CI.

Simple conservative approach

A simple, conservative method for estimating the upper CI of the Neg LR is to use the lower extreme 95% CI for the individual sensitivity and specificity estimates.[16] This would maximize the numerator, *1 – Sensitivity*, and minimize the denominator, *Specificity*, of the Neg LR equation (Figure 1). The 95% CI for sensitivity and specificity can be determined individually from the binomial distribution in general, and can also be estimated for sample sensitivity 100%.[12-14] These computations were performed using the Clopper-Pearson exact CI with StatXact 10 (Cytel Software Corp., Cambridge, MA) and the downloadable PropCI package within the open source R software, version 3.1.2 (R Foundation for Statistical Computing, Vienna, Austria).[13] This approach is overly conservative because it takes the 95% CI extreme for two sampling distributions simultaneously, which is expected to be more extreme than the 95% CI for their ratio as a whole.[16] Any other method for calculating the 95% CI of the Neg LR that yields confidence intervals that are as wide or wider than this result may be considered incorrectly wide.

Commercial and Open Source Methods

Commercial software such as Stata does not formally provide a function for the CI of the ratio of binomials (personal communication with technical support, 10/13). In SAS (version 9.4), the CI for a ratio of binomial proportions was computed with two options. The default option computed the CI by inverting two separate one-sided exact tests,[23] and the "SCORE" option computes the Farrington-Manning standardized Score statistic.[24,25] StatXact 10 software offers three methods: (I) is comparable to the SAS "SCORE," method (II) uses a modification of the Score statistic by inverting a single two-sided test,[26] and (III) is a legacy technique with relatively wide CI's developed by Gart and Namm.[15,27] We use StatXact Method (II). The R PropCI version 0.2-5 software calculates the Wilson Score CI as programmed by Agresti, et al.[28,29]

Testing for appropriate 95% CI coverage

To determine whether each method provides appropriate coverage, we simulated samples from known theoretical populations. We explored the scenarios of population sensitivity ranging from 70% to 99.5%, population specificity 60%, with a total sample size of 40, 80, 200, 1000, and 2000 patients and half of the patients with the condition of interest. Five thousand random samples were drawn from each population. We calculated the NegLR and its 95% CI for each of the 5,000 simulated samples. We then evaluated the number of times the 95% CI covers the population Neg LR. This compute-intensive simulation was performed with bootLR on a high-performance computing cluster. We also determined how often sensitivity =100% when drawing 100,000 samples from each of the specified populations to inform when the binomial sampling technique is expected to significantly affect the bootstrap technique results. As a comparison, we also calculated the coverage for the 95% CI produced by the Conservative Method and the R PropCI package.

Comparing the upper bound of the 95%CI

We calculated the 95% CI from different scenarios of observed samples with specificity fixed at 60%, total sample size varying from 40 to 2000 patients, and half of the patients with the condition of interest. In addition to the case with 100% sensitivity (0 false negative), we allowed for a small number (1-4) of false negatives for each sample and compared the upper bound of the 95% CI of the Neg LR from each method. This was done to assess the transition from very high sensitivity to the specific case of 100% sensitivity using the conventional bootstrap and our binomial sampling technique as compared to other methods.

Results:

We compared coverage by assessing how often the 95% confidence interval included the population negative likelihood ratio (Table 1). bootLR method results are provided in Table 1a. Table 1b informs Table 1a by showing how often, on average, sample sensitivity=100% and the binomial sampling technique described above is used. The lowest coverage of 92-93% in Table 1a occurred when 1-2% of samples were expected to have sensitivity=100% in Table 1b. When more than 2% of samples were expected to have sensitivity=100%, coverage was always 95% or more. The Conservative method was overly conservative, as expected (Table 1c). The R Score method using R PropCI provided good coverage up until high sensitivity of 99% or more where it fell to 91% coverage.

The upper bounds of the Neg LR 95% CI were compared between methods for samples with different sample sizes, with specificity set at 60%, and 0 false negative (100% sensitivity), the

condition of primary interest in this study (Figure 4), and up to 4 false negatives (Figure 5), (Table 2). The 95% CI was also computed for a range of other specificities with a similar pattern of results (not shown). CI's were computed with the six different methods, including the binomial sampling bootstrap technique when there are zero false negatives.

Note in Table 1 and in the first column of Table 2 where there are zero false negatives (100% sensitivity), StatXact provides an estimate for the upper 95% CI that is higher than the simple conservative approach for larger samples. The R Score method yields results intermediate between the StatXact and bootstrapping method. As sample size increases (Figures 5c and 5d) the various Score method results cluster around the simple conservative results. The bootstrap results remain lower, even for samples with size 2,000 and 1 or more false negatives. These samples use only conventional bootstrapping with no binomial sampling and are expected to be most accurate as they approach a large sample asymptotic result.[30] The bootstrap curve has a mild upwards concave deflection moving from the zero to two false negative cases (Figure 5). This is as expected given the choice of the lowest consistent underlying population sensitivity estimate incorporated in the (1-sens) bootstrap for samples with zero false negatives.

Discussion:

The impetus for this work arose when the LR estimates under the condition of extreme sensitivity from commercial software were compared to those obtained using the simple conservative method.[31] It was disconcerting to find that previous versions of commercial software consistently yielded a wider 95% confidence interval than the conservative methodology based on binomial sampling. Available methods have improved, but we were still concerned the confidence intervals produced were overly wide. An alternative pragmatic method

was sought for the special case particularly relevant to emergency and critical care research when sample sensitivity is 100%. Bootstrapping is well suited for analyzing ratios of binomials, and it is used for this purpose in other areas of health care research such as cost effectiveness analysis.[32]

The present study involved developing a methodology for the case where one of the samples is homogeneous (sensitivity 100%), and then comparing it to other available methods across varying sample sizes. The methodology can be adapted for other situations where the sample sensitivity or specificity is homogeneous, and either the negative or positive LR CI is being estimated (Figure 1), or for other metrics with homogeneous results.

The bootstrap result appears to represent a realistic estimate of the metric sought with excellent nominal 95% CI coverage (Table 1). Testing through an array of population sensitivities, specificities, and sample sizes, the bootstrap method coverage is noted to decrease to 92.3% for sample size 200 and population sensitivity of 96%. Because of the marked decrease in bootstrap CI width as sensitivity approaches 100%, the CI of samples with sensitivity approaching and including 100% do not cover the 96% population sensitivity value. This seems primarily to be a result of the inherent challenges in estimating bootstrap confidence intervals rather than our proposed treatment of samples with 100% sensitivity. In fact, only 1.7% of samples with this size from this population would be expected to have sensitivity equal to 100% and invoke the binomial sampling procedure (Table 1b).

The bootstrap method developed generally yields lower values for the upper bound of the Neg LR 95% CI than the other available methods. As noted from the first column of Table 2 with sensitivity of 100%, the differences can be substantial including an increase of up to 24% for the 40 subject sample and 47% with 2000 subjects for the commercial as compared to bootstrapping method.

The SAS commercial package provides both a default and an optional Score method. The default method generally produces wide CIs with these skewed samples with one or more false negatives (Figure 5, Table 2). The Score option seems preferred in this case.

It may seem paradoxical that, compared to the PropCI Score method, the bootstrap technique 95% CI generally covers the true population NegLR more often for populations with high sensitivity (Table 1), but it demonstrates a lower upper 95% CI for samples with 100% sensitivity (Figure 4). We believe this is because the bootstrap technique generally has a smaller lower 95% CI limit for populations with high sensitivity (Fig 5, Table 2) which leads to higher proportions of coverage. In essence, if the large sample bootstrap curves in Figure 5d) are taken as the most realistic coverage, then this suggests an upward bias to both the upper and lower Score confidence interval limits. All of the techniques have a Neg LR lower 95% CI of zero and provide 100% coverage when the population sensitivity is 100%.

Consistent with other reports, the Score method seems to be the best of the previously-available comparators assessed.[19] It generally provides good coverage of the nominal 95% CI, but coverage dips to 91% for population sensitivities over 98% (Table 1). When evaluating

diagnostic tests with anticipated very high sensitivity such as high sensitivity troponins, for example, the bootstrap method may be particularly useful.[33]

The bootstrap methodology requires some sophistication with modeling and the use of R software. The basic code is provided in the appendix. In order to provide easy accessibility for general users, an R package called "bootLR" has been written to simplify and automate the process. The program provides both the positive and negative LR's, and it can be used for all sample sizes and sensitivities and specificities. This includes the homogeneous cases where all test results are positive or negative for patients with or without disease. For example, Perry, et al. found that noncontrast head CT had a sensitivity of 100% (121/121) and specificity of 100% (832/832) for acute subarachnoid hemorrhage when performed within 6 hours of symptom onset.[4] bootLR results for this study are provided in Figure 3. bootLR results for the other studies referenced in the Introduction are listed in Table 3. The NegLR was not reported in these other studies, perhaps in part due to uncertainty in calculating a 95% CI when sample sensitivity is 100%. We hope the bootLR package will ameliorate this.

The "bootLR" package is distributed as free, open-source software via the Comprehensive R Archive Network (CRAN). Users can download and install this package by typing, '*install.packages('bootLR')*' into their R console. Once installed, the "bootLR" package must be loaded at the beginning of each session of use with the dropdown menu for "Packages" in the toolbar or with the command *'library(bootLR)'*. bootLR automatically installs and loads the R "boot" package in order to use the bootstrap command. After loading "bootLR," the help page

can be accessed with the command *'?BayesianLR.test'*. A detailed technical description of the "bootLR" package is forthcoming.

Limitations:

Bootstrapping is inherently dependent on obtaining a representative sample of the intended population for which the diagnostic test will be applied and on the quality of the sample data. Furthermore, a Bayesian decision-making approach is also always limited by the uncertainty in the pretest probability of disease or "prior." Perhaps the greatest limitation of the methodology developed is that by using random sampling based on chance, some instability of results is introduced. This instability is magnified when small sample sizes with rare events are studied. A large number, 10,000 bootstrap samples, was used in the bootstrap procedure, and the procedures were repeated 50 times and averaged, to minimize instability in the results. These single study sample computations were performed within a few seconds with a Xeon microprocessor chip running Windows 7 (Microsoft Corp, Redmond, WA).

A second limitation is the bias found in the conventional bootstrap approach when applied to small and skewed samples. The BCa method compensates largely, but not completely, for this phenomenon.[34] The relative disadvantages of a Monte Carlo approach are balanced against the advantage of a relatively simple intuitive modeling technique that avoids empirical assumptions found in some analytic methods.

Conclusions:

When sample sensitivity is 100%, as is commonly the case in the study of emergency and critical care diagnostic tests, available software packages sometimes yield varying wide 95% CI upper bound results for the Neg LR. An alternative bootstrapping approach that relies on binomial sampling of the lowest population sensitivity likely to yield a sample sensitivity of 100% was developed. The bootstrapping approach yields appropriate results based on actual 95% CI coverage, but when sample sensitivity is 100%, the upper bound of the Neg LR 95% CI is generally lower than that obtained with other available methods.

Our technique and associated software enable calculation of 95% CI upper bound Neg LR limits, and the method can be used to compute CIs for other ratios of binomial proportions that include homogeneous sample results. Utilizing narrower confidence intervals with more appropriate coverage should ensure that studies of diagnostic tests and decision rules to rule out low probability events can be more confident in proclaiming that a negative test result means a patient is truly at low risk. These improved confidence interval characteristics could help reduce unnecessary diagnostic testing and associated costs and adverse effects in our patients.

Funding:

This project was supported by the National Heart, Lung, and Blood Institute of the National Institutes of Health [grant number 5K12HL109068]. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Heart, Lung, and Blood Institute or the National Institutes of Health. Computational resources were provided by the Center for Simulation and Modeling at the University of Pittsburgh.

Conflict of Interest Statement:

The Authors declare that there is no conflict of interest.

Acknowledgements:

The authors acknowledge and are grateful for the substantial beneficial comments and methodologic suggestions provided by the manuscript reviewers over the course of the editorial publication process.

References:

1) Gallagher EJ. Clinical Utility of Likelihood Ratios. *Ann Emerg Med* 1998; 31: 391-397.
2) Hayden S, Brown M. Likelihood ratio: a powerful tool for incorporating the results of a diagnostic test into clinical decisionmaking. *Ann Emerg Med* 1999; 33: 575-580.
3) Deeks JJ, Altman DG. Diagnostic tests 4: likelihood ratios. *BMJ* 2004; 329: 168-169.
4) Perry JJ, Stiell IG, Sivilotti ML, et al. Sensitivity of computed tomography performed within six hours of onset of headache for diagnosis of subarachnoid haemorrhage: prospective cohort study. *BMJ* 2011; 343: d4277.
5) Weber T, Hogler S, Auer J, et al. D-dimer in acute aortic dissection. *Chest* 2003; 123: 1375–1378.
6) Eggebrecht H, Naber CK, Bruch C, et al. Value of plasma fibrin D-dimers for detection of acute aortic dissection. *J Am Coll Cardiol* 2004; 44: 804–809.
7) Szucs-Farkas Z, Christe A, Megyeri B, Rohacek M, Vock P, Nagy EV, Heverhagen JT, Schindera ST. Diagnostic accuracy of computed tomography pulmonary angiography with reduced radiation and contrast material dose: a prospective randomized clinical trial. *Invest Radiol* 2014; 49: 201-208.
8) Bhat PK, Pantham G, Laskey S, Como JJ, Rosenbaum DS. Recognizing cardiac syncope in patients presenting to the emergency department with trauma. *J Emerg Med* 2014; 46: 1-8.
9) Haran JP, Beaudoin FL, Suner S, Lu S. C-reactive protein as predictor of bacterial infection among patients with an influenza-like illness. *Am J Emerg Med* 2013; 31: 137-144.
10) Swenson DW, Lourenco AP, Beaudoin FL, Grand DJ, Killelea AG, McGregor AJ. Ovarian torsion: Case-control study comparing the sensitivity and specificity of ultrasonography and

computed tomography for diagnosis in the emergency department. Eur J Radiol 2014;83:733-738.

11) Esmailian M, Khajouei AS, Eghtedari N, Azarian M, Vaseghi G. Utilization of coronary computed tomography angiography for rapid risk stratification in emergency chest pain units. J Res Med Sci 2014;19:134-138.

12) "Binomial proportion confidence interval," at http://en.wikipedia.org/wiki/Binomial_proportion_confidence_interval Accessed 5/11/14.

13) Scherer R. "Clopper-Pearson exact CI" in "Documentation for package 'PropCIs' version 0.2-4, 2013-08-04," pg. 9 http://cran.r-project.org/web/packages/PropCIs/PropCIs.pdf Accessed 10/15/13.

14) Hanley JA, Lippman-Hand A: If Nothing Goes Wring, Is Everything All Right? *JAMA* 1983; 249: 1743-1745.

15) Gart JJ, Nam J. Approximate interval estimation of the ratio of binomial parameters: a review and corrections for skewness. *Biometrics* 1988; 44: 323-338.

16) Simel DL, Samsa GP, Matchar DB. Likelihood ratios with confidence: sample size estimation for diagnostic test studies. *J Clin Epidemiol* 1991; 44: 763-770.

17) Dann RS, Koch GG. Review and evaluation of methods for computing confidence intervals for the ratio of two proportions and considerations for non-inferiority clinical trials. *J Biopharm Stat* 2005; 15: 85-107.

18) Price RM, Bonett DG. Confidence intervals for a ratio of two independent binomial proportions. *Stat Med* 2008; 27: 5497-5508.

19) Fagerland MW, Lydersen S, Laake P. Recommended confidence intervals for two independent binomial proportions. *Stat Methods Med Res* 2015; 24: 224-254.

20) Efron B, Tibshirani RJ. *An introduction to the bootstrap*. Boca Raton (Fla): Chapman & Hall/CRC; 1993.

21) Haukoos JS, Lewis RJ. Advanced statistics: bootstrapping confidence intervals for statistics with "difficult" distributions. *Acad Emerg Med* 2005; 12: 360-365.

22) DiCiccio T, Tibshirani. Bootstrap confidence intervals and bootstrap approximations. *J Am Stat Assoc* 1987; 82: 163-170.

23) Santner TJ, Snell MK. Small-Sample Confidence Intervals for p1-p2 and p1/p2 in Contingency Tables. *J Am Stat Assoc* 1980; 75, 386–394.

24) Farrington CP, Manning G. Test Statistics and Sample Size Formulae for Comparative Binomial Trials with Null Hypothesis of Non-zero Risk Difference or Non-unity Relative Risk. *Stat Med* 1990; 9: 1447–1454.

25) Chan ISF, Zhang Z. Test-Based Exact Confidence Intervals for the Difference of Two Binomial Proportions. *Biometrics* 1999; 55: 1202–1209.

26) Agresti A, Min Y. On Small-Sample Confidence Intervals for Parameters in Discrete Distributions. *Biometrics* 2001; 57: 963-971.

27) Statxact 10 User Manual. Cambridge (Ma): Cytel Software Corp; 2012; pp 489-90;527-33.

28) Scherer R. "Score confidence interval for the relative risk in a 2x2 table" in "Documentation for package 'PropCIs' version 0.2-4, 2013-08-04," pg. 12 http://cran.r-project.org/web/packages/PropCIs/PropCIs.pdf Accessed 10/15/13.

29) Agresti A. *Categorical Data Analysis*, Hoboken (New Jersey): John Wiley & Sons; 2nd edition, 2002: 73-78.

30) Efron B, Tibshirani RJ "Chapter 12: "Confidence intervals based on bootstrap 'tables'" in *An Introduction to the bootstrap*. Boca Raton (Fla): Chapman & Hall/CRC; 1993

31) Czuczman AD, Thomas LE, Boulanger AB, Peak DA, Senecal EL, Brown DF, Marill KA. Interpreting red blood cells in lumbar puncture: distinguishing true subarachnoid hemorrhage from traumatic tap. *Acad Emerg Med* 2013; 20: 247-256.

32) Briggs AH, Wonderling DE, Mooney CZ. Pulling cost-effectiveness analysis up by its bootstraps: a non-parametric approach to confidence interval estimation. *Health Econ* 1997;6: 327-340.

33) Keller T, Zeller T, Ojeda F, et al. Serial changes in highly sensitive troponin I assay and early diagnosis of myocardial infarction. JAMA 2011;306:2684-2693.

34) Efron B, Tibshirani RJ "Chapter 14: Better bootstrap confidence intervals" in *An Introduction to the bootstrap*. Boca Raton (Fla): Chapman & Hall/CRC; 1993

Appendix:

1) **Computing the population probability where a sample of size 100 has a 50% chance of being entirely positive (100% sensitive)**

```
prob<- exp(log(0.5)/100)
prob
[1] 0.9930925
```

2) **Bootstrap Neg LR (1-Sens/Spec) where sens=100/100 and spec=60/100 (single iteration)**

```
# Lines beginning with # are treated as explanatory
# comments in R, not code

sens<-rbinom(10000, size=100, prob=.9931)/100
spec<-rep(1:0,c(60,40))
specf<- function(spec,i) {return (1/mean(spec[i]))}
specb<- boot(spec, specf, R=10000)
lr<-((1-sens)*specb$t)

# Analyze LR bootstrap finding median, and standard and
# BCa percentile 95% CIs.
# To obtain bca CI on a non-boot result, use a dummy boot
# and replace t and t0 with the results of interest.
dummy<-rep(1:0,c(6,4))
dummyf<- function(dummy,i) {return (mean(dummy[i]))}
dummyb<- boot(dummy, dummyf, R=10000)
b$t<-matrix(lr, nrow=10000, byrow=T)
#(1-.9931)/.6 = .0115
b$t0<-.0115
boot.ci(dummyb, t0=b$t0, t=b$t, conf=.95, type=c("bca"))
```

Figure 1: LR^+, LR^-, and relation to pre- and post-test odds

Figure 2: Schematic flow diagram of the bootstrap procedure for computing the 95% CI of the Negative LR when sensitivity is 100%.

Figure 3: User code for installing, loading, and running the new R package "bootLR," and the pos and neg LR with 95% CI results obtained for the utility of noncontrast head CT to identify subarachnoid hemorrhage within 6 hours of symptom onset (Ref. 4).

Figure 4: Negative LR 95% CI upper bound when sensitivity =100%, specificity=60%, for subject sample sizes of (a) 40, (b) 200, (c) 1,000, and (d) 2,000.

Figure 5: Negative LR and the upper bound for associated 95% CIs as a function of false negatives for subject sample sizes of (a) 40, (b) 200, (c) 1,000, and (d) 2,000.

Table 1: Coverage of the population Neg LR by the 95% CI with varying population sensitivity, fixed specificity = 60%, and varying sample size produced by the (a) bootLR, (b) percentage of samples expected to have sensitivity=100% to inform bootLR result, (c) Conservative method, and (d) R PropCI package.

Table 2: Negative LR and associated 95% CIs as a function of false negatives for 40, 200, 1,000, and 2,000 subjects data.

Table 3: Test characteristics for referenced clinical studies

Figure 1:

$$LR^{+} = \frac{\text{Sensitivity}}{1-\text{Specificity}} \qquad LR^{-} = \frac{1-\text{Sensitivity}}{\text{Specificity}}$$

Post-test odds = LR * Pre-test odds

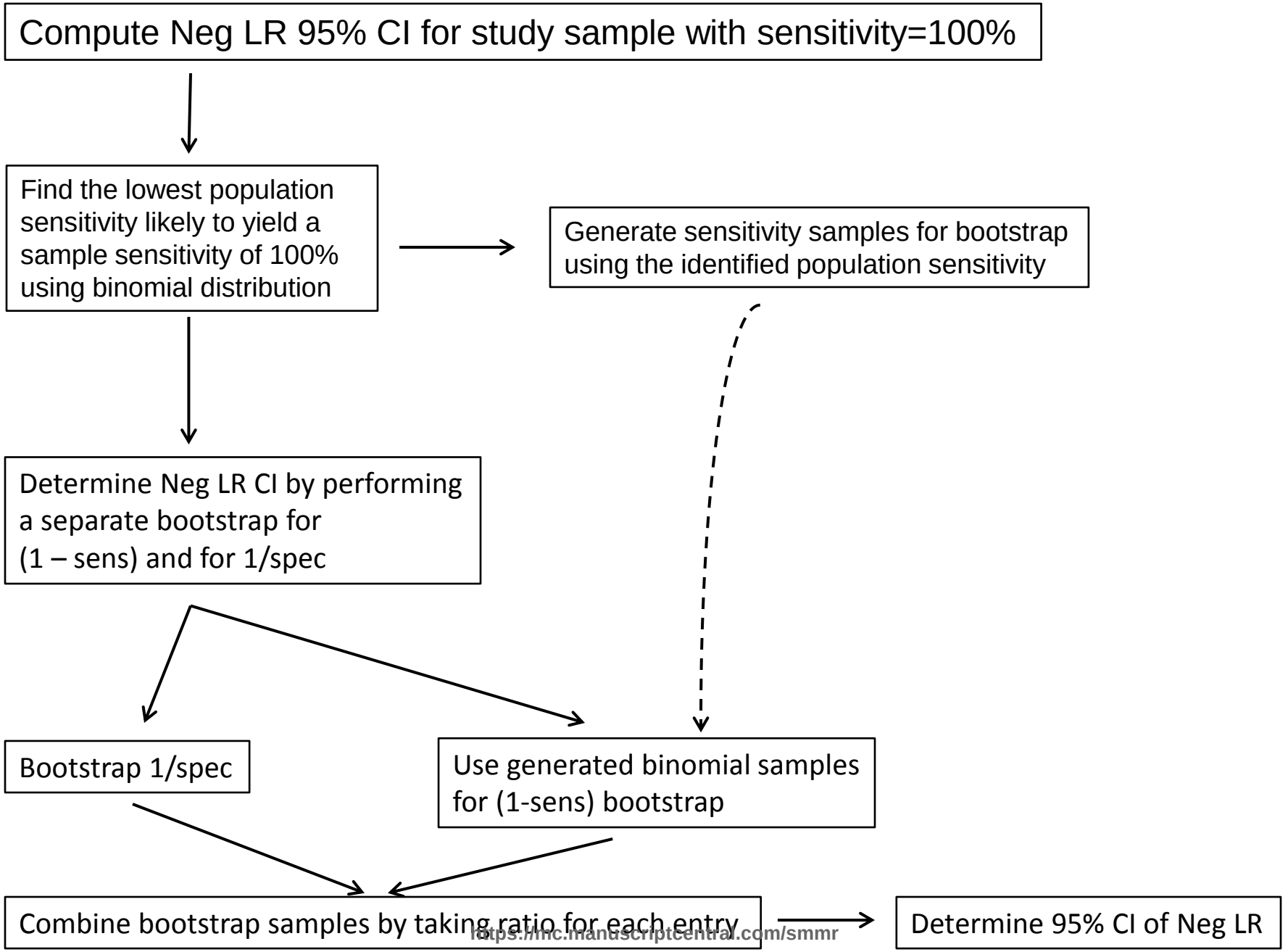
Compute Neg LR 95% CI for study sample with sensitivity=100%
Find the lowest population sensitivity likely to yield a sample sensitivity of 100% using binomial distribution
Generate sensitivity samples for bootstrap using the identified population sensitivity
Determine Neg LR CI by performing a separate bootstrap for (1 – sens) and for 1/spec
Bootstrap 1/spec
Use generated binomial samples for (1-sens) bootstrap
Combine bootstrap samples by taking ratio for each entry
Determine 95% CI of Neg LR

Figure 3:

USER INPUT:

```
> install.packages("bootLR")
> library(bootLR)

> BayesianLR.test(121,121,832,832)
```

RESULT:

```
Likelihood ratio test of a 2x2 table

data:
  truePos  totalDzPos    trueNeg   totalDzNeg
      121         121         832          832
Positive LR: Inf (287.258 - Inf)
Negative LR: 0 (0 - 0.024)
95% confidence intervals computed via BCa bootstrapping.
```

Figure 4: Negative Likelihood Ratio Upper 95% CI When Sample Sensitivity Equals 100%, Specificity Equals 60%

Figure 4a
40 Subjects

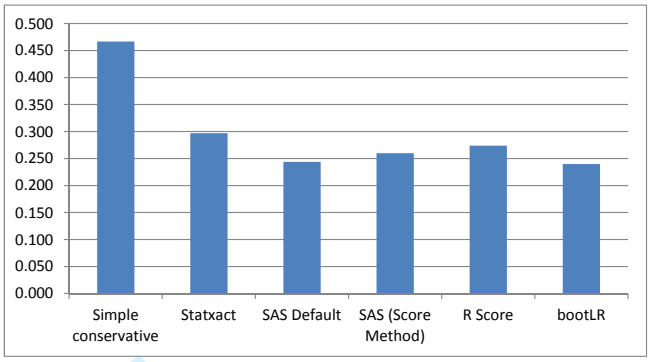

Figure 4b
80 Subjects

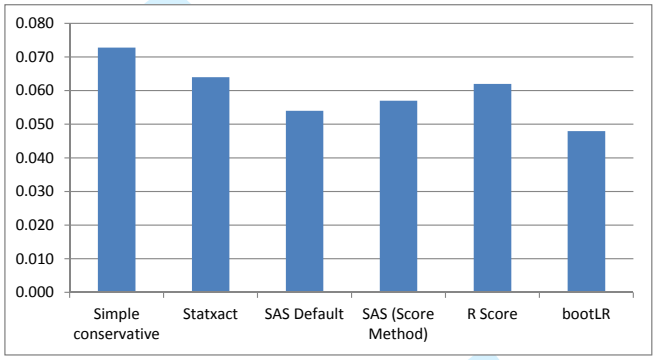

Figure 4c
1000 Subjects

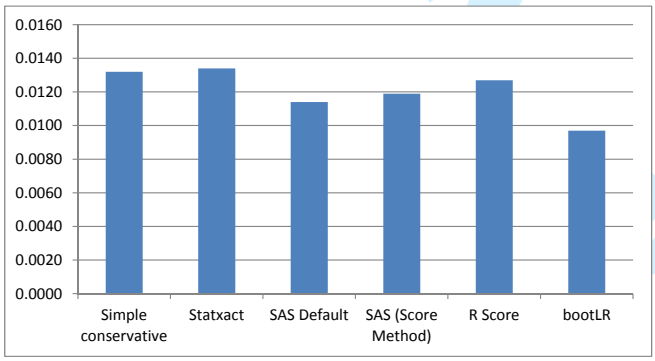

Figure 4d
2000 Subjects

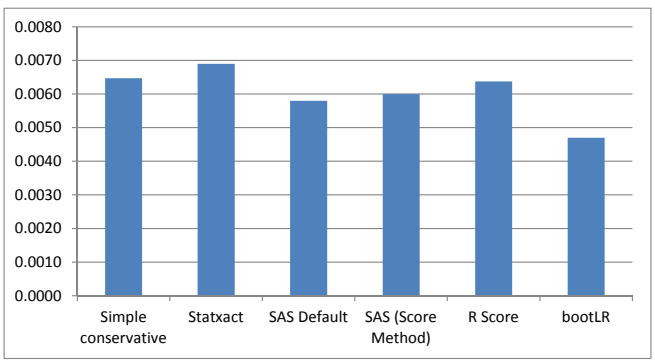

Figure 5: Negative LR and associated 95% CIs as a function of false negatives for subject sample sizes of (a) 40, (b) 200, (c) 1,000, and (d) 2,000

Figure 5a)

	Dz+	Dz-
Test+	20-FN	8
Test-	FN	12

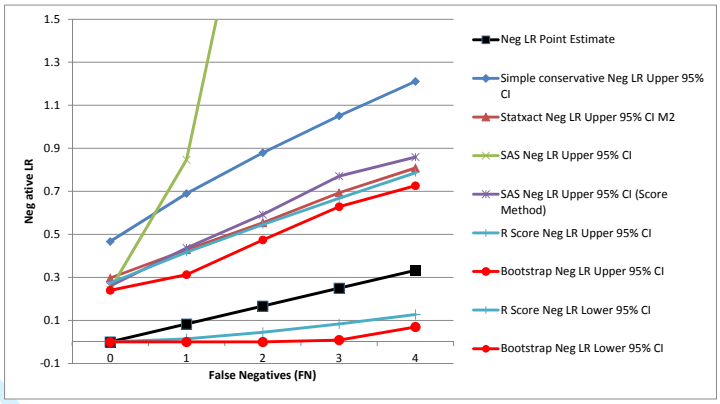

Figure 5b)

	Dz+	Dz-
Test+	100-FN	40
Test-	FN	60

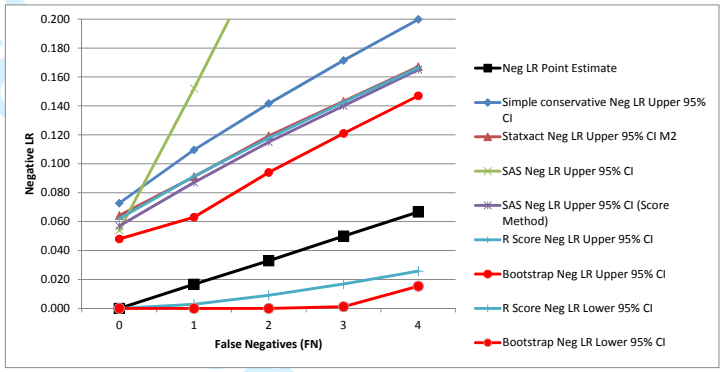

Figure 5c)

	Dz+	Dz-
Test+	500-FN	200
Test-	FN	300

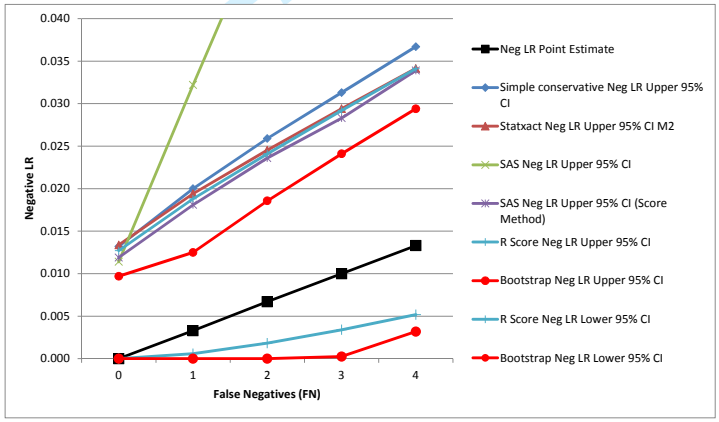

Figure 5d)

	Dz+	Dz-
Test+	1000-FN	400
Test-	FN	600

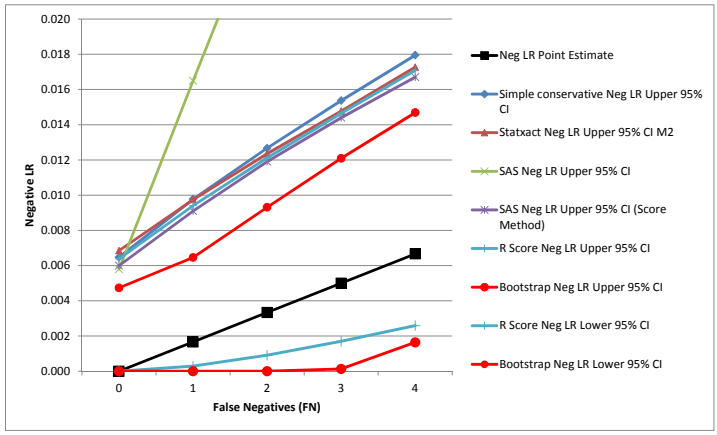

Table 1: How often does the 95% CI include the population Neg LR?

1a: bootLR Method

	Population Sensitivity								
Total Sample Size	70%	80%	90%	95%	96%	97%	98%	99%	99.50%
40	94.90	94.16	98.56	99.26	99.60	99.70	99.88	100.00	100.00
80	94.48	94.56	93.00	99.30	99.38	99.56	99.46	100.00	99.98
200	95.64	95.44	94.04	94.50	92.26	95.30	99.14	99.62	99.86
1000	94.92	94.90	95.46	95.02	94.42	94.58	94.64	94.58	98.60
2000	94.64	94.64	95.48	94.80	95.50	94.36	94.76	94.86	94.58

1b: Percent of random samples from population that have sensitivity=100%

	Population Sensitivity								
Total Sample Size	70%	80%	90%	95%	96%	97%	98%	99%	99.50%
40	0.07%	1.13%	12.11%	35.84%	44.04%	54.36%	66.74%	81.83%	90.56%
80	0.00%	0.01%	1.42%	12.88%	19.67%	29.50%	44.41%	66.85%	81.83%
200	0.00%	0.00%	0.01%	0.57%	1.66%	4.80%	13.25%	36.59%	60.89%
1000	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.66%	8.19%
2000	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.01%	0.67%

Shaded area includes categories where >1% of samples have sensitivity=100%

1c: Conservative Method

	Population Sensitivity								
Total Sample Size	70%	80%	90%	95%	96%	97%	98%	99%	99.50%
40	99.66	99.32	99.56	99.70	99.52	99.78	99.34	99.60	99.66
80	99.72	99.34	99.32	99.50	99.34	99.36	99.32	99.00	99.06
200	99.40	99.36	98.94	98.88	99.20	99.32	99.06	99.18	98.60
1000	99.42	99.06	98.52	98.54	98.00	98.32	98.10	98.12	98.38
2000	99.22	99.20	98.36	97.98	98.14	97.64	97.66	97.36	97.96

Conservative method using the simultaneous upper or lower bounds for the sensitivity and specificity 95% CI computed using the Clopper-Pearson exact CI with PropCI

1d: R Score Method

	Population Sensitivity								
Total Sample Size	70%	80%	90%	95%	96%	97%	98%	99%	99.50%
40	94.80	95.38	97.24	95.34	95.40	95.46	93.94	93.94	91.20
80	95.02	95.26	95.68	96.38	95.66	96.02	95.18	94.10	96.12
200	95.10	94.66	94.66	95.86	95.22	96.32	95.74	95.24	91.60
1000	94.80	95.12	95.00	95.04	95.40	95.34	95.02	95.36	95.72
2000	94.40	95.40	95.56	94.78	95.12	94.20	95.30	95.36	96.34

R Score method using PropCI

Population specificity=60%, number of samples from population=5,000 for all simulations in Tables 1a,c,d

Table 2: Upper and Lower 95% CI as a function of Test Sensitivity

	Number of False Negatives				
40 Subjects	**0**	**1**	**2**	**3**	**4**
Test Sensitivity	20/20 (100%)	19/20 (95%)	18/20 (90%)	17/20 (85%)	16/20 (80%)
Neg LR Point Estimate	0.000	0.083	0.167	0.250	0.333
Simple conservative Neg LR Upper 95% CI	0.467	0.690	0.879	1.051	1.211
Statxact Neg LR Upper 95% CI	0.297	0.429	0.554	0.694	0.809
SAS Neg LR Upper 95% CI	0.244	0.847	2.513	2.518	13.907
SAS Neg LR Upper 95% CI (Score Method)	0.260	0.437	0.592	0.771	0.859
R Score Neg LR Upper 95% CI	0.274	0.417	0.545	0.668	0.787
Lowest population sensitivity with 50% of samples with sensitivity 100%	96.59%				
Bootstrap Neg LR Upper 95% CI	0.240	0.313	0.474	0.629	0.726
R Score Neg LR Lower 95% CI	0.0000	0.0145	0.0448	0.0835	0.1276
Bootstrap Neg LR Lower 95% CI	0.0000	0.0000	0.0000	0.0086	0.0696
200 Subjects					
Test Sensitivity	100/100 (100%)	99/100 (99%)	98/100 (98%)	97/100 (97%)	96/100 (96%)
Neg LR Point Estimate	0.000	0.017	0.033	0.050	0.067
Simple conservative Neg LR Upper 95% CI	0.073	0.110	0.142	0.171	0.200
Statxact Neg LR Upper 95% CI	0.064	0.091	0.119	0.143	0.167
SAS Neg LR Upper 95% CI	0.054	0.152	0.254	0.344	0.454
SAS Neg LR Upper 95% CI (Score Method)	0.057	0.087	0.115	0.140	0.165
R Score Neg LR Upper 95% CI	0.062	0.091	0.118	0.143	0.166
Lowest population sensitivity with 50% of samples with sensitivity 100%	99.31%				
Bootstrap Neg LR Upper 95% CI	0.048	0.063	0.094	0.121	0.147
R Score Neg LR Lower 95% CI	0.0000	0.0029	0.0091	0.0169	0.0258
Bootstrap Neg LR Lower 95% CI	0.0000	0.0000	0.0000	0.0012	0.0154
1000 Subjects					
Test Sensitivity	500/500 (100%)	499/500 (99.8%)	498/500 (99.6%)	497/500 (99.4%)	496/500 (99.2%)
Neg LR Point Estimate	0.0000	0.0033	0.0067	0.0100	0.0133
Simple conservative Neg LR Upper 95% CI	0.0132	0.0200	0.0259	0.0313	0.0367
Statxact Neg LR Upper 95% CI	0.0134	0.0194	0.0245	0.0294	0.0341
SAS Neg LR Upper 95% CI	0.0114	0.0322	0.0525	0.0720	0.0933
SAS Neg LR Upper 95% CI (Score Method)	0.0119	0.0181	0.0236	0.0283	0.0339
R Score Neg LR Upper 95% CI	0.0127	0.0188	0.0241	0.0292	0.0341
Lowest population sensitivity with 50% of samples with sensitivity 100%	99.86%				
Bootstrap Neg LR Upper 95% CI	0.0097	0.0125	0.0186	0.0241	0.0294
R Score Neg LR Lower 95% CI	0.00000	0.00059	0.00183	0.00340	0.00518
Bootstrap Neg LR Lower 95% CI	0.00000	0.00000	0.00000	0.00025	0.00318
2000 Subjects					
Test Sensitivity	1000/1000 (100%)	999/1000 (99.9%)	998/1000 (99.8%)	997/1000 (99.7%)	996/1000 (99.6%)
Neg LR Point Estimate	0.0000	0.0017	0.0033	0.0050	0.0067
Simple conservative Neg LR Upper 95% CI	0.0065	0.0098	0.0127	0.0154	0.0179
Statxact Neg LR Upper 95% CI	0.0069	0.0097	0.0124	0.0148	0.0173
SAS Neg LR Upper 95% CI	0.0058	0.0165	0.0270	0.0370	0.0467
SAS Neg LR Upper 95% CI (Score Method)	0.0060	0.0091	0.0119	0.0144	0.0167
R Score Neg LR Upper 95% CI	0.0064	0.0094	0.0121	0.0147	0.0171
Lowest population sensitivity with 50% of samples with sensitivity 100%	99.93%				
Bootstrap Neg LR Upper 95% CI	0.0047	0.0065	0.0093	0.0121	0.0147
R Score Neg LR Lower 95% CI	0.00000	0.00029	0.00091	0.00170	0.00259
Bootstrap Neg LR Lower 95% CI	0.00000	0.00000	0.00000	0.00013	0.00163

Half of patients do not have the condition, and specificity is a constant 60% for all samples.

Table 3: Test characteristics for referenced clinical studies

Reference #	Diagnostic Test and Disease	Sensitivity	Specificity	Sample NLR	Reported NLR 95% CI	Bootstrap NLR 95% CI
4	Head CT for subarachnoid hemorrhage	100%, 121/121	100%, 832/832	0	0 - .02	0 - 0.024
5	Serum D-Dimer for aortic dissection	100%, 24/24	69%, 24/35	0	Not reported	0 - 0.16
6	Serum D-Dimer for aortic dissection	100%, 16/16	67%, 32/48	0	Not reported	0 - 0.25
7	Low contrast chest CTA for pulmonary embolism	100%, 40/40	97.1%, 200/206	0	Not reported	0 - 0.071
8	Recognition of cardiac syncope in trauma	100%, 24/24	43%, 25/58	0	Not reported	0 - 0.28
9	C-reactive protein for bacterial infection	100%, 41/41	Not reported	0	Not reported	-
10	CT for ovarian torsion, (reader 1)	100%, 20/20	85%, 17/20	0	Not reported	0 - 0.16