Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

HHS Public Access

Author manuscript

*Cancer Epidemiol*. Author manuscript; available in PMC 2019 October 01.

Published in final edited form as:

*Cancer Epidemiol*. 2018 October ; 56: 83–89. doi:10.1016/j.canep.2018.07.014.

# Recommendation to Use Exact P-values in Biomarker Discovery Research in Place of Approximate P-values

**Matthew F. Buas**[1,2], **Christopher I. Li**[3], **Garnet L. Anderson**[4], and **Margaret S. Pepe**[5]

[1]Department of Cancer Prevention and Control, Roswell Park Comprehensive Cancer Center, Buffalo, New York;

[2]Epidemiology Program, Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, Washington;

[3]Translational Research Program, Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, Washington;

[4]Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, Washington;

[5]Biostatiistics and Biomathematics Program, Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, Washington

## Abstract

**Background:** Biomarker candidates are often ranked using *P*-values. Standard *P*-value calculations use normal or logit-normal approximations, which may not be correct for small *P*-values and small sample sizes common in discovery research.

**Methods:** We compared exact *P*-values, correct by definition, with logit-normal approximations in a simulated study of 40 cases and 160 controls. The key measure of biomarker performance was sensitivity at 90% specificity. Data for 3000 uninformative false markers and 30 informative true markers were generated randomly. We also analyzed real data for 2371 plasma protein markers measured in 121 breast cancer cases and 121 controls.

**Results:** In our simulation, using the same discovery criterion, exact *P*-values led to discovery of 24 true and 82 false biomarkers, while logit-normal approximate *P*-values yielded 20 true and 106 false biomarkers. The estimated true discovery rate was substantially off for approximate *P*-values: logit-normal estimated 42 but found 20. The exact method estimated 22, very close to 24, which was the actual number of true discoveries. Although these results are based on one specific

**Corresponding author**: Matthew F. Buas, Department of Cancer Prevention and Control, Roswell Park Comprehensive Cancer Center, Elm & Carlton Streets, Buffalo, NY 14263. Phone: 716-845-4754., matthew.buas@roswellpark.org.

Authorship contribution

Study conception and design: M.S.P. Data analysis: M.S.P. Interpretation: M.S.P., M.F.B. Data acquisition: M.F.B, C.I.L., G.L.A. All authors reviewed the manuscript for intellectual content.

**Publisher's Disclaimer:** This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Conflict of Interest: None

Conflict of Interest: The authors declare no conflicts of interest.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

simulation, qualitatively similar results were obtained from 10 random repetitions. With real data, ranking candidate biomarkers by exact *P*-values, versus approximate P-values, resulted in a very different ordering of these markers.

**Conclusions:** Exact *P*-values, which correspond to permutation tests with non-parametric rank statistics such as empirical ROC statistics, are preferred over approximate P-values. Approximate *P*-values can lead to inappropriate biomarker selection rules and incorrect conclusions.

**Impact:** Exact *P*-values in place of approximate *P*-values in discovery research may improve the yield of biomarkers that validate clinically.

### Keywords

hypothesis tests; ROC curve; biomarkers; arrays; statistical analysis

## Introduction

Biomarker discovery research has yielded few clinically useful biomarkers. Poor methodologies in the statistical design of studies and in the evaluation of studies may be contributing factors (1). With regard to design of discovery studies, guidelines have recently been discussed, including sources and numbers of biological samples for adequate power (2). In this article we address a common and underappreciated issue in the evaluation of biomarker discovery studies.

The classic discovery study entails measuring many biomarkers, perhaps using array-based or other such high-throughput technology, on a set of biological samples from cases and controls. For each biomarker, one calculates a statistic and its *P*-value using the case and control data pertaining to that biomarker. The biomarkers are then ranked according to one or more criteria, such as *P*-value, (average) fold change between cases and controls, sensitivity at a given specificity, area under the curve, biological relevance to the target disease, availability of antibodies for assay development, potential difficulties with targeted assays, and differential expression in publicly available databases. *P*-values are a commonly-used criterion used for ranking biomarker candidates and determining the top set of markers considered for further development and validation. Thus statistical *P*-values can play a fundamental role in the evaluation of biomarker discovery studies.

As an example, consider the “Colocare” study to discover and validate markers to predict colon cancer recurrence in patients diagnosed with stage 1 colon cancer (3). Tissue and blood samples taken at diagnosis from 40 cases with colon cancer recurrence and 160 controls without recurrence will be tested with approximately 3000 autoantibodies. As described in (2), the data analytic plan is to calculate the sensitivity corresponding to 90% specificity for each biomarker and to generate a corresponding standard *P*-value for no association between biomarker and case-control status. We simulated data for 3000 useless biomarkers not associated with case-control status and found that 69 (2.3%) had approximate *P*-values less than 0.01 (see third row of Table 1 in (2)). Since one would expect that approximately 30 markers (1% of markers) would attain *P*-values less than 0.01 if all 3000 biomarkers were useless, i.e. the estimated number of ‘false discoveries’ is 30

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

(=0.01 × 3000), the data analysis suggests that 69−30 = 39 true biomarkers have been discovered. However this conclusion is incorrect since we generated the data in such a way that none of the 3000 markers are predictive of case-control status. The issue here is that standard *P*-value calculations that rely on asymptotic statistical theory are problematic and lead to an erroneous conclusion in this example.

In this paper we demonstrate this phenomenon in more detail and propose an alternative method for calculating *P*-values that is generally correct and robust to the vagaries of biomarker discovery data. This exact *P*-value approach is applicable regardless of the statistic used to rank biomarkers and it is computationally reasonable with modern computing capacities. Most importantly, we show in simulations studies that use of exact *P*-values leads to more reliable conclusions from biomarker discovery data than does use of approximate *P*-values.

## Materials and Methods

In case control studies, the *P*-value associated with a statistic is defined as *P-value = Probability(statistic >= observed data statistic | cases same as controls).* Standard P-value calculations often employ approximations based on an asymptotic normal distribution for a Z-score standardized version of the statistic. Our study was designed to investigate if such standard *P*-value calculations, as commonly performed in case-control studies, are potentially incorrect in practice and if incorrect *P*-value calculations can substantially affect the soundness of conclusions drawn from biomarker discovery studies. To address these questions we simulated biomarker discovery data where the capacities of biomarkers to predict outcome were specified, allowing us to compare conclusions based on data analysis with the specified truth.

Our proposal is to calculate *P*-values exactly without approximation, using this simulated data. This is in fact an old concept for rank statistics such as the Wilcoxon rank sum statistic where published tables have long been available for use with data from studies involving very small sample sizes (7). Modern computing power now makes the approach feasible for studies with larger sample sizes and for any statistic. The idea is to enumerate all the possible values of the statistic for the setting where cases have biomarker values with the same distribution as controls and to evaluate how extreme the observed biomarker data statistic is to calculate its exact P-value.

To demonstrate that the method used to calculate *P*-values in real data analysis can have a substantial effect on conclusions drawn, we also reanalyzed data from an ER/PR positive breast cancer biomarker discovery study reported in (8). A detailed description of our simulation studies, analytic approach, and the ER/PR positive breast cancer discovery study is included in the Methods section of the Supplementary Data file.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

## Results

### Reference Distribution for Calculating Exact P-values

Table 1 shows the reference distribution for estimated sensitivity corresponding to 90% specificity, also known as the empirical estimate of ROC(0.1) ($ROC_{emp}$), based on 40 cases and 160 controls when a biomarker is not informative about case-control status (a false biomarker). This table will be used to calculate exact *P*-values when biomarker data are available from the simulated Colocare study, in which the biomarker positivity threshold is set to the $90^{th}$ percentile of control values so as to guarantee the marker has 90% specificity (Supplementary Data). Possible values for the $ROC_{emp}$ are 0/40, 1/40, 2/40, 3/40, etc. because there are 40 cases and the estimated ROC is the fraction of those 40 cases whose biomarker values exceed the $90^{th}$ percentile of control values (i.e. exceed the $16^{th}$ largest control value). We see that among the 40,000 simulated studies of uninformative markers, in only 1 study did the estimated ROC reach a value of 0.40. Therefore the exact *P*-value corresponding to an $ROC_{emp}$ of 0.40 is 1/40,000 = 0.000025. Correspondingly, in 5 simulations the estimated ROC reached a value of 0.375 or more, so the *P*-value corresponding to 0.375 is 5/40,000 = 0.000125.

### Approximate P-values based on normal distribution with logit transformation can be incorrect

Table 2 demonstrates that *P*-values calculated with the logit-normal approximation method described in Supplemental Methods can be substantially different from the correct exact *P*-values. The data were simulated for a single biomarker discovery study that included 30 true biomarkers and 3000 false biomarkers with all 3030 biomarkers evaluated on 40 case and 160 control samples. Table 2 shows P-values only for the 30 true markers in the simulation study. Although *P*-values calculated with the different methods are often of similar magnitudes that would lead to the same decisions about efforts to validate or not, there are multiple instances where the differences could lead to different decisions with use of logit-normal versus exact *P*-values (see highlighted biomarkers 3, 14 and 27).

### Impact of Approximate P-Values in the Simulated Colocare Study

Differences in *P*-value calculations had a substantial effect on the numbers of biomarkers discovered in the simulated study. The top panel of Table 3 shows the numbers of biomarkers that passed the discovery criterion: *P*-value <=0.0277. We chose this odd threshold since among the finite set of attainable *P*-values that are possible (Table 1) it is closest to 0.02, which was the preferred threshold in the Colocare study, to reduce the number of anticipated false discoveries. If we had chosen say the threshold 0.02, the actual threshold for the exact *P*-value would have been 0.0121 and the comparison between *P*-value methods would have been flawed. Observe from Table 3 that use of different *P*-value algorithms leads to substantially different numbers of markers discovered: 106 for exact, and 126 for logit-normal. Since we simulated the data we know the true and false biomarkers. Use of the exact *P*-value led to discovery of 24 true biomarkers and 82 false biomarkers, while use of logit-normal *P*-values led to 20 true biomarker discoveries and 106 false discoveries.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

We next examined if the incorrect logit-normal *P*-value calculations impacted the conclusions drawn from the discovery study. Discovery data analyses typically report estimates of the true and false discovery rates, that is, the proportion of discovered biomarkers that are likely to be true biomarkers and the proportion that are likely to be false. We illustrate a simple way to estimate these numbers with reference to the logit-normal P-value column in the top panel of Table 3 : By definition of P-value, we expect that 2.77% of false biomarkers will meet the discovery criterion (P-value<0.0277). Therefore, assuming that the vast majority of the 3030 biomarkers evaluated are in fact uninformative, we expect that 0.0277×3030=84 false biomarkers are discovered in this study. That implies that of the 126 biomarkers discovered using the logit-normal P-value criterion, we estimate that 84 are false and therefore the remaining 42 are likely to be true biomarkers.

Comparing the estimated number of false discoveries with the actual numbers of false discoveries, numbers we know because we designed the simulation, in Table 3 we see that the estimate is very close for *P*-exact (84 versus 82), but a substantial under estimate (84 versus 106) for the logit-normal *P*-value. With logit-normal we estimate that 126−84 = 42 (33%) of the 126 discoveries are true discoveries. However, only 20 (16%) of the discoveries are in fact true biomarkers. Therefore, with logit-normal *P*-values we believe we are doing much better than we actually are. In contrast, the exact *P*-value method discovers 24 true biomarkers (23% of total discoveries) and this is in line with the estimated number of true discoveries, namely 106−84 = 22 (21%). In summary, estimates of numbers of true and false discoveries made are much closer to the actual numbers of true and false discoveries when the exact *P*-values are used. In this sense, conclusions drawn from the study are more sound for the exact *P*-value method than for the logit-normal approximation *P*-values.

Results with use of a more stringent *P*-value threshold criterion, namely 0.0121, shown in the lower panel of Table 3 are similar. We repeated the simulation study 10 times to determine if the observations made from the single study shown in Table 3 were found in general. We see from Tables S.1 and S.2 in Supplementary Data that there is a consistent tendency for the logit-normal approximation *P*-values to provide poor estimates of the numbers of true and false discoveries made and that the exact *P*-value method leads to more reliable conclusions. In nine of the 11 total simulation studies conducted, use of exact *P*-values (versus logit-normal *P*-values) led to a smaller number of ‘misclassified’ discoveries. Across all 11 simulations, the number of such misclassifications when using exact *P*-values (n=69) was three-fold lower than when using logit-normal *P*-values (n=217).

When the numbers of cases and controls available are small, investigators often resort to use of more global measures of biomarker performance such as AUC, although pitfalls of using such clinically irrelevant measures are well documented (9, 10). Table 4 investigates performance of exact and logit-normal *P*-values for the AUC statistic when 20 cases and 20 controls are included in the discovery study. Here, logit-normal *P*-values tend to be too large. Interestingly, this is opposite to results for $ROC_{emp}$. The direction in which approximate *P*-values depart from exact *P*-values is likely to depend on the shape of the ROC curve and on the specific performance measure underlying the test. Most importantly, however, we see again that *estimates* of true and false discoveries are much closer to the *actual* numbers of true and false discoveries when using exact rather than logit-normal approximation *P*-values.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

### Application to Real Data for Receptor Positive Breast Cancer Biomarker Candidates

ROC statistics and *P*-values were calculated with antibody array data from the ER-PR positive Women's Health Initiative breast cancer study (8). The top panel of Table 5 shows the top 40 candidates ranked according to exact *P*-value and the bottom panel shows the top 40 candidates ranked according to logit-normal *P*-value. One gets very different impressions of the results depending on which *P*-value method is used. For exact *P*-value, the number of true biomarkers in the top 40 is estimated to be 15.7 and the estimated false discovery rate is 63%. In contrast, the logit-normal *P*-value method estimates 33.8 true markers in the top 40 and a false discovery rate of only 15%. Given the previous simulation results, we believe that the estimates based on the exact *P*-values are more reliable. Note that we estimated the false discovery rate here using the Benjamini-Hochberg method (13) implemented with the qqvalue command and Simes option in the Stata software package (14, 15). The simple intuitive calculation method noted in previous tables gave very similar results (data not shown) but that method does not restrict the FDR to increase with increasing *P*-value as does the Benjamini-Hochberg method.

Another interesting observation in Table 5 is that the ROC values (shown as "Estimated Sensitivity") align pretty well with the *P*-values when using the exact method, i.e. the highest ROC estimates are at the top of the list corresponding to the smallest *P*-values (see also Figure S.1 in Supplementary Data). In contrast, the logit-normal *P*-value method does not align ROC estimates with *P*-values very well. For example, the highest ROC estimate, 0.339, is way down the list at rank 38 according to the logit-normal *P*-value, just above a biomarker with estimated ROC = 0.198.

Considering the biomarker selection criterion '$p<0.05$' for which $0.05 \times 2371 = 118.5$ false biomarkers are expected to be identified, the estimated numbers of true biomarkers selected is 11.5 based on exact *P*-values (130 markers selected in total) and 75.5 with logit-normal *P*-values (194 markers selected in total). These are very different estimates. The biomarker selection criterion $p<0.02$, for which 47.4 false biomarkers are expected to be identified, yields estimated numbers of true biomarkers selected of 11.6 with exact *P*-values (59 markers selected in total) and 73.6 with logit-normal *P*-values (121 in total). Again, given the simulation results above, we have more trust in the estimated numbers of true and false biomarkers based on exact *P*-values than in those based on logit-normal *P*-values.

## Discussion

Exact *P*-values, calculated according to the definition of *P*-value, provide the true probability of observing a statistic as extreme as that observed in the study when case biomarker values are derived from the same distribution as controls. For convenience, approximation *P*-values are typically used in practice. Our results using one classic simulation scenario show that approximations can be substantially off, leading to less reliable conclusions. Additional simulations (Table S.3) show qualitatively similar conclusions when true biomarkers were more diverse than in our classic simulation scenario.

Exact *P*-values can be calculated for any two sample test statistic. Our analyses used nonparametric rank statistics, in particular the empirical sensitivity at fixed 90% specificity

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

and the area under the ROC curve. The issues concerning exact versus approximate *P*-values also apply to parametric non-rank based statistics, including the t-test for example. However, the null hypothesis reference distribution that is needed to calculate exact *P*-values requires estimating the control biomarker distribution and repeatedly generating random case and control simulated study data from it. This can be a complicated exercise for parametric non-rank based statistics particularly in the context of discovery research where automated procedures are needed to deal with diverse data on large numbers of biomarkers. We cannot assess parametric assumptions for each biomarker. Moreover, outliers and non-standard distributions are common in discovery research. Therefore, we prefer rank based non-parametric statistics, and those were the focus of our study. We note that for rank-based nonparametric statistics, p-values from permutation tests are the same as exact *P*-values. Therefore another interpretation of our results is that when one is using rank-based statistics, permutation test *P*-values are preferred over normal approximations. We note, however, that permutation test *P*-values do not correspond with exact *P*-values for parametric non-rank based statistics.

We found two additional advantages from use of exact *P*-values. First, biomarker performance measures align well with exact *P*-values, in that markers with the best estimated performances have the smallest *P*-values. This inverse relationship holds by definition when the same numbers of case and control data points are available for each biomarker. In addition we found the inverse relationship was mostly true in the analysis of the breast cancer data set where data were sporadically missing. However, the analysis that used approximate logit-normal *P*-values led to some major inconsistencies between estimated performance and *P*-value. A second and unexpected advantage to use of exact *P*-values concerns computational effort. When there is no missing biomarker data, the number of case and control data points is the same for each biomarker, and only one reference distribution must be calculated for the entire analysis. For example, exact *P*-values were calculated for each of the 3030 biomarkers in our simulated study using only Table 1. The computation involved to calculate exact *P*-values is therefore very fast once the reference table is created. In contrast, the logit-normal approximation *P*-values required calculation of standard errors for each biomarker separately, a process that was time consuming with use of bootstrap resampling.

We considered two different types of statistical criteria for selecting biomarkers. One approach demonstrated in Tables 3 and 4 was of the form "*P*-value < threshold". Another approach demonstrated in Table 5 was to select the "top K markers" where K was set to 40 in Table 5, similar to the number of biomarkers selected in previous analyses of the same data (8). Yet another criterion is to select markers for which the false discovery rate among markers ranked at or above is below a specified threshold (13). One can see from the breast cancer results in Table 5 that the different *P*-value calculations would lead to very different biomarker selections based on a false discovery rate criterion. For example "false discovery rate<10%" would lead to two markers selected with exact *P*-values but 30 markers selected with the logit-normal *P*-values. Since false discovery rates are functions of *P*-values, it is important to use correct *P*-value calculations when calculating false discovery rates.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

We showed that exact *P*-values can provide more reliable conclusions than standard approximate *P*-values. Specifically, they provide better estimates of true and false discovery rates, key parameters reported in discovery research. Calculation of exact *P*-values instead of approximate *P*-values allows for better conclusions in discovery research where interest is often focused on small *P*-values, sample sizes are often small, and the numbers of biomarkers tested may preclude evaluating data for distributional assumptions. Given that exact *P*-values are correct regardless of sample sizes and biomarker distributions, and can be obtained through minimally burdensome computation, we recommend use of exact *P*-value calculations instead of approximate *P*-value calculations in the analysis of biomarker discovery data. We caution against the assumption that approximate *P*-values are acceptable once a minimum sample size is reached, as results depend on the ROC shape, which will vary across applications. While the focus of this analysis was biomarker discovery, and we did not examine whether exact P values are preferable in all possible applications or scenarios, further studies are warranted to investigate whether it may also be prudent to use exact rather than approximate calculations in biomarker validation research.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Acknowledgements

Data analyzed in this paper was provided by the Women's Health Initiative. A partial list of WHI investigators follows: Program Office: (National Heart, Lung, and Blood Institute, Bethesda, Maryland) Jacques Rossouw, Shari Ludlam, Dale Burwen, Joan McGowan, Leslie Ford, and Nancy Geller; Clinical Coordinating Center: (Fred Hutchinson Cancer Research Center, Seattle, WA) Garnet Anderson, Ross Prentice, Andrea LaCroix, and Charles Kooperberg; Investigators and Academic Centers: (Brigham and Women's Hospital, Harvard Medical School, Boston, MA) JoAnn E. Manson; (MedStar Health Research Institute/Howard University, Washington, DC) Barbara V. Howard; (Stanford Prevention Research Center, Stanford, CA) Marcia L. Stefanick; (The Ohio State University, Columbus, OH) Rebecca Jackson; (University of Arizona, Tucson/Phoenix, AZ) Cynthia A. Thomson; (University at Buffalo, Buffalo, NY) Jean Wactawski-Wende; (University of Florida, Gainesville/Jacksonville, FL) Marian Limacher; (University of Iowa, Iowa City/Davenport, IA) Robert Wallace; (University of Pittsburgh, Pittsburgh, PA) Lewis Kuller; (Wake Forest University School of Medicine, Winston-Salem, NC) Sally Shumaker; Women's Health Initiative Memory Study: (Wake Forest University School of Medicine, Winston-Salem, NC) Sally Shumaker For a list of all the investigators who have contributed to WHI science, please visit: https://www.whi.org/researchers/Documents%20%20Write%20a%20Paper/WHI%20Investigator%20Long%20List.pdf

Financial support: Authors were supported by grants from the National Cancer Institute (NCI) at the NIH (R01 CA152089 to M.S. Pepe; U01 CA152637 to C.I. Li; P30 CA016056 to the Roswell Park Comprehensive Cancer Center). Support for GL Anderson and for data analyzed in this paper was provided by the WHI program that is funded by the National Heart, Lung, and Blood Institute, National Institutes of Health, U.S. Department of Health and Human Services through contracts HHSN268201100046C, HHSN268201100001C, HHSN268201100002C, HHSN268201100003C, HHSN268201100004C, and HHSN271201100004C.

## References

1. Feng Z, Prentice R, Srivastava S. Research issues and strategies for genomic and proteomic biomarker discovery and validation. Pharmacogneomics 2004;5:709–719.
2. Pepe MS, Li CI, Feng Z. Improving the quality of biomarker discovery research: the right samples and enough of them. Cancer Epidemiol Biomarkers Prev 2015;24:944–50. [PubMed: 25837819]
3. ClincalTrials.gov [internet] ColoCare Study — Colorectal Cancer Cohort Clinical Trial NCT02328677 Study Record Detail [verified January 2016]. Available from https://clinicaltrials.gov/ct2/show/NCT02328677

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

4. Pepe M, Longton G, Janes H. Estimation and comparison of receiver operating characteristic curves. Stata J 2009 9(1):1–16. [PubMed: 20161343]
5. Hsieh F, Turnbull BW. Nonparametric and semiparametric estimation of the receiver operating characteristic curve. Ann Stat 1996;24:25–40.
6. Pepe MS. The Statistical Evaluation of Medical Tests for Classification and Prediction. New York: Oxford University Press; 2003.
7. Hollander M and Wolfe DA. Nonparametric Statistical Methods. Wiley, 1999.
8. Buas MF, Rho JH, Chai X, Zhang Y, Lampe PD, Li CI. Candidate early detection protein biomarkers for ER+/PR+ invasive ductal breast carcinoma identified using pre-clinical plasma from the WHI observational study. Br Cancer Res Treat 2015;153:445–54.
9. Pepe MS, Janes HE. Gauging the performance of SNPs, biomarkers and clinical factors for predicting risk of breast cancer JNCI 2008;100(14):978–9. [PubMed: 18612128]
10. Cook NR. Use and Misuse of the Receiver Operating Characteristic Curve in Risk Prediction. Circulation 2007;115:928–935. [PubMed: 17309939]
11. Hanley JA and McNeil BJ. The meaning and use of the area under the ROC curve. Radiology 1982;143:29–36. [PubMed: 7063747]
12. Delong ER, Delong DM and Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988;44:837–45. [PubMed: 3203132]
13. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Royal Stat Soc B 1995;57:289–300.
14. StataCorp. Stata Statistical Software: Release 14. College Station, TX: StataCorp LP,2015.
15. Newson R ALSPAC Study Team. Multiple-test procedures and smile plots. Stata J 2013;3:109–132.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

## Highlights

- Normal and logit-normal approximate *P*-values are frequently used in biomarker studies.
- Approximate *P*-values can lead to inappropriate biomarker selection rules.
- Use of exact *P*-values may improve the yield of biomarkers that validate clinically.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

**Table 1:**

Reference distribution for the sensitivity corresponding to 90% specificity estimated with the empirical ROC when calculated with data for 40 cases and 160 controls[a]. The reference distribution is used to determine exact *P*-values and was generated by 40,000 randomly chosen enumerations[b] of ranks for 200 subjects with the first 40 labelled as cases.

| r | Probability that the estimated sensitivity ≥ r |
|---|---|
| 0.000 | 1.000000 |
| 0.025 | 0.976575 |
| 0.050 | 0.892075 |
| 0.075 | 0.739125 |
| 0.100 | 0.549825 |
| 0.125 | 0.367025 |
| 0.150 | 0.218200 |
| 0.175 | 0.119050 |
| 0.200 | 0.059175 |
| 0.225 | 0.027675 |
| 0.250 | 0.012025 |
| 0.275 | 0.004850 |
| 0.300 | 0.001750 |
| 0.325 | 0.000550 |
| 0.350 | 0.000225 |
| 0.375 | 0.000125 |
| 0.400 | 0.000025 |

[a]Smallest increment for realized values of the estimated sensitivity is 0.025 = 1/40 where 40 is the number of cases.

[b]Smallest increment for probability is 0.000025 = 1/40000 where 40000 is the number of random rank enumerations.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

**Table 2:**

P-values calculated for the 30 true biomarkers in the simulated biomarker study of 40 cases and 160 controls. Calculations use standard logit-normal approximations to the distributions of the estimated sensitivity at 90% specificity or use exact methods. For at least 3 markers, *P*-values are substantially different.

| Biomarker | Exact *P*-value | Logit-Normal *P*-value |
|---|---|---|
| 1 | 0.001750 | 0.0002190 |
| 2 | 0.012025 | 0.0016413 |
| 3 | 0.059175 | 0.0986310 |
| 4 | 0.004850 | 0.0017292 |
| 5 | 0.000550 | 0.0000414 |
| 6 | 0.000225 | 0.0001591 |
| 7 | 0.027675 | 0.0190249 |
| 8 | 0.000550 | 0.0028095 |
| 9 | 0.027675 | 0.0235189 |
| 10 | 0.059175 | 0.0520333 |
| 11 | 0.000025 | 0.0000594 |
| 12 | 0.012025 | 0.0073081 |
| 13 | 0.027675 | 0.0373784 |
| 14 | 0.012025 | 0.0563044 |
| 15 | 0.027675 | 0.0331860 |
| 16 | 0.218200 | 0.1790015 |
| 17 | 0.000125 | 0.0005066 |
| 18 | 0.000550 | 0.0001466 |
| 19 | 0.119050 | 0.1098244 |
| 20 | 0.004850 | 0.0094405 |
| 21 | 0.218200 | 0.1999186 |
| 22 | 0.001750 | 0.0006482 |
| 23 | 0.000550 | 0.0000243 |
| 24 | 0.004850 | 0.0011685 |
| 25 | 0.367025 | 0.3618602 |
| 26 | 0.027675 | 0.0151371 |
| 27 | 0.027675 | 0.0689888 |
| 28 | 0.004850 | 0.0005487 |
| 29 | 0.004850 | 0.0014270 |
| 30 | 0.001750 | 0.0011660 |

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

**Table 3:**

Markers discovered in the simulated study by the selection criterion: biomarker *P*-value < threshold from a dataset with 3,000 uninformative (false) biomarkers and 30 true biomarkers when *P*-values are based on exact calculation or on logit-normal approximation. The test statistic is the sensitivity at 90% specificity estimated with the empirical ROC[a]. Number of study subjects: 40 cases and 160 controls.

| | | Number of Markers | |
|---|---|---|---|
| **Threshold for Sensitivity *P*-value** | | **Exact *P*-value** | **Logit-Normal *P*-value** |
| 0.0277 | **Total Discoveries** | 106 | 126 |
| | **False Discoveries** | | |
| | estimated[b] | 84 | 84 |
| | actual | 82 | 106 |
| | **True Discoveries** | | |
| | estimated[d] (tdr[c]) | 22 (21%) | 42 (33%) |
| | actual(tdr) | 24 (23%) | 20(16%) |
| | | | |
| 0.0121 | **Total Discoveries** | 47 | 75 |
| | **False Discoveries** | | |
| | estimated | 37 | 37 |
| | actual | 29 | 58 |
| | **True Discoveries** | | |
| | estimated(tdr) | 10 (21%) | 38 (51%) |
| | actual(tdr) | 18 (38%) | 17 (23%) |

[a] equivalent to hypothesis testing with the positive predictive value

[b] estimated false discoveries = threshold-p × number of biomarkers

[c] tdr: True discovery rate = number of true discoveries/ number of discoveries

[d] estimated true discoveries = total discoveries – estimated false discoveries

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

**Table 4:**

Markers discovered by the selection criterion: biomarker *P*-value < threshold from a dataset with 3,000 uninformative (false) biomarkers and 30 true biomarkers when *P*-values are based on exact calculation or on logit-normal approximation. The test statistic is the empirical area under the ROC curve (AUC). True biomarkers have AUC = 0.758 (PPV = 0.30) while false (uninformative) biomarkers have AUC=0.50 (PPV = 0.10). Number of study subjects: 20 cases and 20 controls

| | | **Number of Markers** | | |
|---|---|---|---|---|
| **Threshold for AUC *P*-value** | | **Exact *P*-value** | | **Logit-Normal *P*-value**[1,2] |
| 0.0216 | **Total Discoveries** | 94 | | 82 |
| | **False Discoveries** | | | |
| | estimated[3] | 65 | | 65 |
| | actual | 68 | | 56 |
| | **True Discoveries** | | | |
| | estimated | 29 | | 17 |
| | actual | 26 | | 26 |
| | | | | |
| 0.01016 | **Total Discoveries** | 59 | | 47 |
| | **False Discoveries** | | | |
| | estimated[4] | 31 | | 31 |
| | actual | 37 | | 26 |
| | **True Discoveries** | | | |
| | estimated | 28 | | 16 |
| | actual | 22 | | 21 |

[1] standard error calculated using 500 bootstrapped samples of the data

[2] similar results found with standard errors calculated using a large sample theory expression (11, 12)

[3] estimated false discoveries = 3030 × 0.0216 = 65.45

[4] estimated false discoveries = 3030 × 0.01016 = 30.78

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

**Table 5.**

Candidate biomarkers for ER positive PR positive ductal breast cancer measured on preclinical plasma samples from 121 cases and 121 controls in the WHI observation study. The top 40 biomarkers ranked according to *P*-values for sensitivity at 90% specificity are shown. Rankings are with respect to exact *P*-values (top panel) and logit-normal *P*-values (bottom panel).

| Rank by exact *P*-value | Marker | Estimated Sensitivity | *P*-value | False Leads Expected | Estimated True Markers | Estimated[d] False Discovery Rate |
|---|---|---|---|---|---|---|
| 1 | v2621 | 0.305 | 0.000050 | 0.12 | 0.88 | 8.9% |
| 2 | v689 | 0.339 | 0.000075 | 0.18 | 1.82 | 8.9% |
| 3 | v1830 | 0.281 | 0.000150 | 0.36 | 2.64 | 11.9% |
| 4 | V1619 | 0.287 | 0.000250 | 0.59 | 3.41 | 14.8% |
| 5 | V2261 | 0.264 | 0.000325 | 0.77 | 4.23 | 15.4% |
| 6 | V1954 | 0.261 | 0.000825 | 1.96 | 4.04 | 29.6% |
| 7 | V2407 | 0.248 | 0.000875 | 2.07 | 4.93 | 29.6% |
| 8 | V1873 | 0.259 | 0.001050 | 2.49 | 5.51 | 30.3% |
| 9 | V2542 | 0.264 | 0.001150 | 2.73 | 6.27 | 30.3% |
| 10 | V1851 | 0.246 | 0.001300 | 3.08 | 6.92 | 30.8% |
| 11 | V2512 | 0.248 | 0.001875 | 4.45 | 6.55 | 40.4% |
| 12 | V2193 | 0.237 | 0.002225 | 5.28 | 6.72 | 40.4% |
| 13 | V2706 | 0.244 | 0.002250 | 5.33 | 7.67 | 40.4% |
| 14 | V2765 | 0.242 | 0.002475 | 5.87 | 8.13 | 40.4% |
| 15 | V2622 | 0.244 | 0.002650 | 6.23 | 8.72 | 40.4% |
| 16 | V1693 | 0.239 | 0.002725 | 6.46 | 9.54 | 40.4% |
| 17 | V1969 | 0.239 | 0.003550 | 8.41 | 8.58 | 47.4% |
| 18 | V1745 | 0.239 | 0.003600 | 8.54 | 9.46 | 47.4% |
| 19 | V2424 | 0.225 | 0.003925 | 9.31 | 9.69 | 48.6% |
| 20 | V2123 | 0.235 | 0.004100 | 9.72 | 10.28 | 48.6% |
| 21 | V2302 | 0.236 | 0.004700 | 11.14 | 9.86 | 50.3% |
| 22 | V2321 | 0.233 | 0.004850 | 11.50 | 10.50 | 50.3% |
| 23 | V2548 | 0.233 | 0.005225 | 12.39 | 10.61 | 50.3% |
| 24 | V845 | 0.223 | 0.005500 | 13.04 | 10.96 | 50.3% |
| 25 | V1518 | 0.235 | 0.005700 | 13.51 | 11.49 | 50.3% |
| 26 | V2322 | 0.215 | 0.005725 | 13.57 | 13.43 | 50.3% |
| 27 | V2718 | 0.215 | 0.005725 | 13.57 | 12.43 | 50.3% |
| 28 | V2327 | 0.215 | 0.006450 | 15.29 | 12.71 | 51.9% |
| 29 | V2848 | 0.224 | 0.006550 | 15.53 | 13.47 | 51.9% |
| 30 | V2090 | 0.223 | 0.006875 | 16.30 | 14.70 | 51.9% |
| 31 | V2416 | 0.231 | 0.006875 | 16.30 | 13.70 | 51.9% |
| 32 | V2500 | 0.219 | 0.007000 | 16.60 | 15.40 | 51.9% |
| 33 | V2309 | 0.227 | 0.008400 | 19.92 | 13.08 | 59.8% |

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

| | | | | | | |
|---|---|---|---|---|---|---|
| 34 | V2820 | 0.217 | 0.008700 | 20.63 | 13.37 | 59.8% |
| 35 | V2140 | 0.208 | 0.009075 | 21.52 | 14.48 | 59.8% |
| 36 | V2892 | 0.208 | 0.009075 | 21.52 | 13.48 | 59.8% |
| 37 | V2396 | 0.220 | 0.009325 | 22.11 | 14.89 | 59.8% |
| 38 | V2305 | 0.214 | 0.010700 | 25.37 | 12.63 | 62.5% |
| 39 | V1669 | 0.217 | 0.010800 | 25.61 | 13.39 | 62.5% |
| 40 | V1649 | 0.215 | 0.011075 | 26.26 | 15.74 | 62.5% |
| **Rank by logit *P*-value** | **Marker** | **Estimated Sensitivity** | ***P*-value** | **False Leads Expected** | **Estimated True Markers** | **Estimated[a] False Discovery Rate** |
| 1 | v1619 | 0.287 | <.000001 | <0.01 | 1.00 | <0.1% |
| 2 | v1830 | 0.281 | <.000001 | <0.01 | 2.00 | <0.1% |
| 3 | v1518 | 0.235 | 0.000009 | 0.02 | 2.98 | 0.7% |
| 4 | v2512 | 0.248 | 0.000045 | 0.11 | 3.89 | 2.7% |
| 5 | V2416 | 0.231 | 0.000087 | 0.21 | 4.79 | 4.1% |
| 6 | V2765 | 0.242 | 0.000104 | 0.25 | 5.75 | 4.1% |
| 7 | V2309 | 0.227 | 0.000195 | 0.46 | 6.54 | 6.4% |
| 8 | V2622 | 0.244 | 0.000255 | 0.60 | 7.40 | 6.4% |
| 9 | V1873 | 0.259 | 0.000263 | 0.62 | 8.38 | 6.4% |
| 10 | V2090 | 0.223 | 0.000285 | 0.68 | 9.32 | 6.4% |
| 11 | V1704 | 0.208 | 0.000298 | 0.71 | 10.29 | 6.4% |
| 12 | V1954 | 0.261 | 0.000324 | 0.77 | 11.23 | 6.4% |
| 13 | V2621 | 0.305 | 0.000454 | 1.08 | 11.92 | 7.6% |
| 14 | V2123 | 0.235 | 0.000496 | 1.18 | 12.82 | 7.6% |
| 15 | V2407 | 0.248 | 0.000505 | 1.20 | 13.80 | 7.6% |
| 16 | V1914 | 0.214 | 0.000529 | 1.25 | 14.75 | 7.6% |
| 17 | V2436 | 0.202 | 0.000543 | 1.29 | 15.71 | 7.6% |
| 18 | V2302 | 0.236 | 0.000617 | 1.46 | 16.54 | 7.8% |
| 19 | V1847 | 0.203 | 0.000643 | 1.52 | 17.48 | 7.8% |
| 20 | V1648 | 0.215 | 0.000686 | 1.63 | 18.37 | 7.8% |
| 21 | V1669 | 0.217 | 0.000701 | 1.66 | 19.34 | 7.8% |
| 22 | V2321 | 0.233 | 0.000725 | 1.72 | 20.28 | 7.8% |
| 23 | V2548 | 0.233 | 0.000756 | 1.79 | 21.21 | 7.8% |
| 24 | V1740 | 0.198 | 0.000851 | 2.02 | 21.98 | 8.4% |
| 25 | V2828 | 0.200 | 0.000899 | 2.13 | 22.87 | 8.5% |
| 26 | V2706 | 0.244 | 0.000994 | 2.36 | 23.64 | 8.6% |
| 27 | V2193 | 0.237 | 0.000996 | 2.36 | 24.64 | 8.6% |
| 28 | V2426 | 0.202 | 0.001012 | 2.40 | 25.60 | 8.6% |
| 29 | V291 | 0.198 | 0.001066 | 2.53 | 26.47 | 8.7% |
| 30 | V2892 | 0.208 | 0.001176 | 2.79 | 27.21 | 9.3% |
| 31 | V1969 | 0.239 | 0.001363 | 3.23 | 27.77 | 10.2% |

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

| | | | | | | |
|---|---|---|---|---|---|---|
| 32 | V1851 | 0.246 | 0.001373 | 3.25 | 28.75 | 10.2% |
| 33 | V2417 | 0.198 | 0.001592 | 3.77 | 29.22 | 11.4% |
| 34 | V826 | 0.209 | 0.001655 | 3.92 | 30.08 | 11.5% |
| 35 | V2322 | 0.215 | 0.001777 | 4.21 | 30.79 | 12.0% |
| 36 | V2908 | 0.220 | 0.001991 | 4.72 | 31.28 | 12.9% |
| 37 | V2542 | 0.264 | 0.002021 | 4.79 | 32.21 | 12.9% |
| 38 | V689 | 0.339 | 0.002248 | 5.33 | 32.67 | 14.0% |
| 39 | V2814 | 0.198 | 0.002405 | 5.70 | 33.30 | 14.6% |
| 40 | V720 | 0.198 | 0.002612 | 6.19 | 33.81 | 14.9% |

[a] Benjamini-Hochberg estimates (13)