Cancer Epidemiology 56 (2018) 83–89

Contents lists available at ScienceDirect

Cancer Epidemiology

journal homepage: www.elsevier.com/locate/canep

Recommendation to use exact P-values in biomarker discovery research in place of approximate P-values

Matthew F. Buas[a,b,*], Christopher I. Li[c], Garnet L. Anderson[d], Margaret S. Pepe[e]

[a] Department of Cancer Prevention and Control, Roswell Park Comprehensive Cancer Center, Buffalo, NY, United States
[b] Epidemiology Program, Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA, United States
[c] Translational Research Program, Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA, United States
[d] Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA, United States
[e] Biostatiistics and Biomathematics Program, Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA, United States

ARTICLE INFO

Keywords:
Hypothesis tests
ROC curve
Biomarkers
Arrays
Statistical analysis

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 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.

1. 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 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.

* Corresponding author at: Department of Cancer Prevention and Control, Roswell Park Comprehensive Cancer Center, Elm & Carlton Streets, Buffalo, NY 14263, United States.
E-mail address: matthew.buas@roswellpark.org (M.F. Buas).

https://doi.org/10.1016/j.canep.2018.07.014
Received 27 December 2017; Received in revised form 2 July 2018; Accepted 26 July 2018

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 = 1/40,000 where 40,000 is the number of random rank enumerations.

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 (=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.

2. 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 [4–6], and the ER/PR positive breast cancer discovery study is included in the Methods section of Supplementary Data .

3. Results

3.1. 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 90th 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 90th percentile of control values (i.e. exceed the 16th 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.

3.2. 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,