Supplementary Data

Methods

Our study was designed to investigate if standard *P*-value calculations are potentially incorrect in practice and if incorrect *P*-value calculations can substantially affect the reliability 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 truth. The Colocare study provided a context to motivate the simulations.

The purpose of Colocare is to find biomarkers predicting high risk of colon cancer recurrence in stage 1 patients treated with surgery. For each of 40 cases with recurrence and 160 controls without recurrence we simulated data corresponding to 3000 uninformative false biomarkers that were uncorrelated with case-control status and for 30 informative biomarkers that were correlated with case-control status. We use the terminology “true biomarkers” for the 30 informative biomarkers. Each biomarker was generated as standard normal, mean=0 and standard deviation=1, for the 160 controls. The uninformative biomarker data was generated for the 40 cases in the same way as for controls. For the 30 true biomarkers we generated case biomarker data as normal with mean 0.536 and standard deviation 1. The mean was chosen so that true biomarkers would satisfy a performance criterion described below. This is a classic simulation scenario (*1–3*), although in practice biomarker data for cases may have a diverse range of distributions relative to controls. In sensitivity analyses, we also conducted simulations in which the mean value for each of the 30 true biomarkers among cases was a random number ranging from 0.236 to 0.836 with average equal to 0.536.

The key biomarker performance measure of interest in the Colocare study is the recurrence probability among biomarker positive patients. The biomarker positivity threshold is chosen as the 90th percentile of the control values (i.e. the 16th largest of the 160 control values) so as to guarantee the marker has 90% specificity. The recurrence probability among biomarker positive patients in the population (positive predictive value, *PPV*) will be estimated by Bayes formula as

$$\mathit{logit\ (PPV) = logit\ (p) + log\ (sensitivity) – log\ (1\text{-}specificity)}\quad (1)$$

where the *logit* function is *logit(x)=log(x/(1-x))*, *p*=10%=the overall recurrence probability in stage 1 patients (i.e. prevalence), the *specificity* is set to 90% and *sensitivity* is the observed proportion of cases that are biomarker positive. The *sensitivity* is also known as the empirical estimate of ROC(0.1). Testing if a biomarker is uninformative is to test if biomarker-positive individuals have the same prevalence of the outcome as observed in the entire study population, i.e. H_0: PPV =10% . Given the above formula this is equivalent to testing H_0*: ROC(0.1)=0.1 (*i.e. sensitivity=1-specificity, where specificity=0.90), so *P*-values will be based on testing the null hypothesis

$$H_0: ROC(0.1)=0.1$$

using the empirical ROC estimate, for which standard methods are available (*4*). Let obs-ROC_{emp} be the value of the empirical ROC estimate, denoted by ROC_{emp}, calculated with data on a biomarker from our study. The associated one-sided *P*-value is the probability that in repetitions of our study one would observe ROC_{emp} values as large as the one we found assuming that biomarker values for cases in the population have the same distribution as biomarker values for controls in the population.

$$P\text{-}value = Probability(ROC_{emp} >= obs\text{-}ROC_{emp} \mid cases\ same\ as\ controls).$$

Approximate *P*-value calculations use the approximation that ROC_{emp} is normally distributed in large samples

$$standard\ normal\ P\text{-}value = 1\text{-}\ \Phi(Z)$$

where Z = (obs-$\mathrm{ROC_{emp}}$ – 0.1) /se($\mathrm{ROC_{emp}}$)), Φ is the standard normal cumulative distribution function and the standard error, se($\mathrm{ROC_{emp}}$), is estimated by bootstrap resampling (*4*). We used 500 bootstrap samples, separately resampling 40 cases and 160 controls per the study design. Other methods for calculating the standard error are also possible but are more involved because they require estimating probability densities (*5*, *6*). An alternative *P*-value calculation acknowledges that the sensitivity, $\mathrm{ROC_{emp}}$(0.1), can't really be normally distributed since the normal distribution is unrestricted in negative and positive directions while proportions such as $\mathrm{ROC_{emp}}$(0.1) are restricted between 0 and 1. Proportions are often more like normal variables after applying the *logit* transform. This gives rise to the *P*-value calculation

standard logit-normal P-value = 1- Φ(logit-Z)

where *logit-Z = (logit(obs-ROC_{emp}) – logit(0.1)) /se(logit(ROC_{emp}))* and se(logit($\mathrm{ROC_{emp}}$)) is estimated by bootstrapping as above.

Our proposal is to calculate the *P*-value exactly without approximation. 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. For example, in the Colocare study we will have a total of 200 subjects and suppose the cases are labelled as subjects 1-40. If cases have biomarker values with the same population distribution as controls, the study data will be comprised of a random enumeration of ranks for 200 individuals. We calculate the corresponding $\mathrm{ROC_{emp}}$ statistic for a large number of random enumerations (or all 200! possible enumerations) and tabulate the results. Because there are 40 cases, there are at most 40 possible values for ROC_{emp}, so it is easy to tabulate the distribution of ROC_{emp} (Table 1) and report the exact *P*-value corresponding to an observed value of ROC_{emp}

$$\textit{exact P-value = proportion of enumerations with } ROC_{emp} >= \textit{obs-}ROC_{emp}.$$

For example if the empirical ROC estimate calculated for a biomarker is 0.20, the corresponding exact *P*-value is 0.059175 (Table 1). We selected to use 40,000 enumerations at random with replacement since this required far fewer than all 200! enumerations and yet provided reasonably precise *P*-value calculations. In particular the standard errors of the *P*-value estimates are 0.001, 0.00063 and 0.00045 when the *P*-values are 0.05, 0.02 and 0.01, respectively.

To demonstrate that the method used to calculate *P*-values in real data analysis can have a substantial effect on conclusions drawn, we reanalyzed data from an ER/PR positive breast cancer biomarker discovery study reported in (*8*). The study sought to discover early detection biomarkers that might be used to encourage women who do not have easy access to mammography to go for mammography screening. Markers that maximize sensitivity while maintaining at least 90% specificity are preferred for this clinical context. As described in detail in (*8*), preclinical plasma samples from 121 cases and 121 controls from the WHI observational study were interrogated with an array of 3290 antibodies. There were 2467 biomarkers reported in (*8*) after removing technical controls and imposing quality control filters based on coefficient of variation across triplicate spots and a criterion for percent of observations missing data. We only included the subset of 2371 biomarkers where at least 100 controls and at least 100 cases have data. As noted above, and similar to the Colocare study, we focused on the sensitivity corresponding to 90% specificity as the biomarker performance measure of interest. Our analysis approach differed in many respects from that previously reported in (*8*) because our goals were different and more limited. For example, since we just wanted to investigate if different *P*-value calculations provided different rankings and selections of biomarkers, we did not need to split the data into training and test sets as was done in (*8*).

References

1. M. S. Pepe, H. Janes, C. I. Li, Net risk reclassification P values: Valid or misleading? *Journal of the National Cancer Institute*. **106** (2014), doi:10.1093/jnci/dju041.

2. K. Dobbin, R. Simon, Sample size determination in microarray experiments for class comparison and prognostic classification. *Biostatistics (Oxford, England)*. **6**, 27–38 (2005).

3. A. J. Vickers, A. M. Cronin, C. B. Begg, One statistical test is sufficient for assessing new predictive markers. *BMC medical research methodology*. **11**, 13 (2011).

4. M. Pepe, G. Longton, H. Janes, Estimation and Comparison of Receiver Operating Characteristic Curves. *The Stata journal*. **9**, 1 (2009).

5. F. Hsieh, B. W. Turnbull, others, Nonparametric and semiparametric estimation of the receiver operating characteristic curve. *The annals of statistics*. **24**, 25–40 (1996).

6. M. S. Pepe, *The statistical evaluation of medical tests for classification and prediction* (Medicine, 2003).

7. M. Hollander, D. A. Wolfe, E. Chicken, *Nonparametric statistical methods* (John Wiley & Sons, 1999).

8. M. F. Buas *et al.*, Candidate early detection protein biomarkers for ER+/PR+ invasive ductal breast carcinoma identified using pre-clinical plasma from the WHI observational study. *Breast Cancer Res Treat*. **153**, 445–454 (2015).

Supplementary Table S.1: Ten independent replications of the simulation study reported in Table 3. Markers discovered by the selection criterion: biomarker p-value < 0.0277 from a dataset with 3,000 uninformative (false) biomarkers and 30 true biomarkers when p-values are based on exact calculation or on normal approximation with logit transformation. The test statistic is the sensitivity at 90% specificity estimated with the empirical ROC. Number of study subjects: 40 cases and 160 controls.

		False Discoveries Observed		**True Discoveries Estimated[2]: Observed**	
	Estimated[1]	**Exact *P*-value**	**Logit-Normal *P*-value**	**Exact *P*-value**	**Logit-Normal *P*-value**
Replication					
1	84	83	107	25 : 26	48 : 25
2	84	75	92	14 : 23	32 : 24
3	84	85	114	26 : 25	55 : 25
4	84	76	97	18 : 26	38 : 25
5	84	84	107	26 : 26	47 : 24
6	84	82	110	23 : 25	51 : 25
7	84	70	90	11 : 25	28 : 22
8	84	77	94	17 : 24	33 : 23
9	84	74	100	18 : 28	44 : 28
10	84	99	124	41 : 26	64 : 24

[1] estimated false discoveries = (threshold-p) × number of biomarkers. This is the same for all p-value methods

[2] total discoveries = estimated false discoveries + estimated true discoveries

Supplementary Table S.2: Summary results from the primary simulation study (Table 3, “0”) and the ten replications of the simulation study (1-10) described in Supplementary Table S.1. For each simulation, we present the difference between the number of estimated versus observed false discoveries, or estimated versus observed true discoveries, when using exact (E) or logit-normal (L-N) P values. The total number of ‘misclassified’ biomarkers, using either type of P value, is equal to the absolute value of the (E-O) difference for either false or true discoveries (an overestimation of the number of false discoveries corresponds to an equal-magnitude underestimation of true discoveries, and vice versa). For 9 of the 11 simulations (all except #2 and #7), use of exact P values resulted in a smaller number of ‘misclassified biomarkers’. Across all 11 simulations, use of exact P values led to 69 misclassifications, while use of logit-normal P values led to 217 misclassifications.

	False Discoveries			**True Discoveries**	
Simulation	**(Estimated – Observed)**			**(Estimated – Observed)**	
	E	**L-N**		**E**	**L-N**
0	2	-22		-2	22
1	1	-23		-1	23
2	9	-8		-9	8
3	-1	-30		1	30
4	8	-13		-8	13
5	0	-23		0	23
6	2	-26		-2	26
7	14	-6		-14	6
8	7	-10		-7	10
9	10	-16		-10	16
10	-15	-40		15	40

Supplementary Table S.3: Results analogous to Table 3 of the paper in which data for the 3000 uninformative (false) biomarkers is the same as that in Table 3 but data for the 30 true biomarkers are generated anew with variation in separations between cases and controls. Specifically, for each true biomarker, the mean difference between cases and controls is a random uniform number between 0.236 and 0.836 with average equal to 0.536 which is the constant separation for true biomarkers in Table 3.

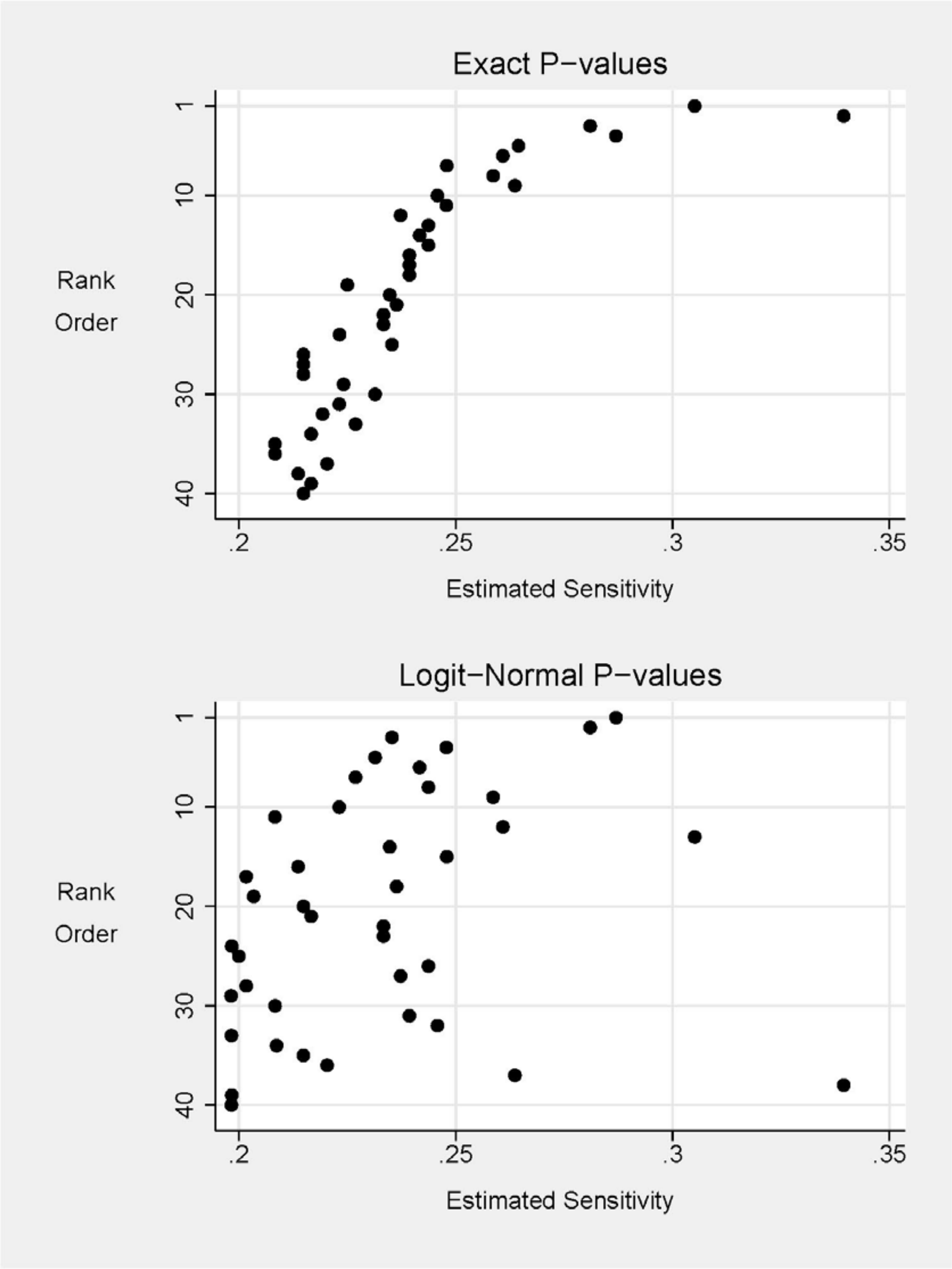
		Number of Markers	
		Exact *P*-value	**Logit-Normal *P*-value**
Threshold for Sensitivity *P*-value			
0.0277	**Total Discoveries**	98	120
	False Discoveries		
	estimated[b]	84	84
	actual	82	106
	True Discoveries		
	estimated[d] (tdr[c])	14 (14%)	36 (30%)
	actual(tdr)	16 (16%)	14 (12%)
0.0121	**Total Discoveries**	43	72
	False Discoveries		
	estimated	37	37
	actual	29	58
	True Discoveries		
	estimated(tdr)	6 (14%)	35 (49%)
	actual(tdr)	14 (33%)	14 (19%)

[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

Supplementary Figure S.1: Rank orders of p-values versus corresponding estimated sensitivity at 90% specificity for the top 40 biomarkers according to exact p-values and for the top 40 biomarkers according to logit-normal p-values.


Exact P−values
Rank Order
Estimated Sensitivity
Logit−Normal P−values
Rank Order
Estimated Sensitivity