Estimating Prevalence, False-Positive Rate, and False-Negative Rate with Use of Repeated Testing When True Responses Are Unknown

To the Editor: We read with great interest the article by Wong et al.[1] We were especially interested in the approach used to estimate false-positive (p_{10}) and false-negative (p_{01}) rates of their copy-number variation (CNV)–detection algorithm.

To estimate p_{10} and p_{01}, Wong et al.[1] performed six repeated experiments in which they applied their CNV-detection algorithm each time on 24,392 clones from a single female versus a male reference sample. Under the assumption that all clones are non-CNV clones and all CNV calls are false, an estimate of the maximum p_{10} is derived (note that this underestimates the true maximum p_{10}), and, by the use of the binomial probability, the probability of calling the same clone twice within six repeated experiments is estimated to be 0.000445. On the basis of this low probability of calling the same (assumed) non-CNV clone twice, Wong et al.[1] assume that any clone called twice or more often is a true CNV. The authors recognize that this assumption is too strong, since a fraction of the single-occurrence calls may represent true CNVs. Under this assumption, they estimate p_{10} to be 0.2323% and p_{01} to be 45.3%.

The problem of estimating p_{10} and p_{01} for this CNV data set is a problem of evaluating diagnostic tests without gold standards.[2] In these situations, the true responses (e.g., the true CNV statuses) are unknown. If, in addition to p_{10} and p_{01}, the population prevalence (θ_1) is unknown, then there are three parameters to be estimated. However, for a single test performed once in a single population, the data provide only 1 df. One way to achieve additional degrees of freedom is to apply the test several times to all individuals (e.g., clones) in the population,[2] as was done by Wong et al.[1]

Dawid and Skene[3] derived the observed data likelihood (ODL) when true responses are unknown and repeated experiments are performed. In our particular case of one type of test (i.e., the CNV-detection algorithm) and six repeated experiments, we have

$$\log(\text{ODL}) = \sum_{i=1}^{n}\left[\log\left(\sum_{d=0}^{1}\theta_d\prod_{t=0}^{1}p_{td}^{n_{it}}\right)\right] ,$$

where n is the number of clones, d is the true CNV status ($d = 0$ for non-CNV clones; $d = 1$ for CNV clones), t is the test result from the CNV-detection algorithm, θ_d is the prevalence of the true status d,

$$p_{td} = P(\text{test result} = t\,|\,\text{CNV status} = d) ,$$

and n_{it} is the number of times clone i has test result t. Let δ_{id} be the indicator variable for the true response of clone i ($\delta_{id} = 1$ if clone i is of true status d). If the true responses are available, then the complete data likelihood (CDL) is

$$\log(\text{CDL}) = \sum_{i=1}^{n}\left\{\log\left[\prod_{d=0}^{1}\left(\theta_d\prod_{t=0}^{1}p_{td}^{n_{it}}\right)^{\delta_{id}}\right]\right\} .$$

The maximum-likelihood estimates (MLEs) of the CDL can be calculated analytically, and the estimators are

$$\hat{p}_{td} = \frac{\sum_{i=1}^{n}\delta_{id}n_{it}}{\sum_{i=1}^{n}\delta_{id}n_{i0} + \sum_{i=1}^{n}\delta_{id}n_{i1}} \qquad (1)$$

and

$$\hat{\theta}_d = \frac{\sum_{i=1}^{n}\delta_{id}}{n} . \qquad (2)$$

The Bayes theorem can be used to get estimates of δ_{id} as probabilities that the true response for clone i is d,

$$P(\delta_{id} = 1\,|\,\text{data}) = \frac{\prod_{t=0}^{1}p_{td}^{n_{it}}\theta_d}{\prod_{t=0}^{1}p_{t0}^{n_{it}}\theta_0 + \prod_{t=0}^{1}p_{t1}^{n_{it}}\theta_1} . \qquad (3)$$

However, the values of δ_{id} are unknown, so, to maximize the ODL, Dawid and Skene[3] suggest using the expectation-maximization (EM) algorithm and proceed as follows:

1. Take initial estimates of δ_{id} (may be indicator variables or probabilities),
2. Use equations (1) and (2) to obtain estimates of ps and θs,
3. Use equation (3) and the estimates of ps and θs from step 2 to calculate new estimates of δ_{id} expressed as probabilities, and
4. Repeat steps 2 and 3 until convergence is achieved (difference in ODL from two successive steps is $<10^{-6}$).

The observed Fisher information matrix of the ODL can be derived analytically; hence, likelihood-based CIs for the parameter estimates are easily calculated, as long as the information matrix is invertible. The likelihood-based CIs will be meaningful if the quadratic approximation of the likelihood is good.

We applied the above EM algorithm to the data of Wong et al.[1] and obtained the estimates $\hat{\theta}_1 = 0.6299\%$ (95% CI 0.5213%–0.7384%), $\hat{p}_{10} = 0.2283\%$ (95% CI 0.2026%–0.2541%), and $\hat{p}_{01} = 48.91\%$ (95% CI 45.37%–52.46%). Additionally, we plotted the likelihood surface for fixed values of θ_1, to make sure the estimates corresponded to global maxima of the ODL. Note that our p_{10} estimate of 0.2283% is close to the estimate of 0.2323% of Wong et

al.,[1] but our estimate of p_{01} of 48.91% is higher than their estimate of 45.3% (which falls just outside our 95% CI).

Because of the symmetric nature of the parameters, the ODL has, in most cases, two equal local maxima; specifically, the set of parameter values $\theta_1 = x_1$, $p_{10} = x_2$, and $p_{01} = x_3$ give the same ODL as the parameter values: $\theta_1 = 1 - x_1$, $p_{10} = 1 - x_3$, and $p_{01} = 1 - x_2$. For the above EM algorithm to result in unique parameter estimates, corresponding to one of the local maxima, it is therefore necessary to assume that the CNV-detection algorithm provides better than random classification of clones. This assumption is similar to but not as strong as the assumption of Wong et al.[1] that all clones called twice or more often represent true CNVs. Additional strengths of the EM algorithm over the more ad hoc procedure of Wong et al.[1] are (1) likelihood-based CIs may be derived, (2) an arbitrary threshold for declaring true CNVs is avoided, (3) the prevalence, θ_1, of CNVs is also estimated (as the proportion of BAC-array clones harboring CNVs detectable at the false-positive rate of p_{10} and false-negative rate of p_{01}), and (4) the EM algorithm works reasonably well in more situations than does the approach of Wong et al.[1]

Regarding the relative performance of the EM algorithm, the estimation procedure of Wong et al.[1] works reasonably well for p_{10} in the special case of low prevalence and large sample size, but it fails when the prevalence increases and the sample size decreases. The convergence of the EM algorithm slows down considerably as the prevalence increases, and the search for a global maximum often needs to be repeated from several starting values.

To illustrate these points, table 1 provides comparative results from the EM algorithm and the procedure of Wong et al.[1] for simulated data. Data sets of three different sample sizes (i.e., number of clones = 500, 1,000, and 25,000) are simulated under three different scenarios (1, 2, and 3). Under all scenarios p_{10} and p_{01} are set equal to the estimates of the EM algorithm for data of the Wong et al.,[1] but the prevalence, θ_1, is varied. Under scenarios 1 and 2, θ_1 is assumed to be high (10% and 20%, respectively). However, for scenario 3, θ_1 is set to equal the estimate derived using the EM algorithm for the data of Wong et al.[1] For each clone, six test results are simulated; for non-CNV clones, six Bernoulli random values are simulated with "success" probability p_{10}; for CNV clones, the Bernoulli values have "success" probability $1 - p_{01}$. Application of the procedure of Wong et al.[1] requires one to set a maximum binomial probability (point 2 in the previous paragraph) for assuming clones called k or more times represent true CNVs. We set this maximum probability to be 0.1%. This means that a clone called $\geq k$ times is assumed to be a true CNV clone only if the binomial probability of calling a non-CNV clone $\geq k$ times is $\leq 0.1\%$; k is taken to be the smallest k for which this condition holds.

As can be seen (table 1), the estimates from the Wong et al.[1] procedure are quite poor in the cases of high prevalence ($\theta_1 = 10\%$ or 20%); the point estimates of p_{10} and p_{01} are far from their true values and outside the 95% CIs.

Table 1. Results from Nine Simulated Data Sets for Three Scenarios, Each with Specific Values of n, θ_1, p_{10}, and p_{01}

Scenario, n, and Parameter	True Value (%)	Point Estimate (%)	
		EM Algorithm (95% CI)	Procedure of Wong et al.[1]
Scenario 1, $k = 5$[a]			
$n = 500$:			
θ_1	20.0000	20.46 (16.64–24.28)	NA
p_{10}	.2283	.07 (−.22 to .37)	8.23
p_{01}	48.9100	46.69 (41.80–51.59)	14.58
$n = 1{,}000$:			
θ_1	20.0000	20.82 (17.97–23.66)	NA
p_{10}	.2283	.07 (−.18 to .33)	8.62
p_{01}	48.9100	50.71 (47.08–54.34)	15.00
$n = 25{,}000$:			
θ_1	20.0000	20.05 (19.50–20.60)	NA
p_{10}	.2283	.24 (.19–.29)	8.39
p_{01}	48.9100	49.07 (48.35–49.79)	14.37
Scenario 2, $k = 4$[a]:			
$n = 500$:			
θ_1	10.0000	9.19 (6.38–11.99)	NA
p_{10}	.2283	.29 (.01–.57)	2.60
p_{01}	48.9100	49.86 (42.00–57.72)	29.17
$n = 1{,}000$:			
θ_1	10.0000	9.57 (7.61–11.53)	NA
p_{10}	.2283	.34 (.14–.54)	2.38
p_{01}	48.9100	46.79 (41.70–51.87)	28.17
$n = 25{,}000$:			
θ_1	10.0000	10.20 (9.79–10.61)	NA
p_{10}	.2283	.19 (.15–.23)	2.70
p_{01}	48.9100	49.68 (48.67–50.69)	27.44
Scenario 3, $k = 2$[a]:			
$n = 500$:			
θ_1	.6299	1.12 (−.26 to 2.51)	NA
p_{10}	.2283	.26 (.03–.49)	.30
p_{01}	48.9100	63.86 (36.42–91.30)	54.17
$n = 1{,}000$:			
θ_1	.6299	.73 (.18–1.28)	NA
p_{10}	.2283	.23 (.11–.35)	.23
p_{01}	48.9100	40.04 (16.91–63.18)	38.10
$n = 25{,}000$:			
θ_1	.6299	.62 (.51–.72)	NA
p_{10}	.2283	.23 (.20–.25)	.23
p_{01}	48.9100	47.56 (43.55–51.57)	44.17

NOTE.—Under each scenario, results from the EM algorithm and the procedure of Wong et al.[1] are given for three sample sizes (n). Under the procedure of Wong et al.,[1] we let the maximum binomial probability be 0.1%; the value of k corresponding to this probability is given in each case. NA = not applicable.

[a] k happened to be the same for all values of n within each scenario.

Under scenario 3 of low prevalence ($\theta_1 = 0.6299\%$) and small number of clones ($n = 500$ or 1,000), the point estimates from the procedure of Wong et al.[1] for p_{01} are quite far from the true value; the Wong et al.[1] estimates of both p_{10} and p_{01} are, however, within the 95% CIs. Regarding the choice of k, we examined five different simulated data sets (results not shown) and found that the Wong et al.[1] estimates strongly depend on k and hence on the choice of the probability threshold. These analyses therefore highlight the arbitrariness in the choice of probability threshold and the importance of using an estimation procedure that is independent of the choice of this threshold,

or k. In contrast, the EM algorithm requires no k at all, and its estimates are quite close to the true values under most scenarios and samples sizes, the 95% CIs cover the true values in all cases, and the point estimates of all parameters improve as the number of clones is increased or the prevalence decreased. In each case, only one data set was simulated, so random variation due to the simulation is not accounted for by averaging over many simulated data sets. Complete evaluation of the performance of the EM algorithm would require data sets to be simulated multiple times under each scenario. However, such an evaluation is outside the scope of this letter.

Our proposed approach does not model any potential dependence between clones on a given BAC array due to variation in, for example, signal-to-noise ratios across BAC arrays. Perhaps random-effects modeling—adding BAC array IDs as a random component to the model—could help account for this additional variation. However, we do not expect that this issue would change the overall conclusions of this letter.

This letter is not a criticism of the work of Wong et al.[1] but rather a note that their procedure will not work well in all cases, so authors facing problems similar to those reported by Wong et al.[1] will need to choose estimation procedures that work in the special cases their data provide. Walter and Irwing[4] and Hui and Zhou[2] provide reviews of methods for evaluation of diagnostic tests without gold standards. Joseph et al.[5] proposed a Bayesian estimation method for parameters of diagnostic tests in the absence of a gold standard; however, it has been suggested[6] that the method of Joseph et al.[5] suffers from lack of good large-sample properties. Shortly after we submitted our letter, we became aware of another letter[7] that describes an alternative statistical approach for tackling the problems addressed here. The procedure is Bayesian and has one very interesting feature, in that it allows one to estimate the number of calls expected in each category (i.e., categories of clones never called, called once, called twice, etc.). However, because it is a Bayesian procedure, prior distributions for p_{10} and p_{01} need to be specified. In any event, as we have shown here, special care needs to be taken when estimating prevalence and false-positive and false-negative rates in the absence of a gold standard.

Johanna Jakobsdottir and Daniel E. Weeks

Acknowledgments

This work was supported by the University of Pittsburgh and by National Eye Institute grant R01EY009859 (to Dr. Michael B. Gorin of the University of California–Los Angeles).

References

1. Wong KK, deLeeuw RJ, Lam WL, Dosanjh NS, Kimm LR, Cheng Z, Horsman DE, MacAulay C, Ng RT, Brown CJ, et al (2007) A comprehensive analysis of common copy-number variations in the human genome. Am J Hum Genet 80:91–104
2. Hui SL, Zhou XH (1998) Evaluation of diagnostic tests without gold standards. Stat Methods Med Res 7:354–370
3. Dawid AP, Skene AM (1979) Maximum likelihood estimation of observer error-rates using the EM algorithm. Appl Statist 28: 20–28
4. Walter SD, Irwing LM (1988) Estimation of test error rates, disease prevalence and relative risk from misclassified data: a review. J Clin Epidemiol 41:923–937
5. Joseph L, Gyorkos TW, Coupal L (1995) Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard. Am J Epidemiol 141:263–272
6. Anderson S (1996) Re "Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard." Am J Epidemiol 144:290
7. Lynch A, Marioni J, Tavaré S (2007) Numbers of copy-number variations and false-negative rates will be underestimated if we do not account for the dependence between repeated experiments. Am J Hum Genet 81:418–420

From the Departments of Biostatistics (J.J.; D.E.W.) and Human Genetics (D.E.W.), Graduate School of Public Health, University of Pittsburgh, Pittsburgh

Address for correspondence and reprints: Johanna Jakobsdottir, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh, 130 DeSoto Street, Pittsburgh, PA 15261. E-mail: jjakobsdottir@hgen.pitt.edu

Am. J. Hum. Genet. 2007;81:1111.

0002-9297/2007/8105-0022$15.00
DOI: 10.1086/521582