Tr. J. of Medical Sciences
29 (1999) 425-430

Ergun KARAAĞAOĞLU

Estimation of the Prevalence of a Disease from Screening Tests

Received: September 14, 1998

Department of Biostatistics, Faculty of Medicine, Hacettepe University, 06100 Ankara-Turkey

Abstract: Since most secreening tests are not 100% accurate, the proportion of subjects screened positive in such a test cannot be used as an estimate of the population prevalence. Methods which take sensitivity and the specificity into consideration should be employed in such circumstances. Estimation of the population prevalence as defined by Gart and Buck may produce results which are outside the range of 0 to 1. A Bayesian approach avoids results of this kind, but requires complicated computations. Lew and Levy proposed an approximation to the Bayesian estimate of the population prevalence. To simplify the computations, I propose a method which requires the evaluation of a logistic function. The coefficients of the function are tabulated for some selected test characteristics and sample sizes. For other values that are not tabulated, coefficients can be interpolated. Although the method is simple it produces very accurate results.

Key Words: Estimation, prevalence, screening tests, sensitivity, specificity.

Introduction

It is often necessary to estimate the prevalence of a disease in the underlying population on the basis of a screening test that cannot discriminate diseased and nondiseased subjects with 100% certainty. Such screening tests are usually far from being a gold standard test, that is, their sensitivities and specificities are less than one, thus yielding false positive and false negative results. When cost or other factors (such as speed and/or risks) are considered, a gold standard test may not be efficient in screening large groups of individuals.

Prevalence estimates are important in planning health services and policies. When such screening tests are administered to a sample of individuals, the proportion of subjects with a positive test result, therefore, cannot be used as an estimate of the true prevalence in the population. The reason is that among those who yield positive test results some are falsely labelled as positive (although they are free of the disease) and among those who are negative on the test, some are actually diseased (false negatives).

Although the main objective of diagnostic tests is to evaluate the positive and negative predictive values, for screening purposes it is also desirable to estimate the disease prevalence.

Estimation of the disease prevalence on the basis of screening tests has long been of interest to many scientists. Gart and Buck (1) derived an estimate of the true population prevalence, π'', by using the sensitivity (S_e) and specificity (S_p) of the test and the proportion of subjects screened positive by that test (R). This estimate is given by:

$$\pi'' = \frac{(R + Sp - 1)}{(Se + Sp - 1)} \qquad (1)$$

Levy and Kass (2), Rogan and Gladen (3) and Gastwirth (4) have also proposed this estimate. Although rare, it can easily be shown that this estimate can yield results which are negative or above unity. For example, if 10% of the persons are screened as positive by a test which has a sensitivity of 0.80 and a specificity of 0.70, the estimated prevalence, π'', would be –0.40 and if 90% of the persons are screened as positive by the same test π'' would be 1.2. Therefore, estimates of the prevalence that always lie between zero and one and have values

close to the estimate defined in (1) in most situations, are of greater concern. Lew and Levy (5) proposed the use of a Bayesian estimate that fulfils the requirements mentioned above. As they indicated, the Bayesian estimator n' requires numerical evaluation of the ratio of integrals as defined below:

$$\pi' = \frac{(d + Sp - 1)}{(Se + Sp - 1)} \quad (2)$$

where,

$$d = \frac{\int_{1-Sp}^{Se} P^{x+1}(1-P)^{n-x}\, dP}{\int_{1-Sp}^{Se} P^{x}(1-P)^{n-x}\, dP} \quad (3)$$

For users without the necessary knowledge or routines to evaluate the above integration they recommended the use of the quadratic function of the observed raw prevalence and obtain π' quickly to a reasonable degree of accuracy for typical values of S_e, S_p and sample sizes from 20 to 100. This procedure, however, requires two functions to be evaluated and the evaluations vary according to the value of p, the observed proportion of positives in the test.

In this article, I propose a very simple and accurate method to obtain the approximate Bayesian estimate of the true population prevalence. For practical purposes, only for some selected sensitivities, specificities and sample sizes are the coefficients tabulated. Approximations for other values that are not tabulated, can be obtained by interpolation.

Methods

Approximate Bayesian Estimate of the True Population Prevalence

As a Bayesian estimate of the population prevalence requires integration, a simple but accurate approximation to this estimate is of interest. When the Bayesian estimate and proportion of positives in a test with a known sensitivity and specificity are plotted for different sample sizes, it becomes apparent that points fit very well to a logistic function in the from of:

$$y = \frac{1}{1 + e^{-(b_0 + b_1 x)}} \quad (4)$$

where y is the Bayesian estimate and x is the proportion of subjects with a positive test result. SPSS Release 6.0 is used to estimate the constants, b_0 and b_1 for samples of sizes 20 to 500 with increments of 10; for sensitivities and specificities from 0.70 to 0.90 with increments of 0.05. This makes a total of 1225 functions. Among the 1225 functions studied, the lowest R^2 was 0.993 and the difference between the calculated Bayesian estimate and its estimated value from (5) never exceeded 4%. Such residual values as high as 4% were very rare and observed either for very low or very high proportions (such as when the proportion of subjects screened positive was 90% or greater), which may not be very common in practice. In most situations the approximations are exact or true within 1–2% of the actual Bayesian estimate. For samples of size greater than 200 the equation given by (1) is recommended. In this case, if π'' is negative it can be assumed to be zero or if it is above unity it can be assumed to be 1. For large samples ($n \geq 200$) the difference between the Bayesian estimate and π'' defined in (1) is negligible and because of its simplicity, use of π'' is more practical. Figure 1 (a) through (c) display the Bayesian estimate of the population prevalence calculated by (2) and (3), the estimate obtained by (1) and the approximate Bayesian estimate calculated by (5) as a function of p, proportion of positives in the test, for three arbitrarily selected sample sizes, sensitivities and specificities.

Since many pages would be necessary to list the constants of the logistic equation for varying sample sizes, sensitivities and specificities, I restricted it to a single page and tabulated the values of b_0 and b_1 for some selected sensitivites (S_e = 0.70, 0.80 and 0.90), specificities (S_p = 0.70, 0.80 and 0.90) and sample sizes (n = 20, 30, 40, 50, 60, 70, 80, 90, 100, 150 and 200). From Table 1, coefficients b_0 and b_1 can be obtained for given values of S_e, S_p and n. By using these coefficients the approximate Bayesian estimate, π_a is:

$$\pi_a = \frac{1}{1 + e^{-(b_0 + b_1 p)}} \quad (5)$$

where p is the observed proportion of subjects screened positive by the test. For screening tests which have sensitivities and specificities between 0.70 and 0.90 and for sample sizes between 20 and 200 but not tabulated, the interpolation method as defined in (6) can be used to obtain the coefficients b_0 and b_1.

Example 1: Suppose that a test which is 0.80 sensitive and 0.90 specific is used to screen a group of 80 subjects

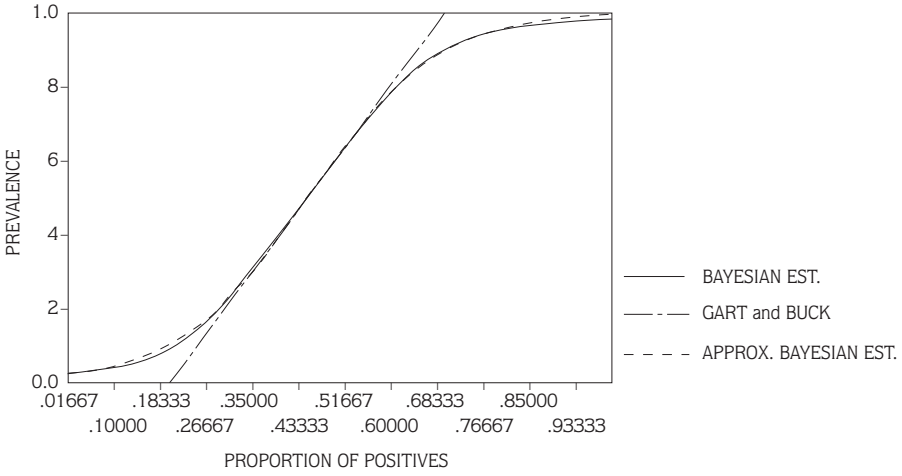

Figure 1.a Se=0.70, Sp=0.80, n=60

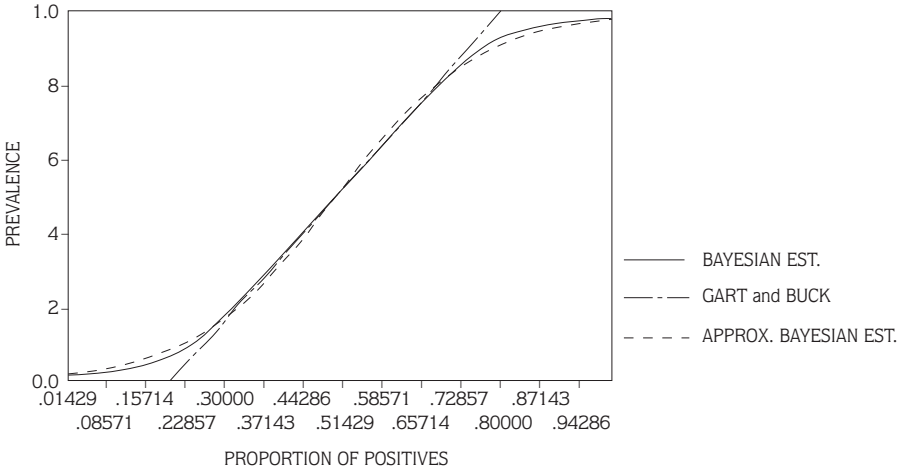

Figure 1.b Se=0.80, Sp=0.80, n=70

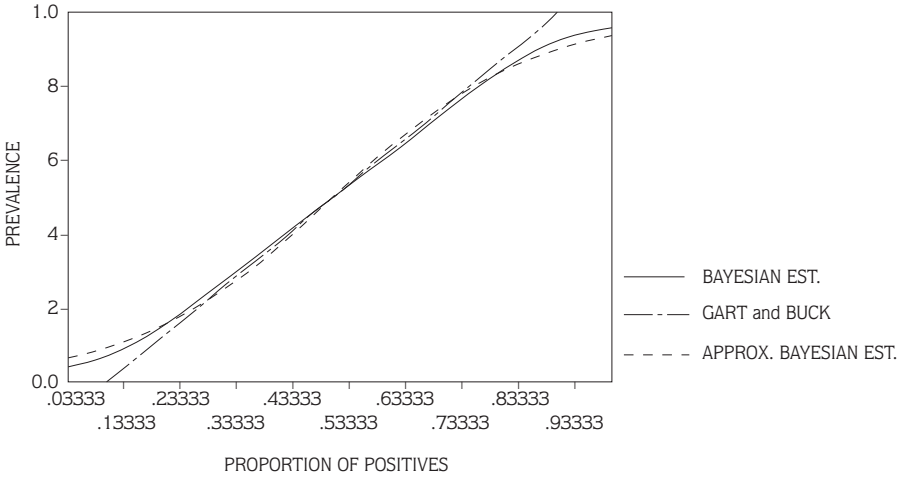

Figure 1.c Se=0.90, Sp=0.90, n=30

Figure 1. (a–c) Comparison of three different methods of estimation; Bayesian, the one derived by Gart and Buck and the approximate Bayesian proposed in this article, for some arbitrarily selected sensitivities, specificities and sample sizes.

for the presence of a certain disease. Assume that 15 subjects were positive in that test (p = 0.1875). The coefficients can be obtained from Table 1 as $b_0 = -3.109$ and $b_1 = 6.920$. Using these coefficients and substituting p = 0.1875 in (5) we get:

$$\pi_a = \frac{1}{1 + e^{-(-3.109 + 6.920 \times 0.1875)}} = 0.14$$

as the approximate Bayesian estimate of the true population prevalence. In such a case the actual Bayesian estimate, found by (2) and (3), would be 0.137.

Interpolation Method

Since the values of b_0 and b_1 are determined by n_1, S_e and S_p, all possible combinations of the three parameters should be considered when interpolating the values. The required steps can be outlined as follows:

A) If $0.7 \le S_e \le 0.9$; $0.7 \le S_p \le 0.9$; $20 \le n \le 200$

Step 1. Find the lower and upper tabulated values of S_e, S_p and n, between which the sensitivity and the specificity of the current test and the sample size lie. Let the subscripts L and U denote the lower and upper tabulated values respectively and subscript C denote the

Table 1. Coefficients of the Logistic Function for Selected S_e, S_p and n Values.

N	S_e	S_p = 0.70		S_p = 0.80		S_p = 0.90	
		b_0	b_1	b_0	b_1	b_0	b_1
	0.70	–2.947	5.888	–2.744	6.089	–2.377	5.962
20	0.80	–3.357	6.105	–2.985	5.967	–2.532	5.646
	0.90	–3.619	6.019	–3.131	5.680	–2.633	5.268
	0.70	–3.603	7.204	–3.203	7.117	–2.671	6.702
30	0.80	–3.918	7.122	–3.349	6.698	–2.763	6.158
	0.90	–4.050	6.736	–3.405	6.179	–2.809	5.620
	0.70	–4.076	8.151	–3.496	7.772	–2.844	7.132
40	0.80	–4.278	7.775	–3.565	7.129	–2.893	6.444
	0.90	–4.303	7.157	–3.559	6.461	–2.905	5.812
	0.70	–4.430	8.858	–3.698	8.221	–2.956	7.412
50	0.80	–4.525	8.224	–3.707	7.413	–2.976	6.628
	0.90	–4.467	7.432	–3.658	6.640	–2.966	5.933
	0.70	–4.701	9.402	–3.844	8.547	–3.035	7.609
60	0.80	–4.704	8.549	–3.806	7.613	–3.034	6.755
	0.90	–4.583	7.625	–3.726	6.765	–3.008	6.017
	0.70	–4.915	9.828	–3.955	8.792	–3.094	7.754
70	0.80	–4.839	8.794	–3.881	7.761	–3.076	6.848
	0.90	–4.668	7.768	–3.776	6.857	–3.038	6.078
	0.70	–5.086	10.171	–4.041	8.984	–3.139	7.866
80	0.80	–4.945	8.986	–3.938	7.875	–3.109	6.920
	0.90	–4.734	7.878	–3.815	6.928	–3.062	6.125
	0.70	–5.226	10.452	–4.110	9.138	–3.175	7.955
90	0.80	–5.029	9.140	–3.983	7.967	–3.135	6.977
	0.90	–4.786	7.966	–3.845	6.984	–3.080	6.161
	0.70	–5.343	10.685	–4.167	9.264	–3.204	8.027
100	0.80	–5.098	9.266	–4.020	8.041	–3.156	7.023
	0.90	–4.829	8.037	–3.870	7.029	–3.095	6.191
	0.70	–5.716	11.432	–4.346	9.661	–3.296	8.252
150	0.80	–5.316	9.662	–4.136	8.273	–3.221	7.166
	0.90	–4.960	8.258	–3.947	7.170	–3.141	6.283
	0.70	–5.917	11.834	–4.440	9.869	–3.343	8.369
200	0.80	–5.430	9.870	–4.917	8.393	–3.255	7.240
	0.90	–5.029	8.374	–3.987	7.243	–3.165	6.331

current values of sensitivity, specificity and the sample size.

Step 2. Write down all possible combinations of S_{eL}, S_{eU}, S_{pL}, S_{pU}, N_L, and N_U and corresponding b_i (i = 0,1) values from Table 1. For each combination find the product of the four terms as defined below and take the sum of these products:

n	S_e	S_p	b_i	Product
n_L	S_{eL}	S_{pL}	$b_i(n_L,S_{eL},S_{pL})$	$b_i(n_L,S_{eL},S_{pL}).(n_U-n_C).(S_{eU}-S_{eC}).(S_{pU}-S_{pC})$
n_L	S_{eL}	S_{pU}	$b_i(n_L,S_{eL},S_{pU})$	$b_i(n_L,S_{eL},S_{pU}).(n_U-n_C).(S_{eU}-S_{eC}).(S_{pC}-S_{pL})$
n_L	S_{eU}	S_{pL}	$b_i(n_L,S_{eU},S_{pL})$	$b_i(n_L,S_{eU},S_{pL}).(n_U-n_C).(S_{eC}-S_{eL}).(S_{pU}-S_{pC})$
n_L	S_{eU}	S_{pU}	$b_i(n_L,S_{eU},S_{pU})$	$b_i(n_L,S_{eU},S_{pU}).(n_U-n_C).(S_{eC}-S_{eL}).(S_{pC}-S_{pL})$
n_U	S_{eL}	S_{pL}	$b_i(n_U,S_{eL},S_{pL})$	$b_i(n_U,S_{eL},S_{pL}).(n_C-n_L).(S_{eU}-S_{eC}).(S_{pU}-S_{pC})$
n_U	S_{eL}	S_{pU}	$b_i(n_U,S_{eL},S_{pU})$	$b_i(n_U,S_{eL},S_{pU}).(n_C-n_L).(S_{eU}-S_{eC}).(S_{pC}-S_{pL})$
n_U	S_{eU}	S_{pL}	$b_i(n_U,S_{eU},S_{pL})$	$b_i(n_U,S_{eU},S_{pL}).(n_C-n_L).(S_{eC}-S_{eL}).(S_{pU}-S_{pC})$
n_U	S_{eU}	S_{pU}	$b_i(n_U,S_{eU},S_{pU})$	$b_i(n_U,S_{eU},S_{pU}).(n_C-n_L).(S_{eC}-S_{eL}).(S_{pC}-S_{pL})$

SP = Sum of products

Step 3. Calculate the value of b_i.

$$b_i = \frac{SP}{(n_U - n_L) \,.\, (S_{eU} - S_{eL}) \,.\, (S_{pU} - S_{pL})} \qquad (6)$$

Step 4. Substitute the values of b_0 and b_1 in (5) to obtain the approximate Bayesian estimate.

Example 2: Suppose that a test which is 0.88 sensitive and 0.76 specific is used to screen a group of 92 subjects for the presence of a certain disease. Assume that 23 subjects were positive in the test (P = 0.25). The sensitivity of the current test is 0.88 (S_{eC} = 0.88) which lies between the two tabulated values 0.80 (S_{eL}) and 0.90 (S_{eU}). Similarly S_{pC} = 0.76, S_{pL} = 0.70 and S_{pU} = 0.80; n_C = 92, n_L = 90 and n_U = 100. To find b_0, the sum of products can be obtained as follows:

n	S_e	S_p	b_0	Product
90	0.80	0.70	–5.029	–5.029 x 8 x 0.02 x 0.04 = –0.0322
90	0.80	0.80	–3.983	–3.983 x 8 x 0.02 x 0.06 = –0.0382
90	0.90	0.70	–4.786	–4.786 x 8 x 0.08 x 0.04 = –0.1225
90	0.90	0.80	–3.845	–3.845 x 8 x 0.08 x 0.06 = –0.1477
100	0.80	0.70	–5.098	–5.098 x 2 x 0.02 x 0.04 = –0.0082
100	0.80	0.80	–4.020	–4.020 x 2 x 0.02 x 0.06 = –0.0097
100	0.90	0.70	–4.829	–4.829 x 2 x 0.08 x 0.04 = –0.0309
100	0.90	0.80	–3.870	–3.870 x 2 x 0.08 x 0.06 = –0.0372

SP = –0.4266

$$b_0 = \frac{-0.4266}{10 \text{ x } 0.1 \text{ x } 0.1} = -4.266$$

To find b_1, replace b_0 values by b_1, in the above table.

n	S_e	S_p	b_1	Product
90	0.80	0.70	9.140	9.140 x 8 x 0.02 x 0.04 = 0.0585
90	0.80	0.80	7.967	7.967 x 8 x 0.02 x 0.06 = 0.0765
90	0.90	0.70	7.966	7.966 x 8 x 0.08 x 0.04 = 0.2039
90	0.90	0.80	6.984	6.984 x 8 x 0.08 x 0.06 = 0.2682
100	0.80	0.70	9.266	9.266 x 2 x 0.02 x 0.04 = 0.0148
100	0.80	0.80	8.041	8.041 x 2 x 0.02 x 0.06 = 0.0193
100	0.90	0.70	8.037	8.037 x 2 x 0.08 x 0.04 = 0.0514
100	0.90	9.80	7.029	7.029 x 2 x 0.08 x 0.06 = 0.0675

SP = 0.7601

$$b_1 = \frac{0.7601}{10 \text{ x } 0.1 \text{ x } 0.1} = 7.601$$

From (5) the approximate Bayesian estimate is:

$$\pi_a = \frac{1}{1 + e^{-(-4.266+7.601 \text{x} 0.25)}} = 0.086$$

B) If S_e and/or S_p > 0.90; 20 ≤ n ≤ 200, use 0.90 for interpolation.

C) Cases where S_e and/or S_p < 0.70 are not very common in practice and will be ignored in this article.

D) If n > 200 use π'' as defined in (1) to find the estimate of the population prevalence.

Discussion

If the sensitivity and the specificity of a test which is used for screening individuals for a particular disease are less than unity, the proportion of subjects screened positive by that test cannot be used as an estimate of the true population prevalence. Estimation procedures with such tests are either complicated or yield undesirable results. The estimation procedure proposed here requires simple calculations. As well as being simple, it produces results that always lie between 0 and 1, and are very close to those obtained by Bayesian techniques. Users who are not familiar with complicated mathematical calculations, can use this method and obtain the required estimates to a high degree of accuracy.

References

1. Gart, J.J. and Buck, A.A. Comparison of a screening test and a reference test in epidemiologic studies. II. A probabilistic model for the comparison of diagnostic tests. Am J Epidemiol 1966; 83: 593–602.
2. Levy, P.S. and Kass, E. H. A three–population model for sequential screening for bacteriuria. Am J Epidemiol 1970; 91: 148–154.
3. Rogan, W.J. and Gladen, B. Estimating prevalence from the results of a screening test. Am J Epidemiol 1978; 107: 71–76.
4. Gastwirth, J.L. The statistical precision of medical screening procedures: Application to polygraph and AIDS antibodies test data. Statistical Science 1987; 2: 213–238.
5. Lew, R.A. and Levy, P.S. Estimation of prevalence on the basis of screening tests. Statistics in Medicine 1989; 8: 1225–1230.