STATISTICS IN MEDICINE
Statist. Med. 2005; **24**:2669–2679
Published online 24 June 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/sim.2131

The probability of failing in detecting an infectious disease at entry points into a country

M. Dell'Omodarme[1] and M. C. Prati[1,2,*,†]

[1]*Scuola Normale Superiore di Pisa, Pisa, Italy*
[2]*INFN, Sezione di Pisa, Italy*

SUMMARY

In a group of N individuals, carrying an infection with prevalence π, the exact probability P of failing in detecting the infection is evaluated when a diagnostic test of sensitivity s and specificity s' is carried out on a sample of n individuals extracted without replacement from the group. Furthermore, the minimal number of individuals that must be tested if the probability P has to be lower than a fixed value is determined as a function of π. If all n tests result negative, confidence intervals for π are given both in the frequentistic and Bayesian approach. These results are applied to recent data for severe acute respiratory syndrome (SARS). The conclusion is that entry screening with a diagnostic test is rarely an efficacious tool for preventing importation of a disease into a country. Copyright © 2005 John Wiley & Sons, Ltd.

KEY WORDS: diagnostic test; SARS; detection failure; prevalence estimation; Bayesian approach

INTRODUCTION

The increased mobility of people causes frequent importation of diseases. Some of them had disappeared many years ago from Western countries and have new outbreaks, some has been recently recognized, like acquired immune deficiency syndrome (AIDS), bovine spongiform encephalopathy (BSE) and severe acute respiratory syndrome (SARS).

A possible method to prevent such importation is to check at entry points into a country passengers coming from zones where a dangerous disease is known to be present. This practice is not always useful because of the ways in which the disease is transmitted (e.g. in the case of AIDS) or it is realizable but very expensive. Sometimes, in the case of outbreaks of a new disease, diagnostic tests are at first not available. Furthermore, even if a diagnostic test is available, there is a positive probability of not detecting the infection because the diagnostic

*Correspondence to: M. C. Prati, Classe di Scienze, Scuola Normale Superiore di Pisa, Piazza dei Cavalieri 7, I-56100 Pisa, Italy.
†E-mail: mc.prati@sns.it

Received 16 December 2003
Accepted 27 October 2004

test is not perfect. Screening tests' results and estimation of disease prevalence have arisen considerable interest in literature. Among other studies one can remember [1–5].

In this paper one considers a finite population of individuals coming from a risk zone where a disease is present with prevalence π, as could be the N passengers of a train, ship or airplane. At the border a number $n \leqslant N$ of them undergo a diagnostic test with sensitivity s and specificity s'. The exact expression of the probability P of failing in detecting the infection (as a function of s, s', n, N and π) is calculated. The limit case of an infinite population is treated. The interesting inverse problem of computing the minimum number of individuals which must be tested if the probability P has to be lower than a certain value, is also studied.

If the n results of the diagnostic tests are all negative, confidence intervals for the prevalence of the disease in the population are given, both in the frequentistic and Bayesian approach. The Bayesian credible interval is especially interesting, because it takes into account prior information of epidemiological and medical character.

The mathematical results of the present study are applied to recent data (October 2003) on a diagnostic test for SARS [6]. The efficacy of border screening for SARS has been evaluated from a clinical point of view in References [7–9].

THE PROBABILITY OF FAILING TO DETECT THE INFECTION

Let N be the size of a population and R the number of subjects carrying a latent infection. We introduce the following notation:

$$I = \text{the individual is infected}, \quad \bar{I} = \text{the individual is not infected}$$

$$\pi = \frac{R}{N} = \text{the prevalence of the infection}$$

A diagnostic test would be carried out on each individual of a sample of size n extracted without replacement from the population. The possible results of the test are

$$T_+ = \text{the test is positive}, \quad T_- = \text{the test is negative}$$

A diagnostic test is perfect only in an ideal case, otherwise there are *false negative* and *false positive* results. The following probabilities are of interest:

$$P(T_+|I) = s, \quad P(T_-|I) = 1 - s$$

$$P(T_+|\bar{I}) = 1 - s', \quad P(T_-|\bar{I}) = s'$$

s is the *sensitivity* of the test, i.e. the probability of a positive test result when the individual is indeed infected (true positive). $1 - s$ is the probability of a false negative result. s' is the *specificity* of the test, i.e. the probability of a true negative result. $1 - s'$ is the probability of a false positive result. One considers also the predictive value of the positive test (PPV) and of the negative test (NPV):

$$\text{PPV} = P(I|T_+), \quad \text{NPV} = P(\bar{I}|T_-)$$

According to Bayes' theorem they can be expressed in terms of the prevalence π, the sensitivity and the specificity as follows:

$$\mathrm{PPV} = \frac{\pi s}{\pi s + (1-\pi)(1-s')}$$

$$\mathrm{NPV} = \frac{(1-\pi)s'}{(1-\pi)s' + \pi\,(1-s)}$$

If in the sample of size n there are k infected individuals ($k = 0, 1, \ldots, m$ where $m = \min(n, R)$), the probability of extracting such a sample without replacement is the hypergeometric probability:

$$h(k, n; R, N) = \frac{\binom{R}{k}\binom{N-R}{n-k}}{\binom{N}{n}}$$

Under the assumption that the results of the tests are independent, the probability that the n diagnostic tests are all negative, i.e. that the infection is not detected, is given by

$$P_n(s, s', R, N) = \sum_{k=0}^{m} h(k, n; R, N)(1-s)^k s'^{n-k} \tag{1}$$

This formula can be generalized in order to find the probability that from the n tests some are positive (see Reference [10]).

In the special case of a *perfect* test ($s = s' = 1$) Equation (1) reduces to the probability that the sample does not contain any infected individual:

$$P_n(1, 1, R, N) = \frac{\binom{N-R}{n}}{\binom{N}{n}}$$

Carrying out the sum in Equation (1) one obtains:

$$P_n(s, s', R, N) = \frac{\binom{N-R}{n}}{\binom{N}{n}} s'^n \; {}_2F_1\left(-n, -R; N-R-n+1; \frac{1-s}{s'}\right) \tag{2}$$

${}_2F_1(\alpha, \beta; \gamma; z)$ is the hypergeometric series, the properties of which are well known and tabulated [11]:

$${}_2F_1(\alpha, \beta; \gamma; z) = 1 + \frac{\alpha\beta}{\gamma \cdot 1} z + \frac{\alpha(\alpha+1)\beta(\beta+1)}{\gamma(\gamma+1)1 \cdot 2} z^2 + \cdots$$

The series terminates if either α or β or both are equal to zero or to a negative integer, as in Equation (2) where the number of terms is $m + 1$.

In the limit $N, R \to \infty$ with fixed finite prevalence π, Equation (2) reduces to the binomial limit:

$$P_n(s, s', \pi) = (\pi(1-s) + (1-\pi)s')^n \tag{3}$$

This result can be obtained either from Equation (1) by some algebra or from the following argument. The result of a test can be negative for two reasons: either the individual is infected but the test shows wrong (with probability $\pi(1-s)$) or the individual is not infected and the

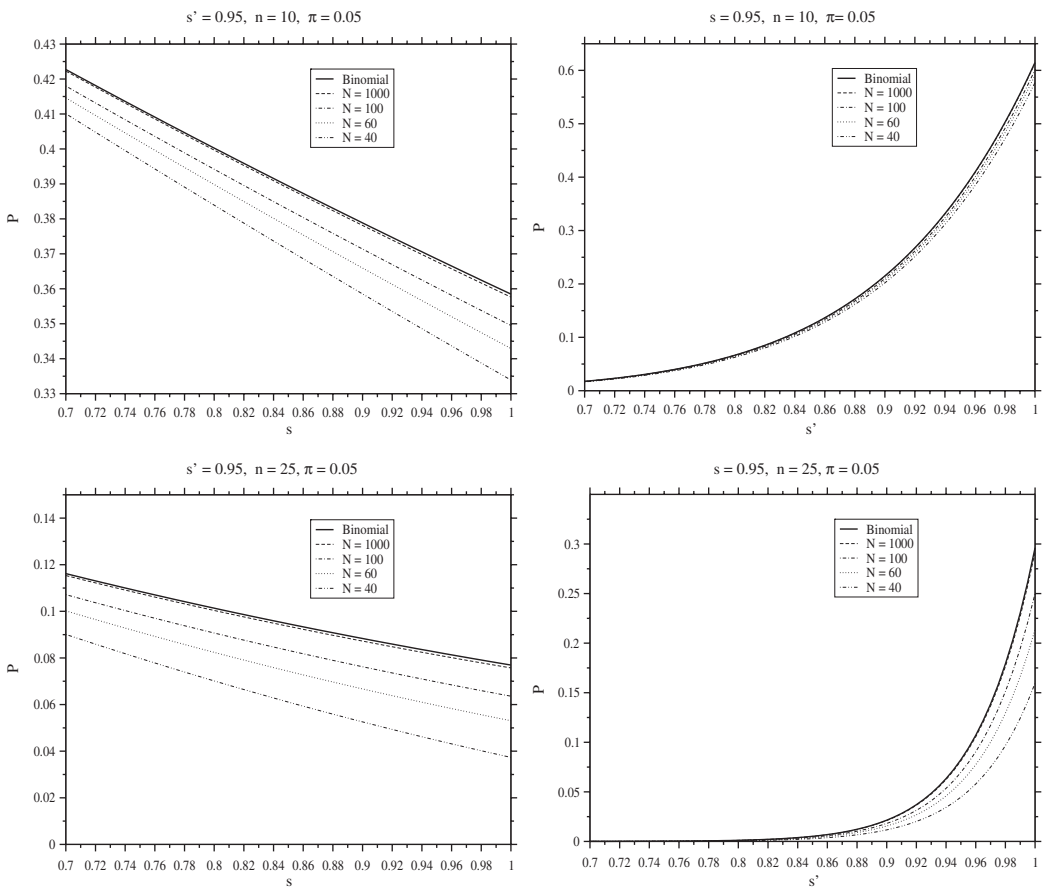

Figure 1. The probability P of not detecting the infection for various sizes N of the population (with fixed prevalence $\pi = 0.05$ and sample size $n = 10,\ 25$) as a function of: (left) the sensitivity s with fixed specificity $s' = 0.95$; (right) the specificity s' with fixed sensitivity $s = 0.95$.

test gives the right result (with probability $(1-\pi)s'$). The fact that the n tests are independent leads directly to Equation (3).

In Equation (3) the probability of missing an infection in a population is computed extracting the sample with replacement (Bernoullian extraction). Carrying out a diagnostic test extraction without replacement is more appropriate. In the limit of an infinite population extraction without replacement is in practice equivalent to the bernoullian one, but for a finite population important differences can arise.

In Figure 1, the exact results obtained from Equation (2) are compared with the binomial limit (3). One can see that, using the binomial limit, the probability P is overestimated and the error is large when the size N of the population is small. The probability P for various sizes N of the population (with fixed prevalence $\pi = 0.05$ and sample size $n = 10, 25$) is plotted as a function of the sensitivity s with fixed specificity $s' = 0.95$ (left side's diagrams) or, vice versa, as a function of s' with fixed $s = 0.95$ (right). From the plots on the left side of the

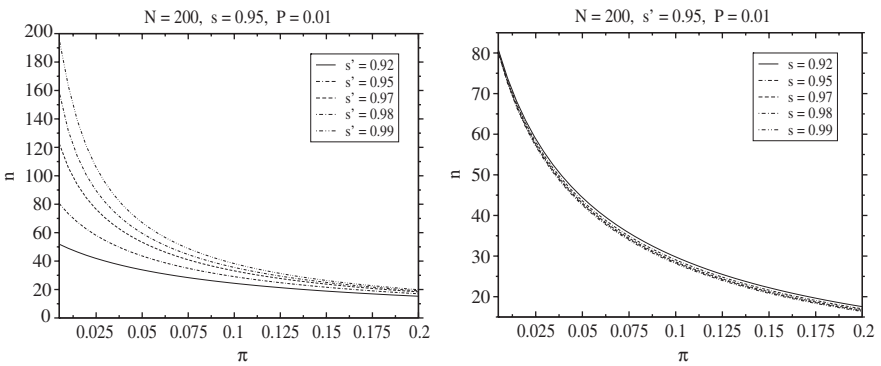

Figure 2. For fixed $N = 200$, the size n of the sample which must be tested in order to have a probability $P = 0.01$ of not detecting the infection, is given in dependence on the prevalence π. The five curves correspond to different values of: (left) the specificity s' with fixed $s = 0.95$; (right) the sensitivity s with fixed $s' = 0.95$.

figure, one can see that P is higher for small values of s, because the probability of false negative results decreases with s. On the right side of the figure one can see that P decreases with s' since the probability of false positive results increases as s' decreases. In the case of infectious diseases a high sensitivity is advisable even at the expense of a lower specificity. However, if the prevalence π is low, the positive predictive value of the test decreases quickly when the specificity gets lower, so that testing is not effective for a clinical diagnosis.

Equation (2) can also be used to solve the inverse problem of computing the size n of the sample needed to have a fixed probability P of not detecting the infection as a function of the unknown prevalence π, given the size N of the population, the sensitivity s and the specificity s' of the diagnostic test. Equation (2) cannot be solved analytically for n so that one has to employ computer facilities.

In Figure 2, for fixed population size $N = 200$ and required probability $P = 0.01$, the size n of the sample is determined as a function of the prevalence π. On the left side of the figure the curves are given for different values of the sensitivity with fixed specificity and vice versa on the right side. As expected from the analysis of Figure 1, the number n of subjects which must be tested in order to reach the required probability P increases with s' when s is fixed. The right side of Figure 2 shows that n does not change appreciably for sufficiently low prevalences in dependence on s, at least for values of s of practical use.

In Figure 3, the difference between the values of n obtained with the approximate binomial expression and the exact values given on left side of Figure 2 is shown as a function of π. In the binomial limit the value of n is always overestimated. The error is larger when the prevalence is low.

AN APPLICATION TO THE RAPID DIAGNOSIS OF SARS

SARS was recognized in China in November 2002, and the culprit agent, a novel strain of coronavirus (SARS-associated CoV), was identified in April 2003 [12, 13]. The 7th of

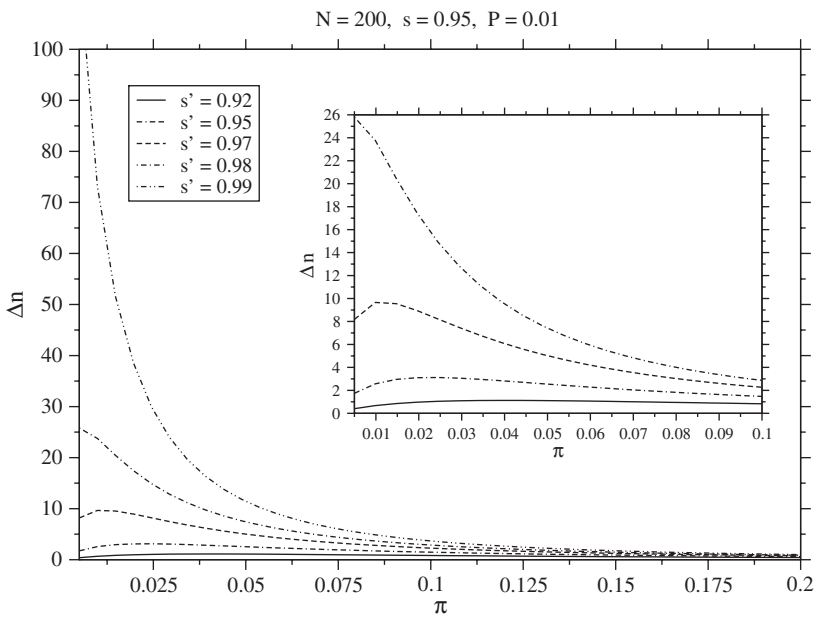

Figure 3. The difference between the exact results for n given in the left side of Figure 2 and the corresponding approximate results obtained in the binomial limit. The insert is a magnification of the low prevalence region of the whole picture.

August 2003, statistics of World Health Organization (WHO) reported that SARS had infected 8422 people and caused 916 deaths worldwide. In Hong Kong it infected 1755 people and killed 300.

Recently, reverse transcription-PCR protocols of two WHO SARS network laboratories were evaluated for the rapid diagnosis of the SARS-associated CoV in Hong Kong [6]. The resulting specificity of these PCR assays was 100 per cent, while for sensitivity the best results for the two laboratories were 71 and 79 per cent, respectively. The low sensitivity is related to the high mutation rate of the coronavirus, which makes difficult to identify its presence. It is a general problem which cannot be easily solved.

Let us suppose that an airplane, carrying $N = 200$ passengers, arrives from a region where SARS prevalence is estimated to be around 3 per cent. A diagnostic test with specificity $s' = 1$ and sensitivity $s = 0.75$ is available (the values are chosen in agreement with Reference [6]). From Figure 4 one can see that, if $n = 80$ passengers are tested, the probability P of missing SARS is estimated from 10 to 20 per cent. Vice versa, requiring P to be around 1 per cent the number n of passengers that must be tested is about 140. If P is required to be less than 0.001, the number of passengers tested must be greater than 180. If the airplane comes from a region where SARS prevalence is 1 per cent, even if all 200 passengers are tested the probability P of not detecting SARS amounts from 5 to 10 per cent. The high values come from the fact that the sensitivity of the diagnostic test is low and many dangerous false negative people cannot be prevented from diffusing the infection.

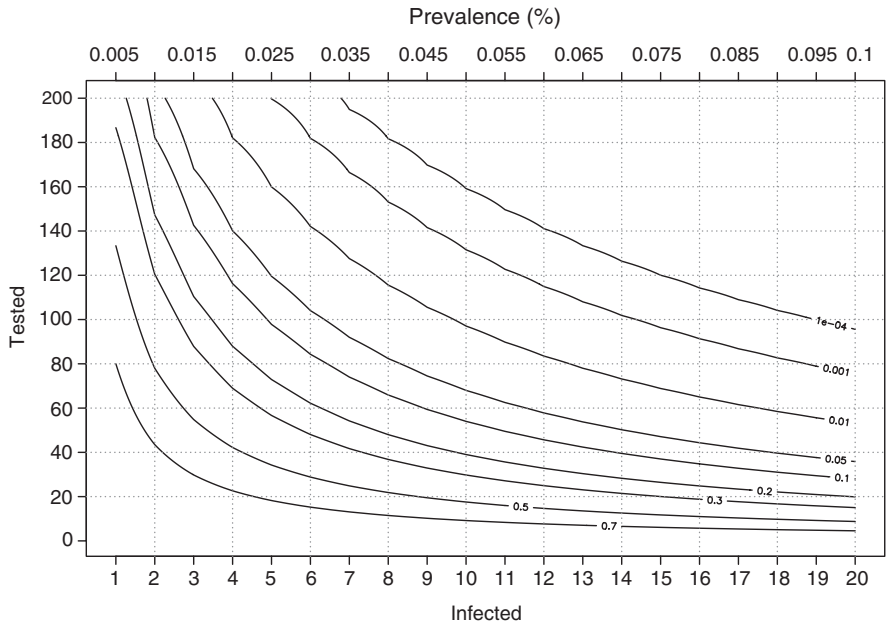

Figure 4. Contour plot of the probability P of not detecting the infection in a population of $N = 200$ individuals, as a function of the number of infected individuals (lower axis) or of the prevalence of the infection (upper axis) and of the size of the sample tested (vertical axis).

One could consider the possibility of testing pools of blood, instead of performing single diagnostic tests. In such a way costs could be saved. However the pooling procedure is not so appealing when the specificity of the diagnostic test is $s' = 1$ [5], as in our case, because in this situation the results of pool testing is always worse than the ones of individual testing. Furthermore, pool testing is very useful in assessing the prevalence of a disease in a population (when this prevalence is supposed to be low), but not for individual diagnosis, which is the target in the case of SARS.

CONFIDENCE INTERVAL FOR THE PREVALENCE: FREQUENTISTIC AND BAYESIAN APPROACH

Once the n diagnostic tests have been carried out and found all negative, the estimated prevalence is $\hat{\pi} = 0$. The confidence interval gives an idea of how reliable is this result. For a generic value of the prevalence π the expression of the confidence interval is given for example in Reference [14, p. 117]. This calculation is not appropriate when the proportion $\hat{\pi}$ is obtained as a result of a non-perfect diagnostic test. In this case the frequentistic confidence interval must be constructed with the correction given in Reference [5]. An alternative approach is the Bayesian credible interval, which allows to take into account prior information of epidemiological and medical type.

Table I. Comparison between the upper limits of the confidence intervals for the prevalence π and the credible intervals, obtained from the Gibbs sampler.

	CI for proportion			CI from Gibbs sampler	
n	95 per cent	99 per cent	median	95 per cent	99 per cent
20	0.225	0.310	0.014	0.070	0.099
40	0.117	0.165	0.011	0.055	0.077
60	0.079	0.113	0.087	0.045	0.064
80	0.060	0.085	0.007	0.038	0.055
100	0.048	0.069	0.006	0.033	0.049
120	0.040	0.058	0.006	0.029	0.042
140	0.035	0.050	0.005	0.026	0.037
160	0.030	0.043	0.004	0.023	0.034
180	0.027	0.039	0.004	0.022	0.032
200	0.024	0.035	0.004	0.020	0.028

Lower limits are 0 in all cases.

One should point out that the meaning of frequentistic and Bayesian intervals is different. A proportion lies actually in its 95 per cent Bayesian interval with probability 95 per cent, while from a frequentistic perspective one assumes that, constructing for many samples the 95 per cent confidence intervals, approximately 95 per cent of them will contain the proportion value.

In the Bayesian approach information available at the start of the study leads to specification of the prior distribution of the parameters. When data are collected and provide new information, Bayes' rule is used in order to compute the posterior distribution. Appropriate quantiles of the posterior probability distribution are used for inference. Direct calculation of posterior distributions can be difficult. The Gibbs sampler (see Appendix A), an iterative Markov chain Monte Carlo technique for approximating posterior densities, is widely used in medical literature.

In this section the 95 and 99 per cent credible intervals obtained from the posterior distribution of the prevalence π (given the informative prior distribution described below) and the frequentistic results for the confidence interval of π are compared.

Let us suppose that the prevalence of SARS in the zone from which air passengers come from is around 3 per cent, with a 95 per cent confidence interval from 0 to 10 per cent. The corresponding informative prior distribution is given by a beta density with $\alpha = 1$ and $\beta = 35$ (see Appendix A). At the airport n passengers are tested and all tests give a negative result. Among them there is a unknown number k of false negatives. Given these data, it is possible to obtain the posterior distribution of the prevalence using the Gibbs sampler algorithm (Appendix A).

In Table I the 95 and 99 per cent credible intervals for π are listed for some values of n, along with the corresponding frequentistic confidence intervals, which coincide with the Bayesian credible intervals computed assuming a non-informative (uniform) prior distribution. As expected, informative credible intervals are narrower. The difference decreases as the number of tests carried out increases, because prior information gets superseded by the likelihood of data.

From this table, one can see that testing $n = 80$ passengers who turn out to be all negative, the prevalence π of SARS is in the credible interval $[0, 0.038]$ with 95 per cent probability. The corresponding confidence interval is $[0, 0.060]$. The frequentistic estimate of the prevalence is 0, while the Bayesian estimate is 0.007, the median of the posterior distribution. In the case when all 200 passengers are tested and found negative, the 95 per cent credible interval is $[0, 0.020]$. The Bayesian estimate of the prevalence is 0.004.

DISCUSSION

Border screening against emerging infectious diseases would be a desirable disease control measure, with a privileged role in preserving public confidence and limiting bad economic consequences. However, before organizing a screening protocol, one should evaluate the possible impact of such a screening on international traffic and trade, the cost of the procedure in terms of personnel and logistics needed, and its real effectiveness.

In the case of the SARS outbreak in 2003, a screening programme was organized at the border entry of several countries, such as Canada, New Zealand, Hong Kong, Australia and Italy [7–9]. Symptomatic passengers, coming from SARS-affected areas (i.e. Vietnam, Taiwan, Singapore, Hong Kong, China, Canada and the Philippines), were sent to a quarantine team and, after further investigations, eventually assessed at hospitals. The efficacy of this programme turned out to be low and the sensitivity and specificity of the testing procedure was not easily assessable.

In this paper the probability of missing an infection was evaluated in the hypothesis that a diagnostic test with known sensitivity and specificity could be used directly on the airport on a random sample of passengers (including pre-symptomatic ones). This probability was found to be high if the prevalence of the disease was low and the diagnostic test used had a sensitivity different from one.

Therefore, before planning an expensive screening program at entry points of a country, one should first of all have a highly sensitive screening test at one's disposal. This is not always the case when a disease outbreaks. Furthermore, the procedure of testing should be the less invasive as possible, in order to prevent a dramatic decrease of tourism with heavy economic consequences, not justified by the real impact of a low prevalence disease. It appears that, in most cases, border entry screening is more effective in keeping public concern low than in stopping the infection to enter a country. A similar conclusion can be found in References [8, 9].

APPENDIX A: THE GIBBS SAMPLER

The Gibbs sampler is an important tool in the Bayesian approach to compute posterior distributions [15]. When the posterior distribution of a proportion π is sought, a beta density is usually taken as prior distribution:

$$f(\pi, \alpha, \beta) = \frac{1}{B(\alpha, \beta)} \pi^{\alpha-1} (1-\pi)^{\beta-1}, \quad 0 \leqslant \pi \leqslant 1,\ \alpha, \beta > 0$$

where $B(\alpha, \beta)$ is the beta function evaluated at (α, β). The result for the posterior distribution is again a beta density with different parameters. The non-informative uniform prior distribution is a particular case with $\alpha = \beta = 1$. If the prior distribution is supposed to be informative, and an expected value for π is given with its 95 per cent confidence interval, the parameters α and β are chosen in such a way that the mean value of the beta distribution equals the expected value of π and the confidence intervals match. For example the case considered in the text, where the expected value of π was 0.03 (95 per cent confidence interval from 0 to 0.10), can be well reproduced with parameters $\alpha = 1$, $\beta = 35$.

Let k be the number of false negatives among the n individuals in the sample. The conditional distributions of k and π, given the values of all other parameters, can be specified as follows [16]:

$$k|n, \pi, s, s' \sim \text{Binomial}\left(n, \frac{\pi(1-s)}{\pi(1-s) + (1-\pi)s'}\right)$$
$$\pi|n, k, \alpha, \beta \sim \text{Beta}(k + \alpha, n - k + \beta)$$

An arbitrary starting value is chosen for π. Then a point is drawn from the conditional distribution of k. This value is used in the conditional distribution of π, from which another value is drawn. The cycle is repeated a large number of times (in our case 20 500), so that the random samples generated for each parameter can be regarded as random samples from their correct unknown marginal distribution [15, 16]. The first 500 points are discarded from the samples, since their aim is to assess convergence.

The Monte Carlo code was run using R version 1.9.1 [17].

ACKNOWLEDGEMENTS

The authors thank the anonymous referees for their useful comments and suggestions, that have contributed to improve this paper.

REFERENCES

1. Gart JJ, Buck AA. Comparison of a screening test and a reference test in epidemiologic studies. *American Journal of Epidemiology* 1966; **85**(3):593–602.
2. Rogan WJ, Gladen B. Estimating prevalence from the results of a screening test. *American Journal of Epidemiology* 1978; **107**(1):71–76.
3. Quade D, Lachenbruch PA, Whaley FS, McClish DK, Haley RW. Effects of misclassifications on statistical inferences in epidemiology. *American Journal of Epidemiology* 1980; **111**(5):503–515.
4. Walter SD, Irwing LM. Estimation of test error rates, disease prevalence and relative risk from misclassified data: a review. *Journal of Clinical Epidemiology* 1988; **41**(9):923–937.
5. Tu XM, Litvak E, Pagano M. On the informativeness and accuracy of pooled testing in estimating prevalence of a rare disease: application to HIV screening. *Biometrika* 1995; **82**:287–297.
6. Yam WC, Chan KH, Poon LL, Guan Y, Yuen KY, Seto WH, Peiris JS. Evaluation of reverse transcription-PRC assays for rapid diagnosis of severe acute respiratory syndrome associated with a novel coronavirus. *Journal of Clinical Microbiology* 2003; **41**(10):4521–4524.
7. Wilder-Smith A, Paton NI, Goh KT. Experience of severe acute respiratory syndrome in Singapore: importation of cases, and defense strategies at the airport. *Journal of Travel Medicine* 2003; **10**(5):259–262.
8. Samaan G, Patel M, Spencer J, Roberts L. Border screening for SARS in Australia: what has been learnt? *The Medical Journal of Australia* 2004; **180**:220–223.
9. Petrosillo N, Puro V, Ippolito G. Border screening for SARS. *The Medical Journal of Australia* 2004; **180**:597.
10. Cameron AR, Baldock FC. A new probability formula for surveys to substantiate freedom from disease. *Preventive Veterinary Medicine* 1998; **34**:1–17.

11. Gradshteyn IS, Ryzhik IM. *Table of Integrals, Series, and Products*. Academic Press: New York, 1980.
12. Drosten C, Gunther S, Preiser W, van der Werf S, Brodt HR, Becker S *et al.* Identification of a novel coronavirus in patients with severe acute respiratory syndrome. *New England Journal of Medicine* 2003; **348**:1967–1976.
13. Ksiazek TG, Erdman D, Goldsmith CS, Zaki SR, Peret T, Emery S *et al.* A novel coronavirus associated with severe acute respiratory syndrome. *New England Journal of Medicine* 2003; **348**:1953–1966.
14. Armitage P, Berry G, Matthews JNS. *Statistical Methods in Medical Research* (4th edn). Blackwell Publishing: Oxford, 2002.
15. Gelfand AE, Smith AFM. Sampling-based approaches to calculating marginal densities. *Journal of American Statistical Association* 1990; **85**:398–409.
16. Joseph L, Gyorkos TW, Coupal L. Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard. *American Journal of Epidemiology* 1995; **141**(3):263–272.
17. R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing Vienna, Austria, 2003.

