Screening programmes and the evaluation of screening tests using Stata and R

B V Girdler-Brown,[1] FFPH (UK), FCPHM (SA), MMed; **R R Bastos,**[2] PhD; **L N Dzikiti,**[1] MSc

[1] *School of Health Systems and Public Health, Faculty of Health Sciences, University of Pretoria, South Africa*
[2] *Department of Statistics, Instituto de Ciências Exatas, Universidade Federal de Juiz de Fora, São Pedro, Brazil*

Corresponding author: *B V Girdler-Brown (Brendan.girdler-brown@up.ac.za)*

This article describes the commonly recognised criteria for establishing a mass screening programme. In addition, the article describes commonly used parameters for assessing the performance of public health screening tests. The calculation of the parameters is described. Stata commands, and R code, are also supplied to assist readers with the estimation of these parameters using Stata or R. This article does not include the evaluation of diagnostic tests.

South Afr J Pub Health 2019;3(3):49-55. https://doi.org/10.7196/SHS.2019.v3.i3.98

In this article we begin by describing the differences between screening tests and diagnostic tests. We then describe desirable features for a mass screening programme, and give a few examples of such programmes. After describing some challenges facing mass screening programmes, we explain how to estimate the commonly used performance parameters of screening tests, along with 95% confidence intervals. The Stata (StataCorp, USA) commands are given, followed by code that may be used if the estimations are performed using R (R Foundation, Austria).[1] A later article, still to be published, will extend the discussion from screening tests to diagnostic tests.

Screening v. diagnostic testing

Screening is said to be carried out when a test is administered on large groups of apparently healthy (i.e. asymptomatic and externally healthy) individuals in order to detect latent disease.[2] In contrast, diagnostic testing is used to describe the process in which a test is carried out for an individual with signs or symptoms of the disease being tested for. This article focuses on the evaluation of screening test parameters.

In either case (screening testing or diagnostic testing), the test is administered to help decide whether to offer further investigations or interventions to those undergoing the tests.

Criteria for a screening programme

Programmes for the screening of apparently healthy individuals may fall into three broad categories:

(*i*) mass screening programmes where testing is offered to large groups of apparently healthy people who are potentially at risk for a disease (e.g. mammography for breast cancer);

(*ii*) focused, or targeted, screening of subpopulations perceived to have an increased risk of disease (e.g. screening for the sickle cell gene among people of west African descent who may be considering marriage); and

(*iii*) opportunistic screening of symptomatic patients for unrelated conditions (such as routine screening for hypertension in someone who is seen in the clinic with a laceration for suturing): this may just be part of 'good medical practice', or it may be an extra routine investigation that is added, perhaps as part of an institutional policy or even a national programme.

The following general classic screening criteria, proposed by Wilson and Jungner,[2] have largely stood the test of time:

- The condition sought should be an important health problem.
- There should be an accepted treatment for patients with recognised disease.
- Facilities for diagnosis and treatment should be available.
- There should be a recognisable latent or early symptomatic stage.
- There should be a suitable test or examination.
- The test should be acceptable to the population.
- The natural history of the condition, including development from latent to declared disease, should be adequately understood.
- There should be an agreed policy on who to treat as patients.
- The cost of case-finding (including diagnosis and treatment of patients diagnosed) should be economically balanced in relation to possible expenditure on medical care as a whole.
- Case-finding should be a continuing process and not a 'once-and-for-all' project.

Andermann *et al.*,[3] who reviewed and reported on development of the Wilson and Jungner[2] criteria over the ensuing 40 years of practice, have added the following criteria to the original list:

- 'The screening programme should respond to a recognised need.
- The objectives of screening should be defined at the outset.

- There should be a defined target population.
- There should be scientific evidence of screening programme effectiveness.
- The programme should integrate education, testing, clinical services and programme management.
- There should be quality assurance, with mechanisms to minimise potential risks of screening.
- The programme should ensure informed choice, confidentiality and respect for autonomy.
- The programme should promote equity and access to screening for the entire target population.
- Programme evaluation should be planned from the outset.
- The overall benefits of screening should outweigh the harm.'

In addition, we suggest that the following points should also be considered when evaluating the desirability or practicability of a screening programme:

- The disease is potentially serious if not detected and treated early on, before symptoms are manifest.
- The disease is of fairly high prevalence and importance in the community.
- The disease may be cheaper and easier to treat if treatment is started early on.
- The prognosis may be improved by an earlier start to treatment.
- The disease should have a long latency period (otherwise screening would need to be carried out more frequently, and this may not be practicable/affordable).
- The screening test/tool should be non-invasive, culturally acceptable and should not cause great discomfort or inconvenience to the person screened.
- There should be an affordable cost per case detected.
- There should be a high ability to detect existing cases.
- There should be a high ability to avoid too many false positives.
- There should be a highly specific confirmatory (diagnostic) test available for those screening positive.
- There should be a successful, acceptable and affordable treatment option available for those confirmed as having the disease.
- There may be ethical and or legal/constitutional mandates that favour the introduction of a programme.

For infectious diseases, one should add that if early detection may lead to an intervention that would prevent new instances of transmission, then screening would be desirable. This is especially so if the disease in question is likely to have serious consequences for those infected, or for families and society more generally. Wilson and Jungner[4] place screening for infectious diseases as a priority in developing countries, where infectious diseases place a high burden on the population and on health services.

Single-stage v. two-stage testing algorithms

In a case where there is an inexpensive, non-invasive, diagnostic test available, that has a high sensitivity (detects a high proportion of cases) and a high specificity (non-cases are highly likely to test negative), the screening test may double as a diagnostic test. An example might be the screening of asymptomatic people to identify and treat pulmonary tuberculosis (TB) using the GeneXpert test.

However, many screening programmes rely on a strategy that involves two-stage testing. The first stage is carried out on the entire population using low-cost and non-invasive tests. These low-cost tests are preferably highly sensitive, but are often not as highly specific.

This high sensitivity/low specificity trade-off is commonly encountered when a test result is obtained on either a continuous or ordinal scale, and a certain cut-off point is used to identify those who are deemed to test positive v. negative. Shifting this cut-point to include more positives (i.e. to obtain higher sensitivity) means that there will also be more false positives, resulting in a lower specificity. This is a good example of the adage that we 'cannot have our cake and eat it'.

In addition, however, specificity may be further depressed as a result of nonspecific reactions that cause false positivity. For example, with the prostate-specific antigen (PSA) test for prostate cancer, nonspecific conditions might result in raised PSA levels. Causes, in the case of the PSA test, might include current urinary tract infection, recent ejaculation, benign prostatic hypertrophy and renal stones. This detracts from the specificity of the test, quite apart from the sensitivity/specificity trade-off (which may also apply when interpreting PSA results, incidentally, as a predetermined cut-off point is used to classify results as either positive or negative).

Tests that are both highly sensitive and highly specific may be more expensive, may require higher levels of expertise to perform and may also be less pleasant for the patient. Higher cost is always an issue when screening whole populations. Since screening is aimed at apparently healthy people, tests that are unpleasant might discourage enthusiastic participation in the screening programme.

As a result, confirmatory tests are often reserved for those testing positive on the screening test. This subgroup of people will have a higher prevalence of the disease, so that the positive predictive value (the proportion of those testing positive who actually have the disease) will be higher.

Other two-test algorithms used in screening and/or diagnostic testing

Two or more tests performed in series on those being screened

In this situation a first, more sensitive test is followed (in those testing positive on the first test) by a second, different kind of test. This second test may be diagnostic. However, it is not uncommon for the second test to be used to narrow down the group of people requiring a diagnostic workup. For example, Australian researchers used a questionnaire with relatively high sensitivity (0.85) to screen a population for diabetes. Those testing positive on the questionnaire were then asked for a fasting blood specimen, and fasting plasma glucose levels were measured. Those with a value >5.4 mmol/L were then referred for a more specific workup at a diagnostic centre.[4]

This 'serial' approach results in a lower sensitivity overall (than if the first test, the questionnaire in this case, were to have been used for the referral decision). However, it increases the specificity, so that fewer workups are performed on people who had false positive responses to the questionnaire. Such an approach might be considered if the system is unable to cope with large numbers of false positive people requiring more careful diagnostic workup.

Two or more tests carried out at the same time on those being screened

In this algorithm, two or more tests are performed on each individual, and he/she is considered 'positive' if at least one of the test results is positive. Sometimes this overall result is considered diagnostic; in other cases, if both/all the screening tests are of low specificity, a more careful diagnostic appraisal is needed on those with an overall positive result.

An example might be the use of two different tests for TB in an area where sputum microscopy services are not widely accessible (but radiography is). One might then use both a chest radiograph and a symptoms questionnaire as the two screening 'tests'. Any individual testing positive with either of these tests would be referred for two separate sputum smear examinations elsewhere.

This approach would result in increased sensitivity (but lower specificity) at the screening stage. As a result, there would be an increased number of false positive individuals referred onwards for further testing, but fewer infectious cases missed.

Limitations of screening programmes

Screening programmes may fail to live up to expectations for a number of reasons, including:

- ill-advised introduction of a poorly thought-through/under-resourced programme
- introduction of a programme that is not evidence-based
- weak administration and poor follow-up rates of those testing positive in the screening stage
- lack of capacity to perform confirmatory investigations
- lack of capacity to treat/intervene for all the cases identified
- poorest rates of access or uptake for those most at risk
- less likelihood of receiving results for those most at risk, which may be a problem if testing positive
- diversion of scarce screening resources from those who need them most to those who need them least, if those who need them least are more likely to access and use screening
- high false negative rates (poor sensitivity) that may give persons testing negative false reassurance
- high false positive rates (lack of specificity) that may cause undue alarm in those testing false positive
- over-enthusiastic marketing of tests that have poor performance parameters, or that may not be cost-effective.

Estimating the parameters of both screening and diagnostic tests

We start by defining the sensitivity and specificity of a test. The sensitivity of a test is the proportion of those who have the disease that will test positive for the disease. The specificity of a test is the proportion of those who are disease-free that will test negative.

The performance parameters of a screening test may be difficult to estimate from cross-sectional studies involving the target population unless suitably large samples are studied. In addition, ideally, one would need to perform highly specific and very sensitive tests (possibly expensive and invasive) at the same time as the screening test on every person included in the study sample. This approach may not be practicable or affordable. Known cases should be excluded from the calculations of the performance parameters, since there would be no point in screening people known to have the condition. Inclusion of known cases may result in higher sensitivity estimates that would not be reproducible under field conditions.

More commonly, the parameters of a test are estimated by studying the test performance in two groups of study participants, using a case-control approach. The one group consists of persons known to definitely have the condition, and the other ('controls') is a group known to definitely be free of the condition ('perfectly healthy'). The groups are usually of equal numbers of participants, leading to a prevalence of the condition among those being tested of 50%. Such studies are more affordable and less time consuming than large cross-sectional studies. Unfortunately, this approach may lead to parameter estimates that are not reproducible under mass screening efforts.

For example, the sensitivity of a new screening test for pulmonary TB may appear to be higher when administered to a group of smear-positive TB patients than when administered to a mixed group of smear-positive and smear-negative patients.

The performance characteristics of a test being evaluated as a screening test, using a cross-sectional study design, with known cases excluded from the study, should never be generalised to the performance of the test if used for diagnostic purposes.

Under diagnostic testing conditions, the patients are symptomatic, hence it is possible that the sensitivities estimated will be higher than would be the case in a screening environment. In addition, the prevalence of the condition among those being tested will likely differ from 50% and, as we will see later in this article, this will affect the positive and negative predictive values of the test.

While the controls in the screening test control group are 'perfectly healthy', this is not usually the case under diagnostic testing conditions. In fact, in clinical practice, all those undergoing a diagnostic test are symptomatic of some or other condition. Therefore, the controls for evaluation of a diagnostic test should not be 'perfectly healthy' people.

For screening tests, we are particularly interested in the sensitivity and specificity of the test. Predictive values (positive and negative) are also of some interest, although of greater interest to clinicians performing diagnostic tests (and in such a case should be determined using study samples that consist of typical controls found in a clinical setting).

The description that follows will present the calculations of both sensitivity and specificity, as well as of positive and negative predictive values. Discussions around likelihood ratios and also the use of receiver operator characteristic (ROC) curves (for test results that are continuous rather than binary) will be described in a later article focusing on diagnostic tests.

Sensitivity and specificity of a test

We begin by summarising, in Table 1, fictitious test results from a case-control study of PSA results (positive taken to be ≥4 mmol/L). These are shown in Table 1.

There were 100 cases of histologically proven cancer of the prostate and 400 healthy and asymptomatic controls with negative prostate biopsy results. Assume that PSA levels were measured before taking the biopsies (the biopsy procedure can, in itself, result in raised PSA levels).

- It can be seen in Table 1 that 85/100 cases tested positive. In other words, the PSA test only identified 85% of those with prostate cancer as having prostate cancer. This proportion, 0.85 (85%), is referred to as the sensitivity of the test. There were 15 false negative results, in the sense that 15 of those with cancer tested negative.
- In addition, it can be seen that of the 400 men who definitely did not have cancer of the prostate, only 120 tested negative with the PSA test. This proportion, the proportion of non-cases testing negative, is known as the specificity. In this case the specificity is 120/400 = 0.30 (30%). Furthermore, there were 280 false positives (those without cancer who, nevertheless, tested positive with the PSA test).

These estimates, of sensitivity and specificity, are based on sample data. Confidence intervals (CIs) should also be calculated in order to obtain 95% CIs for the population parameters.

For this study sample, and for those with the same lifestyle, risk factors and demographic characteristics, the sensitivity and specificity of the PSA test (using a cutoff value of 4 mmol/L) is reasonably assumed to be constant. If the sample size is increased, the estimates will become more precise, and sample estimates may vary slightly, but only as a result of sampling error.

However, it is not correct that, for a given test, sensitivity and specificity are constant across settings.[5] For example, some tests for TB may be highly specific and highly sensitive in an urbanised society with low environmental exposure to non-mycobacterial TB (NMT) organisms and low HIV prevalence. However, in a setting where many people are exposed to cross-reacting environmental NMT species, and where many people have untreated HIV infections, i.e. are not yet taking antiretrovirals, the same test may have lower specificity and lower sensitivity.

The criteria for selection of a good screening test

A good screening test should have a high sensitivity, so that:

- the effort of the screening programme is not wasted by missing too many cases; and
- individuals participating in screening are less likely to be misled by a negative result.

In addition, a screening test should have a reasonable specificity (not too low), so that:

- one does not then have to perform too many more expensive and more invasive tests to rule out the false positives; and
- individuals who participate in the screening are not misled by a false positive result (this may result in unnecessary anxiety while waiting for the results of the more specific confirmatory test).

Finally, the test parameters must be established in the target population before the screening programme adopts the test for use. Sensitivities and specificities obtained elsewhere are only reliable in the community in which they were obtained (or one that is very similar).

Positive and negative predictive values

The positive predictive value (PPV) of a test estimates the probability that a person who tests positive has the disease in question.

The negative predictive value (NPV) of a test estimates the probability that a person who tests negative does not have the disease in question.

These two concepts are of more interest than sensitivity and specificity for those performing diagnostic testing. They may not be estimated directly from a case-control study, unless the ratio of controls to cases has been engineered so that the prevalence of cases in the sample is the same as the prevalence of cases in the population. This is almost never the case.

However, if one has used a case-control study to estimate the sensitivity and specificity of a test, and if one has a large enough sample size to produce very precise estimates of the sensitivity and specificity parameters, then these sensitivity and specificity results may be applied to a hypothetical population with the same prevalence of disease found in the general population, and reasonable estimates may then be obtained for the PPV and NPV under field conditions.

The calculation of sensitivity, specificity, PPV and NPV is illustrated in Table 2 by a fictitious set of PSA results obtained from a large hypothetical cross-sectional study of men aged over 40 years, using a randomly selected sample.

From Table 2, we can see that:

- prevalence = 300/50 000 = 0.006 (or 0.6%)

Table 1. Cross-tabulation of hypothetical PSA results

		Prostate cancer biopsy result		Total
		Positive	Negative	
PSA result	Positive	85	280	365
	Negative	15	120	135
Total		100	400	500

PSA = prostate-specific antigen.

Table 2. Cross-tabulation of PSA results from a cross-sectional study

		Prostate cancer biopsy result		Total
		Positive	Negative	
PSA result	Positive	255	34 790	35 045
	Negative	45	14 910	14 955
Total		300	49 700	50 000

PSA = prostate-specific antigen.

- once again, sensitivity = 255/300 = 0.85
- specificity = 14 910/49 700 = 0.30.

These values for sensitivity and specificity do not change when prevalence is changed, provided that the study population is the same or similar. In addition, from Table 2, we can see that:

- PPV = 255/35 045 = 0.007 (or 0.7%)
- NPV = 14 910/14 955 = 0.997 (or 99.7%).

These results indicate that there are 34 790 false positives in a sample of 50 000. This means that these 34 790 men, none of whom has cancer of the prostate, would all need to undergo further evaluation (usually a prostate biopsy). The lesson is that for a mass screening programme, we should not ignore the specificity of a test completely. Low specificities mean low PPVs, and hence a lot of additional and fruitless follow-up and invasive testing, as well as avoidable anxiety for those who have screened positive.

Software commands and codes

Performing the analyses using Stata and R statistical software, the following Stata commands and R codes may be used.

Stata commands

The Stata commands assume the data layout and cross-tabulation shown in Tables 3 and 4.

Assuming the data format and notations shown in Tables 3 and 4, the Stata commands to be used are listed in Table 5.

All four of the parameter estimates described in Table 5 will be presented as the Stata output, alongside their binomial exact 95% CIs.

There is a simpler alternative available in Stata, where the data need only be entered as two variables, say, 'disease' and 'test', as illustrated by Table 6. This alternative also presents parameter estimates alongside their binomial exact 95% CIs. However, one needs to first install a user-written .ado file called 'diagt'. Once installed it remains installed, so this only has to be done once:

Table 3. Layout of data for Stata computations

disease_present	disease_absent	test_pos	test_neg
1	0	1	0
1	0	1	0
1	0	0	1
0	1	1	0
0	1	0	1
1	0	1	0
0	1	0	1
0	1	0	1
0	1	0	1
0	1	1	0
(etc.)	(etc.)	(etc.)	(etc.)

Table 4. A typical cross-tabulation of screening test results using an a, b, c, d notation

		Disease		Total
		Present	Absent	
Screening test result	Positive	a	b	a+b
	Negative	c	d	c+d
Total		a+c	b+d	

While online, type the following command into the Stata command bar:

<findit diagt>

Select 'sbe36_2' and click on 'install'. The .ado file and the help file will then be installed (this .ado file was written by Seed and Tobias.[6]).

Using the 'diagt' command, all you need is a single command:

<diagt disease test>

If your data are only available in table format, then you may use the following command (refer to Table 4 for the notation):

<diagti a c b d>

The order of entering the numbers is very important; use Table 4 as a guide for the correct order.

The Stata output is illustrated in Fig. 1.

Table 5. A list of Stata commands for estimating the screening test parameters using the a, b, c, d notation from Table 4

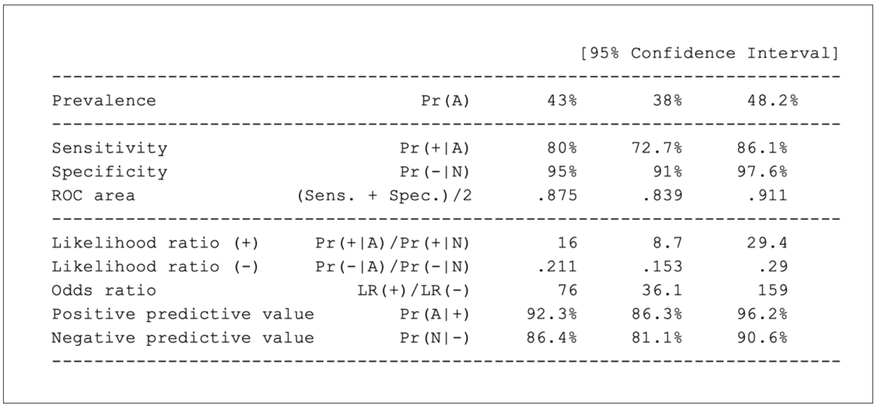
If data are entered in a, b, c, d format	If data are only available in a, b, c, d format
For estimating sensitivity: ci proportion test_pos if disease_present==1	For estimating sensitivity: cii proportion (a+c) a
For estimating specificity: ci proportion test_neg if disease_present==0	For estimating specificity: cii proportion (b+d) d
For estimating PPV: ci proportion disease_present if test_pos==1	For estimating PPV: cii proportion (a+b) a
For estimating NPV: ci proportion disease_absent if test_pos==0	For estimating NPV: cii proportion (c+d) d

Table 6. Alternative data layout if the 'diagt' command is to be used in Stata

disease	test
1	1
1	1
1	0
0	1
0	0
(etc.)	(etc.)

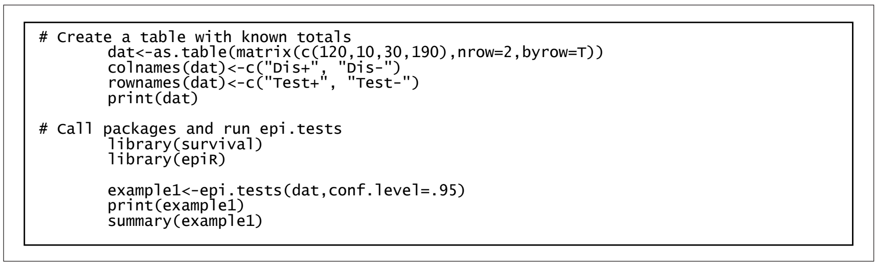
Table 7. A typical cross-tabulation of screening test results

		Disease		Total
		Present	Absent	
Screening test result	Positive	120	10	130
	Negative	30	190	220
Total		150	200	



```
                                                   [95% Confidence Interval]
---------------------------------------------------------------------------
Prevalence                          Pr(A)        43%        38%       48.2%
---------------------------------------------------------------------------
Sensitivity                       Pr(+|A)        80%      72.7%       86.1%
Specificity                       Pr(-|N)        95%        91%       97.6%
ROC area                    (Sens. + Spec.)/2   .875       .839        .911
---------------------------------------------------------------------------
Likelihood ratio (+)      Pr(+|A)/Pr(+|N)         16        8.7        29.4
Likelihood ratio (-)      Pr(-|A)/Pr(-|N)       .211       .153         .29
Odds ratio                        LR(+)/LR(-)     76       36.1         159
Positive predictive value         Pr(A|+)      92.3%      86.3%       96.2%
Negative predictive value         Pr(N|-)      86.4%      81.1%       90.6%
---------------------------------------------------------------------------
```

Fig. 1. Stata output following the use of the 'diagt' command.

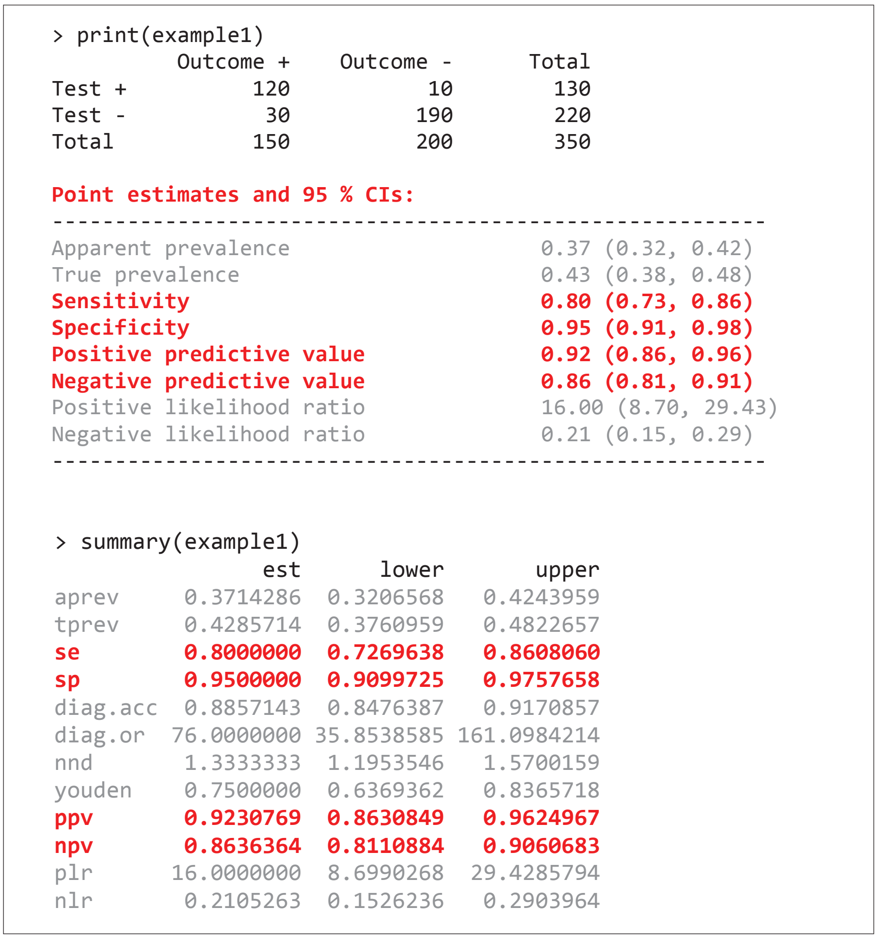


```
# Create a table with known totals
        dat<-as.table(matrix(c(120,10,30,190),nrow=2,byrow=T))
        colnames(dat)<-c("Dis+", "Dis-")
        rownames(dat)<-c("Test+", "Test-")
        print(dat)

# Call packages and run epi.tests
        library(survival)
        library(epiR)

        example1<-epi.tests(dat,conf.level=.95)
        print(example1)
        summary(example1)
```

Fig. 2. R code for the situation where cross-tabulation cell counts are known.



```
> print(example1)
          Outcome +    Outcome -      Total
Test +          120           10        130
Test -           30          190        220
Total           150          200        350

Point estimates and 95 % CIs:
---------------------------------------------------------
Apparent prevalence                     0.37 (0.32, 0.42)
True prevalence                         0.43 (0.38, 0.48)
Sensitivity                             0.80 (0.73, 0.86)
Specificity                             0.95 (0.91, 0.98)
Positive predictive value               0.92 (0.86, 0.96)
Negative predictive value               0.86 (0.81, 0.91)
Positive likelihood ratio               16.00 (8.70, 29.43)
Negative likelihood ratio               0.21 (0.15, 0.29)
---------------------------------------------------------


> summary(example1)
                est       lower       upper
aprev     0.3714286  0.3206568   0.4243959
tprev     0.4285714  0.3760959   0.4822657
se        0.8000000  0.7269638   0.8608060
sp        0.9500000  0.9099725   0.9757658
diag.acc  0.8857143  0.8476387   0.9170857
diag.or  76.0000000 35.8538585 161.0984214
nnd       1.3333333  1.1953546   1.5700159
youden    0.7500000  0.6369362   0.8365718
ppv       0.9230769  0.8630849   0.9624967
npv       0.8636364  0.8110884   0.9060683
plr      16.0000000  8.6990268  29.4285794
nlr       0.2105263  0.1526236   0.2903964
```

Figs 3a and b. R outputs following the (a) print and (b) summary commands.

Using R code

If you have no data set, but the totals are available in a cross-tabulation, as for example in Table 7, use the 'epi.tests' function from epiR package version 0.9-99 (2018-11-06).[7]

You must create an object in table format. As with Stata, the order of the numbers is very important, and you will be provided with binomial exact 95% CIs. Remember to use a text application to write the code, or else type the code directly into RStudio or R. R does not recognise 'smart' inverted commas, for example. The code is shown in Fig. 2.

The print and summary commands will give you print and summary outputs (Figs 3a and b; concentrate on the red values).

On the other hand, if you already have data entered at an individual level, as illustrated in Table 8, then use the same function epi.tests from package epiR, but make sure that you transform the data into a 2 × 2 table. The code is shown in Fig. 4.

If you have the whole data file, e.g. in csv format (mydata.csv), with header and semicolon delimeters, then use the code as shown in Fig. 5.

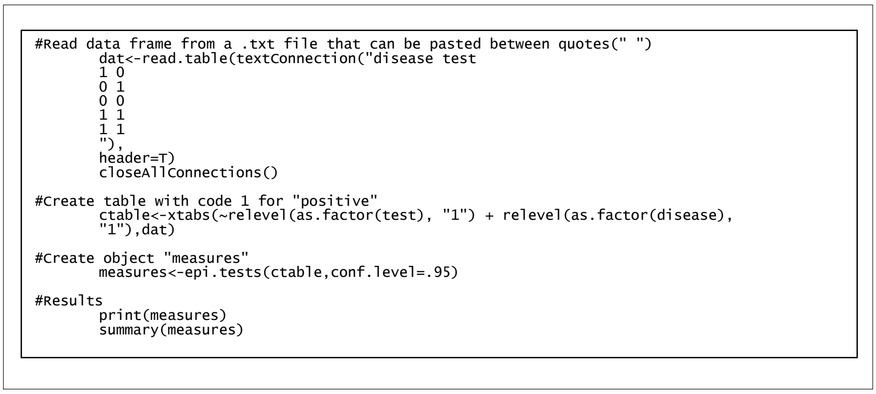
Conclusion

In this article we have focused on the criteria for screening, among groups of apparently non-diseased people, for the presence of disease before signs or symptoms have become apparent. We have also described how screening tests may be evaluated in terms of their sensitivity, specificity and predictive values. Finally, we have presented Stata statistical software commands for the estimation of screening test parameters. We have also presented R code for the same estimations, when using R code rather than Stata software.

We have discussed the importance of using appropriate control groups when estimating the performance characteristics

Table 8. Alternative layout of data using R

disease	test
1	0
0	1
0	0
1	1
1	1
(etc.)	(etc.)



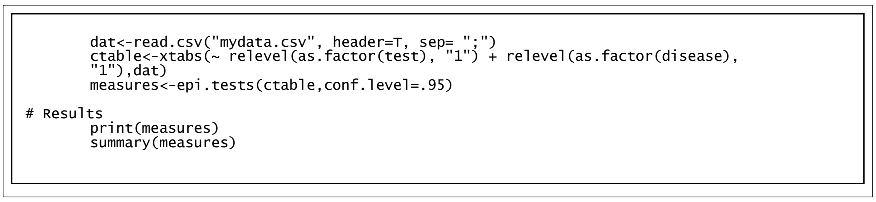
```
#Read data frame from a .txt file that can be pasted between quotes(" ")
        dat<-read.table(textConnection("disease test
        1 0
        0 1
        0 0
        1 1
        1 1
        "),
        header=T)
        closeAllConnections()

#Create table with code 1 for "positive"
        ctable<-xtabs(~relevel(as.factor(test), "1") + relevel(as.factor(disease),
        "1"),dat)

#Create object "measures"
        measures<-epi.tests(ctable,conf.level=.95)

#Results
        print(measures)
        summary(measures)
```

Fig. 4. R code for the situation where individual data values are entered directly.



```
        dat<-read.csv("mydata.csv", header=T, sep= ";")
        ctable<-xtabs(~ relevel(as.factor(test), "1") + relevel(as.factor(disease),
        "1"),dat)
        measures<-epi.tests(ctable,conf.level=.95)

# Results
        print(measures)
        summary(measures)
```

Fig. 5. R code for the situation where the data are available in a csv format file.

for screening tests, and cautioned that these control groups may differ, when estimating the performance characteristics of diagnostic tests. For screening test parameter estimation, both the cases and controls should be apparently healthy individuals; for diagnostic test estimation, controls should be symptomatic patients who do not have the disease being tested for.

Cross-sectional studies are favoured for estimating the parameters of screening tests, whereas case-control studies may be more suitable for estimating parameters for diagnostic tests. Unfortunately, owing to the usual rarity of even so-called common diseases, cross-sectional studies may need to enrol very large numbers of apparently healthy individuals. The result is that there are extra costs involved in studies to determine performance characteristics of screening tests.

Furthermore, since both the screening test and a gold standard test need to be performed on each participant in a screening test study, this, too, adds to the cost of such studies.

Finally, predictive values for a test result depend, in part, on the prevalence of the disease among the study participants. Therefore, estimates obtained from a case-control study should not be assumed to apply to the predictive values for the test when used as a screening tool.

Acknowledgements. None.

Author contributions. BGB wrote the article and provided the Stata commands. RB wrote the R code and also read and commented on the text. LD checked the running of the R code and also read and commented on the text. All three authors checked the final copy for correctness prior to publication.

Funding. None.

Conflicts of interest. None.

1. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2018. https://www.R-project.org/ (accessed 4 March 2019).
2. Wilson JMG, Jungner G. Principles and practice of screening for disease. WHO Chronicle 1968;22(11):473. Public Health Papers, #34. http://www.who.int/bulletin/vo
3. Andermann A, Blancquaert I, Beauchamp S, Déry V. Revisiting Wilson and Jungner in the genomic age: A review of screening criteria over the past 40 years. Bull World Health Org 2008;86(4):317-319. https://doi.org/10.2471/blt.07.050112
4. Dunstan DW, Zimmet P, Welborn TA, De Courten MP, Cameron AJ, Sicree RA. The rising prevalence of diabetes and impaired glucose tolerance. The Australian Diabetes, Obesity and Lifestyle Study. Diabetes Care 2002;25(5):829-834. https://doi.org/10.2337/diacare.25.5.829
5. Usher-Smith JA, Sharp SJ, Griffin SJ. The spectrum effect in tests for risk prediction, screening and diagnosis. BMJ 2016;353:i3139. https://doi.org/10.1136/bmj.i3139
6. Seed PT, Tobias A. Summary statistics for diagnostic tests. Stata Technical Bulletin 2000;59:9-12.
7. Stevenson M, Nunes T, Heuer C, et al. EpiR: An R Package for the Analysis of Epidemiological Data. R package version 9-99, 2018.

Accepted 9 May 2019.