The Stata Journal (2002)
2, Number 3, pp. 267–279

Estimation of sensitivity and specificity arising from validity studies with incomplete designs

Michael E. Reichenheim

Antônio Ponce de Leon

Instituto de Medicina Social
Universidade do Estado do Rio de Janeiro, Brazil

Abstract. This insert introduces `valides`, `validesi`, and `validesu`, a set of commands that allow the calculations of point and precision estimates of sensitivity and specificity obtained in validity studies with incomplete designs.

Keywords: st0019, validity, sensitivity, specificity, incomplete design

1 Description

`valides` estimates sensitivity (Se) and specificity (Sp) arising from multi-stage (phase) study designs, whereby a *new test* (NT) under scrutiny for its concurrent validity is applied first to all subjects and the *reference test* (RT) thereafter on only a subsample of those. This may be suitable, for instance, when the prevalence of an event is relatively rare and the RT is costly, invasive, potentially risky, etc. Very often, the RT is applied to all $NT+$, but to just a fraction of $NT-$, say, on a sample size up to five times that of $NT+$.

`validesi` is the immediate form of `valides`. Here, cell values $\#a$, $\#b$, $\#c$, and $\#d$ comprising a 2×2 contingency table of NT by RT are directly typed as arguments instead of read from a dataset stored in memory. The program is intended for use in the same setting described for `valides`, but is also meant for data originated from multi-phase designs such as, for instance, when the reference procedure is evaluated on (new) test-positives collected in one setting (dataset), whereas independently applied to another (external) sample of test-negatives. Such study designs may be contemplated for logistic, financial, or even convenience reasons.

The population quantities to be estimated in validity studies with complete designs are in Table 1, where $Se = 100(P_A/P_{M_1})$ and $Sp = 100(P_D/P_{M_0})$. Table 2 portrays the structure of the data in classic one-stage designs when all subjects are simultaneously measured through NT and RT, following a random or systematic sampling procedure. Here, direct estimation of sensitivity—$Se = 100(p_a/p_{m_1})$—and specificity—$Sp = 100(p_d/p_{m_0})$—are possible (Streiner and Norman 1995).

Table 1: Population parameters to be estimated.

	$NT+$	$NT-$	
$RT+$	P_A	P_B	P_{M_1}
$RT-$	P_C	P_D	P_{M_0}
	P_{N_1}	P_{N_0}	1

Table 2: Classic one-stage/phase sampling design.

(a) Absolute values

	$NT+$	$NT-$	
$RT+$	a	b	m_1
$RT-$	c	d	m_0
	n_1	n_0	n

$\longleftrightarrow$ *(b)* Proportions

	$NT+$	$NT-$	
$RT+$	p_a	p_b	p_{m_1}
$RT-$	p_c	p_d	p_{m_0}
	p_{n_1}	p_{n_0}	1

Tables 3*(a)*, *(b)*, and *(c)* show a generic situation in which there is an incomplete design due to the breaking up of the measurement procedure. Table 3*(a)* pictures a step whereby NT is first applied to all subjects. Tables 3*(b)* and *(c)* illustrate the sequel of the process, a second step where RT is given to sub-groups of $NT+$ and $NT-$. Usually, n_1' and n_0' are, respectively, subsamples of n_1 and n_0, but sometimes $n_1 = n_1'$.

Table 3: Diagram referring to the incomplete, multi-stage/phase sampling design—*(a)* step for collecting information for the estimation of $p_{n_1} = n_1/n$; *(b)* step for collecting information for the estimation of $p_\alpha = a'/n_1'$ and $p_\delta = d'/n_0'$: absolute values; and *(c) ditto*: proportions.

(a)

	$NT+$	$NT-$	
$RT+$	—	—	
$RT-$	—	—	
	n_1	n_0	n

$\longrightarrow$ *(b)*

	$NT+$	$NT-$
$RT+$	a'	b'
$RT-$	c'	d'
	n_1'	n_0'

$\longleftrightarrow$ *(c)*

$NT+$	$NT-$
p_a	p_b
p_c	p_d

Since sensitivity and specificity cannot be assessed directly from incomplete study designs, `valides`/`validesi` indirectly calculates these quatities by means of

$$Se = \frac{p_{n_1}\ p_\alpha}{p_\alpha\ p_{n_1} + (1 - p_{n_1})\ (1 - p_\delta)} \tag{1}$$

$$Sp = \frac{p_\delta\ (1 - p_{n_1})}{p_{n_1} + p_\delta - p_{n_1}\ p_\alpha - p_{n_1}\ p_\delta} \tag{2}$$

where p_{n_1} is the proportion of positive subjects according to the *test*; p_α is the proportion

of positive subjects detected by the *reference test* among the $NT+$ previously sampled; and p_δ is the proportion of negative subjects detected by the *reference test* among the $NT-$ previously sampled. Equations (1) and (2) are derived from those presented by Choi (1992) a few years ago and are adapted for the purpose of this program.

`valides` and `validesi` also calculate confidence intervals, allowing for the variability occurring in both stages (phases) of the sampling procedure to be taken into account. These can be estimated via maximum likelihood or parametric bootstrap methods. The former is the default. Companion ado-files `_vld_lf.ado` and `_sim_lf.ado` relating, respectively, to each method are needed.

The maximum-likelihood estimation routine uses `ml`'s `lf` method on an internally reshaped data structure and assumes that the number of $NT+$ in the first stage and both $RT+ \mid NT+$ and $RT- \mid NT-$ in the second are binomially distributed. The log-likelihood function is written in terms of sensitivity (Se), specificity (Sp), and the prevalence of the event (P), the three quantities of interest.

Given the equalities

$$p_{n_1} = Se\ Sp + (1 - Sp)\ (1 - P) \tag{3}$$

$$q_{n_1} = 1\ -\ p_{n_1} \tag{4}$$

$$p_\alpha = \frac{Se\ P}{Se\ P + (1 - Sp)\ (1 - P)} \tag{5}$$

$$q_\alpha = 1\ -\ p_\alpha \tag{6}$$

$$p_\delta = \frac{Sp\ (1 - P)}{Sp + P - P\ (Se + Sp)} \tag{7}$$

$$q_\delta = 1\ -\ p_\delta \tag{8}$$

and the reshaped data being

(*Continued on next page*)

$$
\begin{aligned}
X_1 &= \begin{cases} 1, & NT+ \\ 0, & NT- \end{cases} \\
X_2 &= \begin{cases} 1, & NT+,\ RT+ \\ 0, & NT+,\ RT- \\ \cdot\,, & NT+,\ RT_{\text{missing}} \end{cases} \\
X_3 &= \begin{cases} 1, & NT-,\ RT+ \\ 0, & NT-,\ RT- \\ \cdot\,, & NT-,\ RT_{\text{missing}} \end{cases}
\end{aligned}
$$

the log-likelihood function is

$$
\ln f = \begin{cases}
\ln(q_{n_1}) \ + \ X_1 \ * \ \ln(p_{n_1}/q_{n_1}), \\
\qquad \text{for } X_2 = \cdot\,, \quad X_3 = \cdot \\
\ln(q_{n_1}) \ + \ X_1 \ * \ \ln(p_{n_1}/q_{n_1}) \ + \ \ln(q_\alpha) \ + \ X_2 \ * \ \ln(p_\alpha/q_\alpha)\,, \\
\qquad \text{for } X_2 \neq \cdot\,, \quad X_3 = \cdot \\
\ln(q_{n_1}) \ + \ X_1 \ * \ \ln(p_{n_1}/q_{n_1}) \ + \ \ln(q_\delta) \ + \ X_3 \ * \ \ln(p_\delta/q_\delta)\,, \\
\qquad \text{for } X_2 = \cdot\,, \quad X_3 \neq \cdot
\end{cases} \tag{9}
$$

In the optional parametric bootstrap procedure, the means and distributions of the quantities of interest are iteratively calculated in 6 steps and are detailed in Table 4. First, r simulated sensitivity and specificity values are obtained—$Se^{(i)}$ and $Sp^{(i)}$—using, at each i iteration, the simulated proportions $p_{n_1}^{(i)}$, $p_\alpha^{(i)}$, and $p_\delta^{(i)}$ (steps 1 to 5). The desired statistics are then calculated from the r simulated Se and Sp values (step 6). Centiles are of particular interest for the confidence intervals. For example, a 95% CI can be specified by detecting the 2.5 and 97.5 simulated centiles. Alternatively, standard deviates are estimated from the collection of simulated values and are used to parametrically calculate the CI. The two procedures tend to converge as the number of replications increase. The projected prevalence and respective confidence intervals may be directly obtained since $P = (p_{n_1}.p_\alpha)/Se$. The procedure needs `rndbin` to be installed (Hilbe and Linde-Zwirble 1996). Note that if there are any zero-cells arising from the data in memory (or from inputted values when using `validesi`), estimations are not possible and an appropriate warning is issued.

Table 4: Iterative algorithm for the calculation of confidence intervals for sensitivity and specificity obtained from validity studies with incomplete designs.

Step		Procedure
1	a	Simulating n Bernoullis with parameter $p_{n_1} = n_1/n$
	b	Calculating the proportion of $NT+$ in iteration $i \rightarrow p_{n_1}^{(i)}$
2	a	Simulating n_1' Bernoullis with parameter $p_\alpha = a'/n_1'$
	b	Calculating the proportion of $(RT+ \mid T+)$ in iteration $i \rightarrow p_\alpha^{(i)}$
3	a	Simulating n_0' Bernoullis with parameter $p_\delta = d'/n_0'$
	b	Calculating the proportion of $(RT- \mid T-)$ in iteration $i \rightarrow p_\delta^{(i)}$
4		Calculating Se and Sp in iteration i using, respectively, equations (1) and (2), but substituting the terms for $p_{n_1}^{(i)}$, $p_\alpha^{(i)}$ and $p_\delta^{(i)} \rightarrow Se^{(i)}$ and $Sp^{(i)}$
5		Exporting $Se^{(i)}$ and $Sp^{(i)}$ to an "external" reception dataset
		$\downarrow \qquad \downarrow$
6		Detecting the centiles corresponding to the limits of the respective $100(1-\alpha)\%$ confidence intervals

The package also includes `validesu`, which allows a re-analysis of the simulated data previously generated and saved by `valides` or `validesi`. These saved values may also be used to further explore the results, for instance, by visually plotting distributions obtained for different strata of interest.

2 Syntax

`valides` *refvar testvar* [`if` *exp*] [`in` *range*] [, `prv` `notab` `dec2` `level(`#`)`
`simul` `reps(`#`)` `saving(`*filename*`)` `using(`*filename*`)`]

testvar is the variable that identifies the result of the new diagnostic test, and *refvar* is the variable containing the real status of subjects. Note that the lower category must identify the non-exposed or the negative result of NT and RT or the false status of the patient, and must assume value 0. Positive results assume value 1.

`validesi` #*a* #*b* #*c* #*d* [, `ttot(`#`)` `tpos(`#`)` `prv` `notab` `dec2` `level(`#`)`
`simul` `reps(`#`)` `saving(`*filename*`)` `using(`*filename*`)`]

Cells #*a* and #*b* relate to the $NT+$ and $NT-$ among the $RT+$. Conversely, cells #*c* and #*d* relate to the NT status among the $RT-$. Unless the study design is complete, `validesi` needs to be informed of the total NT sample using `ttot()`, as well as the number of detected $NT+$ using `tpos()` (see details below).

`validesu` [, `using(`*filename*`)` `prv` `dec2` `level(`#`)`]

3 Options

`prv` requests the display of the projected prevalence of the event.

`notab` suppresses the display of the 2×2 table containing the *testvar* by *refvar* cross-tabulation.

`dec2` requests 2-digit displays.

`level(#)` specifies the confidence level (%) for the confidence intervals for sensitivity, specificity, and prevalence (if option `prv` is requested). The default is 95%.

`simul` specifies that confidence intervals are to be calculated via simulation (parametric bootstrap) instead of via the default maximum likelihood method.

`reps(#)` specifies the number of simulations to be performed. The default value of 100 is used when the option is not requested. The minimum is 2. `reps()` may only be active if `simul` is also specified.

`saving(`*filename*`)` requests that simulated values be dumped to a file called *filename* located in the `pwd` for further use (e.g., by `validesu`). `saving()` may only be active if `simul` is also specified.

`using(`*filename*`)` requests that the analysis be based on the simulated values previously dumped to an external file by `saving()`. When this option is requested in `valides`, only `level()`, `notab`, and `dec2` remain active. `validesu` should be used for re-analysing data generated by `validesi`. `using()` may only be active if `simul` is also specified.

`ttot(#)` is specific to `validesi`. It provides the total number of subjects undergoing the New Test. If not stated, this is the sum of $\#a$, $\#b$, $\#c$, and $\#d$ values specified by the user, implicitly assuming an analysis for a complete study design.

`tpos(#)` is specific to `validesi`. It provides the total number of positive subjects according to the *new test*. If not stated, this is the sum of values $\#a$ and $\#b$, implicitly assuming that no subsampling of $NT+$ was carried out.

4 Examples

Consider a (fictitious) study to assess the concurrent validity of a new data collection instrument (*test*) developed from an accepted *reference test*, with the aim to gain operational efficiency with minimal loss in validity. For that purpose, the new instrument *test* has been thoroughly reduced in item numbers and has been submitted to a field test.

Referring back to the notation presented in Table 1, assume that the population parameters are $P_A = 0.016$, $P_B = 0.098$, $P_C = 0.004$, and $P_D = 0.882$, with $Se = 0.80$ and $Sp = 0.90$ to be estimated. If the researcher had used a complete study design, the formal analysis using `diagt` (Seed and Tobias 2001) would have been

```
. diagt RT NT
```

Reference Test	New Test Pos.	Neg.	Total
Abnormal	80	20	100
Normal	490	4410	4900
Total	570	4430	5000

```
True abnormal diagnosis defined as RT = 1 (labelled Positive)
                                                          [95% Conf. Inter.]
------------------------------------------------------------------------------
Sensitivity                     Pr( +| D)  80.00%      70.82%   87.33%
Specificity                     Pr( -|~D)  90.00%      89.13%   90.83%
Positive predictive value       Pr( D| +)  14.04%      11.29%   17.16%
Negative predictive value       Pr(~D| -)  99.55%      99.30%   99.72%
------------------------------------------------------------------------------
Prevalence                      Pr(D)       2.00%       1.63%    2.43%
------------------------------------------------------------------------------
```

Alternatively, `ci` could have been employed:

```
. ci NT if RT==1
```

Variable	Obs	Mean	Std. Err.	[95% Conf.	Interval]
NT	100	.8	.0402015	.7202315	.8797685

```
. gen NT_i=1-NT
. ci NT_i if  RT==0
```

Variable	Obs	Mean	Std. Err.	[95% Conf.	Interval]
NT_i	4900	.9	.0042862	.8915972	.9084028

The analysis with `valides` using the default ml-based estimations yields quite similar results, as required:

```
. valides RT NT, prv dec2
```

Reference Test	New Test Positive	Negative	Total
Positive	80	20	100
Negative	490	4410	4900
Total	570	4430	5000

```
    Note: The data and specifications
          assume a complete study design.
True D defined by RT                [95% Conf. Inter.]
                                         ml based
------------------------------------------------------------
Sensitivity   Pr( +| D)    80.00%    72.16%  87.84%
                                     (s_err = 4.00)

Specificity   Pr( -|~D)    90.00%    89.16%  90.84%
                                     (s_err = 0.43)

Prevalence (projected)      2.00%     1.61%   2.39%
                                     (s_err = 0.20)
------------------------------------------------------------
```

Note that `valides` "understands" that the study design is complete. Also note that the projected prevalence has been requested, yielding a result quite consistent with the population parameter.

Minding the costs and/or inconvenience involved in applying the RT to all subjects, the researcher decides to do this for all $NT+$, but only to a fraction of $NT-$. The researcher opts for a 1:2 (sub)sample-size ratio, which still implies reducing to 1/3 the number of subjects to whom the RT needs to be offered. A plausible dataset and the ensuing analysis with `valides` could be as follows:

```
. list id NT RT
             id         NT         RT
   1.         1   Negative   Negative
   2.         2   Negative   Negative
  (output omitted)
1134.      1134   Positive   Negative
1135.      1135   Positive   Negative
  (output omitted)
1625.      1625   Negative   Positive
1626.      1626   Negative   Positive
  (output omitted)
1631.      1631   Positive   Positive
1632.      1632   Positive   Positive
  (output omitted)
1711.      1711   Negative          .
  (output omitted)
5000.      5000   Negative          .
. valides RT NT, prv dec2
 Reference |        New Test
      Test |   Positive   Negative |     Total
-----------+-----------------------+----------
  Positive |         80          6 |        86
  Negative |        490       1134 |      1624
-----------+-----------------------+----------
     Total |        570       1140 |      1710

     Reference Test applied to T+ and T-
     Total sample size for Test = 5000
     Test positives = 570

True D defined by RT              [95% Conf. Inter.]
                                       ml based
-------------------------------------------------------
Sensitivity   Pr( +| D)    77.43%    62.96%  91.90%
                                     (s_err = 7.38)

Specificity   Pr( -|~D)    89.99%    89.15%  90.83%
                                     (s_err = 0.43)

Prevalence (projected)      2.07%     1.56%   2.57%
                                     (s_err = 0.26)
-------------------------------------------------------
```

Inevitably, there was a loss in statistical efficiency since there is now less data available in the incomplete study design. In the case of this example, perhaps the researcher would still be willing to accept the amount of imprecision in the estimates, but a final de-

cision would need to be made in the light of how much the gains in operational efficiency would outweight this relative deficiency. Note that, as before, `valides` "understood" the data, this time as originating from an incomplete study design. In passing, the same output would have been obtained using

```
. validesi 80 6 490 1134, ttot(5000) tpos(570) prv dec2
```

A sub-group analysis could also be requested, for instance, according to severity of the underlying event/disease. The estimations using a 90% CI would be

```
. valides RT NT if D==1, le(90)
(2500 observations deleted)
```

Reference Test	New Test Positive	Negative	Total
Positive	45	2	47
Negative	245	567	812
Total	290	569	859

```
    Reference Test applied to T+ and T-
    Total sample size for Test = 2500
    Test positives = 290

True D defined by RT              [90% Conf. Inter.]
                                       ml based
------------------------------------------------------------
Sensitivity   Pr( +| D)     85.3%     70.4%  100.0%
                                      (s_err =  9.1)

Specificity   Pr( -|~D)     90.0%     89.0%   91.0%
                                      (s_err =  0.6)
------------------------------------------------------------
. valides RT NT if D==0, le(90)
(2500 observations deleted)
```

Reference Test	New Test Positive	Negative	Total
Positive	35	4	39
Negative	245	567	812
Total	280	571	851

```
    Reference Test applied to T+ and T-
    Total sample size for Test = 2500
    Test positives = 280

True D defined by RT              [90% Conf. Inter.]
                                       ml based
------------------------------------------------------------
Sensitivity   Pr( +| D)     69.2%     50.8%   87.7%
                                      (s_err = 11.2)

Specificity   Pr( -|~D)     90.0%     89.0%   91.0%
                                      (s_err =  0.6)
------------------------------------------------------------
```

Once again `valides` acknowledges the data, correctly using the appropriate subsets in the analysis.

Similar results would have been produced when using the parametric bootstrap method. Although admittedly less efficient from a computational stand, an immediate bonus of this option would be the ability to further explore the data using the simulated data dumped to an external file. Perhaps the researcher would be interested in a visual comparison of the *new test*'s properties according to the severity of disease. The following simple do-file would produce Figure 1:

```
use valides.dta, clear
valides RT NT if D==1, sim reps(10000) saving(simfile1)
valides RT NT if D==0, sim reps(10000) saving(simfile0)

use simfile0.dta, clear
rename s_sens s_sens0
rename sens sens0
keep s_numb s_sens0 sens0
sort s_numb
save simf0tmp, replace

use simfile1.dta, clear
rename s_sens s_sens1
rename sens sens1
keep s_numb s_sens1 sens1
sort s_numb
merge using simf0tmp
erase simf0tmp.dta

label var  s_sens1 "BS Sensitivity, Disease=1"
label var  s_sens0 "BS Sensitivity, Disease=0"
label var  sens1 "Sensitivity, Disease=1"
label var  sens0 "Sensitivity, Disease=0"

gr s_sens0 s_sens1 s_numb, xlab ylab s(iii) c(l[_]l)
```

(*Continued on next page*)

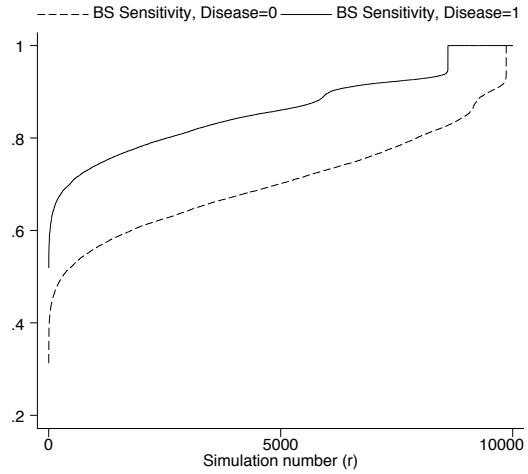

Figure 1: Distribution of BS simulated values of sensitivity according to disease status

For completeness, the output using `valides` with `simul` and `saving()` options on the disease positives and `validesu` applied to previously saved data on the disease negatives is shown next. As expected, estimations are indeed consistent with those obtained with the ml method.

```
. valides RT NT if D==1, sim reps(10000) saving(simfile1) notab
(2500 observations deleted)


     Reference Test applied to T+ and T-
     Total sample size for Test = 2500
     Test positives = 290


Please wait ... calculating confidence intervals


True D defined by RT                          [95% Conf. Inter.]
                                    s_err based        centile based
------------------------------------------------------------------------------
Sensitivity   Pr( +| D)     85.3%     67.9%  100.0%       67.6%  100.0%
                                    (s_err=   8.9)
Specificity   Pr( -|~D)     90.0%     88.8%   91.2%       88.8%   91.2%
                                    (s_err=   0.6)
------------------------------------------------------------------------------
                                                          (reps = 10000)
   Values dumped on 22 Apr 2002 15:39
   to simfile1.dta on the pwd
```

```
. validesu, using(simfile0)

Using data in simfile0
True D defined by RT                                [95% Conf. Inter.]
                                          s_err based        centile based
-------------------------------------------------------------------------------
Sensitivity   Pr( +| D)     69.2%     46.8%   91.7%       49.0%   92.0%
                                     (s_err=   0.1)

Specificity   Pr( -|~D)     90.0%     88.8%   91.2%       88.8%   91.2%
                                     (s_err=   0.0)
-------------------------------------------------------------------------------
                                        (reps in simulation data = 10000)
```

5 Saved Results

`valides`, `validesi`, and `validesu` save in `r()`:

Scalars

`r(sens)`	sensitivity
`r(se_ci_l)`	lower standard-error-based confidence interval for sensitivity
`r(se_ci_u)`	upper standard-error-based confidence interval for sensitivity
`r(se_c_l)`	lower centile-based confidence interval for sensitivity *
`r(se_c_u)`	upper centile-based confidence interval for sensitivity *
`r(se_s_err)`	standard error for sensitivity
`r(spec)`	specificity
`r(sp_ci_l)`	lower standard-error-based confidence interval for specificity
`r(sp_ci_u)`	upper standard-error-based confidence interval for specificity
`r(sp_c_l)`	lower centile-based confidence interval for specificity *
`r(sp_c_u)`	upper centile-based confidence interval for specificity *
`r(sp_s_err)`	standard error for specificity
`r(s_prev)`	prevalence (projected)
`r(pr_ci_l)`	lower standard-error-based confidence interval for prevalence
`r(pr_ci_u)`	upper standard-error-based confidence interval for prevalence
`r(pr_c_l)`	lower centile-based confidence interval for prevalence *
`r(pr_c_u)`	upper centile-based confidence interval for prevalence *
`r(pr_s_err)`	standard error for prevalence
`r(z)`	z score
`r(p_alpha)`	alpha probability *
`r(p_delta)`	delta probability *

Note: those marked with an asterisk are only obtained when option `simul` is used. The others are always provided but relate to the applied estimation method.

6 References

Choi, B. C. K. 1992. Sensitivity and specificity of a single diagnostic test in the presence of work-up bias. *Journal of Clinical Epidemiology* 45: 581–586.

Hilbe, J. and W. Linde-Zwirble. 1996. sg44: Random number generators. *Stata Technical Bulletin* 28: 20–21. In *Stata Technical Bulletin Reprints*, vol. 5, 118–121. College Station, TX: Stata Press.

Seed, P. T. and A. Tobias. 2001. sbe36_1: Summary statistics for diagnostic tests. *Stata Technical Bulletin* 59: 9–12. In *Stata Technical Bulletin Reprints*, vol. 10, 90–93. College Station, TX: Stata Press.

Streiner, D. L. and G. R. Norman. 1995. *Health Measurement Scales. A Practical Guide to Their Development and Use*. 2d ed. Oxford: Oxford University Press.

About the Authors

Michael E. Reichenheim (michael@ims.uerj.br), Departamento de Epidemiologia, Programa de Investigação Epidemiológica em Violência Familiar (PIEVF) / Núcleo de Pesquisa das Violências (NUPEVI), Instituto de Medicina Social (www.ims.uerj.br), Universidade do Estado do Rio de Janeiro, Brasil.

Antônio Ponce de Leon (ponce@ims.uerj.br), Departamento de Epidemiologia, Instituto de Medicina Social, Universidade do Estado do Rio de Janeiro, Brasil.