Psychological Medicine, 1997, **27**, 237–240.

BRIEF COMMUNICATION

Apparent change in caseness in longitudinal studies

ROB SELZER[1] AND JOHN B. CARLIN

From the Centre for Adolescent Health and the Clinical Epidemiology and Biostatistics Unit, Royal Children's Hospital, Melbourne, Victoria, Australia

SYNOPSIS Unless positive responses are verified at a second stage of data collection, a questionnaire-based survey has limited ability to assess change in caseness in a longitudinal study of a condition of low prevalence, assuming imperfect validity of the survey instrument.

The identification of uncommon disorders in a community population is fraught with methodological problems, not the least of which is the likelihood of finding a significant number of false positive subjects. The ability of a test instrument to categorize subjects correctly is reflected by its validity coefficients, sensitivity and specificity, as shown in Table 1. The positive predictive value of an instrument (the proportion of test positives who are true cases) is related to sensitivity, specificity and prevalence by the formula:

$$PPV = \frac{p \,.\, se}{(1-sp)+p(se+sp-1)},$$

where PPV = positive predictive value, p = prevalence, sp = specificity and se = sensitivity (Shrout & Fleiss, 1981). In spite of an exceptionally high sensitivity and specificity, the positive predictive value can become unacceptably low as prevalence decreases. Williams *et al.* (1982) have demonstrated that for a disease with a prevalence of 1%, sensitivity and specificity must be 0·99 before the PPV exceeds 0·90. Few test instruments possess such validity coefficients.

In descriptive studies, if the properties of the test are known then a rearrangement of the above formula allows disease prevalence to be estimated. This means that the proportion of subjects who are likely to be cases can be estimated but individual cases can not be identified with confidence.

As research moves beyond purely descriptive studies to those examining aetiology and natural history, longitudinal surveys are being used more often. In such surveys particular interest lies in those subjects who change case status and in the identification of key variables that predict change. For example, subjects who are initially healthy on testing and then become cases on retesting at some future time can contribute to our understanding of aetiological factors. Those who are originally cases and then become healthy on retesting provide information about the natural history of potentially reversible disorders.

Accurate identification of those who change case status is essential to longitudinal research. This becomes crucial when the prevalence and incidence rates are low as the misclassification of even a few subjects can significantly alter conclusions. We illustrate here how the accuracy of the measurement of change can be markedly

Table 1. *Definition of test validity coefficients in population of* $N = a+b+c+d$ *individuals*

Instrument	Case	Non-case
Positive	a	b
Negative	c	d

Sensitivity (se) = a/(a+c).
Specificity (sp) = d/(b+d).
Positive predictive value (PPV) = a/(a+b).

[1] Address for correspondence: Dr Rob Selzer, Centre for Adolescent Health, Royal Children's Hospital, Parkville Victoria 3052, Australia.

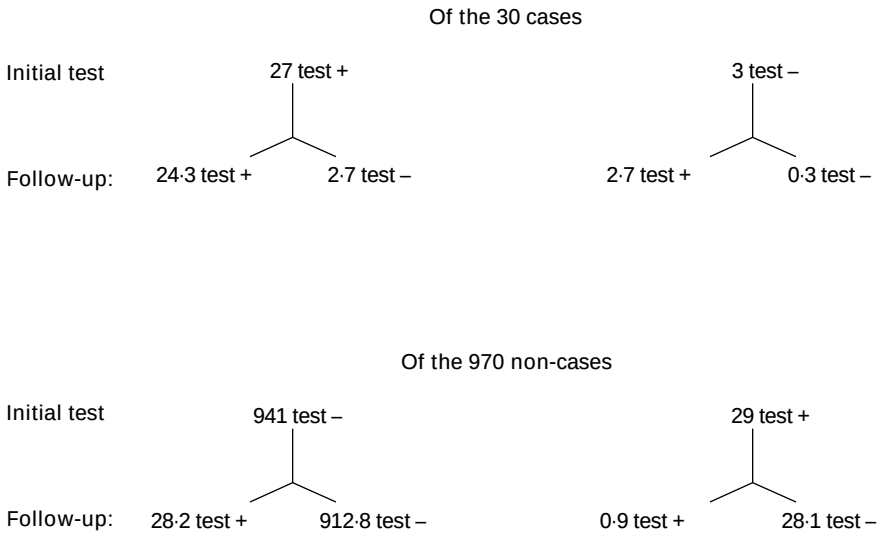

FIG. 1. Demonstration of the unreliability of the one-stage design in determining change of caseness in a longitudinal study. A test with sensitivity of 0·90 and specificity of 0·97 is administered on two occasions to a sample of 1000 subjects including 30 cases of eating disorder. Assuming there is no true change in the subjects and that the test results on the two occasions are independent, of those who were originally test positive 55% (30·8 subjects) change to test negative at follow-up. In fact 61·7 subjects may be expected to change case status.

attenuated by the disease prevalence and the test's validity coefficients.

For example, imagine we are to administer a questionnaire test for eating disorder to 1000 subjects, 30 of whom are true cases. Our test has a sensitivity of 0·90 and a specificity of 0·97. Following the initial administration of the test, 27 of the 30 cases and 29 of 970 normals would be expected to be test positive i.e. 56 positives in all (Fig. 1). The calculated *PPV* (based on the above formula) is 0·48. From this we can estimate that 27 of the 56 positives are cases ($0{\cdot}48 \times 56$), but we do not know which 27. At a designated follow-up time we re-administer the test – and suppose for the purposes of the example that no subject has truly changed case, and that results of the second test are independent of those of the first test. Fig. 1 shows that of our original 56 test positives, 30·8 (55%) would be expected to become test negative, even though in reality they do not change. Drawing conclusions about the natural history of the eating disorder based on those who apparently changed to test negative would be erroneous; similarly so for conclusions concerning aetiology as 30·9 of those who originally tested negative would become positive at follow-up.

The proportion of initial test positives that can be expected to test negative on the second occasion can be calculated for a variety of test coefficients and prevalences, via the formula:

$$D = \frac{p[se(1-se)-sp(1-sp)]+sp(1-sp)}{p(se+sp-1)+1-sp},$$

where D = the proportion of original test positives who can be expected to become test negative, p = prevalence, sp = specificity and se = sensitivity.

In this calculation and the example, the probability of an initial false positive being correctly identified as a true negative at follow-up is assumed to be equal to that of an initial true negative being identified as a true negative at follow-up: in other words, the probability of testing negative is independent of previous results, given that the subject remains truly negative.

In reality, for any instrument, test scores for the same individual on two separate occasions are not completely independent. A more realistic analysis requires consideration of the concept of reliability. The less reliable a test, the more likely a subject is to respond differently at two time points, even if their true status has not changed. Actual tests will behave in a way that lies

https://doi.org/10.1017/S0033291796003546 Published online by Cambridge University Press

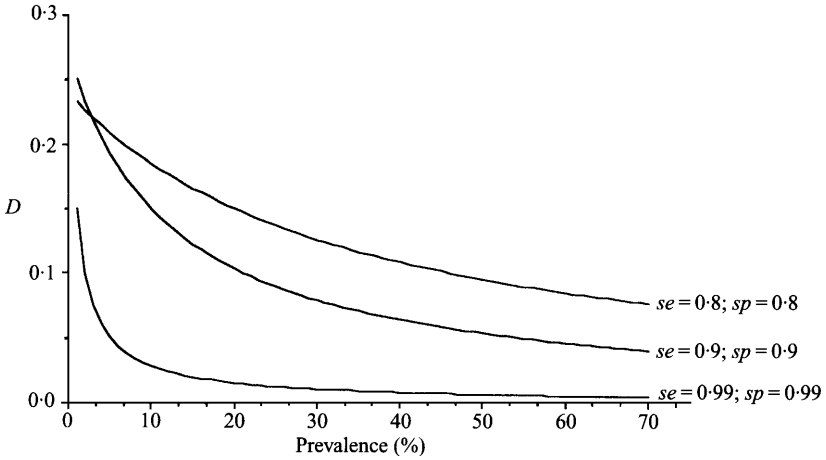

FIG. 2. Graph illustrating the effect of disease prevalence on apparent change in positivity (D) for a test–retest kappa of 0·7 and a variety of validity coefficients (se = sensitivity and sp = specificity).

intermediate between the extremes of complete independence and perfect reliability, although reliability will tend to be lower in more homogeneous test populations. The reliability of a test is unrelated to its validity coefficients and is often described by the kappa (κ) coefficient (Cohen, 1968), which can be used to measure the level of agreement beyond chance between successive applications of the test. It can be shown (see Appendix) that if the same value of κ is assumed to describe the strength of agreement between successive test results for both the true positives and the true negatives, then the value of D, defined above, is reduced by the factor $1-\kappa$. Assuming that typical values of κ for test–retest reliability are 0·7–0·8, the value of D is about one-quarter of that indicated in Fig. 1.

Even with these levels of reliability, however, at low prevalences there is still a substantial false-positive rate and accompanying rate of apparent change in caseness. Fig. 2 illustrates the profound effect even minor changes in test validity coefficients can have on D. As prevalence falls below 10%, D increases substantially, even in the presence of very high sensitivity and specificity.

In order to minimize the distorting effect of random changes in apparent disease positivity, a two-stage survey design can be used. Specifically, all test positives and a random sample of test negatives are examined in depth, usually at interview. Accurate identification of individual cases, at least for most of the true positives, can thus be effected at both initial and follow-up testing. Another lesson to be learned is that, where a second-stage verification is not possible, it may often be more useful to perform analyses based on the original scores or scale measurements rather than on dichotomous classifications based on these scores. Such analyses will be more powerful and less affected by measurement error problems.

APPENDIX; APPARENT CHANGE IN CASE STATUS UNDER DIFFERENT TEST–RETEST RELIABILITIES

We consider the following table of probabilities, describing the joint distribution of test results at two occasions for the same population of individuals, assumed to be either truly positive or truly negative and not to have changed case status between tests.

		Test result: time 2 +	Test result: time 2 −	
Test result: time 1	+	p_{11}	$p-p_{11}$	p
	−	$p-p_{11}$	$1-2p+p_{11}$	$1-p$
		p	$1-p$	1

In the case of the positive subpopulation p represents the sensitivity, while in the case of the negative subpopulation p represents one minus specificity. The probability of testing positive on both occasions is p_{11} and the extent to which this differs

https://doi.org/10.1017/S0033291796003546 Published online by Cambridge University Press

from the value p^2 that would be expected under independence determines the reliability of the test. Kappa is defined as

$$\kappa = \frac{p_{11} - p^2}{p(1-p)}.$$

The probability that an individual tests negative at the second occasion given that they have tested positive at the first (apparent change to negative status) is

$$\frac{p - p_{11}}{p} = \frac{p - [p^2 + \kappa(p(1-p))]}{p} = (1-p)(1-\kappa),$$

which is the (marginal) probability of testing negative multiplied by the factor $1-\kappa$. Using this result for both the true-positive and the true-negative groups and combining to obtain the overall probability of a change in status from positive to negative shows that the value of D defined above is reduced by the factor $1-\kappa$, assuming that κ is the same in both sub-populations.

REFERENCES

Cohen, J. (1968). Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin* **70**, 213–220.

Shrout, P. & Fleiss, J. (1981). Reliability and case detection. In *What is a Case? The Problem of Definition in Psychiatric Community Surveys* (ed. J. Wing, P. Bebbington and L. Robbins), pp. 117–128. Grant McIntyre: London.

Williams, P., Hand, D. & Tarnopolsky, A. (1982). The problem of screening for uncommon disorders – a comment on the Eating Attitudes Test. *Psychological Medicine* **12**, 431–434.

https://doi.org/10.1017/S0033291796003546 Published online by Cambridge University Press