Two-phase epidemiological surveys in psychiatric research

GRAHAM DUNN, ANDREW PICKLES, MICHELE TANSELLA and JOSÉ LUIS VÁZQUEZ-BARQUERO

For the purposes of epidemiological research in psychiatry, diagnostic interviews carried out by a clinician are often too expensive and time-consuming to justify their use in the general population where the great majority will not show any signs of psychopathology. The same is true for epidemiological surveys in other areas of medicine involving laboratory tests, radiological imaging, or invasive diagnostic tools. However, these problems can often be overcome by making use of a survey design that involves the use of an initial screening test such as a questionnaire, that is inexpensive and relatively easy to use in the field, but thought to be less accurate than use of the formal interview. In order to validate the screening questionnaire a sub-sample of the screened participants can be drawn for comparison with the results of an interview. The latter subsample is usually obtained through the use of a 'two-phase' or 'double sampling' design, with the probability of selection at the second phase being dependent on the results of screening and perhaps other information collected at the first phase. Two-phase sampling is a type of stratified design, proposed in psychiatric research as an efficient means for estimating prevalence of psychopathology. Although first proposed as a sampling design by Neyman (1938) its first use in psychiatric epidemiology seems to have been in the 1960s. Its use, however, was given a considerable impetus with the development of the General Health Questionnaire (GHQ; Goldberg & Blackwell, 1970) and by 1978 the method appears to have been well-established (Deming, 1978).

In psychiatry (and in other areas of medical research) the term 'two-stage design' is widely used for two-phase or double sampling. This can lead to confusion since two-stage is an already established name in survey research for a different design (Cochran, 1977). In the latter, the first stage might be a sample of hospitals or clinics, for example, with the second stage being a random sample of patients from within these hospitals or clinics. That is, in a two-stage design the sampling units differ at the two stages; in a two-phase design they do not.

In many situations the relative gains in efficiency from two-phase sampling seem to be rather slight. Two-phase sampling will, however, increase in efficiency (relative to single-phase designs) as the prevalence of disorder gets lower but is likely to be less efficient if the screen costs more than half of the cost of the diagnostic interview. However, Deming (1978) has also pointed out many practical disadvantages of the design, including extra administrative complexity (including problems of greater non-response and non-compliance by the subjects), more complex databases and increased sophistication in the analysis of the results. Nevertheless, the two-phase design is growing in popularity.

The purpose of this paper is to review some of the issues concerning the use of two-phase designs but, in particular, to discuss approaches to the analysis of the data resulting from their use. Our examples will come from epidemiological psychiatry, but the discussion applies equally to surveys conducted in other areas of medical research. Those readers who wish to move on to more technical discussions are referred to Pickles *et al* (1995), Pickles & Dunn (1998) and Clayton *et al* (1998). The latter authors also discuss the more difficult problem of estimating incidence rates from repeated two-phase surveys.

DESIGN

As stated above, the use of a preliminary screening instrument such as the GHQ has many potential advantages when the use of a structured or semi-structured interview is expensive, time-consuming and difficult to carry out. The recent national survey of psychiatric morbidity in the UK avoided a two-phase design for the estimation of the prevalence of non-psychotic illnesses through the use of the Clinical Interview Schedule (CIS-R; Lewis *et al*, 1992), an interview designed for relatively inexpensive lay interviewers, but did, however, use a two-phase strategy to detect psychosis (Jenkins *et al*, 1997).

It may be particularly appealing to be able to exclude the majority of subjects who do not appear to have a problem in order to spend valuable interviewing time on those participants who do. The rarer the psychiatric disorder the more appealing this idea becomes. The screening questionnaire should be acceptable to the survey participants, easy to administer, and cheap (relative to a full interview). It should also be accurate. In the case where the screening questionnaire has a simple cut-off to distinguish potential cases from non-cases, the questionnaire has to have both high sensitivity and high specificity. A single cut-off to partition the first-phase participants into only two strata for sampling in the second phase is not necessary, nor is it necessarily particularly efficient. Clearly, the observed score on a screening questionnaire contains more information about a participant's psychiatric state than simply knowing whether the person is above or below a particular cut-off or threshold. Multiple thresholds provide a simple compromise. In one of the examples discussed in detail below the first-phase sample was partitioned into three strata for subsequent second-phase sampling (see Duncan-Jones & Henderson, 1978, for an early example of this particular design). If a single cut-off is to be used, then Hand (1987) discusses the determination of the cut-off for a screening questionnaire that gives optimal sensitivity and specificity.

Goldberg & Williams (1988) discuss the relative merits of essentially four survey designs for the estimation of the characteristics of the GHQ. These are:

(a) All sampled participants assessed using both the screen and the interview.

(b) Phase two interviews carried out on a stratified random sample of first-phase participants (using two strata defined by the GHQ scores – 'cases' *v.* 'non-cases').

(c) Phase two interviews are carried out on a stratified random sample of first-phase participants (using three strata defined by the GHQ scores).

https://doi.org/10.1192/bjp.174.2.95 Published online by Cambridge University Press

(d) As (b) but one of the strata defined by a zero GHQ score and phase two interviews are only carried out on randomly selected participants from the two non-zero strata. "In order to calculate validity coefficients the zero scorers are therefore (*optimistically*) assumed to be non-cases" (our italics).

Designs (b) and (c) are both special cases of the more general stratified sampling design involving multiple first-phase strata which may arise from the use of more than one screening instrument. In the Spanish example discussed in Pickles *et al* (1995) and Vázquez-Barquero *et al* (1997) the second-phase sampling fractions were determined from both a GHQ score and a general practitioner's assessment. Issues concerning the design of two-phase surveys, including optimality criteria and decisions concerning whether to use a two-phase design at all, are discussed in some detail by Deming (1978), Newman *et al* (1990), Shrout & Newman (1989) and, more recently, by Reilly (1996).

Here we simply comment on design (d) of Goldberg & Williams. It is our view that designs based on such optimistic prior assumptions have sometimes been adopted in circumstances where they lack credibility. None the less, the potential economy of fieldwork in not sampling from the lowest stratum may be significant and may occasionally be justified from external data. In a study of dementia, for example, the scoring of full marks on a Mini-Mental State Examination (Folstein *et al*, 1975) phase one assessment may exclude the possibility of dementia being present. If this assumption is not safe, then it is crucial that not only should the phase two sample contain some of the screen negative stratum but that sufficient should be sampled to allow reliable estimation of the prevalence in that stratum.

METHODS OF ANALYSIS

Estimation of prevalence

The traditional approach to the estimation of prevalence is to take the weighted sum of the prevalences of 'true' cases within each of the strata defined by the first-phase screening questionnaire. The *stratum weight* simply reflects its relative size (that is, first-phase stratum weights sum to 1). The overall prevalence estimate is the sum of the products of the stratum prevalence and stratum weights over all first-phase strata (Cochran, 1977). The resulting proportion should not be considered as a simple binomial proportion for the purposes of variance and standard error estimation, but the formulae provided in Cochran (1977) or Pickles *et al* (1995), for example, should be used for this purpose. If prevalences are to be compared across subgroups, it is also invalid to use simple χ^2 tests and the like. It is surprisingly common to find that authors have correctly estimated the prevalence within each subgroup of a community survey (men and women, for example), have produced valid estimates for the standard errors, but still naively and incorrectly have gone on to compare these subgroups using the χ^2 test.

Now let us approach the estimation problem from a slightly different perspective. In the new approach, which is mathematically equivalent to the one above and gives the same answers, we restrict the analysis to only those survey participants with complete data (that is, the second-phase subjects only). Information arising from the first-phase screen results and the second-phase sampling mechanism is provided by the assignment of a 'sampling weight' to each individual subject, given by the inverse of the phase two sampling fraction. For example, let us assume that we have a first-phase sample of 100 individuals who according to the GHQ can be classified as likely cases and likely non-cases. Let us assume that we have found 70 of the latter. In the second phase of the survey we interview 15 (i.e. 50%) of the likely cases and 14 (i.e. 20%) of the likely non-cases. The sampling weights corresponding to the phase two subjects from the two strata are therefore 2 (i.e. 30/15) and 5 (70/14), respectively. The sampling weight is an indicator of how many phase one subjects are 'represented by' each of the phase two records. Table 1 illustrates the results of our hypothetical survey. Note, first, that it contains data from the second-phase participants only. Second, note that each participant has been assigned a sampling weight that depends on GHQ status (case/non-case) – the value being the reciprocal of the sampling fractions. There are data for 29 subjects (15+14) within Table 1 and the sum of their sampling weights is 100 – that is, these 29 phase two subjects are representing the 100 phase one subjects.

Each subject who is given an interview score of 1 is a 'true' case, a non-case otherwise. The sum of the products of this core and the sampling weights (i.e. 42) gives the estimated number of first-phase 'true' cases represented by the 15 phase two cases. The obvious estimate of morbidity for the first-phase sample (and, therefore, the population from which it was drawn) is 42/100 or 42%[1].

Now we will repeat the example of the use of sampling weights using algebraic notation. The symbol Σ will be used to denote 'the sum of'. Let $y_i = 1$ if the *ith* second-phase subject is a 'true' case, 0 otherwise. Let w_i be the *ith* subject's sampling weight. The estimate of the prevalence, π, is given by:

$$\pi = \Sigma w_i y_i / \Sigma w_i \qquad (1)$$

This is the well-known Horvitz–Thompson estimator from the sampling survey literature (Lehtonen & Pahkinen, 1995). The variance of π can be estimated through the use of the Taylor Series expansion or through bootstrap sampling (Pickles *et al*, 1995; Clayton *et al*, 1998). These methods can also provide large-sample confidence intervals for π. We do not need to discuss any technical details but simply refer the reader to the discussion on statistical software below. Note, however, that the use of this method is quite straightforward for quite complex stratification schemes for the second-phase sampling. In practice, it may be more convenient (and preferable from a theoretical point of view) to estimate prevalence via a logistic model (by just fitting a constant term in the model), produce a symmetrical confidence interval for the regression coefficient, and then reverse the logistic transformation to produce the corresponding interval for the prevalence itself (see below). The latter interval will not be symmetrical about the point estimate but will always stay within the permitted range of a prevalence (0–100%).

Estimation of screen sensitivity and specificity

Although their surveys are primarily concerned with prevalence estimation, many investigators also wish to be able to estimate sensitivity and specificity from their two-phase survey results. We use the sampling weights as in Table 1. First we split the file

1. For comparison, using the traditional approach, the stratum weights for GHQ cases and non-cases are 30/100 and 70/100, respectively. The proportion of true cases in the first stratum is 11/15; that in the second is 4/14. The weighted average for the two strata is therefore [(30/100) × (11/15)]+[(70/100) × (4/14)]=22/100+22/100+20/100+=42/100 or 42%, as before.

https://doi.org/10.1192/bjp.174.2.95 Published online by Cambridge University Press

Table 1 The results of a hypothetical survey of psychiatric morbidity (phase two data only)

Subject (i)	GHQ status[1]	Interview status (y_i)[1]	Sampling weight (w_i)	w_iy_i
1	1	1	2	2
2	1	1	2	2
3	1	0	2	0
4	1	1	2	2
5	1	1	2	2
6	1	1	2	2
7	1	0	2	0
8	1	0	2	0
9	1	1	2	2
10	1	1	2	2
11	1	1	2	2
12	1	0	2	0
13	1	1	2	2
14	1	1	2	2
15	1	1	2	2
16	0	0	5	0
17	0	0	5	0
18	0	0	5	0
19	0	1	5	5
20	0	1	5	5
21	0	0	5	0
22	0	0	5	0
23	0	0	5	0
24	0	0	5	0
25	0	0	5	0
26	0	0	5	0
27	0	0	5	0
28	0	1	5	5
29	0	1	5	5
Total	29	15	100	42

1. 1=case; 0=non-case.

of phase two subjects into two: cases and non-cases according to the interview. In the first file sensitivity is simply the weighted estimate of the proportion of screen positives in the cases. In the second one, the specificity is the weighted estimate of the proportion of non-cases who are screen negative. One can also separate the allocation of the value of the threshold to define caseness according to the screen from the thresholds to define the strata that were used in the survey design (i.e. one can investigate the effect of varying the case-definition after collecting the data using a preset threshold). As in the case of prevalence estimation, it may be preferable to work via logistic modelling (see below).

Modelling: weighted logistic models

The simplest modelling approach is an extension of the weighting estimation of prevalence, using sampling weights. The data are simply the second-phase interview results, relevant covariates and the appropriate sampling weights. The coefficients, β, of a logistic model are then estimated by maximising a modified form of the standard logistic log-likelihood that includes a sampling weight. The parameter covariance matrix (which is used to produce appropriate standard errors, confidence intervals and test statistics) is then obtained through the use of a robust information 'sandwich' or, alternatively, bootstrap sampling might be used to generate a robust parameter covariance matrix (Clayton *et al*, 1998).

Let us have another look at the hypothetical data in Table 1. Using the logit procedure of Stata Release 5 (StataCorp, 1997) to fit a constant (using the command line: logit y [pw=w]) produces an estimate of the intercept of −0.323 with a robust standard error of 0.408. The corresponding 95% CI is (−1.139 to 0.493). The prevalence estimate (that is, exp(−0.323/{1+exp(−0.323)})) is 0.42, or 42%, as before. The corresponding 95% CI is (24–62%). Further analyses of real data involving covariates will be produced for illustrative purposes in the Examples section below.

The above weighted logistic models when applied to those subjects with complete data only (i.e. the second-phase sample) may not be optimally efficient. If there have been covariates measured on subjects at phase one, there are alternative, but technically more demanding, ways of modelling these data. Further details of alternative strategies for modelling two-phase data can be found in Pickles *et al* (1995), Carroll *et al* (1995) and Clayton *et al* (1998). One way of getting extra efficiency from the straightforward weighted logistic models is through the careful choice of sampling weights. One preference is to use the *observed* sampling fraction to calculate the sampling weight, rather than that written into the design. If, for example, it was planned to interview 50% of the screen positive subsample in phase two but, in fact, we obtained data from only 47%, then it can be shown that it is better to use the weight 100/47 in the analysis than 100/50 (Pepe *et al*, 1994). Another improvement can be obtained (in the case of categorical predictors or covariates, at least) by calculating a separate sampling weight for each cell in the phase two data. Subjects might be cross-classified by gender and age group, for instance, and in this case it might be advantageous to calculate sampling weights for each gender–age group combination (Pepe *et al*, 1994). When we have a mix of categorical and quantitative covariates it might be useful to model sampling fractions using a (unweighted) logistic regression on the phase one sample and then to use as the appropriate sampling weight for each phase two subject the reciprocal of the response probability predicted by the model.

https://doi.org/10.1192/bjp.174.2.95 Published online by Cambridge University Press

Statistical software

In general, users of commercial software packages should take great care in the use of any weighting procedures provided. The use of weights within most packages (such as SPSS (SPSS Corporation, 1995), for example) will give correct point estimates of prevalence, regression coefficients from logistic models and the equivalent odds ratios, but, unfortunately, will not usually use the appropriate variance estimator. Estimates of standard errors, confidence intervals and associated significance tests will not be valid. Typically, standard errors will be too small and the corresponding confidence intervals will be too narrow and *P*-values will be too low. Note that this is not a fault in the software. The problem arises from the fact that the weights are typically interpreted as *frequency weights* (an indicator of the number of observations with identical data to that provided by a given record). The package accordingly treats the *ith* subject of the second-phase sample as if it had been observed w_i times. The appropriate use of a *sampling* or *probability weight*, however, recognises that the observation has only occurred once, but that the observed second-phase subject is *representative* of w_i first-phase subjects, all but one of which have not provided second-phase data. Programs such as Stata (StataCorp, 1997) and SUDAAN (Shah *et al*, 1993) will deal satisfactorily with weights which are assumed to be known constants (arising from the design, for instance). If, as is usually the case, we use weights determined from the observed sampling fractions (rather than those actually planned for) then we might wish to allow for sampling variability in the weights. In this case, any software program with macro facilities to allow for the bootstrap estimation of the variance of weighted estimates (Stata, for example) can be used.

In many circumstances it is of no great practical significance to treat the weights as random variables subject to sampling variation rather than known constants – the results will often be very similar (Clayton *et al*, 1998). It is important, however, that the data analyst does not inadvertently use a frequency weight, thinking that it is a sampling or probability weight. Ideally, the manuals and help facilities for statistical software packages should distinguish between the different types of weighting so that the user can be clear about what is being invoked by a particular weight statement.

If one is using bootstrap sampling for variance estimation then it makes no difference to the results whether sampling, analytic or importance weights are used. If not, then it is important that sampling weights are used and correctly specified.

EXAMPLES

The purpose of the present section is to demonstrate that the correct use of weights is vital for valid inference from two-phase surveys. Readers might be tempted to think that we are concerned with technical subtleties that need not be the concern of the clinician. We hope to convince them that this is not the case. We illustrate our point through the analysis of data from a recent survey of psychiatric morbidity in Verona in northern Italy (Piccinelli *et al*, 1995). A similar analysis of a Spanish survey has been discussed by us in some detail (Pickles *et al*, 1995; Vázquez-Barquero *et al*, 1997). Our main concern here is to illustrate the use of a weighted logistic regression to estimate (a) prevalence of a disorder as defined by the second-phase interview and (b) the odds ratio as a measure of the effect of the subject's gender on whether he or she is a case, and (c) the sensitivity and specificity of the first-phase screening instruments.

Prevalence estimation

In the first phase of the Verona survey 1558 subjects were asked to complete the GHQ-12. These subjects were then stratified according to their GHQ score (low, medium or high) and sub-samples of these three strata then interviewed using the Composite International Diagnostic Interview – Primary Care Version, the CIDI–PHC (see Von Korff & Üstün, 1995). Details of the second-phase data are given in Table 2. Table 2 also illustrates how to calculate prevalence using the Horvitz–Thompson estimator. Using the Stata logit command (together with sample or probability weights), we obtain an estimate of the logit of the prevalence of 0.230 (s.e. 0.187). The corresponding weighted estimate provided by SPSS (using the same weights) is 0.229 (s.e. 0.051). The estimates themselves are practically the same but there is almost a fourfold difference in their standard errors. The prevalence of psychiatric disorder is 56%. The 95%CIs for the prevalence of disorder are (47–64%) and (53–58%) as provided by Stata and SPSS, respectively.

Table 2 Second-phase data from the Verona Survey (n=250)

	Male	Female	Total
Stratum 1 (GHQ 0–3): sampling weight=17.48			
Non-case	16	17	33
Case	8	19	27
			60
Stratum 2 (GHQ 4–5): sampling weight=4.94			
Non-case	9	5	14
Case	8	26	34
			48
Stratum 3 (GHQ > 5): sampling weight=1.92			
Non-case	15	8	23
Case	28	91	119
			142

Estimate of overall prevalence: weighted number of cases/first-phase sample size=[(27 × 17.48)+(34 × 4.94) +(119 × 1.92)]/1558=0.56.

The standard errors and CIs (and any associated *P*-values) provided by the naive use of weights in SPSS are far too small.

Odds ratios

Now consider the odds ratio as a measure of the influence of gender on psychiatric morbidity. From the Verona survey, the weighted prevalence estimates for men and women are 39.8% and 65.3%, respectively. A direct estimate of the odds ratio produced using the logistic command of Stata (again using appropriate sampling weights) is 2.852 (s.e. 1.128). The 95% CI for this odds ratio is (1.314–6.191). The corresponding 95% CI from a naive use of SPSS weights within the CROSSTABS procedure is (2.310–3.528). The corresponding *P*-value for the latter is <0.00001, whereas that corresponding to the Stata estimate is 0.008 – at least an 800-fold difference! Although we have not used the SPSS logistic regression procedure here, one can easily get the logistic regression procedure in SPSS to produce essentially the same results as CROSSTABS.

Sensitivity and specificity

Finally, consider the sensitivity and specificity of the screening tests. Table 3 provides the actual GHQ scores for subjects within each of the strata of the second phase of the Verona survey. Let us arbitrarily pick a GHQ score cut-off of between 6 and 7

https://doi.org/10.1192/bjp.174.2.95 Published online by Cambridge University Press

Table 3 Second-phase data from Verona: GHQ score by case

		Interview results	
		Non-case	Case
Stratum 1 (weight=17.48)			
GHQ Score	0	20	8
	1	5	7
	2	4	7
	3	4	7
Stratum 2 (weight=4.94)			
GHQ Score	4	10	21
	5	4	13
Stratum 3 (weight=1.92)			
GHQ score	6	8	37
	7	7	26
	8	5	20
	9	2	20
	10	1	6
	11	0	6
	12	0	4

Characteristics of the GHQ using a cut-off of 6/7: sensitivity=[82 × 1.92]/[(27 × 17.48)+(34 × 4.94)+(119 × 1.92)]=0.18; specificity=[(3 × 17.48)+(14 × 4.94)+(8 × 1.92)]/[(3 × 17.48)+(14 × 4.94)+(23 × 1.92)]=0.96.

in order to classify subjects as possible cases (or not) according to the GHQ. What is the sensitivity and specificity of this criterion with respect to the psychiatric interview? If there were no implications arising from the two-phase design, then sensitivity would be estimated simply as the number of people who were both GHQ positive and CIDI positive, divided by the total number of CIDI positives. Similarly, specificity would be estimated from the number who were both GHQ negative and CIDI negative, divided by the total number of CIDI negatives. In order to take account of the survey design we modify these estimators by introducing the appropriate weights into the two calculations, respectively. This is illustrated at the foot of Table 3. The estimates of sensitivity and specificity are 0.18 and 0.96, respectively. Using the logit command within Stata (together with appropriate sampling weights), as described above, but analysing the subsamples of true positives and true negatives separately, we obtain 95% CIs of (0.14–0.24) and (0.93–0.98), respectively. On the assumption that the message that the naive use of SPSS weights would give an invalid result has, by now, been accepted by the reader, we do not give the results of using SPSS. The choice of a GHQ cut-off of 6/7 is clearly not a very useful one, but it would be quite straightforward to carry out the calculations for another choice.

DISCUSSION

When an investigator chooses to use a two-phase (or multiphase) survey he or she should be aware of the fact that the choice of design has important implications for the analysis and presentation of the results. There are pitfalls for the unwary. An investigator might, for example, choose to use an analysis involving probability or sampling weights (to adjust for biases arising from ignoring the details of the design) and then naively proceed to analyse the data using software that has not been written with this purpose in mind. We have shown one example of an 800-fold difference in *P*-values arising from confusing sampling weights with frequency weights.

What are the implications of these conclusions? The first should be the recognition than the analysis of a complex survey is not a job for an untrained amateur. It is vital that investigators and funding bodies who commit enormous resources to the collection of data using complex survey designs recognise that there should be appropriate levels of resource put into the analysis and interpretation of the resulting data. The second implication is that both the editorial team of a journal and the journal's readership should be aware of the inferential problems surrounding the use of complex survey methodology. Finally, there should be clear standards laid down for the reporting of statistical methods used in the analysis of complex surveys, together with a tightening of the statistical refereeing process to ensure that these standards are being met.

It is beyond the scope of this paper to develop the required reporting standards in detail, but it is worth making two points. First, it cannot be overemphasised that the Methods sections of empirical papers need to be both clear and precise. The key question in asking whether they are clear and precise enough is "Could I repeat this study, including its analysis?" Second, it is vital that investigators provide full details of statistical and other commercial software used in their research. Authors should include the exact version of the software used as well as reporting whether they have tailored their analysis by changing the programs' default settings. Although space is important in scientific journals, editors should be persuaded to make their cuts (if they are really necessary) to the Introduction or Discussion sections of a paper – and not to those parts of the paper (Methods and Results) that are most important.

REFERENCES

Carroll, R. J., Ruppert, D. & Stefanski, L. A. (1995) *Measurement Error in Nonlinear Models*. London: Chapman & Hall.

Clayton, D. G., Spiegelhalter, D., Dunn, G., *et al* (1998) Analysis of longitudinal binary data from multi-phase sampling (with Discussion). *Journal of the Royal Statistical Society Series B*, **60**, 71–102.

Cochran, W. G. (1977) *Sampling Techniques* (3rd edn). New York: Wiley.

Deming, W. E. (1978) An essay on screening, or two-phase sampling, applied to surveys of a community. *International Statistical Review*, **45**, 28–37.

Duncan-Jones, P. & Henderson, S. (1978) The use of a two-phase design in a prevalence survey. *Social Psychiatry*, **13**, 231–237.

Folstein, M., Folstein, S. & McHugh, P. (1975) Mini-mental state: a practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research*, **313**, 1419–1420.

Goldberg, D. P. & Blackwell, B. (1970) Psychiatric illness in general practice. A detailed study using a new method of case identification. *British Medical Journal*, **2**, 439–443.

— & Williams, P. (1988) *A User's Guide to the General Health Questionnaire*. Windsor: NFER–Nelson.

Hand, D. J. (1987) Screening vs prevalence estimation. *Applied Statistics*, **36**, 1–7.

Jenkins, R., Bebbington, P., Brugha, T., *et al* (1997) The National Psychiatric Morbidity Surveys of Great Britain – strategy and methods. *Psychological Medicine*, **27**, 765–774.

Lehtonen, R. & Pahkinen, E. J. (1995) *Practical Methods for Design and Analysis of Complex Surveys*. Chichester: Wiley.

Lewis, G., Pelosi, A. J., Araya, R., *et al* (1992) Measuring psychiatric disorder in the community: A standardised assessment for use by lay interviewers. *Psychological Medicine*, **22**, 465–486.

Neyman, J. (1938) Contribution to the theory of sampling human populations. *Journal of the American Statistical Association*, **33**, 101–116.

Newman, S. C., Shrout, P. E. & Bland, R. C. (1990) The efficiency of two-phase designs in prevalence surveys of mental disorders. *Psychological Medicine*, **20**, 183–193.

Pepe, M. S., Reilly, M. & Fleming, T. R. (1994) Auxiliary outcome data and the mean score method. *Journal of Statistical Planning and Inference*, **42**, 137–160.

https://doi.org/10.1192/bjp.174.2.95 Published online by Cambridge University Press

Piccinelli, M., Pini, S., Bonizzato, P., *et al* (1995) Results from the Verona Centre. In *Mental Illness in General Health Care: An International Study* (eds T. B. Üstün & N. Sartorius). New York: Wiley.

Pickles, A., Dunn, G. & Vázquez-Barquero, J. L. (1995) Screening for stratification in two-phase ('two-stage') epidemiological surveys. *Statistical Methods in Medical Research*, **4**, 75–91.

__ & __ (1998) Prevalence of disease: estimation from screening data. In *Encyclopedia of Biostatistics* (eds P. Armitage & T. Colton), pp. 3484–3490. Chichester: Wiley.

Reilly, M. (1996) Optimal sampling strategies for two-stage studies. *American Journal of Epidemiology*, **143**, 92–100.

Shah, B. V., Folsom, R. E., LaVange, L. M., *et al* (1993) *Statistical Methods and Mathematical Algorithms used in SUDAAN*. Research Triangle Park, NC: Research Triangle Institute.

Shrout, P. E. & Newman, S. C. (1989) Design of two-phase prevalence surveys of rare disorders. *Biometrics*, **45**, 549–555.

GRAHAM DUNN, PhD, Biostatistics Group, School of Epidemiology and Health Sciences, University of Manchester; ANDREW PICKLES, PhD, MRC Child and Adolescent Psychiatry Unit and Department of Biostatistics and Computing, Institute of Psychiatry, London; MICHELE TANSELLA, MD, Institute of Psychiatry, University of Verona, Verona, Italy; JOSÉ LUIS VÁZQUEZ-BARQUERO, MD, Clinical and Social Psychiatry Research Unit, University Hospital "Maqués de Valdecilla", Santander, Spain

Correspondence: Professor G. Dunn, Biostatistics Group, School of Epidemiology and Health Sciences, University of Manchester Stopford Building, Oxford Road, Manchester M13 9PT. Fax: 0161 275 5567; e-mail: g.dunn@man.ac.uk

(First received 9 April 1998, final revision 7 September 1998, accepted 17 September 1998)

SPSS Corporation (1995) SPSS for Windows Version 6.1.3. Chicago, IL: SPSS Inc.

StataCorp (1997) *Stata Statistical Software: Release 5.0*. College Station, TX: Stata Corporation.

Vázquez-Barquero, J. L., Garcia, J., Simón, J. A., *et al* (1997) Mental health in primarycare. An epidemiological study of morbidity and use of health resources. *British Journal of Psychiatry*, **170**, 529–535.

Von Korff, M. & Üstün, T. B. (1995) Methods of the WHO Collaborative Study on 'Psychological Problems in General Health Care'. In *Mental Illness in General Health Care: An International Study* (eds T. B. Üstün & N. Sartorius), pp. 19–38. New York: Wiley.

https://doi.org/10.1192/bjp.174.2.95 Published online by Cambridge University Press