Recalibration methods to enhance information on prevalence rates from large mental health surveys

N. A. TAUB,[1] Z. MORGAN,[1] T. S. BRUGHA,[1] P. C. LAMBERT,[1] P. E. BEBBINGTON,[2] R. JENKINS,[3] R. C. KESSLER,[4] A. M. ZASLAVSKY,[4] T. HOTZ[1]

1 Department of Health Sciences, University of Leicester, UK
2 Department of Mental Health Sciences, University College London, UK
3 Institute of Psychiatry, London, UK
4 Department of Health Care Policy, Harvard Medical School, Boston, USA.

Absract

Comparisons between self-report and clinical psychiatric measures have revealed considerable disagreement. It is unsafe to consider these measures as directly equivalent, so it would be valuable to have a reliable recalibration of one measure in terms of the other. We evaluated multiple imputation incorporating a Bayesian approach, and a fully Bayesian method, to recalibrate diagnoses from a self-report survey interview in terms of those from a clinical interview with data from a two-phase national household survey for a practical application, and artificial data for simulation studies.

The most important factors in obtaining a precise and accurate 'clinical' prevalence estimate from self-report data were (a) good agreement between the two diagnostic measures and (b) a sufficiently large set of calibration data with diagnoses based on both kinds of interview from the same group of subjects. From the case study, calibration data on 612 subjects were sufficient to yield estimates of the total prevalence of anxiety, depression or neurosis with a precision in the region of ±2%. The limitations of the calibration method demonstrate the need to increase agreement between survey and reference measures by improving lay interviews and their diagnostic algorithms.

Key words: mental disorders, diagnostic techniques and procedures, health surveys, statistical methods, anxiety, depression

Introduction

The development and evaluation of policies for addressing mental health problems in the general population benefit from large-scale surveys incorporating lay interview methods (Kessler et al., 1994; Jenkins et al., 1997). Policy decisions require information based on statistics derived from the more feasible and cost-effective lay interviews incorporated into large-scale survey data sets (Department of Health, 1999). Considerable effort has gone into the development of these lay interviews (Regier et al., 1998; Wittchen, et al. 1999), which are intended to reproduce as closely as possible clinical psychiatric diagnoses according to internationally accepted diagnostic criteria (American Psychiatric Association, 1994). However, comparisons between lay measures and systematic clinical psychiatric interviews reveal substantial disagreement both in diagnoses at the individual level and in prevalence estimates (Brugha et al., 1999c; Kessler, 1999), giving rise to expressions of concern among psychiatric epidemiologists (Henderson, 2000). Concerns about these disagreements have led to much debate about the relative merits of clinical and lay interviews (Brugha et al., 1999b; Wittchen et al., 1999). One option that has been considered is to replace lay interviews with clinical interviews carried out by specially trained survey interviewers, which are reliable and less costly than assessments by practising clinicians (Brugha et al., 1999c). However, the financial and logistic impediments to carrying out this approach with nationally

representative samples makes it unlikely to be widely adopted except in parts of the world where interviewing costs are low.

The response to this situation has up to now been largely to continue efforts to improve lay interviews with reference to clinical diagnostic interviews (Kessler et al., 1998). A complementary approach, however, is to make better use of information on the interrelationship between lay and clinician measures when the latter are collected in a subset of survey respondents. This could lead to the development of calibration rules for adjusting the estimates based on lay interviews, to approximate more closely the estimates that would have been obtained if clinical interviews had been used throughout (Dunn, 1998; Kessler, 1999; Frazer and Stram, 2001).

One potentially useful approach is to consider a survey incorporating clinical interviews with a subset of the subjects, according to a prespecified sampling design, as an example of 'data missing by design' or design-incorporated 'item non-response' (de Leeuw, 1999). Appropriate statistical techniques are available for coping with the decreased precision resulting from this incomplete recording of data. Unlike the majority of missing data problems, in such surveys there will usually be a large proportion of missing data and a small proportion of complete data because there will usually be only a small proportion of subjects with both a clinical and a lay interview and a much larger proportion of subjects with only the lay interview. Much of the previous work on missing data has concentrated on calculating statistical models where some study subjects have one or more items of explanatory data missing. In the current context we are considering the situation when the outcome measure may be unknown – however, the methodological principles for dealing with these two types of problem are the same.

Although we use the term 'clinical' to refer to a reference interview, the recalibration methods studied here are applicable to other situations in which (a) a higher cost, intensive method of data collection believed to reflect a key construct more accurately is being compared with (b) a lower cost measure of the kind feasible for large scale data collection. It may seem preferable that policy decisions are made on the basis of epidemiological results based on definitive clinical criteria but their cost may be prohibitive. Treating these designs as a missing data problem may provide a solution to the limitations of relying exclusively on lay interview data.

Background

Numerous statistical methods have been developed to allow for missing data (Little and Rubin, 2002). However, many of the methods have limitations in terms of bias, precision and computational feasibility (Greenland and Finkle, 1995; Zhou et al., 2001). An increasingly popular approach is multiple imputation (Rubin, 1987), where a number of separate imputed data sets are generated, each with missing values replaced with values sampled from a *predictive distribution*. The use of a predictive distribution takes into account the uncertainty in the prediction of the missing values. However, defining the predictive distribution correctly is important. Each generated data set is analysed using standard 'complete case' methods with the results of each analysis then pooled, taking into account both the between- and within-data set variability. Thus, in psychiatric surveys with data missing by design, inferences can be made using the clinical interviews, with data being imputed for those subjects who have only a lay interview.

The problem of missing data fits naturally into a Bayesian framework, as the missing data are treated in the same way as other unknowns (model parameters) (Clogg et al., 1991). From a pragmatic point of view, the use of the Bayesian approach offers the opportunity to incorporate external information through the use of prior distributions. For example, information about the effects of covariates, such as gender, in explaining the relationship between the measures could be taken into consideration by using a prior distribution. This might be particularly useful for analysing data from psychiatric surveys where the information on the interrelationships between such covariates and the diagnoses based on lay and clinical interviews is obtained from existing studies. Other advantages of the Bayesian approach include the use of non-standard distributions such as when there are time-to-event (survival) data (Lambert et al., 1997), and the possibility of including further realistic complexity into the models (Best et al., 1996). In recent years, the development of the WinBUGS computer software package has made fitting such models feasible (Spiegelhalter et al., 2000).

The primary aim of this study is to enhance the decision-making potential of currently available lay interview data on adults by developing and evaluating calibration methods with reference to standard clinical measures. We chose to address this in relation to two common mental conditions with major public health

implications: depression and anxiety. We used multiple imputation (MI) and a fully Bayesian method to assess the prevalence of these disorders in the general population. We combined interviews designed specifically for use by lay survey interviewers with semi-structured interviews carried out by clinicians. The latter requires considerable clinical judgement and was therefore administered by interviewers with extensive clinical experience and training. We then applied statistical methods to adjust for the differences between lay and reference assessments.

The UK Office for National Statistics Household Survey 2000 collected data on psychiatric morbidity, including diagnoses of anxiety, depression and related disorders using both lay and reference instruments; these provide our practical example. Simulation studies were also carried out to provide sensitivity analyses, with computer-generated data samples of various sizes, with various levels for the prevalence, and for the agreement between the survey and reference instruments.

Methods

Sample and measures

The survey used as the main example for the methods described in this paper was the 'Psychiatric Morbidity Among Adults Living in Private Households, 2000' carried out in England and Wales by the Office for National Statistics (ONS) (Singleton, et al. 2001). The 'Clinical Interview Schedule – revised version' (CIS-R) (Lewis et al., 1992) was conducted by lay interviewers on all 8,580 respondents. The second phase of this two-phase design consisted of carrying out a reference assessment on subjects from strata defined according to the perceived likelihood of their suffering from a psychosis or a personality disorder (Table 1). We used the survey form of the Schedules for Clinical Assessment in Neuropsychiatry (SCAN) (Brugha et al., 1999c), a semi-structured clinical assessment administered by clinicians with extensive training in the assessment of psychopathology (Wing et al., 1990). Diagnostic classification was according to ICD-10 in both interviews (World Health Organization, 1993). There was an interval of 3 to 6 months between these assessments. As the analyses performed are for illustrative purposes only, no adjustments have been made for the time lag between the two assessments.

In order to help explain the difference in diagnosis of any ICD-10 study disorder between the CIS-R interview and the SCAN interview, the following available covariates were thought to be predictive of psychiatric conditions and were therefore included in a selection of the analyses: gender, employment status and use of psychotropic medication.

Preliminary analysis

The observed sampling fractions were calculated and are shown in Table 1, while the resulting levels of agreement, between the CIS-R and SCAN diagnoses for the overall category of any anxiety, depressive or other neurotic disorder, by stratum and overall, are shown in Table 2.

The prevalence of the combined study disorders, with confidence intervals, was then calculated using only the SCAN data, first using logistic regression with probability sampling weights (STATA computer package, version 7) to allow for the stratification, and then using CIS-R assessments (Table 4). The prevalence was

Table 1. ONS National Survey 2000: stratified sampling design for SCAN assessment

Stratum	Sampling fraction, by design	Total in survey sample	Number of SCAN assessments included at phase 2
1. Anti-social or borderline personality disorder with psychosis	100%	51	26
2. Anti-social or borderline personality disorder without psychosis	50%	341	70
3. Other personality disorder with psychosis	100%	145	78
4. Other personality disorder without psychosis	7% (i.e. 1/14)	1925	94
5. No personality disorder, but with psychosis	100%	179	94
6. Neither personality disorder nor psychosis	7% (i.e. 1/14)	5939	250
Overall	–	8580	612

also assessed using the CIS-R on all subjects, with a correspondingly higher degree of precision.

Analysis using multiple imputation

The known random sampling mechanism of the second phase of the study means that the unsampled cases are missing by design, independent of both observed and missing values other than the stratification based on the CIS-R assessments (Martin, 1999). Only 56% of the proposed SCAN assessments actually took place, and the association of this non-response with the unmeasured clinical state of the remaining cases is unknown. In our analyses, we assume that these data are missing at random, meaning that missingness

Table 2. ONS National Survey 2000: levels of agreement between CIS-R and SCAN assessments for any anxiety, depressive or other neurotic disorder, with 95% confidence intervals

Stratum	Total Number (No.) of subjects	No. subjects SCAN + CIS-R +	SCAN + CIS-R–	SCAN– CIS-R +	SCAN – CIS-R –	Total proportion agreement	Sensitivity	Specificity	Kappa
1	26	6	1	9	10	0.62 (0.41 to 0.80)	0.86 (0.42 to 1.00)	0.53 (0.29 to 0.76)	0.29 (–0.02 to 0.59)
2	70	1	2	10	57	0.83 (0.71 to 0.91)	0.33 (0.01 to 0.91)	0.85 (0.74 to 0.93)	0.09 (–0.19 to 0.37)
3	78	16	5	22	35	0.65 (0.54 to 0.76)	0.76 (0.53 to 0.92)	0.61 (0.48 to 0.74)	0.30 (0.11 to 0.50)
4	94	3	4	9	78	0.86 (0.78 to 0.92)	0.42 (0.10 to 0.82)	0.90 (0.81 to 0.95)	0.26 (–0.06 to 0.57)
5	94	11	6	8	69	0.85 (0.76 to 0.92)	0.65 (0.38 to 0.86)	0.90 (0.81 to 0.95)	0.53 (0.30 to 0.75)
6	250	4	6	10	230	0.94 (0.90 to 0.96)	0.40 (0.12 to 0.74)	0.96 (0.92 to 0.98)	0.31 (0.04 to 0.58)
Overall	612	41	24	68	479	0.85 (0.82 to 0.88)	0.63 (0.50 to 0.75)	0.88 (0.85 to 0.90)	0.39 (0.29 to 0.49)

Table 3. ONS National Survey 2000: measures of agreement between CIS-R and SCAN diagnoses, with 95% confidence interval, by diagnostic group; without consideration of the stratified survey design

	Any anxiety, depression, neurosis (incl. OCD)[1]	Any anxiety	Depression	Any phobia	Any other neurotic disorder	Non-phobic anxiety
Sensitivity	0.63 (0.50 to 0.75)	0.49 (0.33 to 0.65)	0.35 (0.20 to 0.53)	0.19 (0.05 to 0.53)	0.58 (0.42 to 0.72)	0.47 (0.28 to 0.66)
Specificity	0.88 (0.85 to 0.90)	0.87 (0.84 to 0.90)	0.95 (0.93 to 0.96)	0.94 (0.92 to 0.96)	0.87 (0.84 to 0.90)	0.91 (0.88 to 0.93)
Kappa	0.39 (0.29 to 0.49)	0.23 (0.12 to 0.33)	0.28 (0.14 to 0.42)	0.07 (–0.03 to 0.23)	0.29 (0.19 to 0.40)	0.23 (0.11 to 0.35)
Percentage agreement	0.85 (0.82 to 0.88)	0.85 (0.82 to 0.88)	0.91 (0.89 to 0.93)	0.92 (0.89 to 0.94)	0.85 (0.82 to 0.88)	0.88 (0.86 to 0.91)

[1] OCD = Obsessive-Compulsive Disorder

Table 4. ONS National Survey 2000: precision of estimates of prevalence of anxiety, depression, and neurosis, without and with consideration of the stratified survey design

Method	Prevalence estimate
Using SCAN assessments only, with consideration of the stratified survey design, n = 612	0.056 (0.038 to 0.081)
Using CIS-R assessments only, n = 8,580	0.086 (0.080 to 0.092)
Using SCAN and CIS-R assessments by the MCMC calibration method, without consideration of the stratified survey design or covariate information.	0.076 (0.059 to 0.097)
Using the 'MCMC multiple imputation' calibration method, with consideration of the stratified survey design, but without use of covariate information.	0.055 (0.036 to 0.075)
Using the 'MCMC multiple imputation' calibration method, with consideration of the stratified survey design, adjusting for gender, employment status and psychotropic medication.	0.054 (0.036 to 0.072)
Using 'direct MCMC' calibration method, with consideration of the stratified survey design, but without use of covariate information.	0.054 (0.037 to 0.077)
Using 'direct MCMC' calibration method, with consideration of the stratified survey design, adjusting for gender, employment status and psychotropic medication.	0.055 (0.039 to 0.076)

is unrelated to any unobserved characteristics – although it might be related to the values of the observed measure, CIS-R.

The first stage in our multiple imputation technique was to build a model that predicts the missing binary outcome values (0,1) using the other covariates, based on the complete cases only. Secondly, *m* imputed data sets were generated by drawing *m* parameter vectors (*m* estimates of the parameters) from the Bayesian posterior distribution of the parameters, using Markov Chain Monte Carlo (MCMC) methods, namely the Gibbs Sampling algorithm, implemented by the WinBUGS computer package (Spiegelhalter et al. 2000).

Using these parameter estimates we produced an imputed data set by calculating for each case the probability π_i that the outcome variable (the SCAN assessment) for subject $_i$ was equal to 1. For the incomplete cases, imputed values were then created by imputing a '1' with probability π_i and '0' with probability $1 - \pi_i$. The prevalence estimate from each imputed data set were then calculated using the imputed values for incomplete cases and the observed values for complete cases. Standard formulae were then used to obtain from these separate estimates the multiple imputation estimate of the prevalence, together with a 95% confidence interval that incorporates our uncertainty due to the fact that the values were imputed rather than directly observed (Little and Rubin, 2002: 86).

Although in multiple imputation it is customary to use only between five and 10 imputed data sets, in this study over 90% of data for the SCAN assessment were missing, and the repeated use of only 10 imputed data sets gave highly variable estimates of the standard error, as we would expect. The WinBUGS package makes it easy to produce large numbers of imputed data sets, and for each analysis we therefore used 10,000 imputed data sets, produced in under 10 minutes with a 550 MHz personal computer. No formal tests for convergence were carried out, although examination of the 'trace' sequence of estimates showed no sign of any longer term variability in the estimates. See the appendix for details of the program code used in the analyses.

Analysis using MCMC methods 'directly'

The multiple imputation formulae are based on asymptotic theory, so prevalence estimates were then produced using a method that exploited the advantage of the MCMC approach. MCMC methods for obtaining the joint posterior distributions of the parameter estimates in non-trivial applications (Schafer, 1999) are preferred to asymptotic methods as the latter are

useful and reasonably accurate only when the joint posterior distribution is unimodal and approximately symmetric, a common example being the multivariate normal distribution.

Prevalence estimates were therefore obtained from the imputed data sets, by using imputed values for incomplete cases and probabilities (π_i) sampled from the posterior distribution for complete cases. This method has strong similarities to the multiple imputation approach, in that – at each iteration of the Gibbs sampler – for each individual for whom the SCAN was not assessed, a value is sampled from the predictive distribution for being SCAN positive. The posterior distribution for the prevalence, derived from the calibration data, is combined with the predictive distribution derived from the remaining survey data.

The 'direct MCMC' prevalence estimates presented in Table 4 are the median values of the 10,000 separate prevalence estimates, and the Bayesian Credibility Intervals (CI) are the 2.5 and 97.5 percentiles, corresponding to conventionally calculated confidence intervals. See the appendix for details of the program code used in the analyses.

Examples using artificial data

In addition to the above analyses, carried out on the ONS National Survey 2000 data, calculations using an MCMC-based simulation method were performed on a range of artificially created data sets to investigate the effect of the sizes of the calibration and survey data sets, the prevalence of the condition of interest, and the level of agreement between reference and survey interviews (for the survey considered in this paper, these would be the clinical and lay assessments, respectively) on the estimate of prevalence and, more importantly, the precision of the prevalence estimate. In addition, simulations were performed to explore the effect of the underlying prevalence rate on the estimate and on the precision of the prevalence estimate.

Results

As expected, the prevalence of the combined morbidity group of any anxiety, depression or other neurotic disorder was highest in the strata 1, 3 and 5 where the risk of a psychosis was judged to be highest, and the sensitivity of the CIS-R was also highest in these strata (Table 2).

The CIS-R assessments for the combined morbidity have the best sensitivity, kappa, and overall percentage agreement, but for depression, phobias and non-phobic anxiety there is a higher degree of specificity (Table 3).

It can be seen that the two MCMC-based methods, incorporating multiple imputation and the 'direct' method, give almost identical prevalence estimates and 95% confidence/credibility intervals (Table 4). Moreover, the precision of these prevalence estimates, as indicated by the narrowness of their 95% confidence/credibility intervals, is in the range 7% to 16% lower than that obtained using only the SCAN assessment data. The precision obtained by using the CIS-R assessments is much greater (three times better) than either of these, due to the larger amount of data available, but this estimate is biased because of poor agreement with the reference assessment. The importance of taking account of the stratified design of the survey is also made clear, as the prevalence estimated from the calibration method decreases from 7.6% to 5.4%, which is closer to the value that would generally be expected in community populations on SCAN for this group of diagnoses (Bebbington et al., 1997; Brugha et al., 2001), and also close to the estimate obtained from the SCAN data with allowance for the stratification The precision of this latter estimate (Table 4, first row) is nearly equal to that obtained by the more sophisticated calibration methods.

It is shown that, in this example, the adjustment by gender, employment, and the receipt of psychotropic medication had very little effect on either the point estimate of the prevalence, or on its precision, despite the fact that the level of agreement between CIS-R and SCAN assessments tends to vary with respect to these covariates. Further investigation in larger study populations, and the examination of a wider range of covariates, would be required to draw any conclusive inferences on the impact of covariates on the difference in diagnoses between the two interviews.

We examined the relation between observed and imputed SCAN assessments. The level of agreement between these is similar to that between the observed SCAN and the observed CIS-R assessments. This is to be expected, since the relationship between observed and imputed SCAN assessments is only 'through' their mutual relationship with the CIS-R, as defined in lines 1 and 2 of the WinBUGS code described in the appendix.

Using numerical MCMC-based analyses of 'ideal' data sets without adjustment for any covariates, the results in Table 5 show that the most important factors in increasing the precision of estimates made using

Table 5. Precision of prevalence estimates, calculated using an MCMC-based simulation method, using artificial data covering a range of sizes of calibration and main survey data sets, and of underlying levels of agreement between measures of assessment and of underlying prevalence

			Prevalence 2%		Prevalence 3%		Prevalence 5%		Prevalence 16%		Prevalence 25%	
Survey data set size	Calib. data set size	Agreement (specificity = sensitivity)	SD	95% CI	SD	95% CI	SD	95% CI	SD	95% CI	SD	95% CI
500	50	60%	0.020	0.000 to 0.066	0.025	0.000 to 0.083	0.032	0.000 to 0.012	0.051	0.062 to 0.264	0.061	0.138 to 0.373
500	100	60%	0.014	0.000 to 0.051	0.017	0.000 to 0.068	0.022	0.010 to 0.098	0.036	0.091 to 0.234	0.043	0.169 to 0.336
500	500	60%	0.006	0.008 to 0.033	0.008	0.016 to 0.046	0.010	0.032 to 0.070	0.016	0.129 to 0.192	0.019	0.213 to 0.288
500	50	75%	0.020	0.000 to 0.067	0.024	0.000 to 0.084	0.030	0.000 to 0.114	0.049	0.067 to 0.259	0.056	0.143 to 0.363
500	100	75%	0.014	0.000 to 0.051	0.017	0.000 to 0.067	0.022	0.011 to 0.096	0.035	0.094 to 0.231	0.040	0.173 to 0.329
500	500	75%	0.006	0.008 to 0.033	0.008	0.016 to 0.046	0.010	0.032 to 0.070	0.016	0.129 to 0.192	0.018	0.214 to 0.287
500	50	90%	0.021	0.000 to 0.068	0.024	0.000 to 0.083	0.029	0.000 to 0.112	0.040	0.082 to 0.240	0.043	0.163 to 0.334
500	100	90%	0.014	0.000 to 0.051	0.016	0.000 to 0.065	0.020	0.013 to 0.092	0.028	0.105 to 0.216	0.032	0.190 to 0.312
500	500	90%	0.006	0.009 to 0.033	0.007	0.016 to 0.045	0.009	0.033 to 0.069	0.014	0.132 to 0.189	0.016	0.218 to 0.283
1000	50	60%	0.020	0.000 to 0.064	0.025	0.000 to 0.086	0.031	0.000 to 0.119	0.052	0.061 to 0.266	0.062	0.137 to 0.378
1000	100	60%	0.014	0.000 to 0.050	0.017	0.000 to 0.068	0.022	0.010 to 0.097	0.037	0.092 to 0.236	0.043	0.167 to 0.335
1000	500	60%	0.006	0.008 to 0.034	0.008	0.016 to 0.046	0.010	0.032 to 0.070	0.016	0.128 to 0.193	0.019	0.212 to 0.288
1000	50	75%	0.020	0.000 to 0.068	0.024	0.000 to 0.083	0.030	0.000 to 0.118	0.049	0.069 to 0.260	0.056	0.144 to 0.361
1000	100	75%	0.014	0.000 to 0.052	0.017	0.000 to 0.067	0.021	0.011 to 0.096	0.034	0.097 to 0.231	0.039	0.174 to 0.329
1000	500	75%	0.006	0.008 to 0.033	0.008	0.016 to 0.045	0.010	0.032 to 0.069	0.016	0.131 to 0.191	0.018	0.214 to 0.286
1000	50	90%	0.020	0.000 to 0.067	0.024	0.000 to 0.084	0.029	0.000 to 0.111	0.039	0.084 to 0.239	0.042	0.168 to 0.333
1000	100	90%	0.014	0.000 to 0.050	0.016	0.000 to 0.064	0.020	0.013 to 0.091	0.028	0.107 to 0.215	0.030	0.193 to 0.310
1000	500	90%	0.006	0.009 to 0.032	0.007	0.017 to 0.045	0.009	0.033 to 0.068	0.013	0.134 to 0.186	0.015	0.221 to 0.280
5000	50	60%	0.020	0.000 to 0.065	0.024	0.000 to 0.084	0.031	0.000 to 0.120	0.052	0.064 to 0.271	0.061	0.136 to 0.373
5000	100	60%	0.014	0.000 to 0.051	0.017	0.000 to 0.068	0.022	0.010 to 0.097	0.037	0.090 to 0.234	0.043	0.169 to 0.335
5000	500	60%	0.006	0.008 to 0.033	0.008	0.016 to 0.046	0.010	0.032 to 0.070	0.016	0.129 to 0.193	0.019	0.214 to 0.288
5000	50	75%	0.020	0.000 to 0.068	0.024	0.000 to 0.085	0.031	0.000 to 0.118	0.048	0.071 to 0.260	0.056	0.141 to 0.360
5000	100	75%	0.014	0.000 to 0.052	0.017	0.000 to 0.067	0.021	0.011 to 0.096	0.035	0.095 to 0.231	0.040	0.174 to 0.329
5000	500	75%	0.006	0.008 to 0.033	0.007	0.016 to 0.045	0.010	0.032 to 0.069	0.015	0.131 to 0.191	0.017	0.216 to 0.284
5000	50	90%	0.020	0.000 to 0.065	0.024	0.000 to 0.083	0.028	0.000 to 0.110	0.038	0.085 to 0.236	0.042	0.169 to 0.332
5000	100	90%	0.014	0.000 to 0.051	0.016	0.000 to 0.065	0.019	0.014 to 0.091	0.027	0.107 to 0.213	0.029	0.193 to 0.306
5000	500	90%	0.006	0.009 to 0.032	0.007	0.017 to 0.045	0.009	0.034 to 0.068	0.012	0.136 to 0.184	0.013	0.223 to 0.276

multiple imputation in this way are (a) the size of the calibration data set and (b) the agreement between the two interviews (lay and reference). For a condition with low prevalence, within the ranges presented, the size of the calibration data set dominates all other considerations. This demonstrates the limited value of modelling the relationship between the SCAN and the CIS-R in the complete cases, and that we need to find methods that will predict the SCAN with as much precision as possible. If the link between the two assessments is weak, we should not expect much improvement in precision from the multiple imputation approach.

Extending the simulation analyses to cover low underlying prevalence rates, the effect of having a very large external calibration data set was examined in order to establish the relationship between the two interviews. These analyses show that an increase in agreement between the two instruments did not increase the precision, and in fact for some cases made it worse. Furthermore, if it is assumed that an infinitely large calibration data set is available, so that the probabilities of subjects being SCAN positive conditional on their CIS-R status is known without error, it can be shown that the variance of the prevalence estimate for the survey data set is proportional to the square of the difference between the positive and negative predictive values. As the agreement between the two interviews increases, the precision of the prevalence estimate for the survey dataset therefore decreases.

The simulation exercises showed that for low prevalence rates of 2% and 3%, the precision depends only on the size of the calibration data set, improving with increasing calibration size. It does not at all depend on the level of agreement between the two measures – it appears instead to be totally dependent on the size of the calibration data set. However, for prevalence rates of 5%, 16% and 25%, the precision increases both with increasing size of calibration data set and with increasing level of agreement.

Discussion

The main findings of our work is that the crucial factors in increasing the precision of estimates made using our methods in this way are the size of the calibration data set and the agreement between the two interviews. The relationship between the lay and reference interview is clearly important and it is therefore necessary to investigate whether improved prediction can be obtained by incorporating further covariates. This will require research in larger study populations. However, the application of a large calibration study to external survey data sets, in order to obtain precise estimates of the 'true' prevalence, does carry a number of underlying assumptions. These include the assumptions that the populations in the calibration study and the survey study are the same, and that the relationships between the two interviews are the same in both studies. It is also shown that, in a realistic setting, the 'MCMC multiple imputation' and the 'direct MCMC' methods give very similar results.

The relationship between the lay and reference interview is clearly important. Its further investigation may also explain why there is disagreement between lay and reference interviews for some subgroups, for example older men, who may be less psychologically minded, or less willing than women to acknowledge such difficulties. It is interesting to find that the kappa measure of agreement between SCAN and the CIS-R is highest in Stratum 5, made up of persons selected for clinical interview because they are likely to have psychosis in the absence of personality disorder (Table 2). This is the only stratum in which the number of true positives exceeds the number of false positives. Stratum 6, the stratum selected as being likely to have neither personality disorder nor psychosis, has the highest total proportion of agreement between the two measures – but, as this stratum has the lowest prevalence, we would expect relatively few mistakes in assessment, in either direction, to have been made. Previously Brugha and colleagues (1999b) have suggested that agreement will be better in respondents who are more likely to have had contact with psychiatric services where they could learn more about accepted usage of mental symptom terms. Lay interviews might be further improved by incorporating more instructions to the respondent on the meaning of such terms. For instance they could be asked to choose between example vignettes that contrast clinically significant with psychologically 'normal' mental experiences. For the current data, with a time lag of 3 to 6 months between administrations, the agreement between the instruments may be better for chronic or severe conditions, as opposed to acute or fluctuating conditions, due to possibly greater change in diagnosis over time. Thus the agreement between the two instruments might be improved by reducing the time lag between administration, which for the present study can be considered to be a limiting factor.

For reasons of cost and other resources, it is unlikely to be feasible to carry out a very large calibration study, and it is therefore important to know whether a number of smaller existing calibration data sets may be combined. The similarity between calibration data sets would be expected to depend on the methods of data collection, and the similarity in the aims of the underlying studies that produced these data. Such data sets should be collated from a variety of medical areas, so as to be available for the estimation of prevalence as described in this paper.

Although we have begun to look at how this methodology operates in the context of the low prevalence rates, which are commonly observed for individual psychiatric disorders, further research is needed before any firm conclusions can be made on the application of multiple imputation techniques.

We need also to consider whether the accuracy of prevalence estimation is improved by carrying out these methods on the CIS-R and SCAN assessments as ordinal variables, in which form these may be obtained from their respective algorithms, rather than binary diagnoses. Agreement between lay and clinical interviews appears to be better when ordinal or continuous scaling is compared with binary scaling but examination of ordinal depression measures from the SCAN and CIS-R interviews, revealed poor correlation between the two instruments (Brugha et al., 1999a).

Our findings have cast some light both on the potential and the limitations of the calibration method, particularly in relation to the size of current surveys and the number of clinical reappraisal interviews being carried out. A considerably greater number of such clinical interviews would need to be carried out in order to yield meaningful increases in the precision of estimates. One approach might be to train lay interviewers to conduct clinical interviews within field surveys (Brugha et al., 1999c) but at present this is likely to remain costly because of the considerable training involved and limited by the amount of data that can be collected by a small team of interviewers working over short periods of field work (Singleton et al., 2001). If the calibration method is not effective in addressing the challenges currently facing psychiatric epidemiology, the potential advantages of integrated clinical survey assessment methods in adult population surveys must be addressed all the more vigorously. Given the drawbacks of two-phase designs (Pickles et al., 1995), the advantage of having a more precise single-stage measure in such surveys would be considerable. Progress has begun with the use of such integrated measures to collect child and adolescent mental health information for policy development (Goodman et al., 2000). More work is therefore needed to reduce the economic cost and increase the feasibility of such applied clinical evaluation methods, including the contextual ratings of open-ended survey questions by clinicians working off site, or by incorporating telephone interviewing by clinical interviewers or advances in the development of pattern-recognition algorithms (Brugha et al., 1996).

For these reasons, it is important to ask whether this effort is necessary – that is, whether the differences between lay and reference interviews are of genuine importance? The wish to reproduce the clinical assessment process as closely as possible in lay interviews is seldom questioned, although it might be argued that it is largely theory driven (Brugha et al., 1999b). Although there is evidence that reference (clinical) interviews account for more variation in 'hard outcomes' – such as heritability estimates (Foley et al., 2002) – than do lay measures, more work is needed to explore the extent to which this difference matters not just to science but to policy. With regard to the latter, if reference measures can be shown to provide worthwhile advantages, then we could anticipate a number of potential national policy benefits. The present study provides estimates of anxiety or depression that are reduced following calibration but the method used does not help to define subgroups with greater need. If the present work were to bear significant fruit, the relationship between psychiatric morbidity and service use could be more precisely modelled, and could show stronger associations. Better information on the characteristics of those in the population likely to benefit from services would draw more attention and more resources to them.

References

American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition. Washington DC: APA, 1994.

Bebbington PE, Marsden L, Brewin CR. The need for psychiatric treatment in the general population: the Camberwell Needs for Care survey. Psychological Medicine 1997; 27: 821–34.

Best NG, Spiegelhalter DJ, Thomas A, Brayne C. Bayesian-analysis of realistically complex-models.

Journal of the Royal Statistical Society Series A-Statistics in Society 1996; 159: 323–42.

Brugha TS, Bebbington P, Jenkins R, Meltzer H, Taub NA, Janas M, Vernon J. Cross validation of a household population survey diagnostic interview: a comparison of CIS-R with SCAN ICD-10 diagnostic categories. Psychological Medicine 1999a; 29: 1029–42.

Brugha TS, Bebbington PE, Jenkins R. A difference that matters: comparisons of structured and semi-structured diagnostic interviews of adults in the general population. Psychological Medicine 1999b; 29: 1013–20.

Brugha TS, Jenkins R, Taub NA, Meltzer H, Bebbington P. A general population comparison of the Composite International Diagnostic Interview (CIDI) and the Schedules for Clinical Assessment in Neuropsychiatry (SCAN). Psychological Medicine 2001; 31: 1001–13.

Brugha TS, Nienhuis FJ, Bagchi D, Smith J, Meltzer H. The survey form of SCAN: the feasibility of using experienced lay survey interviewers to administer a semi-structured systematic clinical assessment of psychotic and non psychotic disorders. Psychological Medicine 1999c; 29: 703–12.

Brugha TS, Teather D, Wills KM, Kaul A, Dignon A. Present State Examination by microcomputer: Objectives and experience of preliminary steps. International Journal of Methods in Psychiatric Research 1996; 6: 143–51.

Clogg CC, Rubin DB, Shenker N, Shultz B, Weidman L. Multiple imputation of industry and occupation codes in census public-use samples using Bayesian logistic regression. Journal of the American Statistical Association 1991; 86: 68–78.

De Leeuw E. Item non-response: prevention is better than cure. Survey Methods Newsletter 1999; 19: 40–8.

Department of Health. National Service Frameworks for Mental Health. Modern standards and service models. London: Department of Health, 1999.

Dunn G. Comparative calibration without a gold standard. Statistics in Medicine 1998; 17: 1294–8.

Foley DL, Neale MC, Kendler KS. Genetic and environmental risk factors for depression assessed by subject-rated Symptom Check List versus Structured Clinical Interview. Psychological Medicine 2002; 31: 1413–23.

Frazer GE, Stram GO. Regression calibration in studies with correlated variables measured with error. American Journal of Epidemiology 2001; 154: 836–44.

Goodman R, Ford T, Richards H, Gatward R, Meltzer H. The Development and Well-Being Assessment: description and initial validation of an integrated assessment of child and adolescent psychopathology. J Child Psychol Psychiatry 2000; 41: 645–55.

Greenland S, Finkle WD. A critical look at methods for handling missing covariates in epidemiologic regression analyses. American Journal of Epidemiology 1995; 142: 1255–64.

Henderson S. Epidemiology of mental disorders: the current agenda. Epidemiol Rev 2000; 22: 24–8.

Jenkins R, Bebbington P, Brugha T, Farrell M, Gill B, Lewis G, Meltzer H, Petticrew M. The national psychiatric morbidity surveys of Great Britain – strategy and methods. Psychological Medicine 1997; 27: 765–74.

Kessler RC. The World Health Organization International Consortium in Psychiatric Epidemiology (ICPE): initial work and future directions – the NAPE Lecture 1998. Nordic Association for Psychiatric Epidemiology. Acta Psychiatrica Scandinavica 1999; 99: 2–9.

Kessler RC, McGonagle KA, Zhao S, Nelson CB, Hughes M, Eshleman S, Wittchen H-U, Kendler KS. Lifetime and 12-month prevalence of DSM-III-R psychiatric disorders in the United States. Results from the National Comorbidity Survey. Archives of General Psychiatry 1994; 51: 8–19.

Kessler RC, Wittchen HU, Abelson JM, McGonagle KA, Schwarz N, Kendler KS, Knauper B, Zhao S. Methodological studies of the Composite International Diagnostic Interview (CIDI) in the US National Comorbidity Survey (NCS). International Journal of Methods in Psychiatric Research 1998; 7: 33–55.

Lambert PC, Abrams KR, Sanso B, Jones DR. Synthesis of Incomplete Data using Bayesian Hierarchical Models: An Illustration Based on Data Describing Ssurvival from Neuroblastoma. Leicester: University of Leicester: Epidemiology and Public Health, 1997.

Lewis G, Pelosi AJ, Araya R, Dunn G. Measuring psychiatric disorder in the community: a standardized assessment for use by lay interviewers. Psychological Medicine 1992; 22: 465–86.

Little RJA, Rubin DB. Statistical Analysis with Missing Data. John Wiley: New York, 2002.

Martin J. An overview of imputation methods and their application to survey data. Survey Methods Newsletter 1999; 19: 9–11.

Meltzer H, Gill B, Petticrew M, Hinds K. OPCS Surveys of Psychiatric Morbidity in Great Britain. Report 1: The Prevalence of Psychiatric Morbidity among Adults Living in Private Households. London: HMSO, 1995.

Pickles A, Dunn G, Vazquez-Barquero JL. Screening for stratification in two-phase ('two-stage') epidemiological surveys. Statistical Methods in Medical Research 1995; 4: 73–89.

Regier DA, Kaelber CT, Rae DS, Farmer ME, Knauper B, Kessler RC, Norquist GS. Limitations of diagnostic criteria and assessment instruments for mental disorders. Archives of General Psychiatry 1998; 55: 109–15.

Rubin DB. Multiple Imputation for Nonresponse in Surveys. New York: Wiley, 1987.

Schafer JL. Multiple imputation: a primer. Statistical Methods in Medical Research 1999; 8: 3–15.

Singleton N, Bumpstead R, O'Brien M, Lee A, Meltzer H. Psychiatric Morbidity Among Adults Living in Private Households. London: The Stationary Office, 2001.

Spiegelhalter DJ, Thomas A, Best NG. WinBUGS User

Manual Version 1.3. Cambridge: MRC Biostatistics Unit, 2000.
Wing JK, Babor T, Brugha T, Burke J, Cooper JE, Giel R, Jablensky A, Regier D, Sartorius N. SCAN. Schedules for Clinical Assessment in Neuropsychiatry. Archives of General Psychiatry 1990; 47, 589–93.
Wittchen HU, Üstün B, Kessler RC. Diagnosing mental disorders in the community: a difference that matters? Psychological Medicine 1999; 29, 1021–7.
World Health Organization. The ICD-10 Classification of Mental and Behavioural Disorders: Diagnostic Criteria for Research. Geneva: WHO, 1993.
Zhou XH, Eckert GJ, Tierney WM. Multiple Imputation in Public Health Research. Statistics in Medicine 2001; 20: 1541–9.

Appendix: program code for Bayesian MCMC analysis using the WinBUGS computer package

```
model
{

# Section 1.

 for( j in 1 : 8569 )
# line 1
 {scan.pos[j] ~ dbern(prob.scan.pos[j])
# line 2
 logit(prob.scan.pos[j]) <-
gamma[1,stratum[j]]+gamma[2,stratum[j]]*cisr.pos[j]
# line 3
 scan.imputed[j] <-      scan.pos[j]*scan.miss [j] +
prob.scan.pos[j]*(1-scan.miss [j]) }

# Section 2.

 for (k in 1:6)
 {gamma[1,k] ~ dnorm(0.0,0.0001)
 gamma[2,k] ~ dnorm(0.0,0.0001) }

 prev <-      mean(scan.imputed[])
}
```

Annotation

Section 1

Line 1. For those 612 subjects with a recorded SCAN assessment, the value of 'scan.pos' is fixed as the observed value, for all other subjects it is sampled from a Bernoulli distribution (effectively, a binomial distribution) with the probability 'prob.scan.pos' to be defined in line 2. The values of scan.pos for each iteration represent the prevalence of each imputed dataset and are used in the 'MCMC multiple imputation' method.

Line 2. For each of the 8,569 subjects in the analysis data set, this line defines the relationship between the estimated probability of being SCAN positive (variable 'scan.pos') and the observed CIS-R diagnosis (variable 'cisr.pos'), specific for that subject's stratum. This data set was missing 11 subjects (not given the SCAN) from the total 8,580, due to missing covariate data used in corresponding analyses. In this example, there are just two 'gamma' parameters describing this relation, but there will be a larger number of these if adjustment is made for covariate information thought to be predictive of SCAN status.

Line 3. This line is needed only for the 'direct MCMC' method, which uses the median and 2.5 and 97.5 percentiles of the variable 'scan.imputed'. For those subjects without an observed SCAN assessment (as indicated in the data set by the variable 'scan.miss'=1), the imputed SCAN assessment 'scan.imputed' is taken from the Bernoulli sampling in line 1. For those subjects with an observed SCAN assessment, 'scan.imputed' is instead taken directly from the estimated probability of being SCAN positive – this is in order to reflect the random error inherent in that part of the total prevalence estimated derived from the known assessments.

Section 2

Here, for each stratum, it is specified that the parameters of the relationship between CIS-R and SCAN, for each stratum, are not already known – so that a Bayesian normal 'uninformative prior distribution' with mean zero and standard deviation 100 is used. (The second argument to 'dnorm' is the reciprocal of the variance.)

Finally, for the 'direct MCMC' method, the prevalence estimate is calculated as the mean of the imputed values for the SCAN assessment.

Correspondence: Traolach S Brugha, Department of Health Sciences, University of Leicester, Brandon Mental Health Unit, Leicester General Hospital, Gwendolen Road, Leicester LE5 4PW, UK.
Telephone (+44) 116 2256295.
Fax (+44) 116 2256235.
Email Tsb@le.ac.uc.