Guidelines for evaluating prevalence studies

As stated in the first issue of Evidence-Based Mental Health, we are planning to widen the scope of the journal to include studies answering additional types of clinical questions. One of our first priorities has been to develop criteria for studies providing information about the prevalence of psychiatric disorders, both in the population and in specific clinical settings. We invited the following editorial from Dr Michael Boyle to highlight the key methodological issues involved in the critical appraisal of prevalence studies. The next stage is to develop valid and reliable criteria for selecting prevalence studies for inclusion in the journal. We welcome our readers contribution to this process.

You are a geriatric psychiatrist providing consultation and care to elderly residents living in several nursing homes. The previous 3 patients referred to you have met criteria for depression, and you are beginning to wonder if the prevalence of this disorder is high enough to warrant screening. Alternatively, you are a child youth worker on a clinical service for disruptive behaviour disorders. It seems that all of the children being treated by the team come from economically disadvantaged families. Rather than treating these children on a case by case basis, the team has discussed developing an experimental community initiative in a low income area of the city. You are beginning to wonder if the prevalence of disruptive behaviour disorders is high enough in poor areas to justify such a programme.

Prevalence studies of psychiatric disorder take a sample of respondents to estimate the frequency and distribution of these conditions in larger groups. All of these studies involve sampling, cross sectional assessments of disorder, the collection of ancillary information, and data analysis. Interest in prevalence may extend from a particular clinical setting (a narrow focus) to an entire nation (a broad focus). In the examples given above, the geriatric psychiatrist needs evidence from an institution based study (narrow focus), whereas the child youth worker needs evidence from a general population study (broad focus).

In recent years, concern for the mental health needs of individuals in clinical settings has been broadening to include whole populations. This population health perspective has stimulated numerous prevalence studies of psychiatric disorder which are intended to inform programme planning, evaluation, and resource allocation. In general, the quality of these prevalence studies has been improving as a direct result of drawing on advances in survey methodology. In this note, the guidelines for evaluating prevalence studies arise from criteria applicable to community surveys.

Guidelines for evaluating prevalence studies

The validity of prevalence studies is a function of sampling, measurement, and analysis. Answers to the following questions (see box on page 38) can serve as criteria for assessing these features.

Sampling

(1) DOES THE SURVEY DESIGN YIELD A SAMPLE OF RESPONDENTS REPRESENTATIVE OF A DEFINED TARGET POPULATION?

A valid study enlists a sample that accurately represents a defined target population. Representativeness is a quality associated with the use of statistical sampling methods and careful evaluation of respondent characteristics.

Is the target population defined clearly?

A sample provides the means to obtain information about a larger group, called the target population. The target population must be defined by shared characteristics assessed and measured accurately. Some of these characteristics include age, sex, language, ethnicity, income, and residency. Invariably, subsets of the target population are too expensive or difficult to enlist because, for example, they live in places that are inaccessible to surveys (eg, remote areas, native reserves, military bases, shelters) or they speak languages not accommodated by data collection. These excluded individuals need to be described and their number estimated as a proportion of the target population. The requirements to define the target population and to identify systematic exclusions are necessary to give research consumers a basis for judging the applicability of a study to their question.

Was probability sampling used to identify potential respondents?

Probability sampling relies on the principle of randomisation to ensure that each eligible respondent has a known chance of selection; it requires that members of the target population be identified through a sampling frame or listing of potential respondents. This listing must provide access to all members of the defined target population except for exclusions acknowledged by the study authors. Probability sampling comes in a variety of forms from simple to complex. In simple random sampling, a predetermined number of units (individuals, families, households) is selected from the sampling frame so that each unit has an equal chance of being chosen. More complex methods may include stratified sampling in which a population is divided into relatively homogeneous subgroups, called strata, and samples selected independently and with known probability from each strata; cluster sampling in which a population is divided into affiliated units or clusters such as neighbourhoods or households and a sample of clusters selected with known probability; multistage sampling in which samples are selected with known probability in hierarchical order, for example, a sample of neighbourhoods, then a sample of households, then a sample of individuals; or multiphase sampling in which sampled individuals are screened and subsets are selected with known probability for more intensive assessment. The use of probability sampling is a basic requirement in prevalence studies.

Do the characteristics of respondents match the target population?

Non-response is the failure to enlist sampled individuals. If non-response is extensive and influenced by variables central to study objectives, it can lead to selection bias and estimates that deviate systematically from population values. When information is available on non-respondents, methods exist and should be used to evaluate selection bias.[1] In the absence of such information, sample respresentativeness must be evaluated by comparing the sociodemographic characteris-

Sampling

(1) DOES THE SURVEY DESIGN YIELD A SAMPLE OF RESPONDENTS REPRESENTATIVE OF A DEFINED TARGET POPULATION?

Is the target population defined clearly?

Was probability sampling used to identify potential respondents?

Do the characteristics of respondents match the target population?

Measurement

(2) DO THE SURVEY INSTRUMENTS YIELD RELIABLE AND VALID MEASURES OF PSYCHIATRIC DISORDER AND OTHER KEY CONCEPTS?

Are the data collection methods standardised?

Are the survey instruments reliable?

Are the survey instruments valid?

Analysis

(3) WERE SPECIAL FEATURES OF THE SAMPLING DESIGN ACCOUNTED FOR IN THE ANALYSIS?

(4) DO THE REPORTS INCLUDE CONFIDENCE INTERVALS FOR STATISTICAL ESTIMATES?

tics of respondents with those of the target population derived from a census or other relevant databases.

In clinical studies of treatment, prevention, prognosis, and quality improvement, ≥80% response has become the recommended minimum for follow up.[2] Although apparently fixed, this minimum standard is, in fact, variable because it fails to account for study to study variation in non-response at inception. The threshold for minimally acceptable response in prevalence studies should be set at 70% as long as the report shows that respondents and non-respondents, and/or the study sample and the target population, have similar important sociodemographic characteristics. Without evidence of comparability between respondents and non-respondents and/or the study sample and the target population, the minimum standard should be set at 80%.

Measurement

(2) DO THE SURVEY INSTRUMENTS YIELD RELIABLE AND VALID MEASURES OF PSYCHIATRIC DISORDER AND OTHER KEY CONCEPTS?

A valid study uses instruments that provide reliable and valid measurement. These are qualities that arise from the use of standardised data collection methods and that are confirmed empirically by measurement evaluation studies.

Are the data collection methods standardised?

Prevalence studies collect information for purposes of estimation (eg, frequency and distribution of psychiatric disorder) and hypothesis testing (eg, association between disorder and other variables of interest). To achieve these purposes, identical methods of assessment and data collection must be used with all respondents so that the information for analysis is completely comparable. Any deviation from a standard data collection protocol applicable to all respondents creates the potential for biased comparisons. Standardisation of method refers not only to eliciting information from respondents but also to interviewer training, supervision, enlistment of respondents, and processing of data.

Are the survey instruments reliable?

Reliability establishes the extent to which an instrument can discriminate between individuals. To evaluate reliability, data are collected to separate between individual differences that are real or actual (true variation) from ones that are unreal or artifacts of the measurement process (random variation). An informative empirical test of instrument reliability in prevalence studies is to give the survey instrument on two occasions, about 7–10 days apart (test-retest design), and to examine levels of agreement using κ, for cross classified data, and the intraclass correlation coefficient, for dimensional data.

Instrument reliabilities must be based on a sample derived from, or at least similar to, study respondents; they also need to include effects for all major sources of unwanted random variation. Respondent effects due to temporal fluctuations in memory, mood, and motivation are invariably present. There may also be interviewer effects arising from differences in presentation, competence, and impact and setting effects stemming from variability in the location and circumstances of data collection. If all 3 sources of unwanted variation were applicable in a study, then the test-retest design described above should take them into account.

Although there is no consensus on minimum standards for reliability, a good reason exists for setting them. Random variation in measurement leads to attenuation of effects (bias towards the null). Tolerating large differences in reliability between measures creates an unequal basis for comparing effects, and in the same study, this practice can lead to extremely biased inferences. To prevent the mindless analysis and reporting of associations for poorly measured variables, minimum reliability standards should be set at 0.60 (based on κ) for cross classified data and 0.70 (based on the intraclass correlation coefficient) for dimensional data.

Are the survey instruments valid?

Validity establishes the extent to which an instrument makes discriminations between individuals that are meaningful and useful. Evaluating instrument validity is analogous to testing hypotheses on substantive associations between measured variables, with one important difference: validity testing is done to confirm, not to add to, existing theory and knowledge. In the measurement of psychiatric disorder, this theory and knowledge come from clinical and epidemiological studies that have focused on aetiology, course, and response to treatment. Although the need to present evidence on instrument validity extends to all key variables, it is the assessment of psychiatric disorder which provides the focus here.

Efforts to validate structured interviews for classifying psychiatric disorder have been remarkably circumscribed. This is true in children for a variety of interviews[3] and in adults for the current recommended standard—the Composite International Diagnostic Interview.[4] The best of these studies usually compare assessment data generated by lay interviewers versus clinicians.

There has been no commentary on minimum validity standards for psychiatric instruments used in prevalence studies. The following are recommended here: (1) instrument content for measuring disorder (items and questions) should map into the operational criteria and symptoms contained in existing nosological systems (*International Classification of Diseases* and *Diagnostic and Statistical Manual*); (2) classifications of disorder should be based on compound criteria, including symptoms and evidence of impairment, distress, or disadvantage; (3) the identification of cases should derive from an explicit rationale that includes an external criterion and decision rules for discriminating between test positives and test negatives[5]; and (4) evidence should exist from head to head comparisons

with independent assessment data that the instrument meets specificity criteria (ability to distinguish among different categories of disorder).

Analysis

(3) WERE SPECIAL FEATURES OF THE SAMPLING DESIGN ACCOUNTED FOR IN THE ANALYSIS?

Complex sampling methods mean that eligible respondents will have different probabilities of selection. These methods introduce design effects—a term used by survey reseachers to indicate that the sampling method will have an impact on the calculation of variance estimates for testing hypotheses and determining confidence intervals. Complex sampling methods require the use of special statistical methods to obtain estimates that are unbiased and associated with the correct statistical precision.

(4) DO THE REPORTS INCLUDE CONFIDENCE INTERVALS FOR STATISTICAL ESTIMATES?

A primary objective of prevalence studies is to produce frequency estimates of disorder overall and for population subgroups. The usefulness of these estimates derives from the expected closeness between the unobserved value in the target population and the observed value in the sample. Confidence intervals quantify this closeness by telling us the chance, for example 95%, that the unobserved target population value will fall within a certain range of the observed sample value. Estimates in prevalence studies must be accompanied by confidence intervals or the information needed to calculate them.

Comment

The criteria presented in this commentary identify guidelines to evaluate the basic elements of prevalence studies: sampling, measurement, and analysis. The objective is to help the research consumer make informed judgments about the validity of a particular report. Basic guidelines are set to stimulate debate and further study. Although the criteria arise mostly from experience with prevalence studies done in general population settings, they extend to studies done in clinical settings, with one important caveat. In clinical settings, the question, "does the survey design yield a sample of respondents representative of a defined target population?" is largely unanswerable. It is difficult, if not impossible, to define the target populations that give rise to respondents sampled from clinical settings. The idiosyncracies of referral to mental health services render suspect the general applicability of prevalence estimates from one setting to the next. This issue needs further clarification as it raises an important question about the usefulness of publishing prevalence estimates from studies done in clinical settings.

MICHAEL H BOYLE, PHD
Department of Psychiatry, McMaster University
Hamilton, Ontario, Canada

1 Boyle MH. Sampling in epidemiological studies. In: Verhulst FH, Koot HM, editors. *The epidemiology of child and adolescent psychopathology.* Oxford: Oxford University Press, 1995.
2 Purpose and procedure. *Evidence-Based Mental Health* 1998 Feb;**1**:2–3.
3 Hodges K. *J Child Psychol Psychiatry* 1993;**34**:49–68.
4 Robins LN, Sartorius N. *International Journal of Methods in Psychiatric Research* 1993;**3**:63–141.
5 Zarin DA, Earls F. *Am J Psychiatry* 1993;**150**;197–206.

Some useful concepts and terms used in articles about treatment

One of the important principles of practising evidence-based mental health is that the results of research studies should be used to influence clinical decisions about a particular patient. The best quality evidence for making decisions about treatment comes from randomised controlled trials or from overviews of several randomised controlled trials such as a meta-analysis. The reason that a randomised controlled trial provides the best evidence is that, in most circumstances, randomisation avoids any systematic tendency to produce an unequal distribution of prognostic factors between the experimental and control treatments that could influence the outcome. It is important to remember that not all methods of allocation which are described as random are truly random, and even with true randomisation there may still be important differences at baseline between the groups due to small sample sizes. It is also important that those who are assessing outcome are blind to whether the patient received the experimental or control treatments. If there is a statistically significant difference in the rate of a favourable outcome or in change scores from baseline in the experimental group compared with the control group, then it is concluded that the treatment is "effective".

Evidence-Based Mental Health will only abstract treatment studies if the method of allocation is random, if there was adequate follow up of subjects entered into the trial, and if clinically important outcomes were reported. Unfortunately, there may not be a randomised controlled trial for each clinical question. If that is the case, then clinical decisions must be made on the basis of the best available evidence taking all relevant factors into account. Frequent replication of the intervention using different samples and outcome tools can add to the weight of the evidence in non-experimental designs.

Statistical significance versus clinical importance

Given evidence from a randomised controlled trial, statistical significance is not the only criterion for deciding whether to apply the results of a study. Statistical significance depends on the size of the difference between the groups, the amount of variation in outcome within the groups, and on the number of patients. Clinically trivial differences can be statistically significant if the sample size is sufficiently large. Conversely, clinically important differences can be statistically non-significant if the sample size is too small—that is, if the study lacks power. Clinicians need to evaluate statistical significance and clinical importance in interpreting the results of randomised controlled trials and meta-analyses.

Measures of clinical importance

How does one measure clinical importance? The usual estimate of clinical importance is the effect size; the size of the difference between the experimental and control groups. Whether the outcome is measured in a categorical way (eg, the prevention or treatment of "disorders" or the appearance of specific side effects) or in a continuous way (eg, mean symptom scores), the effect size reflects the difference between the experimental and control groups. Effect sizes tend to be smaller in randomised controlled trials than in non-experimental designs and smaller when there is adequate blinding or concealment of the intervention from the assessors of outcome.[1]

A common way of expressing effect size for categorical data is the relative risk (RR) or relative benefit (RB—depending on whether one is assessing a negative or positive outcome). The study by Kendall *et al* in this issue of *Evidence-Based Mental Health* (p 43) provides a good illustration of these points. Kendall *et al* report the results of a randomised controlled trial comparing cognitive behaviour therapy (CBT) with a waiting list control for children with anxiety disorders. 60 children were randomised to CBT and 34 to a waiting list control. After 8 weeks of treatment, 53% (32 of 60) of the children receiving CBT no longer met diagnostic criteria for their primary anxiety disorder compared with 6% (2 of 34) in the control group ($p < 0.001$). This difference is certainly statistically significant, but the p value tells us nothing about its clinical importance. One measure of clinical importance is the RB; that is, the probability of being free of anxiety disorder after 8 weeks of CBT compared with the probability of being free of anxiety disorder in the control group. Using data from the article, we can calculate that the RB is 9.1 or $32/60 \div 2/34 = 9.1$. In other words, anxious children receiving CBT are 9.1 times more likely to be free of anxiety disorder than children on the waiting list after 8 weeks. An alternative but similar statistic is to calculate the relative benefit increase (RBI) which is the proportional increase in rates of a good outcome between the experimental and control patients in the trial. It is calculated as the experimental group event rate (EER) minus the control group event rate (CER) divided by the CER or $(EER - CER) \div CER$. In this case, the RBI is $(32/60 - 2/34) \div 2/34$ or 8.07. In other words, there is roughly an 8 fold increase (800%) in rates of being free of anxiety disorder in the experimental compared with the control group. Attentive readers will notice that $RBI = RB - 1$ (the difference due to rounding), a relation that always holds.

This is a legitimate and popular way of reporting effect sizes but it has one serious limitation; it ignores the base rates in a study which could have a profound influence on the clinical application. Consider a situation in which the rate of improvement in CBT was 9% compared with 1% in the control group. With a large enough sample size, this difference could be statistically significant. The RB still equals 9 and the RBI is still roughly 8 or 800%. However, most would agree that the magnitude of the difference between the experimental and control groups is quite trivial, particularly if the treatment was expensive, difficult to deliver, or required considerable training (as CBT does, for example). In view of these limitations, it has been argued that RB and the similar RBI are not user friendly and do not provide the most clinically important information.

An alternative to these statistics that does take account of base rates is to consider absolute benefit increase (ABI) and, from this, the number needed to treat (NNT). The ABI is the absolute arithmetic difference in rates of good outcomes between the experimental and control patients and refers to the number of patients who benefit per 100 treated. It is simply calculated as rate of a good outcome in the experimental group minus the rate in the control group (in the study by Kendall *et al*, the ABI is $53\% - 6\% = 47\%$). Going one step further, the reciprocal of the ABI is the NNT—that is, the number of patients who need to be treated to achieve 1 additional good outcome. It is calculated as 1/ABI and in the study by Kendall *et al* the NNT is $1/.47 = 2.13$ which is rounded up to 3. In other words, 3 children need to be treated with CBT to achieve 1 additional good outcome over having a patient on a waiting list.

In *Evidence-Based Mental Health* (as in *Evidence-Based Medicine*), we have preferred to use terms such as ABI and NNT to capture the essence of clinically important differences.

What is a clinically important NNT?

The answer to this depends on the burden of suffering of the disorder as measured by prevalence, morbidity, and outcome; the economics and difficulty of the treatment procedure; and, finally, the cost of not treating the disorder. It is useful to compare the NNT in the study by Kendall *et al* with values obtained in other areas in medicine and in mental health. For example, in a meta-analysis by Hotopf *et al*,[2] 42 patients need to be treated with a serotonin specific reuptake inhibitor to prevent 1 additional discontinuation of treatment with a tricyclic antidepressant presumably due to side effects. Based on the data from Essali *et al*,[3] 37 patients need to be treated with clozapine to prevent 1 additional relapse on a typical neuroleptic, however only 6 patients had to be treated with clozapine to have 1 additional patient experience a "clinically important improvement". Thus the NNT in the study by Kendall *et al* is really quite impressive and if replicated means that an effective form of psychotherapy is now available for children with anxiety disorders.

So far we have just considered ways of expressing effect size using categorical data. Most treatment studies in mental health report changes in symptoms over time and between patient groups. With continuous data, the issue is more complicated but it is still possible to convert continuous measures into NNT. (More about this in a forthcoming issue of the glossary.)

Uncertainty and confidence intervals

One final point needs to be made. The statistics outlined above to estimate effect sizes are just that; they are estimates derived from a particular sample. The true value may or may not be exactly the same as the estimated value. There is a degree of uncertainty associated with these estimates and we can quantify that degree of uncertainty using confidence intervals. Altman provides a useful definition of confidence intervals as "the range of values within which we can be 95% sure that the population value lies".[4] In the example used above from the study by Kendall *et al* we can be 95% certain that the true NNT is between 2 and 4 to produce one more child free of anxiety disorder using CBT. In the study by Hotopf *et al* the 95% CI is between 24 and 148.[2] Because the degree of uncertainty is such an important variable in comparing results from different studies, we will also provide CIs around estimates of ABI and NNT even if these are not provided in the article itself.

We hope that these tools derived from clinical epidemiology will be helpful to clinicians in translating the results of treatment interventions into clinical practice. Future issues of the notebook will explain terms used in prognosis studies and studies of causation and cost effectiveness among others. We welcome the feedback of our readers on these and other topics.

PETER SZATMARI, MD
Editor, Evidence-Based Mental Health

1 Schulz KF, Chalmers I, Hayes RJ, *et al. JAMA* 1995;**273**:408–12.
2 *Evidence-Based Mental Health* 1998 Feb;**1**:21. Abstract of: Hotopf M, Hardy R, Lewis G. Discontinuation rates of SSRIs and tricylcic antidepressants: a meta-analysis and investigation of heterogeneity. *Br J Psychiatry* 1997 Feb; **170**:120–7.
3 *Evidence-Based Mental Health* 1998 Feb;**1**:17. Abstract of: Essali MA, Rezk E, Wahlbeck K, *et al.* Clozapine *v* "typical" neuroleptic medication for schizophrenia. In: *Cochrane Database of Systematic Reviews*, [updated 04 March 1997]. In the Cochrane Library [database on disk and CD-ROM]. The Cochrane Collaboration; issue 2. Oxford: Update Software, 1997.
4 Altman DG. *Evidence-Based Medicine* 1996 May-Jun;**1**:102–4.