Overcoming obstacles to routine outcome measurement

The nuts and bolts of implementing clinical audit*

ISAAC MARKS

The call for evidence-based health care is widespread. It fuels Cochrane reviews of randomised controlled studies (RCTs). How far the results of RCTs apply to everyday care, however, cannot be judged without the regular measurement of outcomes in daily practice too. Few mental health care providers gather data regularly concerning the outcome of routine care as opposed to research treatment, analyse the data, and apply the results to improve practice (Walter *et al*, 1996). Few purchasers of care engage in the dialogue with clinicians required to evolve workable clinical audit that is meaningful to both clinicians and purchasers. Few research funders, editors and referees prioritise its study.

Measurement of the outcome of everyday care is a poorly funded area that is looked down on compared with the supposed gold standard of RCTs. RCTs are indeed vital. So is the study of the outcome of routine care. The results of RCTs and of routine care are two sides of the same gold coin. Each deserves equal scientific status and funding to yield its own kind of essential information. Both are needed to practise efficient evidence-based health care.

Many obstacles block the routine rating and display of outcome in regular care. Some of the barriers can be reduced by hard work from individual clinicians and managers (Smith *et al*, 1997*a*,*b*). Dealing with other hurdles needs action by professional and government organisations. I would like to outline some steps which have to be taken before evidence-based mental health care can become the rule, rather than the exception, for the average practitioner who does not have the luxury of research personnel to gather the data required. It draws on hard-won lessons from decades of practice of the boring business of implementing audit.

A NON-MEDICAL EXAMPLE OF OVERCOMING OBSTACLES TO AUDIT

Issues in industrial audit bear on clinical audit (Toole, personal communication). A small company supplies bolts to an aircraft manufacturer. Each bolt is numbered. The bolt supplier records for the aircraft company an audit trail for each bolt. The trail starts with the smelting of the metal out of which the bolt was made and leads on to the cutting of the bolt out of a particular metal rod, through to other stages to delivery of the numbered bolt to the aircraft assembly line. Each bolt must have its testing documented by international standards – non-destructive tests for the numbered bolt and destructive tests for the batch it comes from.

The audit trail of each bolt crosses five companies: (a) the metal smelter and rod manufacturer; (b) the metal rod stockholder; (c) the company that cuts a bolt out of a rod and cuts a thread on it to international standards; (d) another company that heat-treats (anneals) the threaded bolt; and (e) the aircraft-assembling factory itself. Each bolt is traceable through a computerised audit process that is facilitated by a specially written complex manual. Implementing the computerised audit process took four person-years of work and many intercontinental journeys.

Many bolts fail their tests and are discarded. An untested unaudited bolt costs just a few pence. A bolt with a strict audit history, however, costs far more to satisfy the need for aircraft safety. In an aeroplane crash any of its parts must be traceable. That is why aeroplanes are usually safe.

Evaluation in the aircraft industry has agreed standards that cost much to implement. Companies which fail to meet the exacting rules go out of business. The mental health care industry, in contrast, has few widely agreed standards, spends very little on outcome evaluation, and levies no charge on clinicians who do not evaluate outcome. Small wonder that good clinical audit is so rare.

The obstacles to measuring clinical outcome can be divided into two kinds: those impeding the introduction of any type of outcome measurement regardless of whether it is done by pen and paper or by computer, and other obstacles which mainly concern ratings aided by a computer.

REDUCING OBSTACLES TO CARRYING OUT ANY OUTCOME RATINGS

Few agreed, simple and valid outcome criteria

There are few widely agreed measures of clinical efficacy and efficiency that are valid and quick to make. Getting a consensus on this requires a push towards international agreement despite the different ideological perspectives that yield differing criteria of clinical outcome and how to cost its attainment. Today's ease of phoning the other end of the world only became possible after many technical experts met repeatedly over decades to agree standards for telephony.

Consensus is required on the minimum measures to use with all patients and with particular kinds of patients. It may be possible to agree on simple ratings of work and social disability and of global impression that have proved their worth. Others such as Health of the Nation Outcome Scales (HoNOS) (Taylor & Wilkinson, 1997) are widely used in the UK and Australia.

Regarding the cost of achieving a clinical outcome, clinician time spent per patient is not hard to estimate. After any activity concerning a patient's care (e.g. face to face or phone interview with that patient or a relative, or writing a letter about the patient) it is possible to note down how long it took and add that time to the previous times spent on that patient's care; this yields a cumulative clinician time for that patient. Clinician time spent on a patient is easily

*Based on a paper originally presented at the 'Second "Getting Better: Outcome Measurement and Resource Allocation in Psychiatry" Conference', Sydney, 28 March 1998.

converted to a rough cost. Similarly lawyers and accountants add up time spent to know how much to charge their clients.

For costs other than time, we can create and consult simple lists of the cost of various drugs, investigations and overheads (ignoring pence and paper clips) and from these quickly jot down the rough costs of these for a given patient to date. Adding these to the rough cost of clinician time yields a useful estimate of the cost to achieve a given clinical outcome for a given patient. All this, however, assumes that we are willing to spend a minute recording such information each time we have done something for a patient. Audit has a cost.

Before we can rate outcomes meaningfully many seemingly obvious terms need defining. Is 'pretreatment' when a patient was first given an appointment or screened or accepted for treatment or began the first treatment session? What is a 'session'? A brief phone chat explaining procedures or a few minutes of advice or a 50-minute interview? What is a 'mid-treatment' rating? It may be after two days of alcohol detoxification or after a month for depression or obsessive–compulsive disorder (OCD) or never for chronic schizophrenia or Alzheimer's disease, where care never ends. How does 'discharge' differ from 'follow-up'? Defining such terms is a chore. Without it, however, sound clinical audit is impossible.

The time that clinicians need to rate and record outcome information for just one patient on one occasion is one reason why so few do it as a routine. It takes over a minute to reliably rate and store even a single-item global impression scale. To rate and store a set of 10 items may take 15 minutes. To get reasonable validity clinical status has to be rated at least twice – just before and just after treatment, and if possible a third time at follow-up. Repeating even brief sets of ratings three times per patient from the start of treatment to the end of follow-up may add 45 minutes (3×15 minutes) to the clinician's total treatment time. A further 45 minutes may be eaten up in plotting successive ratings on 'psychiatric temperature charts' to show outcome over time at a glance.

Lack of incentives to measure outcome

Because even the simplest way of rating outcome and its cost costs time and therefore money, incentives are needed to do so. Rewards to measure outcome are fewer than the exhortations to base health care on evidence. Some rewards are easier to arrange than others.

Senior staff can encourage juniors to make simple outcome ratings on every patient and use those ratings to help decide about patients' ongoing care. When senior staff stop asking for clinical outcome data then junior staff tend to stop collecting them, as has been found repeatedly.

The main reason why my own staff bother to measure outcome and cost is that from the time they join the unit they are asked to do it as a routine and are trained how to do it. 'Psychiatric temperature charts' showing the outcome ratings for a given patient form a part of clinical decision-making at every in-patient ward round and out-patient review. When discussing a patient the keyworker shows that patient's outcome charts to the team gathered round. The outcome charts help us decide on further care of the patient. Discrepancies between what the outcome charts show and what staff report can suggest ideas for further care.

The Royal College of Psychiatrists and its equivalents in other countries give trainees little credit for learning about and performing outcome measurement. Unlike basic science, phenomenology, diagnosis, treatment and medico-legal issues, outcome measurement is not taught or tested in examinations. It would work wonders if outcome measurement became an obligatory part of professional training requirements and examinations.

Solo practitioners and clinical teams would feel more inclined to go to the trouble of rating outcome if they were rewarded for doing so. A New Zealand health insurance organisation now pays a practitioner in Christchurch for each clinical outcome and resource measure (CORM) rating chart supplied. Australian and UK purchasers of mental health care are increasingly asking providers for evidence that their treatment is effective, but do not pay clinicians or units for the work needed to collect that evidence. Even simple outcome measurement can add 10% to a clinician's total time spent per patient. The cost of clinical audit is well worth it if staff act on its findings so as to increase efficiency. In the long-run not evaluating is more expensive than evaluating because inefficient care is then harder to spot.

A crucial and as yet unsolved question is how to persuade funders to reward clinicians for producing data about the outcome and cost of their treatments. Once funders give such rewards, ongoing random checks would have to be made to verify the validity of the ratings being supplied by clinicians, say on a 1% sample of patients.

After clinicians are persuaded to make outcome and cost ratings in day-to-day practice, how do they store, retrieve and analyse the growing mountain of data that accumulates in their files? Over the 17 years from 1972–1989 my unit used hand-written forms to collect outcome and cost data as a routine. By 1989 we had data on 2500 patients. A per-patient mean of four rating occasions over six months, each with five paper sheets of ratings containing a total of 40 items rated, yielded a total of $4 \times 5 \times 2500 = 50\,000$ pieces of paper and $4 \times 40 \times 2500 = 400\,000$ numbers to store, retrieve, enter into a computer and analyse. The task grew ever more formidable as the years went by, requiring major organisation, time and money. Worse still, this led to years of delay in getting feedback of the outcomes to act on and hone clinical practice.

The tedious task of analysing my unit's written results did yield meaningful data affecting clinical practice. One example is a finding that changing the unit's way of treating obsessive ruminations from 1987 onwards led to more improvement than had been obtained before (Lovell *et al*, 1994). This justified continuation of the new method of helping obsessions.

Another finding, that self-exposure treatment was as good as therapist-accompanied exposure, allowed an in-patient unit to shed its night staff and become a hostel unit staffed only by day (Thornicroft *et al*, 1991). Analysis of the routine outcome ratings that continued to be made thereafter found that patient outcome remained as good as before despite the major reduction of staffing costs.

REDUCING OBSTACLES TO RATING OUTCOME WITH A COMPUTER AID

The tasks of rating outcome and storing, retrieving and analysing the ratings are repetitive and so should be capable of being speeded up by using computers. Computer systems, however, only help when we know what we want from them and design them to fit the way we work. In trying to design that fit we discover that what we do in

practice differs from what we think we do. Without even thinking about it we take many little steps in making even the simplest handwritten ratings. A computer aid must take each little step into account. Moreover, each clinical unit works in its own way. The computer system has the daunting task of having to be a tool for routine care by clinicians who are practising in varying settings and have few incentives to learn about and measure outcomes.

Only when clinicians have met the conditions for success of non-computerised audit is it sensible for them to consider using computers to speed the process. Staff should have first agreed which measures to use and at which times before, during and after treatment, been trained and given time to rate their patients, used the ratings for clinical decision-making, and if students, shown their ratings to seniors when discussing care plans. Even the best computer system will languish unused if clinicians do not use the ratings made on it to help patients' ongoing care or for reimbursement or some other reward. Computer systems are less forgiving than handwritten ones. They only print out useful reports once ratings have been entered properly by clinicians who understand what the data are needed for.

It takes a clinical unit at least a year to implement outcome measurement to the point where clinicians do it as a routine and regard it not as a non-clinical nuisance, but rather as being as much a part of clinical work as is an initial assessment interview. Implementation of any audit imposes a way of working and values implicit in the measures chosen and rated.

Units which have implemented computer-aided audit successfully have had an ongoing dialogue between clinical and information technology (IT) support staff. Such dialogue may involve contacts face to face, by individual and conference phone calls, by e-mail, and by Fax. Success took over 50 clinician–IT staff contacts before and after the CORM system (Marks, 1996) was installed in some units. Dialogue was possible even when the clinical and the IT staff were at opposite ends of the world. Distance is no longer critical. Dialogue is also needed within the clinical unit itself, as my in- and out-patient staff found when clarifying misunderstandings about what, when and how to rate outcome.

Though computer-aided outcome measurement is a clinical tool, many clinicians are prevented from using it by their hospital's IT department which perceives its own non-clinical needs. This turf issue is vital. Clinical computing systems are tools for clinical use that must be clinically led, not IT led, despite their need for IT support.

Some problems can be bypassed if a clinical unit can access funds to buy and support its computer system independent of its hospital's IT department. Against stiff opposition from my hospital's then IT director I bypassed a start-up problem on realising that starting computer-aided outcome rating was a clinical-research issue that could be initially paid for and have outside IT support on my research budget outside the hospital's control. Thereafter, ongoing negotiations with a new enlightened hospital IT director eased the situation. IT versus clinician battles and other conflicts impede the advance of clinical computing systems in many units and countries. Delicate negotiation may resolve some of these conflicts when the warring parties come to see the advantages of cooperation.

Staff will not use a computer unless they know what to use it for, are trained to use it, get time to use it and can get to it quickly to enter ratings while seeing the patient. If staff encounter difficulties they need quick access to help to set things right. Staff should not be asked to duplicate the entry of data that have already been entered elsewhere. The unit needs to arrange for the data to be backed up frequently and stored outside the unit. The unit has to guard against strikes by lightning, theft of computers and unauthorised access to data.

The audit process itself requires audit in two ways. First, someone in the unit has to be given the responsibility of checking at intervals that staff are entering all the desired data properly. Second, the validity of samples of the data needs testing from time to time against other information obtained from interviews, the case notes and case discussions.

The computer system must work well and simply, fit the unit's way of working, and yield data in a format that staff understand easily. It should not increase the time that staff spend in rating and recording the outcome of their patients over and above the time they would spend doing the same by pen and paper.

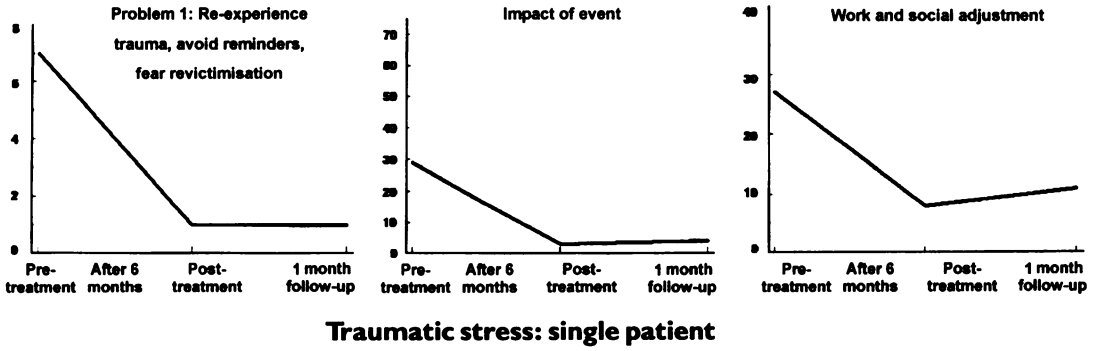

Fig. 1 CORM 'psychiatric temperature chart' from I.M's Traumatic Stress Service showing outcome (lower scores=improvement) of a single patient on a patient-specific measure (Problem 1, assessor-rated), a syndrome-specific measure for traumatic stress (Impact of Event Scale, self-rated) and a generic measure given to all patients (work/social adjustment, self-rated).

BENEFITS OF COMPUTER-AIDED RATINGS

The grind of rating outcome of individual patients is eased somewhat by entering the ratings directly into suitable computer-aided systems rather than rating by pen and paper. Use of the CORM system in my unit slightly speeds up reratings and printing out of the outcome graphs (charts).

Other benefits flow from computer-aided audit with individual patients too. Some staff find that letting a patient see the outcome charts on screen or printed out focuses discussions on the care plan, especially if the measures include problems and goals tailored to that patient's needs (see Fig. 1). Patients occasionally ask to take the charts home to show their family and use as a guide to doing self-treatment homework. Some patients confide more honest ratings to a computer than to a human regarding sensitive areas.

Having data from many patients amassed in a computer can greatly speed their analysis if it has been planned for. Instead of being on thousands of sheets of paper the data are in one or a few machines only and are thus more easily aggregated for analysis. Analysts of data sets can learn to produce means, standard errors and graphs, for example, how much did the last 50 out-patients with major depression improve, and how much did it cost to obtain that improvement?

Data are rarely entered completely cleanly into a computer and need cleaning before analysis. Moreover, the report of aggregate output, as from individual patients, must be interpreted with an understanding of what the data mean and how raters entered or did not enter them (Fig. 2 shows ways of portraying outcome of aggregated patients in my Traumatic Stress Unit). CORM data can be accessed by SPSS, SAS, Microsoft Excel etc. Whoever does the detailed analysis must learn to work with large databases. Ratings stored appropriately on a computer can be analysed within days or weeks rather than the months or years it took previously when they were made on paper and special personnel had to be paid to enter and clean them before analysis could proceed.

CURRENT STATUS OF COMPUTER-AIDED OUTCOME-RATING

There are many systems to allow mental health ratings to be entered into a computer (Kobak *et al*, 1996). Most merely allow the rating of one or two measures, for example, the Beck or the Hamilton Depression Scale (Beck & Beamcoderifer, 1974; Hamilton, 1967). Others are engines to allow the rating and display of a wide variety of measures.

The majority of computer-aided rating systems are for standalone computers. Some are robust enough to use in local area networks. Few are yet used on wide area networks. The wider its use on a network the more robust the system must be and the greater the cost of maintaining the network.

The CORM system allows the rating and display of many kinds of measure. It is used at the Bethlem Maudsley Hospital Trust in London: (a) in my behavioural anxiety disorders unit on a network of 14 computers for out-patients and on another network of five computers for in-patients so that all clinicians and patients can quickly make outcome ratings on computers in patient-interviewing rooms as needed (the unit's network worked well on its own, but linking it to the Trust network created problems that took enormous effort to sort out); (b) in two traumatic stress units (one behavioural, one psychodynamic) in each on a standalone computer; (c) it is about to be installed in a unit for patients who commit self-harm.

In Australasia CORM is available in Christchurch, New Zealand on a network of six machines in a counselling/general practitioner centre for somatisation and other disorders, in Goulburn, New South Wales, Australia on a network of seven machines for a community team treating serious mental illness, and in Adelaide on two standalone machines for anxiety disorders and compulsive gambling. In Holland CORM has been translated into Dutch and psychologists are piloting it in two out-patient units in Nijmegen. In Germany a hospital piloted CORM in English and there are plans to translate it into German.

The use of CORM has met the same political tangles as the use of any other audit and computer system. Its very first site was a private hospital in London and staff in two units were trained to use it. The hospital won a big clinical contract partly because it was rating outcomes. After CORM had been used for a year the

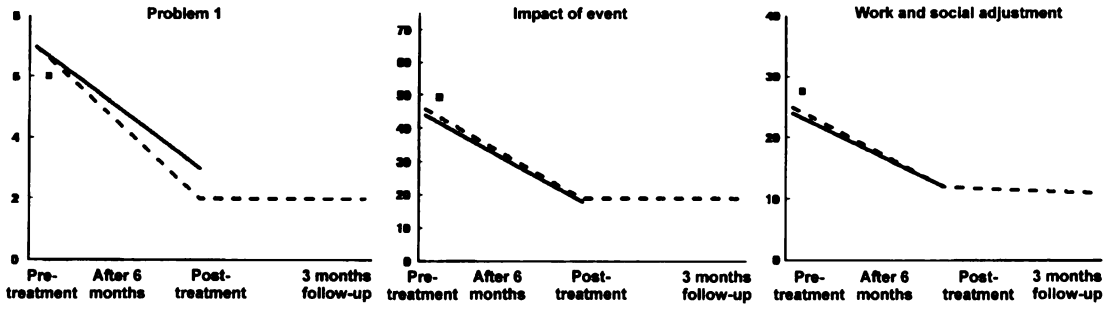

Fig. 2 CORM-type chart of outcome (lower scores=improvement) of all patients completing treatment in I.M.'s Traumatic Stress Service. Measures are the same as in Fig. 1 – *patient*-specific (Problem 1), *syndrome*-specific (Impact of Event Scale), and *generic* (work). $P < 0.001$ for improvement since baseline. Score 0=normal, highest score=severe. —, completed treatment but not follow-up (n=34) - - -, completed treatment and follow-up (n=23); ■, drop-outs (n=18).

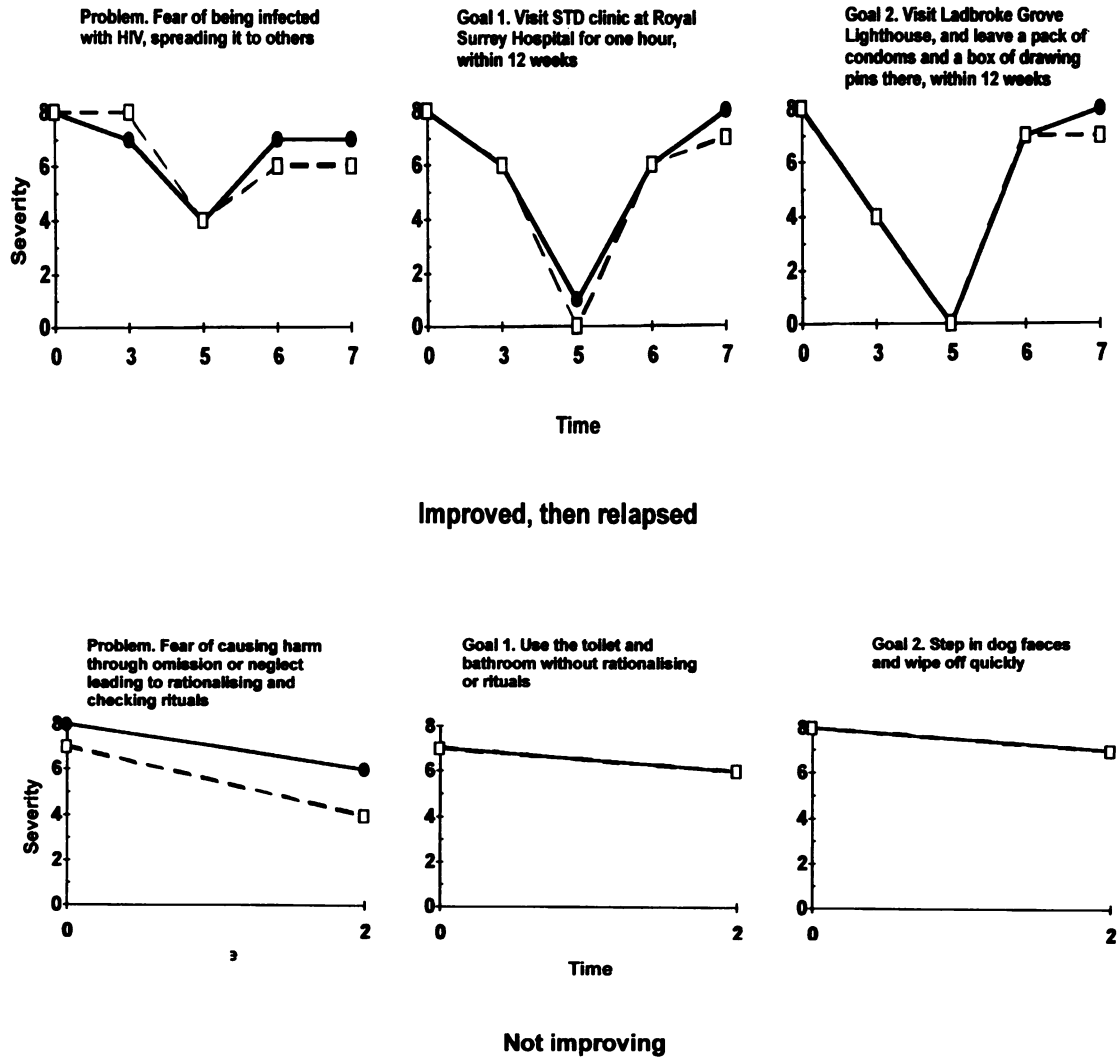

Fig. 3 CORM-type charts show at a glance whether a patient has relapsed (scores go up, as in top three graphs) or has not improved (scores change little, as in bottom three graphs). Each patient in the top and bottom three graphs had obsessive–compulsive disorder. Lowest score=normal, highest score=severe. □- -□, patient's rating; ●–●, clinician's rating.

political agenda changed and the hospital suddenly transferred elsewhere all the clinicians who had been trained to use it. No-one was left who knew how to use the system. CORM was also bought by a clinician in charge of audit in a London teaching hospital, but that hospital's IT department forbade its use.

At least two computerised forms of HoNOS are available in the UK and Australia but are not yet widely used to track outcome in routine clinical practice. Further computer-aided systems to rate one or two measures are in use in many countries.

Certain computer-aided rating systems are used in North America and Europe that are accessed by phone using interactive voice response (IVR). The clinician and/or the patient use a touchtone phone to call the computer (Greist *et al*, 1997, 1998). The user's keypresses on the telephone keypad drive prerecorded segments of natural-voice speech to produce a computer interview. A US-managed care organisation is using IVR screening of its patients before they see a doctor, and gives doctors the resulting printed output when they see a patient. A UK company facilitates IVR ratings for drug trials in many countries. IVR outcome-tracking systems were built into IVR self-help systems to reduce smoking (Schneider *et al*, 1995), OCD (Greist *et al*, 1998) and depression (Osgood-Hynes *et al*, 1998). The IVR system for OCD is called BTSTEPS. It is now in daily use in a regular National Health Service clinic at the Maudsley Hospital in London and in an eight-site RCT in the US and Canada.

The spread of computer-aided measurement of outcome is opening up a tantalising possibility – electronic data exchange across sites and countries. An international clearing-house of clinical outcomes and their cost is within sight. It could help clinicians from different places to compare clinical outcomes and costs with similar types of patients and to refine their practice in the light of those results.

IN BRIEF

The regular tracking of patient outcome and its cost is still a rare activity in everyday mental health care but is slowly spreading with the aid of computer systems. Many obstacles impede the spread of such outcome measurement, whether handwritten or aided by a computer. Some of the barriers can be reduced by the actions of individual clinicians. Overcoming other hurdles requires action by professional and government organisations. Patients will be more likely to receive cost-effective treatment once clinicians regularly measure and analyse outcomes and costs in routine care.

ACKNOWLEDGEMENTS

We thank Stuart Toole, who made many valuable comments on the manuscript and, along with Paul Byrne, played a leading role in developing CORM. We also thank Jonathan Ash, Brendan McLoughlin and Drs Candida Graham, Nigel McKenzie and Susan Shaw for help in implementing computer-aided audit in my unit, to Maria Livanou for the outcome charts from my Traumatic Stress Unit, to the many staff who over the years entered outcome data and overcame teething troubles and to Gus Heafield for much help with my unit's network.

ISAAC MARKS, FRCPsych, Director, Clinical Computing Unit, Institute of Psychiatry, Bethlem Maudsley Hospital Trust, London SE5 8AF

(First received 18 May 1998, final revision 23 June 1998, accepted 23 June 1998)

REFERENCES

Beck, A. T. & Beamcoderifer, A. (1974) Assessment of depression: the depression inventory. In *Psychological Measurements in Psychopharmacology: Modern Problems of Pharmacopsychiatry*, Vol. 7 (eds P Pichot & R. Olivier-Martin). Basel: Karger.

Greist, J. H., Jefferson, J. W., Wenzel, K. W., *et al* (1997) Telephone Assessment Program: Efficient patient monitoring and clinician feedback. *MD Computing*, **14**, 382–387.

—, Marks, I. M., Baer, L., *et al* (1998) Home self-exposure therapy of OCD by a manual plus computer-driven phone interview: A US–UK Study. *MD Computing*, **15**, 149–157.

Hamilton, M. (1967) Development of a rating scale for primary depressive illness. *British Journal of Social and Clinical Psychology*, **6**, 278–296.

Kobak, K. A., Greist, J. H., Jefferson, J. W., *et al* (1996) Computer-administered clinical rating scales: A review. *Psychopharmacology*, **127**, 291–301.

Lovell, K., Marks, I. M., Noshirvani, H., *et al* (1994) Should treatment distinguish anxiogenic from anxiolytic OC ruminations? Results of a pilot controlled study and of clinical audit. *Psychotherapy and Psychosomatics*, **61**, 150–155.

Marks, I. M. (1996) A computerized system to measure treatment outcomes and cost. *Psychiatric Services*, **47**, 811–812.

Osgood-Hynes, D., Baer, L., Greist, J., *et al* (1998) Computer-aided telephone administered behavioral treatment of depression. *Journal of Clinical Psychiatry* (in press).

Schneider, S. J., Schwartz, M. D. & Fast, J. (1995) Computerized, telephone-based health promotion: Smoking cessation program. *Computers in Human Behavior*, **11**, 135–148.

Smith, G. R., Manderscheid, R. W., Flynn, L. M., *et al* (1997*a*) Principles for assessment of patient outcomes in mental health care. *Psychiatric Services*, **48**, 1033–1036.

—, Fischer, E. P., Nordquist, O. R., *et al* (1997*b*) Implementing outcomes management systems in mental health settings. *Psychiatric Services*, **48**, 364–368.

Taylor, J. R. & Wilkinson, G. (1997) HoNOS *v.* GP opinion in a shifted out-patient clinic. *Psychiatric Bulletin*, **21**, 483–485.

Thornicroft, G., Colson, L. & Marks, I. M. (1991) An in-patient behaviour therapy unit: Description and audit. *British Journal of Psychiatry*, **158**, 362–367.

Walter, G., Kirkby, K., Marks, I. M., *et al* (1996) Outcome measurement sharing experiences in Australia. *Australasian Psychiatry*, **4**, 316–318.