

The quality and reliability of health information on the internet remains of paramount concern in Europe, as elsewhere. Self regulatory codes of ethics for health websites abound, yet the quality and practices of many are highly questionable.

Little progress seems to have been made, moreover, in assuring consumers that the information they share with health websites will not be misused. Several US studies have already concluded that websites' privacy practices do not match their proclaimed policies.⁵ In an attempt to counter this erosion of trust in Europe, the European Commission's guidelines for quality criteria for health related websites have recognised that there is no shortage of legislation in the field of privacy and security.⁶ They have drawn specific attention to a new recommendation regarding online data collection adopted in May 2001 that explains how European directives on issues such as data protection should be applied to the most common processing tasks carried out via the internet.⁷

The challenge facing Europe's health professionals and policymakers is to carefully craft the development of new approaches to the supervision of medical and pharmaceutical practice. Their ultimate goal is to raise

consumers' confidence in online healthcare. They must ensure that the mechanisms are put in place whereby health professionals themselves can benefit from using the internet, while still ensuring the highest standards of medical practice.

Avienda was formerly known as the Centre for Law Ethics and Risk in Telemedicine.

Competing interests: None declared.

- 1 http://news.bbc.co.uk/1/hi/english/uk/england/newsid_1752000/1752670.stm (accessed 5 Feb 2002).
- 2 Case C-322/01: Reference for a preliminary ruling by the Landgericht Frankfurt am Main by order of that court of 10 August 2001 in the case of Deutscher Apothekerverband e.V. against DocMorris NV and Jacques Waterval. *Official Journal of the European Communities No C* 2001 December 8:348/10.
- 3 Council Directive 1992/28/EEC of 31 March 1992 on the advertising of medicinal products for human use. (Articles 1(3) and 3(1)). *Official Journal of the European Communities No L* 1995 11 February:32/26.
- 4 Directive 2000/31/EC on mutual recognition of primary medical and specialist medical qualifications and minimum standards of training. *Official Journal of the European Communities No L* 2001 July 31:206/1-51.
- 5 Schwartz J. Medical websites faulted on privacy. *Washington Post* 2000 February 1.
- 6 http://europa.eu.int/information_society/eeurope/ehealth/quality/draft_guidelines/index_en.htm (accessed 5 Feb 2002).
- 7 European Commission. Recommendation 2/2001 on certain minimum requirements for collecting personal data on-line in the European Union. Adopted on 17 May 2001. http://europa.eu.int/comm/internal_market/en/dataprot/wpdocs/wp43en.htm (accessed 25 Jan 2002).

Statistics Notes

Validating scales and indexes

J Martin Bland, Douglas G Altman

Papers p 569

Department of
Public Health
Sciences, St
George's Hospital
Medical School,
London SW17 0RE
J Martin Bland
*professor of medical
statistics*

Cancer Research
UK Medical
Statistics Group,
Centre for Statistics
in Medicine,
Institute for Health
Sciences, Oxford
OX3 7LF

Douglas G Altman
*professor of statistics
in medicine*

Correspondence to:
Professor Bland
mbland@sghms.ac.uk

BMJ 2002;324:606-7

An index of quality is a measurement like any other, whether it is assessing a website, as in today's *BMJ*,¹ a clinical trial used in a meta-analysis,² or the quality of a life experienced by a patient.³ As with all measurements, we have to decide whether it measures what we want it to measure, and how well.

The simplest measurements, such as length and distance, can be validated by an objective criterion. The earliest criteria must have been biological: the length of a pace, a foot, a thumb. The obvious problem, that the criterion varies from person to person, was eventually solved by establishing a fundamental unit and defining all others in terms of it. Other measurements can then be defined in terms of a fundamental unit. To define a unit of weight we find a handy substance which appears the same everywhere, such as water. The unit of weight is then the weight of a volume of water specified in the basic unit of length, such as 100 cubic centimetres. Such measurements have *criterion validity*, meaning that we can take some known quantity and compare our measurement with it.

For some measurements no such standard is possible. Cardiac stroke volume, for example, can be measured only indirectly. Direct measurement, by collecting all the blood pumped out of the heart over a series of beats, would involve rather drastic interference with the system. Our criterion becomes agreement with another indirect measurement. Indeed, we sometimes have to use as a standard a method which we know produces inaccurate measurements.

Some quantities are even more difficult to measure and evaluate. Cardiac stroke volume does at least have an objective reality; a physical quantity of blood is pumped out of the heart when it beats. Anxiety and depression do not have a physical reality but are useful artificial constructs. They are measured by questionnaire scales, where answers to a series of questions related to the concept we want to measure are combined to give a numerical score. Website quality is similar. We are measuring a quantity which is not precisely defined, and there is no instrument with which we can compare any measure we might devise. How are we to assess the validity of such a scale?

The relevant theory was developed in the social sciences in the context of questionnaire scales.⁴ First we might ask whether the scale looks right, whether it asks about the sorts of thing which we think of as being related to anxiety or website quality. If it appears to be correct, we call this *face validity*. Next we might ask whether it covers all the aspects which we want to measure. A phobia scale which asked about fear of dogs, spiders, snakes, and cats but ignored height, confined spaces, and crowds would not do this. We call appropriate coverage of the subject matter *content validity*.

Our scale may look right and cover the right things, but what other evidence can we bring to the question of validity? One question we can ask is whether our score has the relationships with other variables that we would expect. For example, does an anxiety measure

distinguish between psychiatric patients and medical patients? Do we get different anxiety scores from students before and after an examination? Does a measure of depression predict suicide attempts? We call the property of having appropriate relationships with other variables *construct validity*.

We can also ask whether the items which together compose the scale are related to one another: does the scale have *internal consistency*? If not, do the items really measure the same thing? On the other hand, if the items are too similar, some may be redundant. Highly correlated items in a scale may make the scale over-long and may lead to some aspects being overemphasised, impairing the content validity. A handy summary measure for this feature is Cronbach's alpha.⁵

A scale must also be repeatable and be sufficiently objective to give similar results for different observers. If a measurement is repeatable, in that someone who has a high score on one occasion tends to have a high score on another, it must be measuring something. With physical measurements, it is often possible for the same observer (or different observers) to make repeated measurements in quick succession. When there is a subjective element in the measurement the observer can be blinded from their first measurement,

and different observers can make simultaneous measurements. In assessing the reliability of a website quality scale, it is easy to get several observers to apply the scale independently. With websites, repeat assessments need to be close in time because their content changes frequently (as does *bmj.com*). With questionnaires, either self administered or recorded by an observer, repeat measurements need to be far enough apart in time for the earlier responses to be forgotten, yet not so far apart that the underlying quantity being measured might have changed. Such data enable us to evaluate *test-retest reliability*. If two measures have comparable face, content, and construct validity the more repeatable one may be preferred for the study of a given population.

- 1 Gagliardi A, Jadad AR. Examination of instruments used to rate quality of health information on the internet: chronicle of a voyage with an unclear destination. *BMJ* 2002;324:569-73.
- 2 Jüni P, Altman DG, Egger M. Assessing the quality of controlled clinical trials. *BMJ* 2001;323:42-6.
- 3 Muldoon MF, Barger SD, Flory JD, Manuck B. What are quality of life measurements measuring? *BMJ* 1998;316:542-5.
- 4 Streiner DL, Norman GR. *Health measurement scales: a practical guide to their development and use*. 2nd ed. Oxford: Oxford University Press, 1996.
- 5 Bland JM, Altman DG. Cronbach's alpha. *BMJ* 1997;314:572.

Honour a physician with the honour due unto him

A few years ago my general practitioner told me that anyone aged over 40 with upper abdominal discomfort needed investigating. At the local teaching hospital, a pleasant young doctor did a gastroscopy, which showed a mass in my stomach wall. I was sent for a barium meal. A consultant radiologist took the x ray films, instructing me briskly to turn this way and that but not otherwise paying me any attention. He told me to wait a few minutes while he checked the films to see if all the views were satisfactory. I sat alone in the room for about five minutes.

From the moment the consultant re-entered I could see that he was slightly agitated. "I'm terribly sorry," he called out as he came through the door at the far end. And then again, "I'm terribly sorry." Perhaps these words of regret, coupled with the concern on his face, might not have had the effect they did had I not been a man with an abdominal mass on his mind. At this moment of truth and reckoning, certain visions swam before my eyes.

Three strides later, he was in front of me and looking me full in the face: "I'm terribly sorry, I hadn't realised you were a doctor." In his hand was the request form, and I could see that my general practitioner had written "ex-SR here" in one corner. He must have spotted this when checking the form as he looked at the preliminary plates. Though no further x rays were needed, he proceeded a little breathlessly to deliver three or four minutes of almost a caricature of caring, empathic interest in a patient. What branch of medicine was I in, and where did I work? Good heavens, that must be tough. Is that an Australian accent I hear? A St Mary's old boy, ah yes. What did I think about...?

I don't mean to imply that this was insincere, merely splendidly different from his earlier matter of factness and economy of word. I had thought nothing of this at the time: in such a bread and butter procedure I had

no more reason to expect the doctor to engage with me as a person than I would the phlebotomist taking a routine blood sample. Clearly, this consultant saw things similarly as a rule, but when the patient was a doctor the aesthetics of the encounter changed. He had apologised three times for what he felt was a lapse on his part, arising from his failure to notice what was written in the corner of the request form. Perhaps he thought I knew that my general practitioner had written this and that I expected this of a medical referral, and thus expected to be recognised by him not just as a patient but also as a colleague. He seemed to see this as my due. (As it happens, I didn't.)

I had forgotten this incident, but it was brought back to me by the aftermath of the Bristol cardiac surgery debacle, and by the publicity surrounding other recent medical scandals. These have all put a spotlight on relations between doctors, who seem to offer each other acknowledgement and empathy, as my consultant had sought belatedly to do to me. The general public may be coming to suspect that this collegiate solidarity is somehow not in their interests, associating it with mutual protectiveness and thus with cover-ups of medical malpractice. It is too soon to say how the profession will react, but my consultant was an older man and my guess is that, with younger generations of doctors, we will see the waning of a tradition whose roots lie with Hippocrates. For it was his oath that bound doctors to look well on each other (and not charge each other for their services).

It's another story, but I found out later that the mass was the gastroscopy instrument itself distorting the stomach wall, misdiagnosed by an inexperienced registrar. No special treatment there, anyway.

Derek Summerfield *consultant psychiatrist, CASCAID, South London and Maudsley NHS Trust, London*